Apostolos Antonacopoulos ·
Subhasis Chaudhuri · Rama Chellappa ·
Cheng-Lin Liu · Saumik Bhattacharya ·
Umapada Pal (Eds.)

LNCS 15314

Pattern Recognition

27th International Conference, ICPR 2024
Kolkata, India, December 1–5, 2024
Proceedings, Part XIV

14 Part XIV

Springer

MOREMEDIA

Lecture Notes in Computer Science **15314**

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Apostolos Antonacopoulos ·
Subhasis Chaudhuri · Rama Chellappa ·
Cheng-Lin Liu · Saumik Bhattacharya ·
Umapada Pal
Editors

Pattern Recognition

27th International Conference, ICPR 2024
Kolkata, India, December 1–5, 2024
Proceedings, Part XIV

Springer

Editors
Apostolos Antonacopoulos
University of Salford
Salford, UK

Subhasis Chaudhuri
Indian Institute of Technology Bombay
Mumbai, India

Rama Chellappa
Johns Hopkins University
Baltimore, MD, USA

Cheng-Lin Liu
Chinese Academy of Sciences
Beijing, China

Saumik Bhattacharya
IIT Kharagpur
Kharagpur, India

Umapada Pal
Indian Statistical Institute Kolkata
Kolkata, India

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-78340-1 ISBN 978-3-031-78341-8 (eBook)
https://doi.org/10.1007/978-3-031-78341-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

President's Address

On behalf of the Executive Committee of the International Association for Pattern Recognition (IAPR), I am pleased to welcome you to the 27th International Conference on Pattern Recognition (ICPR 2024), the main scientific event of the IAPR.

After a completely digital ICPR in the middle of the COVID pandemic and the first hybrid version in 2022, we can now enjoy a fully back-to-normal ICPR this year. I look forward to hearing inspirational talks and keynotes, catching up with colleagues during the breaks and making new contacts in an informal way. At the same time, the conference landscape has changed. Hybrid meetings have made their entrance and will continue. It is exciting to experience how this will influence the conference. Planning for a major event like ICPR must take place over a period of several years. This means many decisions had to be made under a cloud of uncertainty, adding to the already large effort needed to produce a successful conference. It is with enormous gratitude, then, that we must thank the team of organizers for their hard work, flexibility, and creativity in organizing this ICPR. ICPR always provides a wonderful opportunity for the community to gather together. I can think of no better location than Kolkata to renew the bonds of our international research community.

Each ICPR is a bit different owing to the vision of its organizing committee. For 2024, the conference has six different tracks reflecting major themes in pattern recognition: Artificial Intelligence, Pattern Recognition and Machine Learning; Computer and Robot Vision; Image, Speech, Signal and Video Processing; Biometrics and Human Computer Interaction; Document Analysis and Recognition; and Biomedical Imaging and Bioinformatics. This reflects the richness of our field. ICPR 2024 also features two dozen workshops, seven tutorials, and 15 competitions; there is something for everyone. Many thanks to those who are leading these activities, which together add significant value to attending ICPR, whether in person or virtually. Because it is important for ICPR to be as accessible as possible to colleagues from all around the world, we are pleased that the IAPR, working with the ICPR organizers, is continuing our practice of awarding travel stipends to a number of early-career authors who demonstrate financial need. Last but not least, we are thankful to the Springer LNCS team for their effort to publish these proceedings.

Among the presentations from distinguished keynote speakers, we are looking forward to the three IAPR Prize Lectures at ICPR 2024. This year we honor the achievements of Tin Kam Ho (IBM Research) with the IAPR's most prestigious King-Sun Fu Prize "for pioneering contributions to multi-classifier systems, random decision forests, and data complexity analysis". The King-Sun Fu Prize is given in recognition of an outstanding technical contribution to the field of pattern recognition. It honors the memory of Professor King-Sun Fu who was instrumental in the founding of IAPR, served as its first president, and is widely recognized for his extensive contributions to the field of pattern recognition.

The Maria Petrou Prize is given to a living female scientist/engineer who has made substantial contributions to the field of Pattern Recognition and whose past contributions, current research activity and future potential may be regarded as a model to both aspiring and established researchers. It honours the memory of Professor Maria Petrou as a scientist of the first rank, and particularly her role as a pioneer for women researchers. This year, the Maria Petrou Prize is given to Guoying Zhao (University of Oulu), "for contributions to video analysis for facial micro-behavior recognition and remote biosignal reading (RPPG) for heart rate analysis and face anti-spoofing".

The J.K. Aggarwal Prize is given to a young scientist who has brought a substantial contribution to a field that is relevant to the IAPR community and whose research work has had a major impact on the field. Professor Aggarwal is widely recognized for his extensive contributions to the field of pattern recognition and for his participation in IAPR's activities. This year, the J.K. Aggarwal Prize goes to Xiaolong Wang (UC San Diego) "for groundbreaking contributions to advancing visual representation learning, utilizing self-supervised and attention-based models to establish fundamental frameworks for creating versatile, general-purpose pattern recognition systems".

During the conference we will also recognize 21 new IAPR Fellows selected from a field of very strong candidates. In addition, a number of Best Scientific Paper and Best Student Paper awards will be presented, along with the Best Industry Related Paper Award and the Piero Zamperoni Best Student Paper Award. Congratulations to the recipients of these very well-deserved awards!

I would like to close by again thanking everyone involved in making ICPR 2024 a tremendous success; your hard work is deeply appreciated. These thanks extend to all who chaired the various aspects of the conference and the associated workshops, my ExCo colleagues, and the IAPR Standing and Technical Committees. Linda O'Gorman, the IAPR Secretariat, deserves special recognition for her experience, historical perspective, and attention to detail when it comes to supporting many of the IAPR's most important activities. Her tasks became so numerous that she recently got support from Carolyn Buckley (layout, newsletter), Ugur Halici (ICPR matters), and Rosemary Stramka (secretariat). The IAPR website got a completely new design. Ed Sobczak has taken care of our web presence for so many years already. A big thank you to all of you!

This is, of course, the 27th ICPR conference. Knowing that ICPR is organized every two years, and that the first conference in the series (1973!) pre-dated the formal founding of the IAPR by a few years, it is also exciting to consider that we are celebrating over 50 years of ICPR and at the same time approaching the official IAPR 50th anniversary in 2028: you'll get all information you need at ICPR 2024. In the meantime, I offer my thanks and my best wishes to all who are involved in supporting the IAPR throughout the world.

September 2024

Arjan Kuijper
President of the IAPR

Preface

It is our great pleasure to welcome you to the proceedings of the 27th International Conference on Pattern Recognition (ICPR 2024), held in Kolkata, India. The city, formerly known as 'Calcutta', is the home of the fabled Indian Statistical Institute (ISI), which has been at the forefront of statistical pattern recognition for almost a century. Concepts like the Mahalanobis distance, Bhattacharyya bound, Cramer–Rao bound, and Fisher–Rao metric were invented by pioneers associated with ISI. The first ICPR (called IJCPR then) was held in 1973, and the second in 1974. Subsequently, ICPR has been held every other year. The International Association for Pattern Recognition (IAPR) was founded in 1978 and became the sponsor of the ICPR series. Over the past 50 years, ICPR has attracted huge numbers of scientists, engineers and students from all over the world and contributed to advancing research, development and applications in pattern recognition technology.

ICPR 2024 was held at the Biswa Bangla Convention Centre, one of the largest such facilities in South Asia, situated just 7 kilometers from Kolkata Airport (CCU). According to ChatGPT "Kolkata is often called the 'Cultural Capital of India'. The city has a deep connection to literature, music, theater, and art. It was home to Nobel laureate Rabindranath Tagore, and the Bengali film industry has produced globally renowned filmmakers like Satyajit Ray. The city boasts remarkable colonial architecture, with landmarks like Victoria Memorial, Howrah Bridge, and the Indian Museum (the oldest and largest museum in India). Kolkata's streets are dotted with old mansions and buildings that tell stories of its colonial past. Walking through the city can feel like stepping back into a different era. Finally, Kolkata is also known for its street food."

ICPR 2024 followed a two-round paper submission format. We received a total of 2135 papers (1501 papers in round-1 submissions, and 634 papers in round-2 submissions). Each paper, on average, received 2.84 reviews, in single-blind mode. For the first-round papers we had a rebuttal option available to authors.

In total, 945 papers (669 from round-1 and 276 from round-2) were accepted for presentation, resulting in an acceptance rate of 44.26%, which is consistent with previous ICPR events. At ICPR 2024 the papers were categorized into six tracks: Artificial Intelligence, Machine Learning for Pattern Analysis; Computer Vision and Robotic Perception; Image, Video, Speech, and Signal Analysis; Biometrics and Human-Machine Interaction; Document and Media Analysis; and Biomedical Image Analysis and Informatics.

The main conference ran over December 2–5, 2024. The main program included the presentation of 188 oral papers (19.89% of the accepted papers), 757 poster papers and 12 competition papers (out of 15 submitted). A total 10 oral sessions were held concurrently in four meeting rooms with a total of 40 oral sessions. In total 24 workshops and 7 tutorials were held on December 1, 2024.

The plenary sessions included three prize lectures and three invited presentations. The prize lectures were delivered by Tin Kam Ho (IBM Research, USA; King Sun

Fu Prize winner), Xiaolong Wang (University of California, San Diego, USA; J.K. Aggarwal Prize winner), and Guoying Zhao (University of Oulu, Finland; Maria Petrou Prize winner). The invited speakers were Timothy Hospedales (University of Edinburgh, UK), Venu Govindaraju (University at Buffalo, USA), and Shuicheng Yan (Skywork AI, Singapore).

Several best paper awards were presented in ICPR: the Piero Zamperoni Award for the best paper authored by a student, the BIRPA Best Industry Related Paper Award, and the Best Paper Awards and Best Student Paper Awards for each of the six tracks of ICPR 2024.

The organization of such a large conference would not be possible without the help of many volunteers. Our special gratitude goes to the Program Chairs (Apostolos Antonacopoulos, Subhasis Chaudhuri, Rama Chellappa and Cheng-Lin Liu), for their leadership in organizing the program. Thanks to our Publication Chairs (Ananda S. Chowdhury and Wataru Ohyama) for handling the overwhelming workload of publishing the conference proceedings. We also thank our Competition Chairs (Richard Zanibbi, Lianwen Jin and Laurence Likforman-Sulem) for arranging 12 important competitions as part of ICPR 2024. We are thankful to our Workshop Chairs (P. Shivakumara, Stephanie Schuckers, Jean-Marc Ogier and Prabir Bhattacharya) and Tutorial Chairs (B.B. Chaudhuri, Michael R. Jenkin and Guoying Zhao) for arranging the workshops and tutorials on emerging topics. ICPR 2024, for the first time, held a Doctoral Consortium. We would like to thank our Doctoral Consortium Chairs (Véronique Eglin, Dan Lopresti and Mayank Vatsa) for organizing it.

Thanks go to the Track Chairs and the meta reviewers who devoted significant time to the review process and preparation of the program. We also sincerely thank the reviewers who provided valuable feedback to the authors.

Finally, we acknowledge the work of other conference committee members, like the Organizing Chairs and Organizing Committee Members, Finance Chairs, Award Chair, Sponsorship Chairs, and Exhibition and Demonstration Chairs, Visa Chair, Publicity Chairs, and Women in ICPR Chairs, whose efforts made this event successful. We also thank our event manager Alpcord Network for their help.

We hope that all the participants found the technical program informative and enjoyed the sights, culture and cuisine of Kolkata.

October 2024

Umapada Pal
Josef Kittler
Anil Jain

Organization

General Chairs

Umapada Pal	Indian Statistical Institute, Kolkata, India
Josef Kittler	University of Surrey, UK
Anil Jain	Michigan State University, USA

Program Chairs

Apostolos Antonacopoulos	University of Salford, UK
Subhasis Chaudhuri	Indian Institute of Technology, Bombay, India
Rama Chellappa	Johns Hopkins University, USA
Cheng-Lin Liu	Institute of Automation, Chinese Academy of Sciences, China

Publication Chairs

Ananda S. Chowdhury	Jadavpur University, India
Wataru Ohyama	Tokyo Denki University, Japan

Competition Chairs

Richard Zanibbi	Rochester Institute of Technology, USA
Lianwen Jin	South China University of Technology, China
Laurence Likforman-Sulem	Télécom Paris, France

Workshop Chairs

P. Shivakumara	University of Salford, UK
Stephanie Schuckers	Clarkson University, USA
Jean-Marc Ogier	Université de la Rochelle, France
Prabir Bhattacharya	Concordia University, Canada

Tutorial Chairs

B. B. Chaudhuri	Indian Statistical Institute, Kolkata, India
Michael R. Jenkin	York University, Canada
Guoying Zhao	University of Oulu, Finland

Doctoral Consortium Chairs

Véronique Eglin	CNRS, France
Daniel P. Lopresti	Lehigh University, USA
Mayank Vatsa	Indian Institute of Technology, Jodhpur, India

Organizing Chairs

Saumik Bhattacharya	Indian Institute of Technology, Kharagpur, India
Palash Ghosal	Sikkim Manipal University, India

Organizing Committee

Santanu Phadikar	West Bengal University of Technology, India
SK Md Obaidullah	Aliah University, India
Sayantari Ghosh	National Institute of Technology Durgapur, India
Himadri Mukherjee	West Bengal State University, India
Nilamadhaba Tripathy	Clarivate Analytics, USA
Chayan Halder	West Bengal State University, India
Shibaprasad Sen	Techno Main Salt Lake, India

Finance Chairs

Kaushik Roy	West Bengal State University, India
Michael Blumenstein	University of Technology Sydney, Australia

Awards Committee Chair

Arpan Pal	Tata Consultancy Services, India

Sponsorship Chairs

P. J. Narayanan	Indian Institute of Technology, Hyderabad, India
Yasushi Yagi	Osaka University, Japan
Venu Govindaraju	University at Buffalo, USA
Alberto Bel Bimbo	Università di Firenze, Italy

Exhibition and Demonstration Chairs

Arjun Jain	FastCode AI, India
Agnimitra Biswas	National Institute of Technology, Silchar, India

International Liaison, Visa Chair

Balasubramanian Raman	Indian Institute of Technology, Roorkee, India

Publicity Chairs

Dipti Prasad Mukherjee	Indian Statistical Institute, Kolkata, India
Bob Fisher	University of Edinburgh, UK
Xiaojun Wu	Jiangnan University, China

Women in ICPR Chairs

Ingela Nystrom	Uppsala University, Sweden
Alexandra B. Albu	University of Victoria, Canada
Jing Dong	Institute of Automation, Chinese Academy of Sciences, China
Sarbani Palit	Indian Statistical Institute, Kolkata, India

Event Manager

Alpcord Network

Track Chairs – Artificial Intelligence, Machine Learning for Pattern Analysis

Larry O'Gorman	Nokia Bell Labs, USA
Dacheng Tao	University of Sydney, Australia
Petia Radeva	University of Barcelona, Spain
Susmita Mitra	Indian Statistical Institute, Kolkata, India
Jiliang Tang	Michigan State University, USA

Track Chairs – Computer and Robot Vision

C. V. Jawahar	International Institute of Information Technology (IIIT), Hyderabad, India
João Paulo Papa	São Paulo State University, Brazil
Maja Pantic	Imperial College London, UK
Gang Hua	Dolby Laboratories, USA
Junwei Han	Northwestern Polytechnical University, China

Track Chairs – Image, Speech, Signal and Video Processing

P. K. Biswas	Indian Institute of Technology, Kharagpur, India
Shang-Hong Lai	National Tsing Hua University, Taiwan
Hugo Jair Escalante	INAOE, CINVESTAV, Mexico
Sergio Escalera	Universitat de Barcelona, Spain
Prem Natarajan	University of Southern California, USA

Track Chairs – Biometrics and Human Computer Interaction

Richa Singh	Indian Institute of Technology, Jodhpur, India
Massimo Tistarelli	University of Sassari, Italy
Vishal Patel	Johns Hopkins University, USA
Wei-Shi Zheng	Sun Yat-sen University, China
Jian Wang	Snap, USA

Track Chairs – Document Analysis and Recognition

Xiang Bai	Huazhong University of Science and Technology, China
David Doermann	University at Buffalo, USA
Josep Llados	Universitat Autònoma de Barcelona, Spain
Mita Nasipuri	Jadavpur University, India

Track Chairs – Biomedical Imaging and Bioinformatics

Jayanta Mukhopadhyay	Indian Institute of Technology, Kharagpur, India
Xiaoyi Jiang	Universität Münster, Germany
Seong-Whan Lee	Korea University, Korea

Metareviewers (Conference Papers and Competition Papers)

Wael Abd-Almageed	University of Southern California, USA
Maya Aghaei	NHL Stenden University, Netherlands
Alireza Alaei	Southern Cross University, Australia
Rajagopalan N. Ambasamudram	Indian Institute of Technology, Madras, India
Suyash P. Awate	Indian Institute of Technology, Bombay, India
Inci M. Baytas	Bogazici University, Turkey
Aparna Bharati	Lehigh University, USA
Brojeshwar Bhowmick	Tata Consultancy Services, India
Jean-Christophe Burie	University of La Rochelle, France
Gustavo Carneiro	University of Surrey, UK
Chee Seng Chan	Universiti Malaya, Malaysia
Sumohana S. Channappayya	Indian Institute of Technology, Hyderabad, India
Dongdong Chen	Microsoft, USA
Shengyong Chen	Tianjin University of Technology, China
Jun Cheng	Institute for Infocomm Research, A*STAR, Singapore
Albert Clapés	University of Barcelona, Spain
Oscar Dalmau	Center for Research in Mathematics, Mexico

Tyler Derr	Vanderbilt University, USA
Abhinav Dhall	Indian Institute of Technology, Ropar, India
Bo Du	Wuhan University, China
Yuxuan Du	University of Sydney, Australia
Ayman S. El-Baz	University of Louisville, USA
Francisco Escolano	University of Alicante, Spain
Siamac Fazli	Nazarbayev University, Kazakhstan
Jianjiang Feng	Tsinghua University, China
Gernot A. Fink	TU Dortmund University, Germany
Alicia Fornes	CVC, Spain
Junbin Gao	University of Sydney, Australia
Yan Gao	Amazon, USA
Yongsheng Gao	Griffith University, Australia
Caren Han	University of Melbourne, Australia
Ran He	Institute of Automation, Chinese Academy of Sciences, China
Tin Kam Ho	IBM, USA
Di Huang	Beihang University, China
Kaizhu Huang	Duke Kunshan University, China
Donato Impedovo	University of Bari, Italy
Julio Jacques	University of Barcelona and Computer Vision Center, Spain
Lianwen Jin	South China University of Technology, China
Wei Jin	Emory University, USA
Danilo Samuel Jodas	São Paulo State University, Brazil
Manjunath V. Joshi	DA-IICT, India
Jayashree Kalpathy-Cramer	Massachusetts General Hospital, USA
Dimosthenis Karatzas	Computer Vision Centre, Spain
Hamid Karimi	Utah State University, USA
Baiying Lei	Shenzhen University, China
Guoqi Li	Chinese Academy of Sciences, and Peng Cheng Lab, China
Laurence Likforman-Sulem	Institut Polytechnique de Paris/Télécom Paris, France
Aishan Liu	Beihang University, China
Bo Liu	Bytedance, USA
Chen Liu	Clarkson University, USA
Cheng-Lin Liu	Institute of Automation, Chinese Academy of Sciences, China
Hongmin Liu	University of Science and Technology Beijing, China
Hui Liu	Michigan State University, USA

Jing Liu	Institute of Automation, Chinese Academy of Sciences, China
Li Liu	University of Oulu, Finland
Qingshan Liu	Nanjing University of Posts and Telecommunications, China
Adrian P. Lopez-Monroy	Centro de Investigacion en Matematicas AC, Mexico
Daniel P. Lopresti	Lehigh University, USA
Shijian Lu	Nanyang Technological University, Singapore
Yong Luo	Wuhan University, China
Andreas K. Maier	FAU Erlangen-Nuremberg, Germany
Davide Maltoni	University of Bologna, Italy
Hong Man	Stevens Institute of Technology, USA
Lingtong Min	Northwestern Polytechnical University, China
Paolo Napoletano	University of Milano-Bicocca, Italy
Kamal Nasrollahi	Milestone Systems, Aalborg University, Denmark
Marcos Ortega	University of A Coruña, Spain
Shivakumara Palaiahnakote	University of Salford, UK
P. Jonathon Phillips	NIST, USA
Filiberto Pla	University Jaume I, Spain
Ajit Rajwade	Indian Institute of Technology, Bombay, India
Shanmuganathan Raman	Indian Institute of Technology, Gandhinagar, India
Imran Razzak	UNSW, Australia
Beatriz Remeseiro	University of Oviedo, Spain
Gustavo Rohde	University of Virginia, USA
Partha Pratim Roy	Indian Institute of Technology, Roorkee, India
Sanjoy K. Saha	Jadavpur University, India
Joan Andreu Sánchez	Universitat Politècnica de València, Spain
Claudio F. Santos	UFSCar, Brazil
Shin'ichi Satoh	National Institute of Informatics, Japan
Stephanie Schuckers	Clarkson University, USA
Srirangaraj Setlur	University at Buffalo, SUNY, USA
Debdoot Sheet	Indian Institute of Technology, Kharagpur, India
Jun Shen	University of Wollongong, Australia
Li Shen	JD Explore Academy, China
Chen Shengyong	Zhejiang University of Technology and Tianjin University of Technology, China
Andy Song	RMIT University, Australia
Akihiro Sugimoto	National Institute of Informatics, Japan
Qianru Sun	Singapore Management University, Singapore
Arijit Sur	Indian Institute of Technology, Guwahati, India
Estefania Talavera	University of Twente, Netherlands

Wei Tang	University of Illinois at Chicago, USA
Joao M. Tavares	Universidade do Porto, Portugal
Jun Wan	NLPR, CASIA, China
Le Wang	Xi'an Jiaotong University, China
Lei Wang	Australian National University, Australia
Xiaoyang Wang	Tencent AI Lab, USA
Xinggang Wang	Huazhong University of Science and Technology, China
Xiao-Jun Wu	Jiangnan University, China
Yiding Yang	Bytedance, China
Xiwen Yao	Northwestern Polytechnical University, China
Xu-Cheng Yin	University of Science and Technology Beijing, China
Baosheng Yu	University of Sydney, Australia
Shiqi Yu	Southern University of Science and Technology, China
Xin Yuan	Westlake University, China
Yibing Zhan	JD Explore Academy, China
Jing Zhang	University of Sydney, Australia
Lefei Zhang	Wuhan University, China
Min-Ling Zhang	Southeast University, China
Wenbin Zhang	Florida International University, USA
Jiahuan Zhou	Peking University, China
Sanping Zhou	Xi'an Jiaotong University, China
Tianyi Zhou	University of Maryland, USA
Lei Zhu	Shandong Normal University, China
Pengfei Zhu	Tianjin University, China
Wangmeng Zuo	Harbin Institute of Technology, China

Reviewers (Competition Papers)

Liangcai Gao	Da-Han Wang
Mingxin Huang	Yang Xue
Lei Kang	Wentao Yang
Wenhui Liao	Jiaxin Zhang
Yuliang Liu	Yiwu Zhong
Yongxin Shi	

Reviewers (Conference Papers)

Aakanksha Aakanksha
Aayush Singla
Abdul Muqeet
Abhay Yadav
Abhijeet Vijay Nandedkar
Abhimanyu Sahu
Abhinav Rajvanshi
Abhisek Ray
Abhishek Shrivastava
Abhra Chaudhuri
Aditi Roy
Adriano Simonetto
Adrien Maglo
Ahmed Abdulkadir
Ahmed Boudissa
Ahmed Hamdi
Ahmed Rida Sekkat
Ahmed Sharafeldeen
Aiman Farooq
Aishwarya Venkataramanan
Ajay Kumar
Ajay Kumar Reddy Poreddy
Ajita Rattani
Ajoy Mondal
Akbar K.
Akbar Telikani
Akshay Agarwal
Akshit Jindal
Al Zadid Sultan Bin Habib
Albert Clapés
Alceu Britto
Alejandro Peña
Alessandro Ortis
Alessia Auriemma Citarella
Alexandre Stenger
Alexandros Sopasakis
Alexia Toumpa
Ali Khan
Alik Pramanick
Alireza Alaei
Alper Yilmaz
Aman Verma
Amit Bhardwaj
Amit More
Amit Nandedkar
Amitava Chatterjee
Amos L. Abbott
Amrita Mohan
Anand Mishra
Ananda S. Chowdhury
Anastasia Zakharova
Anastasios L. Kesidis
Andras Horvath
Andre Gustavo Hochuli
André P. Kelm
Andre Wyzykowski
Andrea Bottino
Andrea Lagorio
Andrea Torsello
Andreas Fischer
Andreas K. Maier
Andreu Girbau Xalabarder
Andrew Beng Jin Teoh
Andrew Shin
Andy J. Ma
Aneesh S. Chivukula
Ángela Casado-García
Anh Quoc Nguyen
Anindya Sen
Anirban Saha
Anjali Gautam
Ankan Bhattacharyya
Ankit Jha
Anna Scius-Bertrand
Annalisa Franco
Antoine Doucet
Antonino Staiano
Antonio Fernández
Antonio Parziale
Anu Singha
Anustup Choudhury
Anwesan Pal
Anwesha Sengupta
Archisman Adhikary
Arjan Kuijper
Arnab Kumar Das

Arnav Bhavsar
Arnav Varma
Arpita Dutta
Arshad Jamal
Artur Jordao
Arunkumar Chinnaswamy
Aryan Jadon
Aryaz Baradarani
Ashima Anand
Ashis Dhara
Ashish Phophalia
Ashok K. Bhateja
Ashutosh Vaish
Ashwani Kumar
Asifuzzaman Lasker
Atefeh Khoshkhahtinat
Athira Nambiar
Attilio Fiandrotti
Avandra S. Hemachandra
Avik Hati
Avinash Sharma
B. H. Shekar
B. Uma Shankar
Bala Krishna Thunakala
Balaji Tk
Balázs Pálffy
Banafsheh Adami
Bang-Dang Pham
Baochang Zhang
Baodi Liu
Bashirul Azam Biswas
Beiduo Chen
Benedikt Kottler
Beomseok Oh
Berkay Aydin
Berlin S. Shaheema
Bertrand Kerautret
Bettina Finzel
Bhavana Singh
Bibhas C. Dhara
Bilge Gunsel
Bin Chen
Bin Li
Bin Liu
Bin Yao
Bin-Bin Jia
Binbin Yong
Bindita Chaudhuri
Bindu Madhavi Tummala
Binh M. Le
Bi-Ru Dai
Bo Huang
Bo Jiang
Bob Zhang
Bowen Liu
Bowen Zhang
Boyang Zhang
Boyu Diao
Boyun Li
Brian M. Sadler
Bruce A. Maxwell
Bryan Bo Cao
Buddhika L. Semage
Bushra Jalil
Byeong-Seok Shin
Byung-Gyu Kim
Caihua Liu
Cairong Zhao
Camille Kurtz
Carlos A. Caetano
Carlos D. Martã-Nez-Hinarejos
Ce Wang
Cevahir Cigla
Chakravarthy Bhagvati
Chandrakanth Vipparla
Changchun Zhang
Changde Du
Changkun Ye
Changxu Cheng
Chao Fan
Chao Guo
Chao Qu
Chao Wen
Chayan Halder
Che-Jui Chang
Chen Feng
Chenan Wang
Cheng Yu
Chenghao Qian
Cheng-Lin Liu

Chengxu Liu
Chenru Jiang
Chensheng Peng
Chetan Ralekar
Chih-Wei Lin
Chih-Yi Chiu
Chinmay Sahu
Chintan Patel
Chintan Shah
Chiranjoy Chattopadhyay
Chong Wang
Choudhary Shyam Prakash
Christophe Charrier
Christos Smailis
Chuanwei Zhou
Chun-Ming Tsai
Chunpeng Wang
Ciro Russo
Claudio De Stefano
Claudio F. Santos
Claudio Marrocco
Connor Levenson
Constantine Dovrolis
Constantine Kotropoulos
Dai Shi
Dakshina Ranjan Kisku
Dan Anitei
Dandan Zhu
Daniela Pamplona
Danli Wang
Danqing Huang
Daoan Zhang
Daqing Hou
David A. Clausi
David Freire Obregon
David Münch
David Pujol Perich
Davide Marelli
De Zhang
Debalina Barik
Debapriya Roy (Kundu)
Debashis Das
Debashis Das Chakladar
Debi Prosad Dogra
Debraj D. Basu
Decheng Liu
Deen Dayal Mohan
Deep A. Patel
Deepak Kumar
Dengpan Liu
Denis Coquenet
Désiré Sidibé
Devesh Walawalkar
Dewan Md. Farid
Di Ming
Di Qiu
Di Yuan
Dian Jia
Dianmo Sheng
Diego Thomas
Diganta Saha
Dimitri Bulatov
Dimpy Varshni
Dingcheng Yang
Dipanjan Das
Dipanjyoti Paul
Divya Biligere Shivanna
Divya Saxena
Divya Sharma
Dmitrii Matveichev
Dmitry Minskiy
Dmitry V. Sorokin
Dong Zhang
Donghua Wang
Donglin Zhang
Dongming Wu
Dongqiangzi Ye
Dongqing Zou
Dongrui Liu
Dongyang Zhang
Dongzhan Zhou
Douglas Rodrigues
Duarte Folgado
Duc Minh Vo
Duoxuan Pei
Durai Arun Pannir Selvam
Durga Bhavani S.
Eckart Michaelsen
Elena Goyanes
Élodie Puybareau

Emanuele Vivoli
Emna Ghorbel
Enrique Naredo
Enyu Cai
Eric Patterson
Ernest Valveny
Eva Blanco-Mallo
Eva Breznik
Evangelos Sartinas
Fabio Solari
Fabiola De Marco
Fan Wang
Fangda Li
Fangyuan Lei
Fangzhou Lin
Fangzhou Luo
Fares Bougourzi
Farman Ali
Fatiha Mokdad
Fei Shen
Fei Teng
Fei Zhu
Feiyan Hu
Felipe Gomes Oliveira
Feng Li
Fengbei Liu
Fenghua Zhu
Fillipe D. M. De Souza
Flavio Piccoli
Flavio Prieto
Florian Kleber
Francesc Serratosa
Francesco Bianconi
Francesco Castro
Francesco Ponzio
Francisco Javier Hernández López
Frédéric Rayar
Furkan Osman Kar
Fushuo Huo
Fuxiao Liu
Fu-Zhao Ou
Gabriel Turinici
Gabrielle Flood
Gajjala Viswanatha Reddy
Gaku Nakano
Galal Binamakhashen
Ganesh Krishnasamy
Gang Pan
Gangyan Zeng
Gani Rahmon
Gaurav Harit
Gennaro Vessio
Genoveffa Tortora
George Azzopardi
Gerard Ortega
Gerardo E. Altamirano-Gomez
Gernot A. Fink
Gibran Benitez-Garcia
Gil Ben-Artzi
Gilbert Lim
Giorgia Minello
Giorgio Fumera
Giovanna Castellano
Giovanni Puglisi
Giulia Orrù
Giuliana Ramella
Gökçe Uludoğan
Gopi Ramena
Gorthi Rama Krishna Sai Subrahmanyam
Gourav Datta
Gowri Srinivasa
Gozde Sahin
Gregory Randall
Guanjie Huang
Guanjun Li
Guanwen Zhang
Guanyu Xu
Guanyu Yang
Guanzhou Ke
Guhnoo Yun
Guido Borghi
Guilherme Brandão Martins
Guillaume Caron
Guillaume Tochon
Guocai Du
Guohao Li
Guoqiang Zhong
Guorong Li
Guotao Li
Gurman Gill

Haechang Lee
Haichao Zhang
Haidong Xie
Haifeng Zhao
Haimei Zhao
Hainan Cui
Haixia Wang
Haiyan Guo
Hakime Ozturk
Hamid Kazemi
Han Gao
Hang Zou
Hanjia Lyu
Hanjoo Cho
Hanqing Zhao
Hanyuan Liu
Hanzhou Wu
Hao Li
Hao Meng
Hao Sun
Hao Wang
Hao Xing
Hao Zhao
Haoan Feng
Haodi Feng
Haofeng Li
Haoji Hu
Haojie Hao
Haojun Ai
Haopeng Zhang
Haoran Li
Haoran Wang
Haorui Ji
Haoxiang Ma
Haoyu Chen
Haoyue Shi
Harald Koestler
Harbinder Singh
Harris V. Georgiou
Hasan F. Ates
Hasan S. M. Al-Khaffaf
Hatef Otroshi Shahreza
Hebeizi Li
Heng Zhang
Hengli Wang
Hengyue Liu
Hertog Nugroho
Hieyong Jeong
Himadri Mukherjee
Hoai Ngo
Hoda Mohaghegh
Hong Liu
Hong Man
Hongcheng Wang
Hongjian Zhan
Hongxi Wei
Hongyu Hu
Hoseong Kim
Hossein Ebrahimnezhad
Hossein Malekmohamadi
Hrishav Bakul Barua
Hsueh-Yi Sean Lin
Hua Wei
Huafeng Li
Huali Xu
Huaming Chen
Huan Wang
Huang Chen
Huanran Chen
Hua-Wen Chang
Huawen Liu
Huayi Zhan
Hugo Jair Escalante
Hui Chen
Hui Li
Huichen Yang
Huiqiang Jiang
Huiyuan Yang
Huizi Yu
Hung T. Nguyen
Hyeongyu Kim
Hyeonjeong Park
Hyeonjun Lee
Hymalai Bello
Hyung-Gun Chi
Hyunsoo Kim
I-Chen Lin
Ik Hyun Lee
Ilan Shimshoni
Imad Eddine Toubal

Imran Sarker
Inderjot Singh Saggu
Indrani Mukherjee
Indranil Sur
Ines Rieger
Ioannis Pierros
Irina Rabaev
Ivan V. Medri
J. Rafid Siddiqui
Jacek Komorowski
Jacopo Bonato
Jacson Rodrigues Correia-Silva
Jaekoo Lee
Jaime Cardoso
Jakob Gawlikowski
Jakub Nalepa
James L. Wayman
Jan Čech
Jangho Lee
Jani Boutellier
Javier Gurrola-Ramos
Javier Lorenzo-Navarro
Jayasree Saha
Jean Lee
Jean Paul Barddal
Jean-Bernard Hayet
Jean-Philippe G. Tarel
Jean-Yves Ramel
Jenny Benois-Pineau
Jens Bayer
Jerin Geo James
Jesús Miguel García-Gorrostieta
Jia Qu
Jiahong Chen
Jiaji Wang
Jian Hou
Jian Liang
Jian Xu
Jian Zhu
Jianfeng Lu
Jianfeng Ren
Jiangfan Liu
Jianguo Wang
Jiangyan Yi
Jiangyong Duan
Jianhua Yang
Jianhua Zhang
Jianhui Chen
Jianjia Wang
Jianli Xiao
Jianqiang Xiao
Jianwu Wang
Jianxin Zhang
Jianxiong Gao
Jianxiong Zhou
Jianyu Wang
Jianzhong Wang
Jiaru Zhang
Jiashu Liao
Jiaxin Chen
Jiaxin Lu
Jiaxing Ye
Jiaxuan Chen
Jiaxuan Li
Jiayi He
Jiayin Lin
Jie Ou
Jiehua Zhang
Jiejie Zhao
Jignesh S. Bhatt
Jin Gao
Jin Hou
Jin Hu
Jin Shang
Jing Tian
Jing Yu Chen
Jingfeng Yao
Jinglun Feng
Jingtong Yue
Jingwei Guo
Jingwen Xu
Jingyuan Xia
Jingzhe Ma
Jinhong Wang
Jinjia Wang
Jinlai Zhang
Jinlong Fan
Jinming Su
Jinrong He
Jintao Huang

Jinwoo Ahn
Jinwoo Choi
Jinyang Liu
Jinyu Tian
Jionghao Lin
Jiuding Duan
Jiwei Shen
Jiyan Pan
Jiyoun Kim
João Papa
Johan Debayle
John Atanbori
John Wilson
John Zhang
Jónathan Heras
Joohi Chauhan
Jorge Calvo-Zaragoza
Jorge Figueroa
Jorma Laaksonen
José Joaquim De Moura Ramos
Jose Vicent
Joseph Damilola Akinyemi
Josiane Zerubia
Juan Wen
Judit Szücs
Juepeng Zheng
Juha Roning
Jumana H. Alsubhi
Jun Cheng
Jun Ni
Jun Wan
Junghyun Cho
Junjie Liang
Junjie Ye
Junlin Hu
Juntong Ni
Junxin Lu
Junxuan Li
Junyaup Kim
Junyeong Kim
Jürgen Seiler
Jushang Qiu
Juyang Weng
Jyostna Devi Bodapati
Jyoti Singh Kirar
Kai Jiang
Kaiqiang Song
Kalidas Yeturu
Kalle Åström
Kamalakar Vijay Thakare
Kang Gu
Kang Ma
Kanji Tanaka
Karthik Seemakurthy
Kaushik Roy
Kavisha Jayathunge
Kazuki Uehara
Ke Shi
Keigo Kimura
Keiji Yanai
Kelton A. P. Costa
Kenneth Camilleri
Kenny Davila
Ketan Atul Bapat
Ketan Kotwal
Kevin Desai
Keyu Long
Khadiga Mohamed Ali
Khakon Das
Khan Muhammad
Kilho Son
Kim-Ngan Nguyen
Kishan Kc
Kishor P. Upla
Klaas Dijkstra
Komal Bharti
Konstantinos Triaridis
Kostas Ioannidis
Koyel Ghosh
Kripabandhu Ghosh
Krishnendu Ghosh
Kshitij S. Jadhav
Kuan Yan
Kun Ding
Kun Xia
Kun Zeng
Kunal Banerjee
Kunal Biswas
Kunchi Li
Kurban Ubul

Lahiru N. Wijayasingha
Laines Schmalwasser
Lakshman Mahto
Lala Shakti Swarup Ray
Lale Akarun
Lan Yan
Lawrence Amadi
Lee Kang Il
Lei Fan
Lei Shi
Lei Wang
Leonardo Rossi
Lequan Lin
Levente Tamas
Li Bing
Li Li
Li Ma
Li Song
Lia Morra
Liang Xie
Liang Zhao
Lianwen Jin
Libing Zeng
Lidia Sánchez-González
Lidong Zeng
Lijun Li
Likang Wang
Lili Zhao
Lin Chen
Lin Huang
Linfei Wang
Ling Lo
Lingchen Meng
Lingheng Meng
Lingxiao Li
Lingzhong Fan
Liqi Yan
Liqiang Jing
Lisa Gutzeit
Liu Ziyi
Liushuai Shi
Liviu-Daniel Stefan
Liyuan Ma
Liyun Zhu
Lizuo Jin
Longteng Guo
Lorena Álvarez Rodríguez
Lorenzo Putzu
Lu Leng
Lu Pang
Lu Wang
Luan Pham
Luc Brun
Luca Guarnera
Luca Piano
Lucas Alexandre Ramos
Lucas Goncalves
Lucas M. Gago
Luigi Celona
Luis C. S. Afonso
Luis Gerardo De La Fraga
Luis S. Luevano
Luis Teixeira
Lunke Fei
M. Hassaballah
Maddimsetti Srinivas
Mahendran N.
Mahesh Mohan M. R.
Maiko Lie
Mainak Singha
Makoto Hirose
Malay Bhattacharyya
Mamadou Dian Bah
Man Yao
Manali J. Patel
Manav Prabhakar
Manikandan V. M.
Manish Bhatt
Manjunath Shantharamu
Manuel Curado
Manuel Günther
Manuel Marques
Marc A. Kastner
Marc Chaumont
Marc Cheong
Marc Lalonde
Marco Cotogni
Marcos C. Santana
Mario Molinara
Mariofanna Milanova

Markus Bauer
Marlon Becker
Mårten Wadenbäck
Martin G. Ljungqvist
Martin Kampel
Martina Pastorino
Marwan Torki
Masashi Nishiyama
Masayuki Tanaka
Massimo O. Spata
Matteo Ferrara
Matthew D. Dawkins
Matthew Gadd
Matthew S. Watson
Maura Pintor
Max Ehrlich
Maxim Popov
Mayukh Das
Md Baharul Islam
Md Sajid
Meghna Kapoor
Meghna P. Ayyar
Mei Wang
Meiqi Wu
Melissa L. Tijink
Meng Li
Meng Liu
Meng-Luen Wu
Mengnan Liu
Mengxi China Guo
Mengya Han
Michaël Clément
Michal Kawulok
Mickael Coustaty
Miguel Domingo
Milind G. Padalkar
Ming Liu
Ming Ma
Mingchen Feng
Mingde Yao
Minghao Li
Mingjie Sun
Ming-Kuang Daniel Wu
Mingle Xu
Mingyong Li
Mingyuan Jiu
Minh P. Nguyen
Minh Q. Tran
Minheng Ni
Minsu Kim
Minyi Zhao
Mirko Paolo Barbato
Mo Zhou
Modesto Castrillón-Santana
Mohamed Amine Mezghich
Mohamed Dahmane
Mohamed Elsharkawy
Mohamed Yousuf
Mohammad Hashemi
Mohammad Khalooei
Mohammad Khateri
Mohammad Mahdi Dehshibi
Mohammad Sadil Khan
Mohammed Mahmoud
Moises Diaz
Monalisha Mahapatra
Monidipa Das
Mostafa Kamali Tabrizi
Mridul Ghosh
Mrinal Kanti Bhowmik
Muchao Ye
Mugalodi Ramesha Rakesh
Muhammad Rameez Ur Rahman
Muhammad Suhaib Kanroo
Muming Zhao
Munender Varshney
Munsif Ali
Na Lv
Nader Karimi
Nagabhushan Somraj
Nakkwan Choi
Nakul Agarwal
Nan Pu
Nan Zhou
Nancy Mehta
Nand Kumar Yadav
Nandakishor Nandakishor
Nandyala Hemachandra
Nanfeng Jiang
Narayan Hegde

Narayan Ji Mishra
Narayan Vetrekar
Narendra D. Londhe
Nathalie Girard
Nati Ofir
Naval Kishore Mehta
Nazmul Shahadat
Neeti Narayan
Neha Bhargava
Nemanja Djuric
Newlin Shebiah R.
Ngo Ba Hung
Nhat-Tan Bui
Niaz Ahmad
Nick Theisen
Nicolas Passat
Nicolas Ragot
Nicolas Sidere
Nikolaos Mitianoudis
Nikolas Ebert
Nilah Ravi Nair
Nilesh A. Ahuja
Nilkanta Sahu
Nils Murrugarra-Llerena
Nina S. T. Hirata
Ninad Aithal
Ning Xu
Ningzhi Wang
Niraj Kumar
Nirmal S. Punjabi
Nisha Varghese
Norio Tagawa
Obaidullah Md Sk
Oguzhan Ulucan
Olfa Mechi
Oliver Tüselmann
Orazio Pontorno
Oriol Ramos Terrades
Osman Akin
Ouadi Beya
Ozge Mercanoglu Sincan
Pabitra Mitra
Padmanabha Reddy Y. C. A.
Palaash Agrawal
Palaiahnakote Shivakumara
Palash Ghosal
Pallav Dutta
Paolo Rota
Paramanand Chandramouli
Paria Mehrani
Parth Agrawal
Partha Basuchowdhuri
Patrick Horain
Pavan Kumar
Pavan Kumar Anasosalu Vasu
Pedro Castro
Peipei Li
Peipei Yang
Peisong Shen
Peiyu Li
Peng Li
Pengfei He
Pengrui Quan
Pengxin Zeng
Pengyu Yan
Peter Eisert
Petra Gomez-Krämer
Pierrick Bruneau
Ping Cao
Pingping Zhang
Pintu Kumar
Pooja Kumari
Pooja Sahani
Prabhu Prasad Dev
Pradeep Kumar
Pradeep Singh
Pranjal Sahu
Prasun Roy
Prateek Keserwani
Prateek Mittal
Praveen Kumar Chandaliya
Praveen Tirupattur
Pravin Nair
Preeti Gopal
Preety Singh
Prem Shanker Yadav
Prerana Mukherjee
Prerna A. Mishra
Prianka Dey
Priyanka Mudgal

Qc Kha Ng
Qi Li
Qi Ming
Qi Wang
Qi Zuo
Qian Li
Qiang Gan
Qiang He
Qiang Wu
Qiangqiang Zhou
Qianli Zhao
Qiansen Hong
Qiao Wang
Qidong Huang
Qihua Dong
Qin Yuke
Qing Guo
Qingbei Guo
Qingchao Zhang
Qingjie Liu
Qinhong Yang
Qiushi Shi
Qixiang Chen
Quan Gan
Quanlong Guan
Rachit Chhaya
Radu Tudor Ionescu
Rafal Zdunek
Raghavendra Ramachandra
Rahimul I. Mazumdar
Rahul Kumar Ray
Rajib Dutta
Rajib Ghosh
Rakesh Kumar
Rakesh Paul
Rama Chellappa
Rami O. Skaik
Ramon Aranda
Ran Wei
Ranga Raju Vatsavai
Ranganath Krishnan
Rasha Friji
Rashmi S.
Razaib Tariq
Rémi Giraud
René Schuster
Renlong Hang
Renrong Shao
Renu Sharma
Reza Sadeghian
Richard Zanibbi
Rimon Elias
Rishabh Shukla
Rita Delussu
Riya Verma
Robert J. Ravier
Robert Sablatnig
Robin Strand
Rocco Pietrini
Rocio Diaz Martin
Rocio Gonzalez-Diaz
Rohit Venkata Sai Dulam
Romain Giot
Romi Banerjee
Ru Wang
Ruben Machucho
Ruddy Théodose
Ruggero Pintus
Rui Deng
Rui P. Paiva
Rui Zhao
Ruifan Li
Ruigang Fu
Ruikun Li
Ruirui Li
Ruixiang Jiang
Ruowei Jiang
Rushi Lan
Rustam Zhumagambetov
S. Amutha
S. Divakar Bhat
Sagar Goyal
Sahar Siddiqui
Sahbi Bahroun
Sai Karthikeya Vemuri
Saibal Dutta
Saihui Hou
Sajad Ahmad Rather
Saksham Aggarwal
Sakthi U.

Salimeh Sekeh
Samar Bouazizi
Samia Boukir
Samir F. Harb
Samit Biswas
Samrat Mukhopadhyay
Samriddha Sanyal
Sandika Biswas
Sandip Purnapatra
Sanghyun Jo
Sangwoo Cho
Sanjay Kumar
Sankaran Iyer
Sanket Biswas
Santanu Roy
Santosh D. Pandure
Santosh Ku Behera
Santosh Nanabhau Palaskar
Santosh Prakash Chouhan
Sarah S. Alotaibi
Sasanka Katreddi
Sathyanarayanan N. Aakur
Saurabh Yadav
Sayan Rakshit
Scott McCloskey
Sebastian Bunda
Sejuti Rahman
Selim Aksoy
Sen Wang
Seraj A. Mostafa
Shanmuganathan Raman
Shao-Yuan Lo
Shaoyuan Xu
Sharia Arfin Tanim
Shehreen Azad
Sheng Wan
Shengdong Zhang
Shengwei Qin
Shenyuan Gao
Sherry X. Chen
Shibaprasad Sen
Shigeaki Namiki
Shiguang Liu
Shijie Ma
Shikun Li
Shinichiro Omachi
Shirley David
Shishir Shah
Shiv Ram Dubey
Shiva Baghel
Shivanand S. Gornale
Shogo Sato
Shotaro Miwa
Shreya Ghosh
Shreya Goyal
Shuai Su
Shuai Wang
Shuai Zheng
Shuaifeng Zhi
Shuang Qiu
Shuhei Tarashima
Shujing Lyu
Shuliang Wang
Shun Zhang
Shunming Li
Shunxin Wang
Shuping Zhao
Shuquan Ye
Shuwei Huo
Shuyue Lan
Shyi-Chyi Cheng
Si Chen
Siddarth Ravichandran
Sihan Chen
Siladittya Manna
Silambarasan Elkana Ebinazer
Simon Benaïchouche
Simon S. Woo
Simone Caldarella
Simone Milani
Simone Zini
Sina Lotfian
Sitao Luan
Sivaselvan B.
Siwei Li
Siwei Wang
Siwen Luo
Siyu Chen
Sk Aziz Ali
Sk Md Obaidullah

Sneha Shukla
Snehasis Banerjee
Snehasis Mukherjee
Snigdha Sen
Sofia Casarin
Soheila Farokhi
Soma Bandyopadhyay
Son Minh Nguyen
Son Xuan Ha
Sonal Kumar
Sonam Gupta
Sonam Nahar
Song Ouyang
Sotiris Kotsiantis
Souhaila Djaffal
Soumen Biswas
Soumen Sinha
Soumitri Chattopadhyay
Souvik Sengupta
Spiros Kostopoulos
Sreeraj Ramachandran
Sreya Banerjee
Srikanta Pal
Srinivas Arukonda
Stephane A. Guinard
Su O. Ruan
Subhadip Basu
Subhajit Paul
Subhankar Ghosh
Subhankar Mishra
Subhankar Roy
Subhash Chandra Pal
Subhayu Ghosh
Sudip Das
Sudipta Banerjee
Suhas Pillai
Sujit Das
Sukalpa Chanda
Sukhendu Das
Suklav Ghosh
Suman K. Ghosh
Suman Samui
Sumit Mishra
Sungho Suh
Sunny Gupta
Suraj Kumar Pandey
Surendrabikram Thapa
Suresh Sundaram
Sushil Bhattacharjee
Susmita Ghosh
Swakkhar Shatabda
Syed Ms Islam
Syed Tousiful Haque
Taegyeong Lee
Taihui Li
Takashi Shibata
Takeshi Oishi
Talha Ahmad Siddiqui
Tanguy Gernot
Tangwen Qian
Tanima Bhowmik
Tanpia Tasnim
Tao Dai
Tao Hu
Tao Sun
Taoran Yi
Tapan Shah
Taveena Lotey
Teng Huang
Tengqi Ye
Teresa Alarcon
Tetsuji Ogawa
Thanh Phuong Nguyen
Thanh Tuan Nguyen
Thattapon Surasak
Thibault Napolãon
Thierry Bouwmans
Thinh Truong Huynh Nguyen
Thomas De Min
Thomas E. K. Zielke
Thomas Swearingen
Tianatahina Jimmy Francky Randrianasoa
Tianheng Cheng
Tianjiao He
Tianyi Wei
Tianyuan Zhang
Tianyue Zheng
Tiecheng Song
Tilottama Goswami
Tim Büchner

Tim H. Langer
Tim Raven
Tingkai Liu
Tingting Yao
Tobias Meisen
Toby P. Breckon
Tong Chen
Tonghua Su
Tran Tuan Anh
Tri-Cong Pham
Trishna Saikia
Trung Quang Truong
Tuan T. Nguyen
Tuan Vo Van
Tushar Shinde
Ujjwal Karn
Ukrit Watchareeruetai
Uma Mudenagudi
Umarani Jayaraman
V. S. Malemath
Vallidevi Krishnamurthy
Ved Prakash
Venkata Krishna Kishore Kolli
Venkata R. Vavilthota
Venkatesh Thirugnana Sambandham
Verónica Maria Vasconcelos
Véronique Ve Eglin
Víctor E. Alonso-Pérez
Vinay Palakkode
Vinayak S. Nageli
Vincent J. Whannou De Dravo
Vincenzo Conti
Vincenzo Gattulli
Vineet Padmanabhan
Vishakha Pareek
Viswanath Gopalakrishnan
Vivek Singh Baghel
Vivekraj K.
Vladimir V. Arlazarov
Vu-Hoang Tran
W. Sylvia Lilly Jebarani
Wachirawit Ponghiran
Wafa Khlif
Wang An-Zhi
Wanli Xue
Wataru Ohyama
Wee Kheng Leow
Wei Chen
Wei Cheng
Wei Hua
Wei Lu
Wei Pan
Wei Tian
Wei Wang
Wei Wei
Wei Zhou
Weidi Liu
Weidong Yang
Weijun Tan
Weimin Lyu
Weinan Guan
Weining Wang
Weiqiang Wang
Weiwei Guo
Weixia Zhang
Wei-Xuan Bao
Weizhong Jiang
Wen Xie
Wenbin Qian
Wenbin Tian
Wenbin Wang
Wenbo Zheng
Wenhan Luo
Wenhao Wang
Wen-Hung Liao
Wenjie Li
Wenkui Yang
Wenwen Si
Wenwen Yu
Wenwen Zhang
Wenwu Yang
Wenxi Li
Wenxi Yue
Wenxue Cui
Wenzhuo Liu
Widhiyo Sudiyono
Willem Dijkstra
Wolfgang Fuhl
Xi Zhang
Xia Yuan

Xianda Zhang
Xiang Zhang
Xiangdong Su
Xiang-Ru Yu
Xiangtai Li
Xiangyu Xu
Xiao Guo
Xiao Hu
Xiao Wu
Xiao Yang
Xiaofeng Zhang
Xiaogang Du
Xiaoguang Zhao
Xiaoheng Jiang
Xiaohong Zhang
Xiaohua Huang
Xiaohua Li
Xiao-Hui Li
Xiaolong Sun
Xiaosong Li
Xiaotian Li
Xiaoting Wu
Xiaotong Luo
Xiaoyan Li
Xiaoyang Kang
Xiaoyi Dong
Xin Guo
Xin Lin
Xin Ma
Xinchi Zhou
Xingguang Zhang
Xingjian Leng
Xingpeng Zhang
Xingzheng Lyu
Xinjian Huang
Xinqi Fan
Xinqi Liu
Xinqiao Zhang
Xinrui Cui
Xizhan Gao
Xu Cao
Xu Ouyang
Xu Zhao
Xuan Shen
Xuan Zhou
Xuchen Li
Xuejing Lei
Xuelu Feng
Xueting Liu
Xuewei Li
Xueyi X. Wang
Xugong Qin
Xu-Qian Fan
Xuxu Liu
Xu-Yao Zhang
Yan Huang
Yan Li
Yan Wang
Yan Xia
Yan Zhuang
Yanan Li
Yanan Zhang
Yang Hou
Yang Jiao
Yang Liping
Yang Liu
Yang Qian
Yang Yang
Yang Zhao
Yangbin Chen
Yangfan Zhou
Yanhui Guo
Yanjia Huang
Yanjun Zhu
Yanming Zhang
Yanqing Shen
Yaoming Cai
Yaoxin Zhuo
Yaoyan Zheng
Yaping Zhang
Yaqian Liang
Yarong Feng
Yasmina Benmabrouk
Yasufumi Sakai
Yasutomo Kawanishi
Yazeed Alzahrani
Ye Du
Ye Duan
Yechao Zhang
Yeong-Jun Cho

Yi Huo
Yi Shi
Yi Yu
Yi Zhang
Yibo Liu
Yibo Wang
Yi-Chieh Wu
Yifan Chen
Yifei Huang
Yihao Ding
Yijie Tang
Yikun Bai
Yimin Wen
Yinan Yang
Yin-Dong Zheng
Yinfeng Yu
Ying Dai
Yingbo Li
Yiqiao Li
Yiqing Huang
Yisheng Lv
Yisong Xiao
Yite Wang
Yizhe Li
Yong Wang
Yonghao Dong
Yong-Hyuk Moon
Yongjie Li
Yongqian Li
Yongqiang Mao
Yongxu Liu
Yongyu Wang
Yongzhi Li
Youngha Hwang
Yousri Kessentini
Yu Wang
Yu Zhou
Yuan Tian
Yuan Zhang
Yuanbo Wen
Yuanxin Wang
Yubin Hu
Yubo Huang
Yuchen Ren
Yucheng Xing
Yuchong Yao
Yuecong Min
Yuewei Yang
Yufei Zhang
Yufeng Yin
Yugen Yi
Yuhang Ming
Yujia Zhang
Yujun Ma
Yukiko Kenmochi
Yun Hoyeoung
Yun Liu
Yunhe Feng
Yunxiao Shi
Yuru Wang
Yushun Tang
Yusuf Osmanlioglu
Yusuke Fujita
Yuta Nakashima
Yuwei Yang
Yuwu Lu
Yuxi Liu
Yuya Obinata
Yuyao Yan
Yuzhi Guo
Zaipeng Xie
Zander W. Blasingame
Zedong Wang
Zeliang Zhang
Zexin Ji
Zhanxiang Feng
Zhaofei Yu
Zhe Chen
Zhe Cui
Zhe Liu
Zhe Wang
Zhekun Luo
Zhen Yang
Zhenbo Li
Zhenchun Lei
Zhenfei Zhang
Zheng Liu
Zheng Wang
Zhengming Yu
Zhengyin Du

Zhengyun Cheng
Zhenshen Qu
Zhenwei Shi
Zhenzhong Kuang
Zhi Cai
Zhi Chen
Zhibo Chu
Zhicun Yin
Zhida Huang
Zhida Zhang
Zhifan Gao
Zhihang Ren
Zhihang Yuan
Zhihao Wang
Zhihua Xie
Zhihui Wang
Zhikang Zhang
Zhiming Zou
Zhiqi Shao
Zhiwei Dong
Zhiwei Qi
Zhixiang Wang
Zhixuan Li
Zhiyu Jiang
Zhiyuan Yan
Zhiyuan Yu
Zhiyuan Zhang
Zhong Chen
Zhongwei Teng
Zhongzhan Huang
Zhongzhi Yu
Zhuan Han
Zhuangzhuang Chen
Zhuo Liu
Zhuo Su
Zhuojun Zou
Zhuoyue Wang
Ziang Song
Zicheng Zhang
Zied Mnasri
Zifan Chen
Žiga Babnik
Zijing Chen
Zikai Zhang
Ziling Huang
Zilong Du
Ziqi Cai
Ziqi Zhou
Zi-Rui Wang
Zirui Zhou
Ziwen He
Ziyao Zeng
Ziyi Zhang
Ziyue Xiang
Zonglei Jing
Zongyi Xu

Contents – Part XIV

Detection of Intracranial Hemorrhage for Trauma Patients

Antoine P. Sanner[1,2(✉)], Nils F. Grauhan[2], Merle Meyer[2], Laura Leukert[2], Marc A. Brockmann[2], Ahmed E. Othman[2], and Anirban Mukhopadhyay[1]

[1] Department of Computer Science, Technical University of Darmstadt, Darmstadt, Germany
antoine.sanner@gris.tu-darmstadt.de

[2] Department of Neuroradiology, University Medical Center Mainz, Mainz, Germany

Abstract. Whole-body CT is used for multi-trauma patients in the search of any and all injuries. Since an initial assessment needs to be rapid and the search for lesions is done for the whole body, very little time can be allocated for the inspection of a specific anatomy. In particular, intracranial hemorrhages are still missed, especially by clinical students. In this work, we present a Deep Learning approach for highlighting such lesions to improve the diagnostic accuracy. While most works on intracranial hemorrhages perform segmentation, detection only requires bounding boxes for the localization of the bleeding. In this paper, we propose a novel Voxel-Complete IoU (VC-IoU) loss that encourages the network to learn the 3D aspect ratios of bounding boxes and leads to more precise detections. We extensively experiment on brain bleeding detection using a publicly available dataset, and validate it on a private cohort, where we achieve 0.877 AR_{30}, 0.728 AP_{30} and 0.653 AR_{30}, 0.514 AP_{30} respectively. These results constitute a relative +5% improvement in Average Recall for both datasets compared to other loss functions. Finally, as there is little data currently publicly available for 3D object detection and as annotation resources are limited in the clinical setting, we evaluate the cost of different annotation methods, as well as the impact of imprecise bounding boxes in the training data on the detection performance.

Keywords: 3D voxel object detection · Intracranial Hemorrhage · Multi-Trauma

1 Introduction

Trauma remains one of the leading causes of death around the world, despite recent improvement in treatment protocols [3]. The latest protocol consists of doing a whole-body CT of multi-trauma patients. However, this also means that the radiologist has more anatomies to review and less time can be allocated for specific body parts.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 1–14, 2025.
https://doi.org/10.1007/978-3-031-78341-8_1

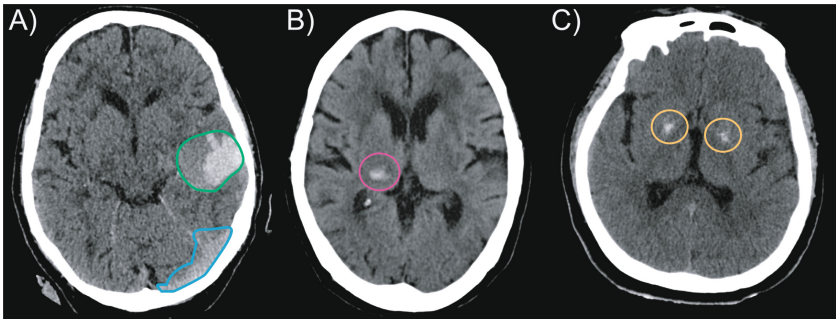


Fig. 1. Insights into the challenges of brain hemorrhage detection with examples of hard cases. A) The patient suffers both from a bleeding and a thrombosis further below. While the thrombosis has a similar appearance to the bleeding, the blood is still in a blood vessel and is not considered a bleeding. B and C) Both images show hyperdensities near the ventricle system. While the spot in image B is a bleeding, the two spots in C are simple calcifications of the brain tissue.

Medical students are often left alone to do these readings due to staff shortages and will often miss lesions [2] or will struggle on hard cases (see Fig. 1). In particular, missed intracranial hemorrhages can have devastating effects on the patient's odds of survival and future physical autonomy [14]. As a result, there is a clear utility in having expert tools that can focus on a specific part of the whole-body CT and highlight relevant areas.

When it comes to stroke, the majority of previous work on 3D data has been dedicated to semantic segmentation [13]. Segmentation masks take a long time to annotate and only experts have the required skills. Further, given the diversity of strokes in terms of lesion size, location, and shape [15], state-of-the-art models only achieve 0.6 to 0.7 Dice score, which is not yet clinically sufficient [9].

Additionally, Dice score as a metric is ill-suited for detection, as it only captures the proportion of blood volume detected. Similarly, a high Hausdorff distance could be due to a badly segmented bleeding or by a missed one. Both metrics provide no information on how many bleedings are completely missed. In contrast, bleeding detection aims to detect and localize individual bleeding to aid the clinician in finding relevant injuries or complication in a patient. This is particularly important for multi-trauma patients, as they will frequently suffer from multiple ailments and hemorrhages. As there are multiple solutions for annotating 3D voxel data, we take a deep dive into the pros and cons of different methods. As this process is often tedious, one can be tempted to choose a fast but inaccurate solution. As such, we also analyze the effect of various annotation errors on the model performance.

Detecting bleeding of vastly different size, shape, and position requires a multiscale approach. Inspired by Feature Pyramid Network (FPN) strategy of Retina-Net [1,10], we introduce an anisotropic-resolution-aware method leveraging features from 5 different, axis-independent scales. Additionally, we introduce

a family of anchors based on the distribution of the bleeding aspect ratio in training data, that also respect the anatomical structure of the brain. However, the network still needs to be able to learn these bleeding-specific shapes. To this end and inspired by the Complete Iou loss [17], we propose a novel Voxel-Complete IoU (**VC-IoU**) loss that encourages the network to learn the **3D aspect ratios** of bounding boxes. This loss specifically penalizes deviations in aspect ratios along all three planes in the volume. We evaluate our proposed method on the publicly available INSTANCE2022 dataset [9], as well as on a private cohort for external validation. Our contributions in this work include:

1. A novel loss (**VC-IoU**) for object detection in 3D voxel images[1], which leads to more precise detections.
2. An in depth comparison of annotation effort of multiple methods and an assessment of the impact of noisy bounding box annotation on the model's performance. These insights can help research groups to select the most adequate solution for other applications.

2 Methodology

In this section, we introduce our method for bleeding detection. More precisely, we introduce our backbone architecture, then define our novel Voxel-Complete IoU (VC-IoU) loss and finally go over challenges with sampling.

2.1 Backbone Architecture

The proposed 3D object detection method consists of a 3D Retina-Net with a Resnet-50-based FPN. This architecture has already proved its usefulness for 3D medical imaging [8] as its FPN allows to leverage multiscale features. In particular, it is very flexible regarding the pyramid levels used. **We choose to not only use levels P2 to P5, but to also include P6 for detection** (see Fig. 2). This is crucial as the volume of the bleeding can range from only 0.1 cm^3 to more than 100 cm^3. Additionally thanks to the convolutional nature of this architecture, we can customize the in-slice and depth-wise resolution scaling used to compute the next pyramid level. For instance, we use an in-slice downscaling factor of 2 from P0 to P1 and from P1 to P2. So if input slices have a 512×512 px size, their shape will already be shrunk to 128×128 px at P2. We only reduce the depth-wise resolution from P3 to P4 and from P4 to P5, since the slice thickness is $10\times$ greater than the in-slice resolution.

One also has to consider the great diversity in bleeding shape when designing the shape of the anchors at a 5mm slice thickness. Indeed, small bleeding may appear in one up to four slices. This range in object depth is problematic for matching, as the Intersection over Union (IoU) of a potential match with a ground truth box can be excessively penalized by an inadequately chosen depth for the designed anchors. While the ATSS matching algorithm [16] does not

[1] Code available at https://github.com/MECLabTUDA/VoxelSceneGraph.

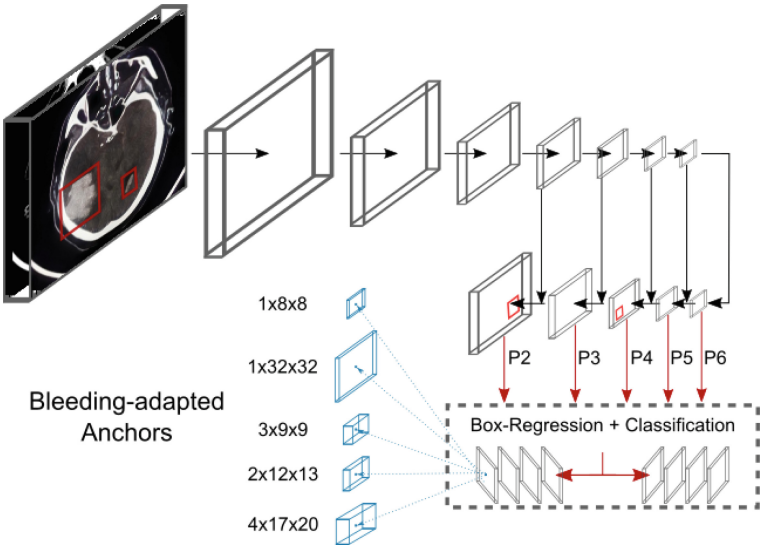


Fig. 2. Our method for bleeding detection. Anchor sizes are given for level P2.

require a hard IoU threshold for matching, the convergence rate of the box regression head can still suffer from misshapen anchors. Increasing the number of anchors can potentially improve matching results, but also increases the memory needed for computation. We have iteratively selected a family of 5 anchors. These are representative of the diverse 3D aspect ratios of bleeding within the datasets, while limiting the hardware requirements.

2.2 Voxel-Complete IoU Loss

Learning these specific shapes is not trivial. For 2D boxes, the aspect ratio is commonly defined as the proportion between the box's width and its height. However, no single ratio can be defined for a given 3D box (see Fig. 3). Inspired by Zheng et al. [17], we design a novel Voxel-Complete IoU (**VC-IoU**) loss by measuring the consistency of all three aspect ratios for a predicted box $B = (w, h, d)$ and a ground truth box $B^{gt} = (w^{gt}, h^{gt}, d^{gt})$:

$$v = \frac{4}{\pi^2} \begin{bmatrix} (\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h})^2 \\ +(\arctan \frac{h^{gt}}{d^{gt}} - \arctan \frac{h}{d})^2 \\ +(\arctan \frac{d^{gt}}{w^{gt}} - \arctan \frac{d}{w})^2 \end{bmatrix}$$

Similarly to the C-IoU loss [17], we define a trade-off factor α as

$$\alpha = \frac{v}{1 - IoU + v}$$

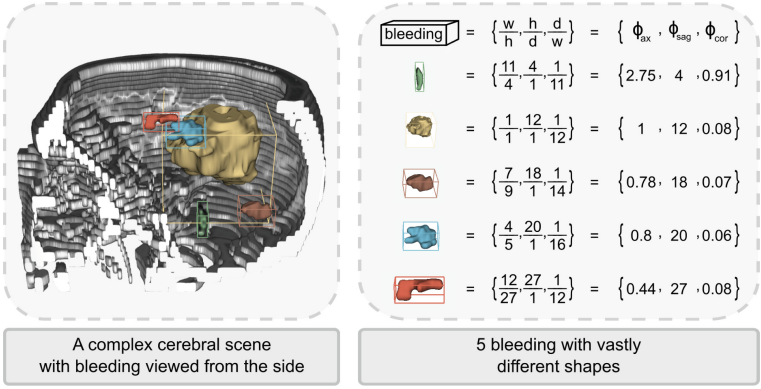


Fig. 3. In 2D, the aspect ratio ϕ gives a scale-independent representation of a given bounding box. However, no such representation exists for 3D bounding boxes. We can instead consider the three aspect ratios $\{\phi_{ax}, \phi_{sag}, \phi_{cor}\}$ in the three natural planes of the 3D space (axial, sagittal and coronal). In particular, bleeding have a wide variety of shapes even within a single individual.

to ensure that the overlap factor is given priority. The VC-IoU loss can then be defined as

$$\mathcal{L}_{VC-IoU} = \mathcal{L}_{DIoU} + \alpha v$$

where $\mathcal{L}_{DIoU}$ is the Distance IoU [17]. There are of course multiple options to aggregate the regularization term for all three aspect ratios when defining v. We choose to compute a sum rather than a mean, as we consider each of these penalties to be as important as the single penalty that would be observed in a 2D setting. Besides, the acceptable IoU threshold is typically much lower in 3D applications than in 2D ones (e.g. 10% vs 50%). So, having higher values of v allows the trade-off factor α to be in a comparable range compared to a 2D CIoU loss.

2.3 Sampling of Bleeding

The sampling algorithm for the classification head has to be tuned to avoid detecting all hyperdensities within the head. In particular, the transverse sigmoid sinuses are large veins along the back of the skull that has a texture close to bleeding in CT imaging. It is critical that a model learns to distinguish such structures. As such, it is critical to select negative samples from the most confidently predicted false positives. Qualitative examples are shown in Fig. 1.

3 Data

While whole-body CT are acquired for trauma patients, the focus for this application is exclusively on the patient's head. Handling the head region of whole-body CT or directly using a head CT are equivalent, as converting the former

to the latter only requires a rough localization of the patient's neck. Since it is not a hard task, we consider it out of scope of this study and directly use head CTs for this study.

3.1 Source Images

The first dataset is the publicly available INSTANCE2022 challenge dataset [9]. It contains 130 non-contrast head CTs with $0.42 \times 0.42 \times 5\,\mathrm{mm}^3$ voxel spacing from patients diagnosed with intracranial hemorrhage. Additionally, we use a private dataset for the purpose of external validation. It is constituted of 18 head CTs of patients diagnosed with intracranial hemorrhage between 2021 and 2022. As shown in Fig. 4, this dataset contains out-of-distribution cases, both in terms of bleeding size and number. These images are high-resolution head CTs and have a native voxel spacing of $0.41 \times 0.41 \times 0.3\ \mathrm{mm}^3$. As such, we resampled them to a voxel spacing of $0.41 \times 0.41 \times 5\,\mathrm{mm}^3$ for inference (see Fig. 5), to match the resolution of the training data.

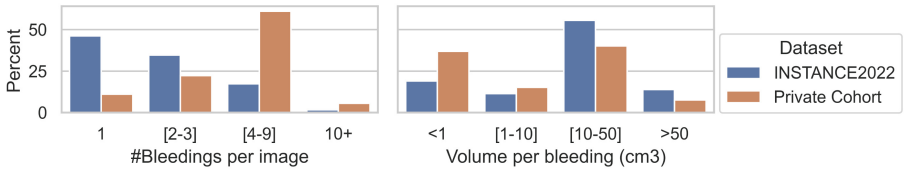


Fig. 4. Distribution shift of the number of bleeding per image (left) and volume per bleeding (right) between the INSTANCE2022 dataset and our private cohort.

3.2 Data Annotation

Of the 130 images from INSTANCE2022, 100 were released as training cases along with their corresponding segmentation of the bleeding. We derived the object annotation by generating the 3D connected components from the masks and computed the bounding box of each component. All other images were annotated internally by a senior radiologist from the University Medical Center Mainz using 3D Slicer [4]. In particular, we chose to annotate the objects as rough label maps (see Sect. 5.4).

4 Experimental Setup

In this section, we introduce the base images used and their annotation process. We then describe our evaluation setup for the different stages of our method.

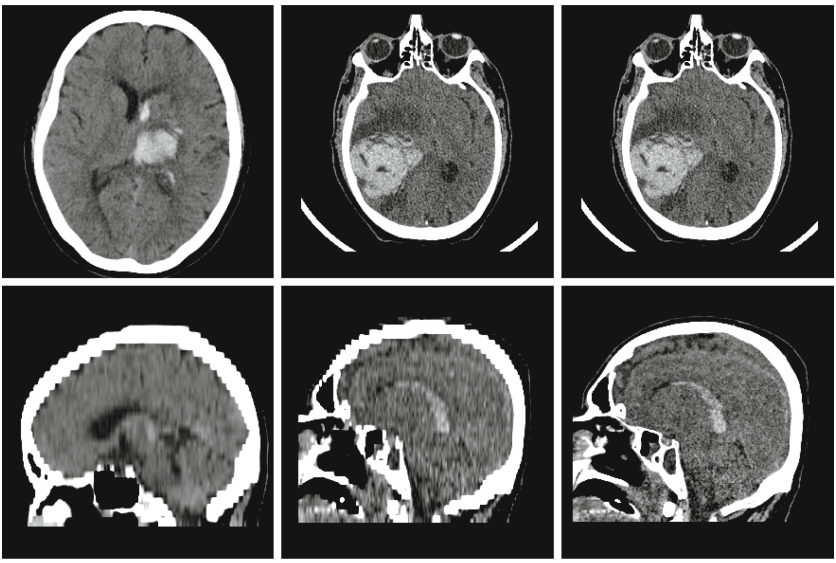


Fig. 5. Visualization of head CTs, axial (top) and sagittal (bottom). INSTANCE2022 data (left). Downscaling (middle) of the original private cohort data (right).

4.1 Bounding Box Prediction

We compare our method to the state-of-the-art nnDetection framework [1]. To train the detection models, we split the first 100 cases from the INSTANCE2022 dataset [9] for training and validation in an 80/20 fashion. The 30 remaining cases are used for in-distribution testing. Finally, the private cohort is used for external validation.

4.2 Segmentation for Object Detection

To evaluate segmentation for bleeding detection, we use a nnU-Net [7] trained with 5-fold cross-validation. This architecture performed well in the test phase of INSTANCE2022 with a Dice score of 0.69 [9]. We then utilize the sister framework nnDetection [1] to ensemble and convert the predicted segmentations to bounding boxes. The in-distribution testing and external validation are done in the same manner as for bounding box prediction.

4.3 Metrics

We evaluate methods for bleeding detection using average precision (AP) and average recall (AR) at IoU thresholds of 10% and 30%. Prior studies [1] suggest that using a threshold of 10% IoU is sufficient, as IoU is more penalizing in 3D and clinical applications only require a coarse localization. However, having such a low threshold is problematic for bleeding detection, as their volume can vary

greatly. In particular, a bounding box predicted for a given bleeding could easily overlap with a smaller neighboring bleeding. At a low 10% IoU threshold, this bounding box can often sufficiently overlap the second bleeding to be counted as a positive match for both objects. This can be particularly troublesome if the first larger bleeding was in fact a false positive (see Fig. 1A)). To mitigate this issue, we also provide all results at a 30% IoU threshold.

4.4 Implementation Details

Our framework is implemented in Python 3.10 and PyTorch 1.13 [12]. All configuration files including hyperparameters and data splits will be made available with the code. All models are trained on a single RTX3090 GPU. The training for our method takes approximately 7 h and requires up to 16 GB of VRAM. As a comparison, *nnDetection* [1] takes 24 h to train over 5 folds and requires 24 GB of VRAM. This increased memory usage compared to our method can mainly be attributed to the higher number of anchors that *nnDetection* uses.

5 Results

In this section, we first compare our method to the existing state-of-the-art methods. We then further evaluate our method through ablation studies. Finally, since annotating data is performed under heavy resource constraints in the clinical world, we perform a comparison of different annotation solutions and study the impact of imprecise boxes on model performances.

5.1 Bleeding Detection

Table 1. Detection rates at 10% and 30% IoU using our method, nnDetection [1] and [1]nnU-Net for detection [1].

Method	INSTANCE2022				Private Cohort			
	AR_{10} ↑	AR_{30} ↑	AP_{10} ↑	AP_{30} ↑	AR_{10} ↑	AR_{30} ↑	AP_{10} ↑	AP_{30} ↑
nnU-Net[1]	0.789	0.600	0.708	0.522	0.561	0.378	0.416	0.257
nnDetection	0.815	0.631	0.672	0.549	0.704	0.439	0.494	0.263
Ours	**0.892**	**0.877**	**0.760**	**0.728**	**0.724**	**0.653**	**0.590**	**0.514**

We first evaluate our method against state-of-the-art solutions. As shown in Table 1, **our method significantly outperforms both nnDetection and nnU-Net on both datasets**. nnU-Net comes last, showing how segmentation is inadequate for object detection. While there is a drop in performance on the private cohort for all methods, ours is more robust against distribution shifts. The FROC analysis (Fig. 6) sheds some further light on the difference in performance between datasets for each method. Additionally, we provide some qualitative results for our method in Fig. 7.

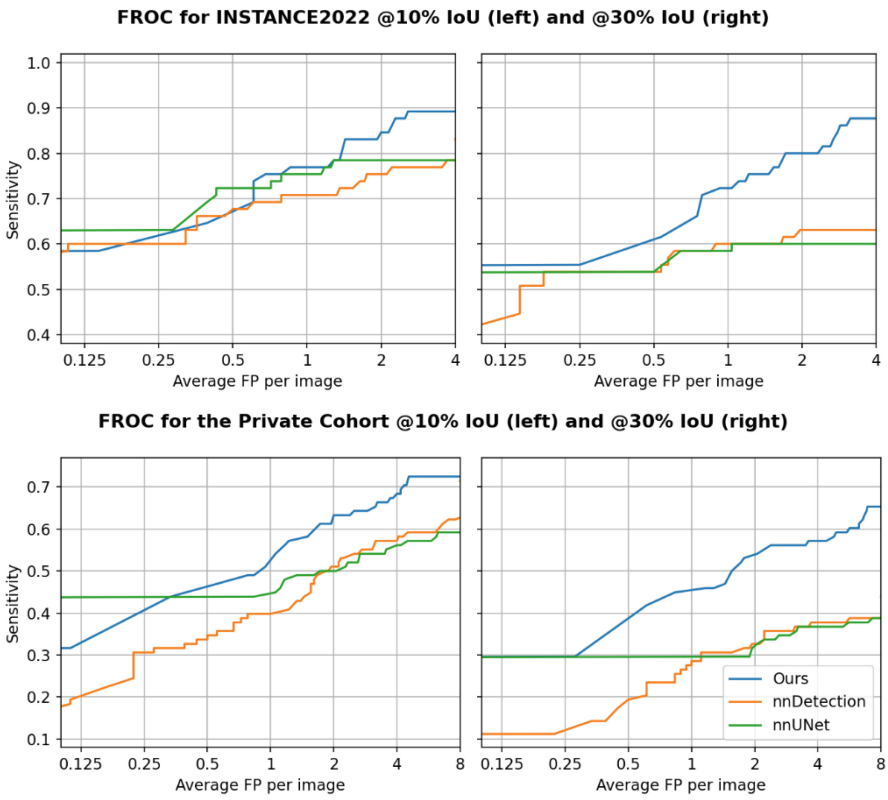


Fig. 6. FROC analysis of our method, nnDetection [1] and [1]nnU-Net for detection [1] on the INSTANCE2022 dataset (top), and our private cohort (bottom) at 10% IoU (left) and 30% IoU (right).

5.2 Ablation: Loss Function

To detail the influence of the additional loss term in our VC-IoU loss, we plug commonly-used losses in our model architecture and compare the results. Table 2 confirms that our VC-IoU allows the network to detect more bleeding more precisely, with a relative increase of 5% Average Recall for both IoU thresholds and for both datasets.

5.3 Ablation: Bleeding Size

This ablation aims to shed light on the drop of performance of our method on the private cohort. In particular, one might recall that the bleeding size distribution is significantly different from the distribution in the training set. As such, we additionally give the performance of our method on each size group (Table 3). We can observe the steepest drop in performance from INSTANCE2022 to the

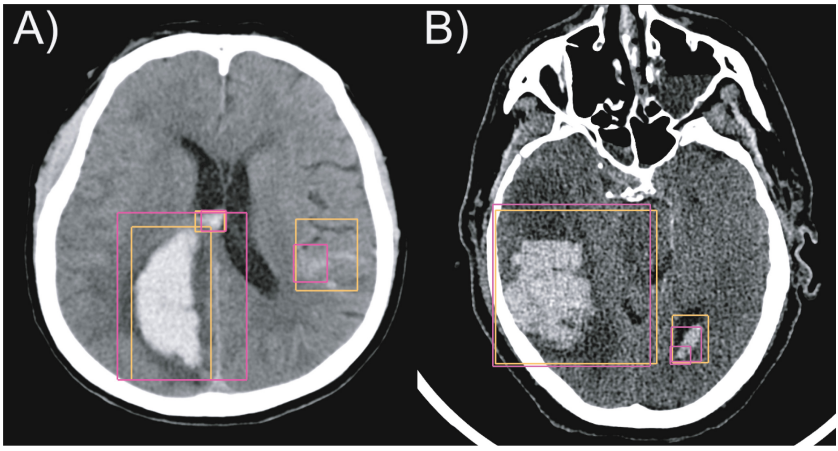


Fig. 7. Detection examples using our method on A) INSTANCE2022 and B) the private cohort. Ground truth boxes are also provided for comparison. In A), the model correctly identifies the three bleeding present. In particular, the smaller center bleeding is in the patient's ventricle system, which can cause serious complications if left untreated [5]. The model also manages to detect the diffuse bleeding to the right. In B), the model again detects relevant bleeding. Please note that ground truth boxes appear larger than the bleeding within this slice, as these expand further in neighboring slices. The bleeding to the bottom right also highlights the challenges of non-maxima suppression, as it gets detected twice but with little overlap between predictions.

Table 2. Ablation: Detection rates at 30% IoU using our method and different loss functions for bounding box regression.

Loss	INSTANCE2022				Private Cohort			
	AR_{10} ↑	AR_{30} ↑	AP_{10} ↑	AP_{30} ↑	AR_{10} ↑	AR_{30} ↑	AP_{10} ↑	AP_{30} ↑
Smooth L1 [6]	0.846	0.754	0.660	0.532	**0.724**	0.531	0.460	0.311
DIoU [17]	0.877	0.815	**0.758**	0.692	0.684	0.622	**0.592**	**0.515**
VC-IoU (Ours)	**0.892**	**0.877**	**0.760**	**0.728**	**0.724**	**0.653**	**0.590**	**0.514**

private cohort for both the smallest and largest bleeding. The former can be expected, as smaller bleeding are harder to detect and are more prevalent in the private cohort. However, the latter phenomenon is due to a strong distribution shift of the cerebral scene. Indeed, the private cohort also contains trauma patients with severe hemorrhages but also fractures or even an open skull. No similar cases are present in the training set.

Table 3. Detection rates at 10% and 30% IoU using our method for each bleeding size group.

Bleeding size (cm^3)	INSTANCE2022				Private Cohort			
	AR_{10} ↑	AR_{30} ↑	AP_{10} ↑	AP_{30} ↑	AR_{10} ↑	AR_{30} ↑	AP_{10} ↑	AP_{30} ↑
all	0.892	0.877	0.760	0.728	0.724	0.653	0.590	0.514
<1	0.824	0.765	0.605	0.585	0.611	0.528	0.521	0.481
[1−10]	0.895	0.895	0.857	0.836	0.788	0.727	0.713	0.637
[10−50]	0.867	0.867	0.784	0.771	0.571	0.571	0.475	0.466
>50	0.929	0.929	0.896	0.865	0.588	0.471	0.497	0.373

5.4 Comparison of Annotation Methods

Annotating 3D medical images is a time intensive task. Tools like 3D Slicer [4] offer multiple solutions, both for bounding box annotation and segmentation. We consider the 4 following methods:

1. Directly placing bounding boxes within 3D Slicer.
2. Segmenting a rough and sparse label map, from which bounding boxes are computed for each segment. This method does not require the complete segmentation of a bleeding, only a bleeding's extremities have to be segmented for the resulting bounding box to be accurate.
3. Precisely segmenting the bleeding.
4. Using a Deep Learning model to pre-segment the images and manually correcting the resulting masks.

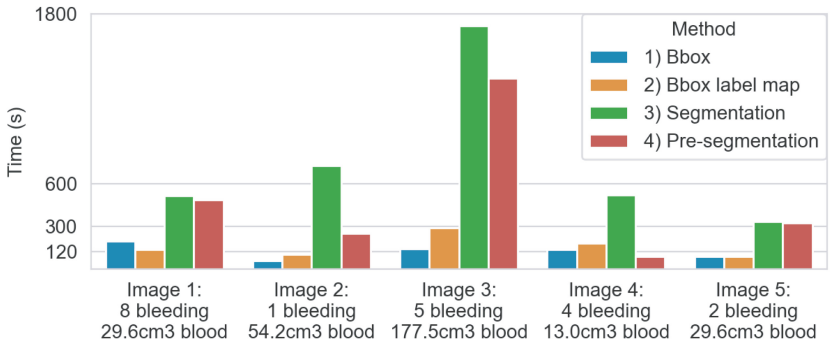


Fig. 8. Annotation effort in seconds to annotate 5 downsampled images from our private cohort using 3D Slicer with 4 different methods. 1) Annotating bounding boxes directly 2) Annotating bounding boxes through rough label maps 3) Segmenting from scratch 4) Refining of automatically-generated pre-segmentation.

We evaluate these 4 methods on 5 images of our private cohort with 5mm slice-thickness and use the model described in Sect. 4 for the pre-segmentation.

It is critical to choose a single method before starting the annotation process, as discrepancies can arise, resulting in lower detection rates [11]. As shown in Fig. 8, bounding boxes are 2 to 6 times faster to annotate than segmentation masks. Annotating a label map is slightly slower but more precise, as object bounds are still accurately annotated. Additionally, it provides a clearer overview of which objects have been annotated.

In contrast, segmenting images is always much slower, even for smaller bleeding. However, it is now unrealistic to expect the expert to annotate images from scratch, when Deep Learning models can pre-segment to an extent. The most tedious part of segmenting is making sure that the object boundaries are accurate. As such, the speed-up offered by the pre-segmentation is not determined by the volume of the pre-segmented mask. The pre-segmentation of image 3 is a good example of this. Even though 70% of the bleeding volume was already pre-segmented, many boundaries still had to be adjusted.

5.5 Impact of Noisy Annotation

Annotating images is a task often done after work hours. As such, it is unsurprising when some noise is introduced in the annotated bounding boxes. We evaluate its impact on the detection rate by simulating the following scenarios:

1. Bounding boxes are too small. We randomly shrink all bounding boxes by up to 10%. The resulting boxes have on average 78.5% IoU with their original counterparts.
2. Bounding boxes are too large. We randomly enlarge all bounding boxes by up to 10%. The resulting boxes have on average 82.2% IoU with their original counterparts.
3. Bounding boxes are off center. We randomly move all bounding boxes' center by up to 10% of their size. The resulting boxes have on average 79.9% IoU with their original counterparts.
4. Some bleeding are not annotated, especially smaller ones. We randomly remove 20% of bleeding under 1 cm^3 and 10% of bleeding under 10 cm^3.

Table 4. Detection rates at 10% and 30% IoU using our method under different annotation noise regimes in the training data (bottom, see Sect. 5.5).

Noise	INSTANCE2022				Private Cohort			
	AR_{10} ↑	AR_{30} ↑	AP_{10} ↑	AP_{30} ↑	AR_{10} ↑	AR_{30} ↑	AP_{10} ↑	AP_{30} ↑
None	0.892	0.877	0.760	0.728	0.724	0.653	0.590	0.514
Smaller	0.862	0.815	0.763	0.706	0.735	0.643	0.586	0.463
Larger	0.862	0.785	0.767	0.682	0.704	0.602	0.584	0.464
Moved	0.862	0.831	0.758	0.710	0.724	0.622	0.605	0.500
Missing	0.862	0.831	0.766	0.732	0.643	0.571	0.528	0.453

The results in Table 4 show that even slight annotation errors in bounding box size will hinder the network from the learning precise detections. Additionally, missing bleeding during the annotation process is most detrimental to the robustness of the network. Nevertheless, these models still outperform both nnDetection and nnU-Net [1].

In particular, if we consider which annotation method to use, one can better understand the trade-off of annotating bounding boxes directly. While images can be swiftly annotated, even slight errors can be detrimental to the final model performance. In contrast, annotating rough label maps solves these risks with only little effort overhead.

6 Conclusion

Detecting bleeding of vastly different size, shape, and position in 3D voxel imaging requires a multiscale approach. We introduce an anisotropic-resolution-aware method and a family of anchors that respect the anatomical structure of the brain. Additionally, we propose a novel Voxel-Complete IoU (VC-IoU) loss that encourages the network to learn the 3D aspect ratios of bounding boxes, which we evaluate our method on two datasets for brain hemorrhage detection. We demonstrate that our model significantly outperforms state-of-the-art methods [1] and that our loss yields a relative increase in Average Recall of 5% compared to other loss functions. Our method has the potential to provide a second level of security to the neuro-radiologist when reading the CT scan of a newly admitted multi-trauma patient and to help ensure that no bleeding is missed.

As little data is currently publicly available for 3D object detection, training this method for any new applications would require to annotate new data. As annotation resources are limited in the clinical setting, we evaluate the cost of different annotation methods, as well as the impact of imprecise bounding boxes in the training data on the detection performance. These results can help shed light on whether segmentation or pure object detection is the better approach for new applications. With this work, we pave the way towards automated tools to offer strong decision support.

Compliance with Ethical Standards. This study was performed in line with the principles of the Declaration of Helsinki. The retrospective evaluation of imaging data from the University Medical Center Mainz was approved by the local ethics boards (Project 2021-15948-retrospektiv). Ethical approval was not required, as confirmed by the license attached with the open access data.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H.: nnDetection: a self-configuring method for medical object detection. In: de Bruijne, M., et al. (eds.)

MICCAI 2021. LNCS, vol. 12905, pp. 530–539. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_51
2. Cohn, S.M.: The impending severe shortage of trauma/critical care surgeons in the USA. World J. Surg. **34**, 2355–2358 (2010). https://api.semanticscholar.org/CorpusID:1468536
3. Çorbacıoğlu, ŞK., Aksel, G.: Whole body computed tomography in multi trauma patients: review of the current literature. Turk. J. Emerg. Med. **18**(4), 142–147 (2018)
4. Fedorov, A., et al.: 3D slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging **30**(9), 1323–1341 (2012)
5. Garton, T., et al.: Intraventricular hemorrhage: the role of blood components in secondary injury and hydrocephalus. Transl. Stroke Res. **7**(6), 447–451 (2016)
6. Girshick, R.: Fast R-CNN (2015)
7. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
8. Jaeger, P.F., et al.: Retina U-net: embarrassingly simple exploitation of segmentation supervision for medical object detection. CoRR abs/1811.08661 (2018). http://arxiv.org/abs/1811.08661
9. Li, X., et al.: The state-of-the-art 3D anisotropic intracranial hemorrhage segmentation on non-contrast head CT: the instance challenge (2023)
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2018)
11. Murrugarra-Llerena, J., Kirsten, L., Jung, C.R.: Can we trust bounding box annotations for object detection? In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4812–4821 (2022). https://doi.org/10.1109/CVPRW56347.2022.00528
12. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library (2019)
13. Wagner, D.T., et al.: Artificial intelligence in neuroradiology: a review of current topics and competition challenges. Diagnostics **13** (2023). https://api.semanticscholar.org/CorpusID:260916853
14. Wang, Y.F., Wu, J.S., Mao, Y., Chen, X.C., Zhou, L.F., Zhang, Y.: The optimal time-window for surgical treatment of spontaneous intracerebral hemorrhage: result of prospective randomized controlled trial of 500 cases. Acta Neurochir. Suppl. **105**, 141–145 (2008)
15. Wardlaw, J.M.: The impact of delays in computed tomography of the brain on the accuracy of diagnosis and subsequent management in patients with minor stroke. J. Neurol. Neurosurg. Psychiatry **74**(1), 77–81 (2003)
16. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection (2020)
17. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: faster and better learning for bounding box regression (2019)

Improving the Robustness of Time Series Neural Networks from Adversarial Attacks Using Time Warping

Yoh Yamashita(✉) and Brian Kenji Iwana

Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
yoh.yamashita@human.ait.kyushu-u.ac.jp, iwana@ait.kyushu-u.ac.jp

Abstract. Time series neural networks have been shown to be weak against adversarial attacks. This study aims to enhance the robustness of time series neural networks in order to defend against such attacks. To do so, we introduce a new defense method called a Random Warping Self-Ensemble (RWSE). The RWSE has two main components. First, a novel random time warping layer to add randomness to trained models in order to disrupt the adversarial attack. Second, the use of self-ensembling increases robustness and maintains the accuracy of the network. The proposed RWSE does not require any special or extra training, can be used with most time series neural networks, including already trained ones, and does not require any extra trainable parameters. We demonstrate that the RWSE is effective in helping reduce the effects of four gradient-based adversarial attacks on five time series datasets.

Keywords: Adversarial Attacks · Robust Neural Networks · Time Warping

1 Introduction

Deep neural networks have had many successes in pattern recognition and classification [29], including time series recognition and classification [2,35], forecasting [33], and signal processing [20]. However, recent research has shown that neural networks are weak against adversarial examples [32], including time series [12]. Adversarial examples are patterns that would normally be correctly classified by a trained recognition system but are subjected to adversarial attacks which intentionally cause misclassifications. These adversarial attack algorithms create such adversarial examples by adding adversarial noise or perturbations that are typically hardly perceivable to humans. Since the presence of an attack is difficult to detect, there are security concerns for the use of neural networks in security-sensitive systems [23].

This work was partially supported by MEXT-Japan (Grant No. 23K16949).

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 15–30, 2025.
https://doi.org/10.1007/978-3-031-78341-8_2

In response to adversarial attacks on time series neural networks, various defense mechanisms have been proposed. These algorithms aim to automatically protect neural networks from such attacks, regardless of whether there is knowledge of an attack or not. One prevalent approach involves employing ensemble networks as a defense strategy [15,31]. However, utilizing ensemble networks often demands multiple trained models, which might pose constraints on resource-limited systems and may not be universally applicable across all network architectures. Therefore, methods such as Random Self-Ensemble (RSE) [25] have been proposed to create ensembles but using only one network. To do this, RSE uses noise layers so that the output of the network changes and can be ensembled.

However, most defense methods, including RSE, are proposed for image-based neural networks. Little research exists on adversarial attack and defense algorithms for time series neural networks [12]. Time series and sequence recognition are important due to their wide range of applications, such as signals, biometrics, speech, etc. Accordingly, it is important to ensure the robustness of time series recognition models.

In this paper, we propose a new defense algorithm, specifically suitable for time series called Random Warp Self-Ensemble (RWSE). Given adversarial examples created from adversarial perturbation, the proposed RWSE can add robustness to trained models. Specifically, we introduce a new time warping-based layer that adds randomness to the features by warping the time steps of the internal representation of each input. Not only does this remove some of the effects of the specifically designed adversarial noise due to the randomness, but it also allows for the network to be used with self-ensembling. We show that the proposed RWSE makes temporal neural networks more robust to adversarial noise without sacrificing accuracy on non-attacked data.

Furthermore, the proposed method is only applied in the test step. It does not require specific training and can be used for any trained neural network with a fully connected layer. It also does not increase the number of trainable parameters. Thus, the proposed method can be widely applicable to time series neural networks.

The contributions of this paper are as follows.

- We propose a new warping layer that randomly warps the time steps of the internal representation within a time series neural network. The warping layer is only used during the test time and can be widely used with many neural networks, including already trained networks.
- We use the proposed random warping layer to add randomness for a self-ensemble network. The randomness allows for different outputs to be predicted given the same network with the same weights.
- We demonstrate that the proposed RWSE outperforms similar defense algorithms on four time series datasets. The RWSE is able to protect a temporal Convolutional Neural Network (CNN) from well-known adversarial attacks.
- The code for the proposed method can be found at https://github.com/uchidalab/random-warping-self-ensemble.

2 Related Work

2.1 Adversarial Attacks

Most of the research on adversarial attacks has focused on image recognition. Szegedy et al. [32] demonstrated the vulnerability of well-known image benchmarks to adversarial attacks. They generated adversarial examples by minimizing the perturbation needed to misclassify images. Similarly, DeepFool [27] perturbs the image toward the nearest class of hyperplanes. Other methods include using gradient information, such as the Fast Gradient Sign Method (FGSM) [14], Basic Iterative Method (BIM) [22], and Projected Gradient Descent (PGD) [26]. There are also ensemble attacks such as AutoAttack [9] that use multiple attack methods.

However, there is much less research on adversarial attacks in the field of time series recognition. Carlini et al. [4] showed that it is possible to encode hidden commands in speech recognition systems. Fawaz et al. [12] assess the performance of FGSM and BIM techniques using the time series classification datasets. There has also been work to create adversarial examples in time series classification without neural networks [28].

2.2 Defense Against Attacks

Defense algorithms are designed to enhance the robustness of neural networks against attackers. There are various methods for defending against adversarial attacks, each with differing levels of success [23]. For instance, training with adversarial examples can improve robustness [14]. Another approach is to limit the impact of adversarial perturbations by using defensive techniques such as defense distillation [16], feature squeezing [36], and denoising [24]. Additionally, employing ensembles and modular networks can be effective in avoiding attacks trained for specific gradients [15,25,31,37].

3 Adversarial Attacks

3.1 Threat Model

This paper focuses on *white-box* attacks, or attacks with full knowledge of the trained model. Specifically, given a neural network model $f(\mathbf{x})$ with input $\mathbf{x}$, a white-box attack has access to all information about $f(\mathbf{x})$, including its parameters, gradients, etc.

The objective of the attack is to find an adversarial sample $\mathbf{x}_{adv}$ that is similar to $\mathbf{x}$ but is misclassified by $f(\mathbf{x})$. Furthermore, the similarity between $\mathbf{x}_{adv}$ and $\mathbf{x}$ must be within a budget ϵ, in order to be undetectable by an untrained eye.

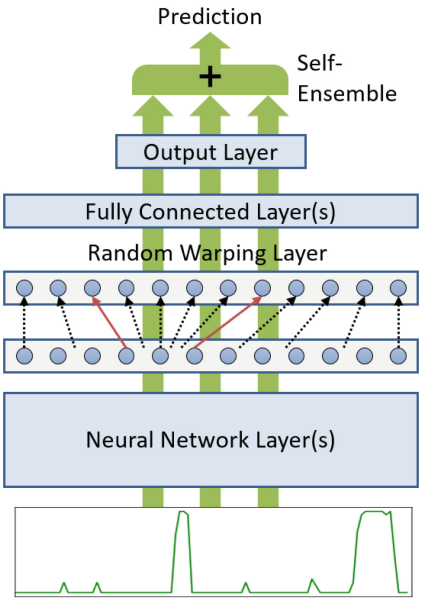


Fig. 1. A diagram of the proposed method. In this example, the input is fed to the network three times, and each time the random warping layer warps the time steps differently. The output of each prediction added to form the self-ensemble. The red arrows in the random warping layer represent the knots. (Color figure online)

3.2 Gradient-Based Adversarial Attacks

A gradient-based attack is a type of adversarial attack on neural networks where the attacker manipulates the model's input data by adding a small amount of carefully crafted adversarial noise. The noise is created by utilizing information about the gradient of the model's loss function.

In our experiments, we use the five most popular gradient-based attacks: FGSM [14], BIM [22], PGD [26], CW [5], and AutoAttack [9]. FGSM perturbs the input data in the direction of the sign of the gradient of the loss function with respect to the input. BIM takes this idea and uses FGSM in an iterative manner. PGD is similar to BIM, but it aims to minimize perceptibility and CW formulates the adversarial noise as an optimization problem between moving the classification and ensuring the perturbed noise is within a constraint. The CW attack is generally respected as one of the most difficult attacks for neural networks [25].

Finally, the last adversarial attack used is AutoAttack. AutoAttack uses an ensemble of four attacks including, two different Auto-PGD methods, Square Attack [1], and Fast Adaptive Boundary (FAB) attack [8].

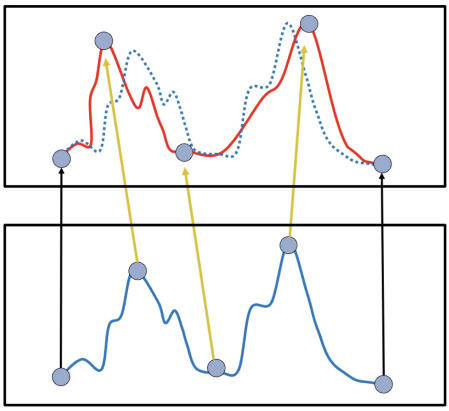


Fig. 2. Processing image of time warp layer.

4 Random Warping Self-Ensemble (RWSE)

The proposed method has two main components. The first is the use of a proposed random warping layer. The random warping layer is a layer that can be included in the structure of a trained neural network. The purpose of the layer is to randomly warp the time steps of the internal representation. Second, we use networks with the random warping layer in a self-ensemble [25]. An overview of the proposed model is shown in Fig. 1.

4.1 Random Warping Layer

Since the adversarial perturbation added to $\mathbf{x}$ is based on the specific gradient of the loss given $\mathbf{x}$, we propose a nonlinear random time warping layer in an attempt to reduce the impact of the adversarial perturbations. The idea is that the adversarial perturbations are specific to the gradients of the network. By warping the internal representation, it is possible to disturb the alignment between the perturbations in the input and the attacked gradients. At the same time, we aim to only warp in the time dimension enough to disturb the alignment without affecting the overall accuracy.

The proposed random warping layer is based on the time warping data augmentation technique [34], but unlike data augmentation, the time warping occurs within the first fully-connected layer of the network. The input features of the random warping layer are treated as a time series and a time warping algorithm is used to warp the time steps based on a smooth curve with I number of knots, or:

$$\mathbf{x}' = x_{\tau(1)}, \ldots, x_{\tau(t)}, \ldots, x_{\tau(T)}, \tag{1}$$

where $\tau(\cdot)$ is a warping function defined by a cubic spline $S(\mathbf{u})$ with knots $\mathbf{u} = u_1, \ldots, u_i, \ldots, u_I$. The magnitude of the knots u_i are taken from $u_i \sim \mathcal{N}(1, \sigma^2)$.

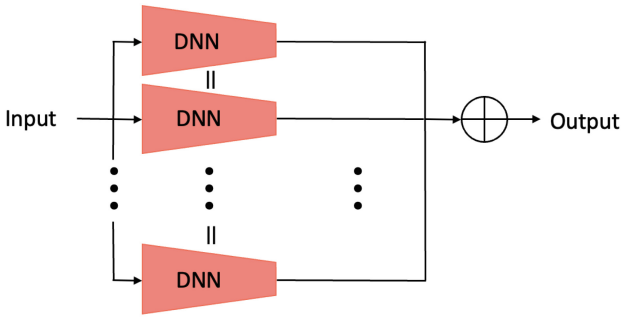


Fig. 3. A diagram of a self-ensemble method. Self-ensemble uses the same input with the same model, and ensembles the results. However, due to a random factor in the model, the output of each model is different for the ensemble.

Using the knots, the time steps $x_{\tau(t)}$ are moved. Then, the rest of the times steps are resampled based on a smooth cubic spline-based curve (Fig. 3).

The result is a warped time series with a smooth transition between the knots as shown in the Fig. 2. This random warping layer is used between the convolution and the flattening before the first fully connected layer of the network. The purpose of the random warping is to disrupt the adversarial perturbations. In gradient-based adversarial attacks, adversarial perturbations are added to the data in the original input space. However, due to the warping within the representation, the gradients specific to the input no longer correspond directly to the input elements. Furthermore, while warping is occurring, the characteristics of the local features of original data remains intact.

During training, any standard temporal neural network can be used and the random warping layer is only used during testing. The neural network can be any architecture as long as the features are represented structurally for the fully connected layer (i.e. the model does not remove the spatial information using Global Average Pooling).

4.2 Self-ensemble

Ensemble methods combine multiple different classifiers to form the prediction. By combining multiple predictions, the robustness can be increased over any of the single classifiers [11]. Furthermore, ensemble methods have been shown to help defend against adversarial attacks [31]. To have a meaningful ensemble, classifiers should be able to produce different predictions from each other. In order to do this, ensemble methods can be constructed using different classifiers or models [21] or models can be constructed using bagging [3].

However, using multiple models increases the computational requirements. Having multiple models means that each model must be trained and the weights

stored. Consequently, self-ensemble methods were proposed. self-ensemble methods are methods that use a single classifier but with added randomness to create distinct predictions. The previous self-ensemble methods use test time Gaussian noise within the networks [25,31].

The random warping layer adds test time randomness which allows for the proposed method to utilize self-ensembling. This is because given the same trained network, different predictions is possible. Specifically, during testing, the same input is fed to the RWSE N number of times. Each time, the random warping layer randomly warps the time steps of the features and a slightly different prediction will be obtained.

Finally, as shown in Algorithm 1, the output of each prediction is summed, or:

$$\hat{y} = \text{argmax}\left(\sum_{n=1}^{N} f_n(\mathbf{x})\right), \tag{2}$$

where $f_n(\cdot)$ is the post-softmax output of the network and $\mathbf{x}$ is the input that may or may not be attacked. It should be noted that we use the addition of the outputs instead of voting like some ensemble methods.

Algorithm 1. Random Warping Self Ensemble

function RWSE($\mathbf{x}$)
  $sum \leftarrow 0$
  for $n = 1, N$ **do**
    $\hat{y}_n \leftarrow \text{DNN}_{\text{RW}}(\mathbf{x})$
    $sum \leftarrow sum + \hat{y}_n$
  end for
  $\hat{y} \leftarrow \text{argmax}(sum)$
  return $\hat{y}$
end function

function $\text{DNN}_{\text{RW}}(\mathbf{x})$
  $\mathbf{z} \leftarrow f(\mathbf{x})$ ▷ Feature Extraction Layers (e.g. Convolutional Layers)
  $\mathbf{u} \leftarrow u_1, \ldots, u_i, \ldots, u_I | u_i \sim \mathcal{N}(1, \sigma^2)$
  $\tau \leftarrow S(\mathbf{u})$ ▷ Cubic Spline
  $\tau' \leftarrow \text{clip}(\tau(1), \ldots, \tau(t), \ldots, \tau(T), 0, T)$
  $\mathbf{z}' \leftarrow z_{\tau'(1)}, \ldots, z_{\tau'(t)}, \ldots, z_{\tau'(T)}$
  $\mathbf{z}' \leftarrow \text{FC}(\mathbf{z}')$ ▷ Fully Connected Layers
  $\hat{y}_n \leftarrow \text{softmax}(\mathbf{z}')$
  return $\hat{y}_n$
end function

5 Experimental Results

5.1 Datasets

To evaluate the proposed method, experiments were conducted on five datasets, ElectricDevices [10], FordA [10], InsectSound [6], RightWhaleCalls [7], and FruitFlies [13]. The datasets were selected as time series datasets that had sufficiently large training sets suitable for neural networks and that represent a range of signals and time series.

Table 1. Comparison of the Datasets

Dataset	Type	Class	Length	Train	Test
ElectricDevices [10]	Device	7	96	8,926	7,711
FordA [10]	Sensor	2	500	3,601	1,320
InsectSound [6]	Audio	10	600	25,000	25,000
RightWhaleCalls [7]	Audio	2	4,000	10,934	1,962
FruitFlies [13]	Audio	3	5,000	17,259	17,259

5.2 Architecture and Settings

In the experiments, a temporal CNN is used as the backbone for defense methods. This temporal CNN consists of four 1D convolutions, each followed by batch normalization [17], rectified linear unit (ReLU) activation, and max pooling. The first block has 64 filters, and the subsequent blocks have 128 filters. Following the convolutional layers, two fully connected layers are used. The first fully connected layer comprises 512 nodes with ReLU activation, while the second serves as the output layer with a number of nodes equal to the number of classes and softmax activation. Between the two fully connected layers, dropout with a probability of 0.5 is applied.

To train all networks, we use the Adam optimizer [19] with an initial learning rate of 0.001. The network is trained for 10,000 iterations with a batch size of 256. A single CNN is trained, and six test sets (original test set, test sets with AutoAttack, FGSM, BIM, PGD, and CW attacks) are used. The trained CNN is used with defense methods only during testing.

5.3 Adversarial Attacks

For the adversarial attack methods, we used AutoAttack, FGSM, BIM, PGD, and CW. As for the hyperparameters, FGSM and BIM have a maximum distortion of $\epsilon = 0.2$, while for BIM and PGD use a step size of $\alpha = 0.05$ with $I = 10$ iterations.

Table 2. Accuracy (%) Without Attacks

Method	ElectricDevices	FordA	InsectSound	RightWhaleCalls	FruitFlies
No Defense	67.2	93.9	76.6	87.0	96.4
Input Noise	65.3	93.9	**77.2**	87.0	96.4
RSE	56.0	93.6	55.6	86.9	93.6
TTA	62.9	91.9	72.9	**87.3**	96.5
Median Filter	45.0	93.9	68.7	73.2	95.9
RW (Proposed)	**67.4**	**95.2**	74.7	85.9	96.1
RWSE (Proposed)	**67.4**	94.9	76.0	85.6	**96.7**

5.4 Defense Methods

To evaluate the proposed method, we compare it to similar methods to defend against adversarial attacks. The following models were used in the evaluation:

- *No Defense*: This trial is the original CNN that is attacked by the adversarial attacks.
- *Input Noise*: This evaluation adds Gaussian noise to the input in an attempt to cancel the adversarial noise [31]. A standard deviation of $\sigma = 0.15$ was used for the noise.
- *Random Self-Ensemble (RSE)*: Adapted from an image-based defense method, RSE [25] adds a noise layer after every convolutional layer in the CNN. For the self-ensemble, five networks are used. For the experiments, a standard deviation of $\sigma = 0.15$ was used.
- *Test Time Augmentation (TTA)*: TTA [30] is an ensemble of networks with each network using a different augmentation method during test time. To match the proposed method and RSE, five networks in the ensemble is used. The inputs of the five networks are the original time series features, jittering, magnitude warping, time warping, and window warping with the parameters suggested by Iwana and Uchida [18].
- *Median Filter*: The median filter is a smoothing filter that removes noise by taking the median of a sliding window. In the experiment, the filter size was set to 3.
- *Random Warping (RW)*: Random Warping uses the proposed random warping layer but no ensemble. This trial demonstrates the benefits of self-ensembling. The time warping path is defined by a smooth cubic spline-based curve with four knots with random magnitudes with a $\sigma = 0.2$.
- *Random Warping Self-Ensemble (RWSE)*: RWSE uses five networks with random warping layers and ensembles the results through addition.

In all of the evaluations, a normal CNN is used for training. The defense methods are only applied in the testing step.

5.5 Results

Accuracy in the no-attack case is shown in Table 2. Ideally, defense methods should not decrease the accuracy of the unmodified test set in the ideal case.

Table 3. Accuracy (%) Under Adversarial Attacks

Method	ElectricDevices	FordA	InsectSound	RightWhaleCalls	FruitFlies
Under FGSM Attack					
No Defense	46.4	59.2	50.0	22.5	66.5
Input Noise	46.5	63.7	41.2	22.9	66.7
RSE	44.7	63.7	41.2	22.9	71.6
TTA	**48.1**	63.9	52.2	44.9	74.1
Median Filter	38.3	60.5	48.7	32.5	68.8
RW (Proposed)	47.8	63.9	52.2	44.9	**76.1**
RWSE (Proposed)	**48.1**	**72.6**	**56.1**	**57.5**	75.9
Under BIM Attack					
No Defense	35.8	50.5	20.3	21.8	57.8
Input Noise	36.0	51.4	21.0	22.4	58.0
RSE	**40.8**	57.5	31.4	21.8	64.7
TTA	38.8	56.0	29.5	38.5	**68.3**
Median Filter	32.4	52.1	**35.2**	27.3	61.1
RW (Proposed)	37.5	62.9	32.4	48.0	67.9
RWSE (Proposed)	37.8	**63.6**	32.7	**50.5**	67.7
Under PGD Attack					
No Defense	33.6	50.5	18.7	22.1	57.0
Input Noise	33.9	51.1	19.4	22.3	57.1
RSE	**39.8**	61.0	30.7	21.9	63.8
TTA	37.9	56.1	28.4	38.2	**67.8**
Median Filter	30.6	52.1	**35.2**	27.4	60.5
RW (Proposed)	35.6	64.3	31.2	49.9	67.4
RWSE (Proposed)	35.7	**64.7**	31.9	**51.4**	67.3
Under CW Attack					
No Defense	33.4	12.2	30.8	26.1	6.94
Input Noise	38.8	30.6	62.0	21.0	59.3
RSE	**46.8**	72.2	42.7	60.1	73.7
TTA	46.0	56.5	58.3	73.1	81.9
Median Filter	39.9	50.2	63.4	55.2	80.8
RW (Proposed)	46.6	80.3	63.6	81.0	86.1
RWSE (Proposed)	46.6	**84.2**	**68.1**	**83.0**	**87.7**
Under AutoAttack					
No Defense	24.0	53.8	12.9	22.4	16.3
Input Noise	25.2	53.6	13.2	22.4	19.9
RSE	26.9	55.5	13.9	21.6	43.4
TTA	25.4	54.6	13.5	22.2	20.2
Median Filter	**27.3**	52.7	15.7	25.0	23.2
RW (Proposed)	24.4	56.7	16.4	**49.5**	50.4
RWSE (Proposed)	24.3	**56.8**	**16.7**	48.5	**50.9**

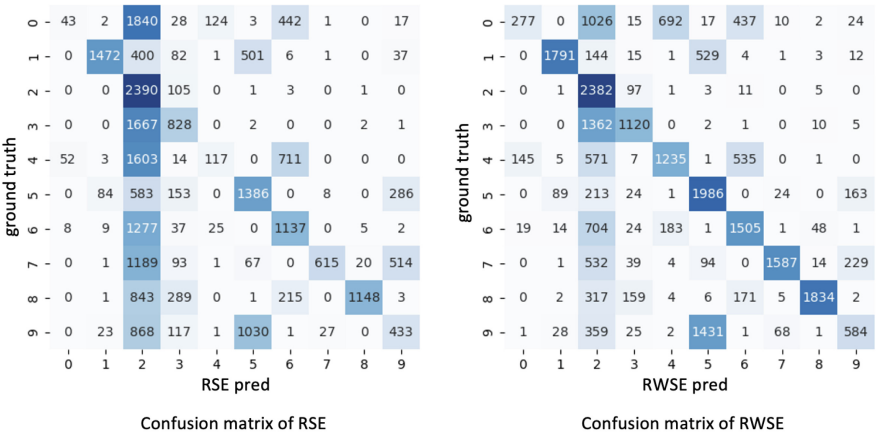


Fig. 4. Confusion matrix of RSE and RWSE for dataset InsectSound under an FGSM attack.

The experimental results show that the accuracy remains comparable to that of the CNN under normal conditions. This is a good result, because in the problem set, it is unknown whether an attack is taking place or not. Thus, it is important to maintain the accuracy on normal data.

Next, we conduct AutoAttack, FGSM, BIM, PGD, and CW attacks, and the results are shown in Table 3. All four sections of the table have the same method, and for these sections, a robust method would close the gap between the accuracy of the CNNs under attack and those not under attack. As a result, for all attacks, all datasets show an improvement in accuracy with the proposed method over the normal case. Overall, the proposed RWSE outperforms the main competitor, RSE, for all datasets in each attack except ElectricDevices. Furthermore, RWSE outperforms RW without the ensemble. This indicates that the self-ensemble method also contributes to the improvement in robustness. Attempts to remove the adversarial noise in the input space directly, such as Input Noise and Median Filter, rarely improved accuracy by significant amounts in normal conditions. These operations proved ineffective against adversarial attacks.

5.6 Discussion

From the results shown in Tables 2 and 3, it can be concluded that the proposed RWSE is effective as a defense method against time-series data, since it can improve the accuracy against attacks while maintaining the original accuracy. In particular, the InsectSound dataset showed a significant improvement in accuracy over the no-protection condition, and was more accurate than the RSE of the competing method. The confusion matrices of RSE and the proposed method RWSE for the FGSM attack are shown in Fig. 4 in the confusion matrix of RSE, most of the model's predictions output Class 2. On the other hand, the RWSE confusion matrix shows a only slight increase in Class 2. This

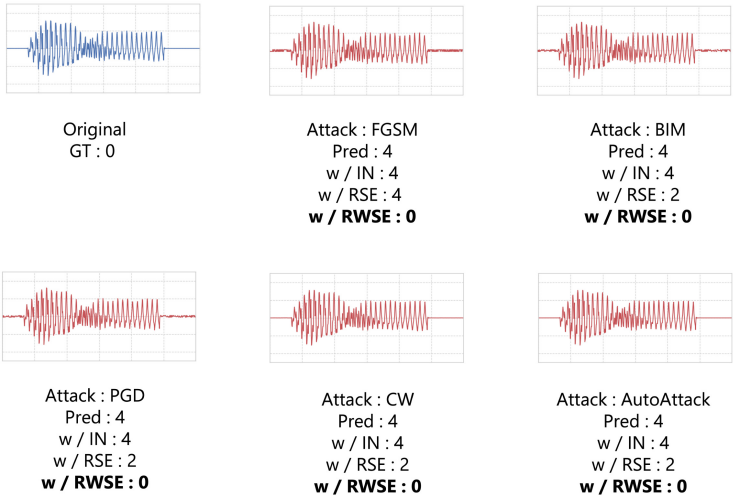


Fig. 5. Example data from InsectSound for each attack and their predictions with robustness measures.

shows one instance that warping instead of noise is better suited at disrupting the adversarial perturbations that would cause the network to overfit to Class 2.

Figure 5 shows the model predictions for each defense method for the attack data. In comparison methods such as Input Noise, RSE, and TTA, the misclassified results are nearly identical to the output of the no attack condition. This indicates that processing in the input space alone leaves residual effects of the attack. The importance of the warp layer inserted before the FC layer can be seen in the reduction of the effect of the perturbation caused by the attack.

Next, in order to see the difference between the proposed RWSE and the main competitor, RSE, we visualized the results by dividing them into those that could be correctly classified only by RSE, those that could be correctly classified only by RWSE, those that could be correctly classified by both, and those that could not be correctly classified by either, as shown in Fig. 6. This figure shows that the data that neither RSE nor RWSE could protect, were generally shorter regions of interest in length. On the other hand, the cases that had longer region of interests were correctly protected by the defenses. The data that could be correctly classified using only RSE tended to be relatively short in length, and the data that could be correctly classified using only RWSE included data of various lengths. This indicates that while RWSE can handle a large range of region of interests, the RSE model was able to handle smaller patterns slightly better.

5.7 Limitations

The proposed method did not work well with ElectricDevices. For ElectricDevices, the main competitor, RSE, was more accurate than RWSE for all attacks

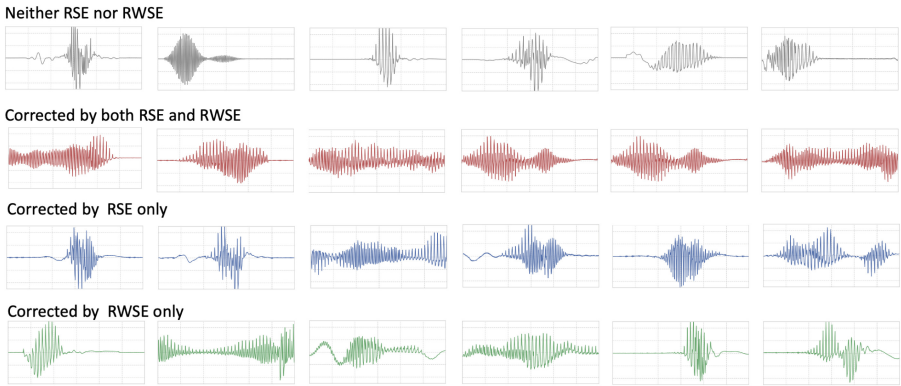


Fig. 6. Examples of data from InsectSound where models have been improved by RSE or RWSE under a CW attack.

except the FGSM attack. ElectricDevices has unusually noisy Class 2 data. This perturbation would have caused most of the data to be classified as class 2, which would have prevented the proposed method from destroying the perturbation. Another factor is the length of the data. As shown in Table 1, the ElectricDevices data set is made of time series that are shorter than the other data sets. In this experiment, all parameters in the warp layer were the same. Therefore, the same warping process used for the short data set as for the other long data sets may have resulted in a loss of original characteristics.

Adversarial attacks add noise to change the class. Therefore, the ensemble may have contributed to erroneous output as a result of a large number of data distributed near the class boundaries. By ensembling, a more generalized prediction is possible.

5.8 Ablation

In this section, we examine the effects that the parameters have on the proposed method. Specifically, we conducted experiments on the relationship between the number of networks and accuracy in self-ensemble, and on the parameters of random warping used in the proposed method.

First, we conducted experiments on the parameter σ in the time warping process. This parameter corresponds to the degree of warp movement in the time warp process. In other words, if σ is small, it is equivalent to warping to data that is close to the adversarial data, so no improvement in accuracy can be expected. Figure 7 shows the results of the experiment. For all datasets, the lowest accuracy was observed for small values of σ. After an upward trend, the accuracy decreased again for larger values of σ. Each attack had the highest accuracy when $\sigma = 0.2$ for all data sets. This indicates that $\sigma = 0.2$ is the best for the time warp used in the proposed RWSE method, regardless of the attack method.

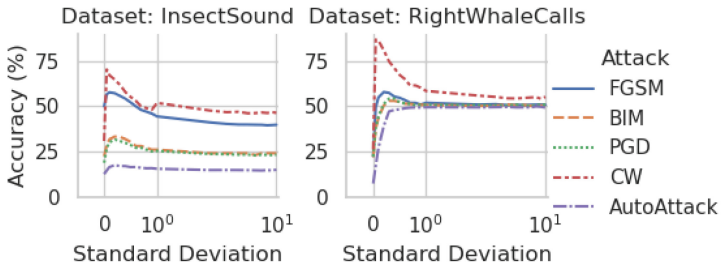


Fig. 7. Effect of the amount of time warping.

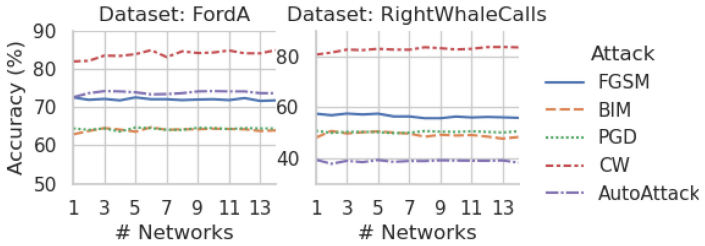


Fig. 8. Accuracy of the ensemble with different number of networks.

Experiments were then conducted on the relationship between the number of networks used in the RWSE self-ensemble and accuracy. Figure 8 shows the results. The number of networks corresponds to RW when the number of networks is 1. Figure 8 shows that for each attack, the number of networks tends to increase slightly up to 5, while it does not change much for larger numbers. Although the ensemble method increased accuracy, the improvement in accuracy was small. This is because the ensemble method increases the confidence of ambiguous data, such as those at class boundaries, by majority voting. The ensemble did not function strongly because data distributed near the boundary of another class in the feature space due to adversarial attacks were pulled back to the original class domain due to the destruction of adversarial perturbations by the time warp.

6 Conclusion

In this paper, we proposed a method to improve the robustness of time series neural networks against adversarial perturbations, called a Random Warping Self-Ensemble (RWSE). The RWSE uses a self-ensemble of networks with a random warping layer. It does not require training models or parameters and can be easily incorporated into any already trained time series neural network that

contains a fully connected layer. To evaluate the proposed RWSE, we compared it to other test-time defenses against five adversarial attacks, AutoAttack, FGSM, BIM, PGD, and CW on five time series datasets.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12368, pp. 484–501. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58592-1_29
2. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
3. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
4. Carlini, N., et al.: Hidden voice commands. In: USENIX, pp. 513–530 (2016)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE (2017)
6. Chen, Y., Why, A., Batista, G., Mafra-Neto, A., Keogh, E.: Flying insect classification with inexpensive sensors. J. Insect Behav. **27**(5), 657–677 (2014). https://doi.org/10.1007/s10905-014-9454-4
7. Cornell Research Foundation, Inc.: The marinexplore and cornell university whale detection challenge (2022). https://www.kaggle.com/competitions/whale-detection-challenge/data
8. Croce, F., Hein, M.: Minimally distorted adversarial examples with a fast adaptive boundary attack. In: ICLR, pp. 2196–2205 (2020)
9. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML, pp. 2206–2216 (2020)
10. Dau, H.A., et al.: Hexagon-ML: the UCR time series classification archive (2018). https://www.cs.ucr.edu/~eamonn/time_series_data_2018/
11. Dietterich, T.G.: Ensemble methods in machine learning. In: IWMCS, pp. 1–15 (2000)
12. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Adversarial attacks on deep neural networks for time series classification. In: IJCNN (2019). https://doi.org/10.1109/ijcnn.2019.8851936
13. Flynn, M.: Classifying dangerous species of mosquito using machine learning. Ph.D. thesis, University of East Anglia (2022)
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
15. Guo, C., Rana, M., Cisse, M., Van Der Maaten, L.: Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117 (2017)
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)
17. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML, pp. 448–456 (2015)
18. Iwana, B.K., Uchida, S.: An empirical survey of data augmentation for time series classification with neural networks. PLoS ONE (2021). https://doi.org/10.1371/journal.pone.0254841
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

20. Kiranyaz, S., Ince, T., Abdeljaber, O., Avci, O., Gabbouj, M.: 1-D convolutional neural networks for signal processing applications. In: ICASSP, pp. 8360–8364 (2019)
21. Kolen, J., Pollack, J.: Back propagation is sensitive to initial conditions. In: NeurIPS, vol. 3 (1990)
22. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016)
23. Liang, H., He, E., Zhao, Y., Jia, Z., Li, H.: Adversarial attack and defense: a survey. Electronics **11**(8), 1283 (2022). https://doi.org/10.3390/electronics11081283
24. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: CVPR, pp. 1778–1787 (2018)
25. Liu, X., Cheng, M., Zhang, H., Hsieh, C.J.: Towards robust neural networks via random self-ensemble. In: ECCV, pp. 381–397 (2018). https://doi.org/10.1007/978-3-030-01234-2_23
26. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
27. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: CVPR (2016). https://doi.org/10.1109/cvpr.2016.282
28. Oregi, I., Ser, J.D., Perez, A., Lozano, J.A.: Adversarial sample crafting for time series classification with elastic similarity measures. In: Del Ser, J., Osaba, E., Bilbao, M., Sanchez-Medina, J., Vecchio, M., Yang, X.S. (eds.) IDC 2018. SCI, vol. 798, pp. 26–39. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99626-4_3
29. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
30. Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J.: When and why test-time augmentation works. arXiv preprint arXiv:2011.11156 (2020)
31. Strauss, T., Hanselmann, M., Junginger, A., Ulmer, H.: Ensemble methods as a defense to adversarial perturbations against deep neural networks. arXiv preprint arXiv:1709.03423 (2017)
32. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
33. Torres, J.F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., Troncoso, A.: Deep learning for time series forecasting: a survey. Big Data **9**(1), 3–21 (2021). https://doi.org/10.1089/big.2020.0159
34. Um, T.T., et al.: Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks. In: ACM ICMI, pp. 216–220 (2017). https://doi.org/10.1145/3136755.3136817
35. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: a strong baseline. In: IJCNN, pp. 1578–1585 (2017). https://doi.org/10.1109/ijcnn.2017.7966039
36. Xu, W., Evans, D., Qi, Y.: Feature squeezing: detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017)
37. Yu, Y., Yu, P., Li, W.: AuxBlocks: defense adversarial examples via auxiliary blocks. In: IJCNN, pp. 1–8 (2019)

SIGN-Diffusion: Generating User Specific Online Signature for Digital Verification

Anurag Pandey[1(✉)], Pushap Deep Singh[1], Arnav Bhavsar[1], Aditya Nigam[1], Divya Acharya[2], and Basu Verma[2]

[1] Indian Institute of Technology, Mandi, Mandi, HP, India
{erpd1903,erpd2201}@students.iitmandi.ac.in,
{arnav,aditya}@iitmandi.ac.in
[2] HCLTech, Noida, India
divya.acharya@hcl.com, basu.verma@hcltech.com

Abstract. Online signature analysis plays a vital role in today's digital landscape, where the number of digital transactions and the need for identity verification are constantly increasing. The signature encompasses various aspects of an individual's unique characteristics, including both structural and behavioural elements. Researchers have been intrigued by the intricacies surrounding signature verification and generation for quite some time. The main objective of this study is to present a fresh approach for creating online signatures that improves security in digital transactions and signature verification systems. This is accomplished by utilising the proposed Sign-Diffusion framework. Traditional approaches to signature recognition often lack the required resilience, as they are trained on datasets that include fraudulent attempts to replicate the original signature. To overcome the challenges and limitations generated because of manual mimicking capability, we propose a solution that has conditional diffusion as a building block aided with a state space model to capture long-term structural forecasting for online signature generation. This method tackles the task of extracting detailed features that capture the intricate spatial and temporal characteristics of signature dynamics. It also ensures adaptability to various signing styles. The approach we have developed introduces a foundational model that is capable of generating near-equal user dependent online signatures. This will also help to identify deepfakes in the area of online signature generation. The above mentioned facts highlights the prospects for further exploration in the area of online signature verification methods, specifically for detection of system-generated forgeries that are more advanced and sophisticated than manually created forgeries.

Keywords: Diffusion · Biometric · GenAI · Identity Verification · Signature

1 Introduction

The emergence of online signature generation has become a crucial technology in the ever-expanding digital era, bringing about substantial implications.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 31–45, 2025.
https://doi.org/10.1007/978-3-031-78341-8_3

Generating online signatures involves the creation of electronic versions of handwritten signatures. This approach is widely used for various purposes, including document authentication, digital agreements, and identity verification [11]. This technology holds immense importance, as it has the potential to enhance security and optimize various aspects of our digital existence. It guarantees the integrity of digital documents, safeguarding against unauthorized modifications or counterfeits. Online signature generation enhances security by integrating with identity verification systems. This ensures that the signer's identity is verified and compared, reducing the risk of fraudulent activities. The growing prominence of online transactions, remote work, and e-commerce has underscored the need for a reliable and secure method of verifying identities and authorizing documents. Furthermore, this technology not only offers convenience, but also finds relevance in crucial sectors like finance, legal affairs, and healthcare [3]. By implementing a robust online signature generation system, the risk of fraudulent activities can be significantly reduced. This strengthens forgery detection systems and reduces the need for traditional paper-based procedures, ultimately accelerating the pace of digital transformation.

The process of online signature involves capturing structural information in (x, y) coordinates, along with behavioral information like the pressure (p) applied to each coordinate and the time elapsed (τ) during the signing process. Existing online signature verification systems heavily rely on datasets that contain both genuine signatures and forgeries, which are created by individuals with varying levels of expertise in i mitating genuine signatures. The current system has certain limitations when it comes to detecting manual forgeries and is not very effective in identifying system-generated forgeries. It is crucial to create a system that can detect both manual and system-generated forgeries. In order to accomplish this, it is crucial to possess a system that is capable of generating genuine online signature data.

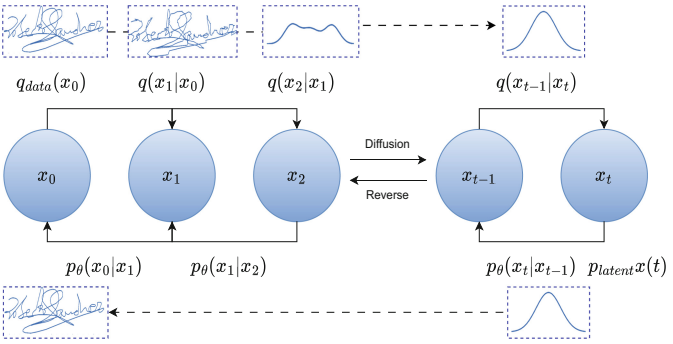


Fig. 1. Diffusion process for generating structural online signature data as x and y coordinates.

This research work aims to introduce a new framework called Sign-Diffusion, which utilizes a conditional diffusion model to generate online signatures for

structural forecasting. The implementation of diffusion and reverse process on input data i.e., x and y coordinates is illustrated in Fig. 1, providing a comprehensive flow of online signature generation. The system can generate structural data, which is represented by coordinates x and y. Sign-Diffusion explores the complete range of signature dynamics by utilizing conditional diffusion.

The length of online signature data can vary significantly, as it is influenced by the individual's preference for shorter or longer signatures. In order to tackle this issue, we have implemented an approach called conditional diffusion, as proposed by Kong et al. [14]. Additionally, we have incorporated a state space model (SSM), as suggested by Gu et al. [5], to effectively handle long temporal dependencies. In this paper, we introduce the Sign-Diffusion framework, which is designed for online signature generation. Through our experiments, we showcase the remarkable performance of the proposed framework in comparison to two other models: Sign-Diffusion-1, which includes one transformer and one SSM block, and Sign-Diffusion-2, which consists of two transformer blocks without an SSM block. The results of our study demonstrate that Sign-Diffusion, when combined with SSM, surpasses other models in terms of performance. This emphasizes the effectiveness and reliability of Sign-Diffusion in generating online signatures. In order to assess the system's robustness, we have conducted a thorough analysis of the mean square error between the original signatures and both manually forged signatures and system-generated forged signatures. Our findings indicate that the system-generated signatures closely resemble the original signatures.

Our network has been trained using publicly available datasets, including MCYT [19], e-BioSign(EBDS1) [22], BioSecure-DS2 (EBDS2) [20], Biosecure-ID [2]. To evaluate its performance, we have calculated various similarity and distance-based metrics. The model is trained using a limited number of original signatures from the datasets, and its accuracy is evaluated by providing initial prompts of different lengths to estimate its generation accuracy. Thus, the system exemplifies impressive capabilities in few shot learning.

2 Related Work

Through an extensive review of the literature, the goal is to gain a thorough understanding of the methodologies that influence the generation of online signatures. In the arena of handwriting generation, the pioneering work was carried out by Graves et al. [4]. They employed a Recurrent Neural Network (RNN) [21] to synthesize online handwriting, marking a significant milestone in this area. On the contrary, when it comes to generating offline handwriting, the emphasis has been given to employing a generative adversarial network (GAN) [1]. RNN-based models, such as Variational RNN, have been utilized by Kotani et al. [15] in their work. These models, incorporating a Variational Autoencoder [12], are known for their simplicity in training and sampling. However, a major challenge they encounter is the necessity for the writer-style data to be in an online format. Recording the user's pen strokes during sampling presents a challenge when it comes to incorporating stylistic information. In their respective

studies, Kang et al. [10] utilize the GAN framework to enhance the quality of generated samples compared to the work done in the literature. Moreover, both studies demonstrate the ability to condition the generation process on specific writer stylistic features. In their study, Lian et al. [16] explore the application of Chinese fonts to create a wide range of handwriting styles using only a small number of samples.

This study presents a new approach called Sign-Diffusion, which aims to generate online signature data. The approach captures both the structural x and y coordinates, as well as the behavioral pen tip pressure data, using a specific framework.

To the best of our knowledge following are the contribution of this research work:

- To best of authors knowledge this is the first attempt to create a foundational framework to generate online signatures using conditional diffusion aided with state space model.
- We validated the generation using publicly available online signature datasets for various similarity measures with original signature and found generated signatures are near equal to original signatures.

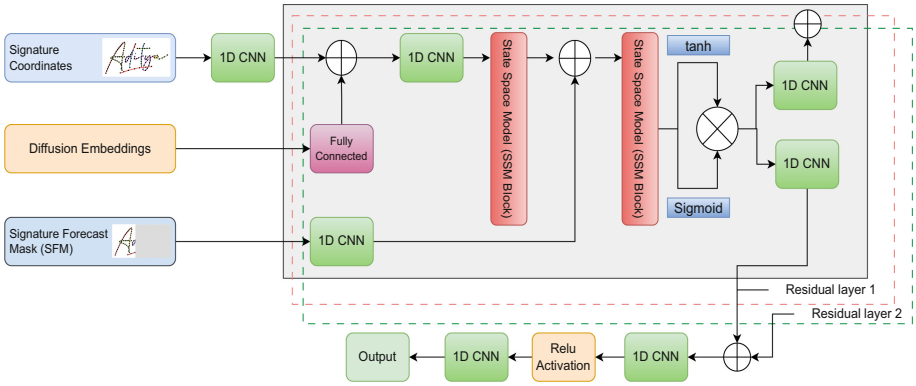


Fig. 2. Sign-Diffusion architecture is proposed for generating online signature data, specifically focusing on the x and y coordinates. Each coordinate is trained independently using mean square error loss.

3 Proposed Methodology

In this method, a conditional diffusion model aided with SSM is used to cater the long temporal dependencies present in online signature dataset as shown in Fig. 2. Unlike other approaches of 1D Convolution [13] and bidirectional dilated convolution [9], which may struggle in capturing long-range dependencies effectively, the SSM offers a more nuanced approach for handling the long temporal dependencies.

The Sign-Diffusion framework further enhances the temporal modeling capabilities of the system. It improves the generation process by allowing the model

to iteratively refine its predictions based on the context of the entire signature sequence. This iterative refinement process is crucial for generating realistic and coherent signatures, especially when dealing with complex and varied signature patterns. Additionally, Sign-Diffusion provides a probabilistic framework for generating signatures, allowing for uncertainty estimation in the generated samples. This is particularly useful in online signature generation, where the model needs to produce diverse and natural-looking signatures while maintaining the individuality of the user's style.

In this approach we have taken signature data from signature time-series captures on e-pads. Coordinates $x(t)$ and $y(t)$ are utilized after removing the redundant information weighted by $p(t)$. Here zero/non-zero pressure value indicates that the pen tip is up/down from the e-pad. After passing the signature coordinates $x(t)$ and $y(t)$ through the 1D convolution, they are concatenated with diffusion embeddings and subsequently a signature forecast mask (SFM) as a condition to be executed through SSM. Here SFM refrains the system to look forward into the future. This mask can have forward masking, backward masking, and imputation masking thus, making the system work for forward forecasting, backward forecasting, and imputation generation. The network is trained using the mean square error loss function to get the best-desired outcomes.

By combining the strengths of the conditional diffusion model and SSM mechanism, the proposed approach achieves prominent performance in online signature generation tasks. The ability of conditional diffusion model to cater complex temporal dependencies, coupled with the SSM's mechanism, enables the system to generate high-quality signatures that closely resemble the input data, regardless of the signature length or complexity.

Table 1. Datasets details used for training the proposed SignGPT network

Dataset	MCYT	EBDS1	EBDS2	BiosecureID
User Count	231	30	46	268
Original Sample Per User	25	40	8	16

3.1 Dataset Preparation

This approach utilizes four online signature datasets: MCYT [19], EBDS1 [22], EBDS2 [20], and Biosecure-ID [2]. The specific information regarding these datasets can be found in Table 1. The data captured in online signature records includes the (x, y) coordinates, the pressure at each coordinate, and the corresponding time of capture. During the data preparation process, any coordinates with zero pressure values are removed. This is because zero pressure values indicate that the pen tip is not in contact with the e-pad, resulting in the absence of a signature stroke. In addition, the e-pads have a high sampling frequency of 200 Hz, which can lead to repetitive coordinate capture. However, this issue has been addressed to minimize redundancy in the signature data.

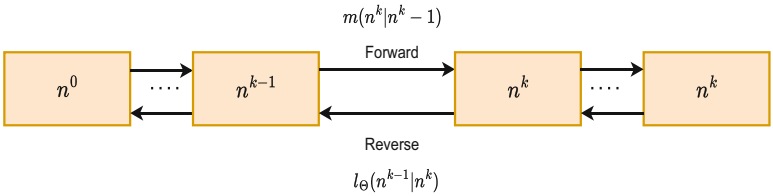


Fig. 3. An illustration of forward reverse diffusion probabilistic model.

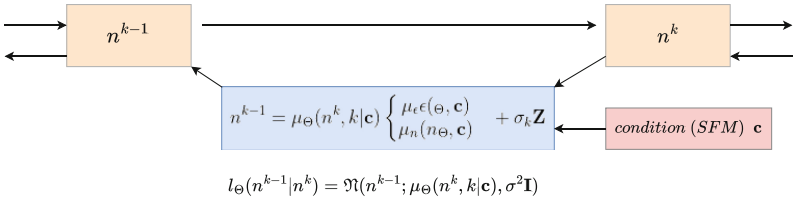


Fig. 4. An illustration of forward reverse conditional (SFM) diffusion probabilistic model

3.2 Conditional Diffusion

Generative models known as diffusion models have demonstrated exceptional performance across various modalities such as images, videos, and speech [23]. These models aim to optimize their performance by operating on the latent space of the signal and effectively eliminating any unwanted noise during the backward process. The noise is incorporated into the diffusion process in a Markovian manner, as described by Gurvich et al. [7]. The process described is independent of any conditions and can be represented by the following Eq. 1:

$$m(n_1, ...n_T|n_0) = \prod_{t=1}^{T} m(n_t|n(t-1)) \tag{1}$$

where $m(n_t|n_{t-1} = \mathfrak{N}(\sqrt{1-\beta_t}n_{t-1}, \beta_t 1)[n_t]$ and the (fixed or learnable) forward-process variances β_t adjust the noise level. Equivalently, n_t can be expressed in closed form as $n_t = \sqrt{\alpha_t}n_0 + (1-\alpha_t)\epsilon$ for $\epsilon \sim \mathfrak{N}(0,1)$ where $\alpha_t = \sum_{i=1}^{t}(1-\beta_t)$. The backward process is parameterized as Eq. 2:

$$l_\Theta(n_0, ...n_{t-1}|n_T) = l(n_T)\prod_{t=1}^{T} l_\Theta(n_{t-1}|n_t) \tag{2}$$

where n_t $\mathfrak{N}(0,1)$. Again, $l_{Theta}(n_{t-1}|nt)$ is assumed as normal-distributed (with diagonal covariance matrix) with learnable parameters. In the context of signature generation x and y coordinates. In addition to the aforementioned process, the backward process incorporates additional condition information, resulting in a conditional diffusion model [18]. Figure 3, shows the illustration of forward

reverse diffusion probabilistic model flow. Figure 4, describes forward reverse conditional (SFM) diffusion probabilistic model flow.

Our approach involves incorporating SFM into the concatenation of input data, effectively preventing the model from accessing future information. This mask has the capability to generate signatures in both directions, making it bidirectional. These can also fill in the gaps when added to the signature, creating similar strokes in the process.

3.3 Signature Forecast Mask (SFM)

Masking can be defined as Random Missing (RM), Random Batch missing (RBM), and Forecasting (FM). These masks when included with input data are treated as conditions while generating the signature sequences hence making generation a conditional diffusion process. In the context of online signature generation, we used FM known as Signature Forecast Mask (SFM). This plays a crucial role in determining which parts of the sequence to generate. The input signature is first processed to create the SFM, which involves modifying the signature with Gaussian noise to introduce variability. This modified signature is then multiplied with the generated mask, resulting in the SFM. During the generation process, the SFM is used as a binary mask, where ones indicate values to condition on for forecasting, and zeros indicate values to forecast. By focusing the generation process on relevant parts of the sequence, the SFM helps enhance the model's ability to generate realistic and accurate online signatures.

3.4 State Space Model (SSM)

In contrast to the use of bidirectional directional convolution [5], we have introduced SSMs to address the long-range dependencies present in signature data. Signatures, as a form of behavioral biometrics, can consist of a few hundred to several thousand points. In essence, SSMs can be described as linear state space transition equations that map a 1D input sequence $a(t)$ to an n-dimensional output sequence $c(t)$ before projecting it to a 1D output sequence $b(t)$. This mapping is represented by Eq. 3 and Eq. 4:

$$c'(t) = A\,c(t) + B\,a(t) \tag{3}$$

$$b(t) = C\,c(t) + D\,a(t) \tag{4}$$

In this paper, we present the representation of transition matrices obtained through gradient descent as A, B, C, and D. In this approach SSM is used as a black box in a discrete convolution manner as shown in Eq. 5, 6 and 7. Its purpose is to capture long-term dependencies, in line with the HiPPO theory [6] for online signature generation.

$$c_k = \overline{CA}^k \overline{B} a_0 + \overline{CA}^{k-1} \overline{B} a_1 + \cdots + \overline{CAB} a_{k-1} + \overline{CB} a_k \tag{5}$$

Table 2. Sign-Diffusion Hyperparameters

Hyperparameter	Value	Hyperparameter	Value
Diffusion embedding dim. 1, 2, 3	128, 512, 512	Training Epochs	10000
Diffusion steps T	200	Loss Function	MSE
$\beta 0, \beta$ 1	0.0001, 0.02	Learning Rate	0.0002
S4 Dropout	0.01	Batch Size	8
S4-Lmax	180	Optimizer	Adam

$$a = K * a \tag{6}$$

$$K \in \mathbb{R}^L \tag{7}$$

In our online signature generation approach, the initial step involves augmenting the input signature coordinate data with noise. This noisy data is then fed into the diffusion model, where it undergoes processing in each residual layer. Within each residual layer, the input is combined with the condition mask and diffusion hyperparameters. These layers consist of a 1D convolutional layer followed by a State Space Model (SSM) layer.

The SSM is a key component that maps the input signature sequence to an n-dimensional output sequence, as described above, capturing the intricate patterns and dynamics of the signature data. This mapping process involves bidirectional layers with layer normalization, no dropout, and an internal state with a gated linear unit in each layer. After the diffusion embedding is added, a convolutional layer is applied to double the channel dimension of the input, preparing it for the subsequent SSM block computation. Notably, a single three-layer diffusion embedding is utilized in all residual layers. The diffusion process is applied selectively to the signature coordinate sequence that is intended to be generated, ensuring that the model focuses on relevant parts of the sequence for accurate generation.

By incorporating these components into our approach, we are able to generate online signatures that closely resemble real-world signatures.

Table 3. Online Signature Data Generation Result Across Different Metrics for Different Initial Coordinate Lengths

Metrics	MSE		ED		Cosine	
Initial Points	50	100	50	100	50	100
MCYT	X: 2.438 Y: 4.837	X: 2.124 Y: 4.256	0.9854	0.9763	0.9926	0.9996
EBDS1	X: 2.638 Y: 4.168	X: 2.467 Y: 4.013	1.134	1.081	0.993	0.9992
EBDS2	X: 2.534 Y: 4.790	X: 2.345 Y: 4.621	1.142	1.098	0.9995	0.9994
Biosecure-ID	X: 2.376 Y: 4.721	X: 2.267 Y: 4.653	0.9938	0.9728	0.9997	0.9998

3.5 Evaluation Strategy

Understanding the quality of the generated signatures is crucial in justifying the model's decision. It is crucial that the generated signatures are of high quality and closely resemble the original signatures. This aligns with the concept of developing a stronger system in the future to detect forgeries generated by the system. In order to accomplish this, the generated signature data was compared with the manual forgery data available in different datasets using similarity-based metrics.

Original vs Generated: The evaluation of the model generation performance involved the utilization of Cosine and Euclidean as similarity metrics. In addition, we have computed the mean square error (MSE) between the original and generated signature.

Original vs Manual Forgery: Our evaluation of the signature generation process involved comparing it to manual forgery data found in datasets. We found that the signature generation was highly realistic and closely resembled the original signature. In order to highlight the model decision process, we have conducted calculations on the Euclidean Distance and Cosine similarity.

The computed values are standardized based on the signing area resolution of the e-pad for both the x and y coordinates.

Evaluation Metrics. In this section, we will discuss the evaluation metrics that were employed to assess the effectiveness of the proposed approach in handling online signature generation.

Mean Squared Error (MSE). Mean Squared Error [8] quantifies the average squared difference between the original signature points and the generated signatures. The formula for MSE between two sets of values, Y and $\hat{Y}$, each comprising n observations, is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{8}$$

where Y_i and $\hat{Y}_i$ denote the i-th observations in the original and generated signature sets respectively.

Fréchet Inception Distance (FID). FID measures similarity between two data distribution sets, like original and generated signature points, using:

$$FID = \|\mu_1 - \mu_2\|^2 + \text{Tr}(\Sigma_1 + \Sigma_2 - 2\sqrt{\Sigma_1\Sigma_2}) \tag{9}$$

Lower FID values suggest higher similarity.

Euclidean Distance (ED). ED computes the straight-line distance between two points in Euclidean space, such as original and generated signature points, using:

$$\text{ED} = \sqrt{\sum_{i=1}^{n}(P_{1i} - P_{2i})^2} \tag{10}$$

where P_{1i} and P_{2i} are the coordinates of points P_1 and P_2 along the i-th dimension respectively.

Cosine Similarity (CS). CS measures similarity between two vectors, like original and generated signature points, using:

$$\text{CS} = \frac{A \cdot B}{\|A\| \times \|B\|} \tag{11}$$

Normalized values consider the e-pad's signing area resolution for x and y coordinates and the maximum pressure value for p.

4 Empirical Evaluation

In this section, we will discuss the experimental setup and provide a comprehensive analysis of the benchmarks using publicly available datasets. These datasets include MCYT [19], EBDS1 [22], EBDS2 [20], Biosecure-ID [2] dataset. The following sections will cover the training and testing protocols, as well as the results obtained.

4.1 Experiment Protocol

We collected five unique signatures from each user in the MCYT, EBDS1, EBDS2, and Biosecure-ID datasets to train the Sign-Diffusion model. The dataset addresses the issue of variability between sessions by including data from multiple sessions conducted at different times to capture the unique characteristics of each user. During the system training process, the training set is divided into two parts: 80% for training and 20% for validation. The model undergoes extensive training over a significant number of epochs, employing the mean square error loss function. The hyperparameters are finalized after a series of experiments as shown in Table 2. We trained the model from scratch using random weight initialization and these hyperparameters.

The model was developed in the Python (3.9) environment and experimented on a state-of-the-art Nvidia DGX-1 supercomputer. This supercomputer features the advanced NVIDIA Volta architecture and cuDNN, a GPU-accelerated library tailored for deep neural networks. This makes it easier to deploy deep learning models whose complexity is increasing at a rapid pace and to provide solutions in a fair amount of time. The signatures are generated by giving different initial coordinate values as 50 and 100.

4.2 Experimental Results

This section presents the evaluation results obtained across various metrics for analyzing the generated online signatures in comparison to the original signatures.

Figures 5, 6, 7, 8 depicts the generation of x, y coordinates w.r.t original x, y coordinates of signature datasets. In all of the above-mentioned Fig. 5, 6, 7, 8 orange stroke shows initial points given to the model and blue strokes is

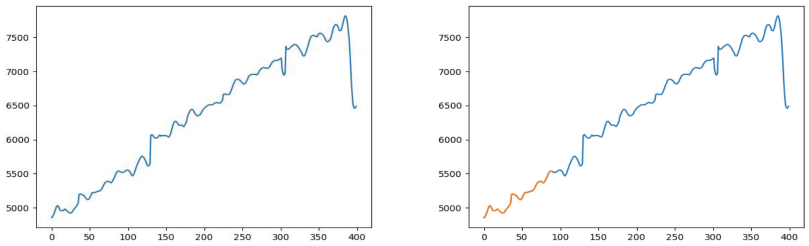


Fig. 5. Illustration of x coordinates generated through Sign-Diffusion.

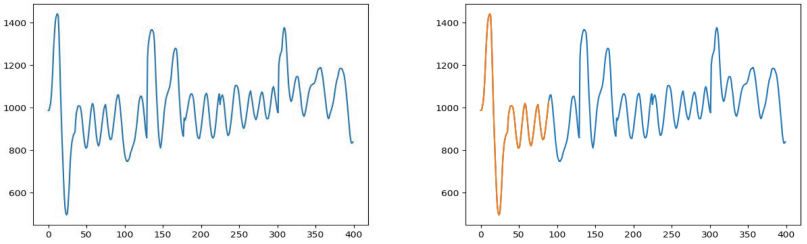


Fig. 6. Illustration of y coordinates generated through Sign-Diffusion.

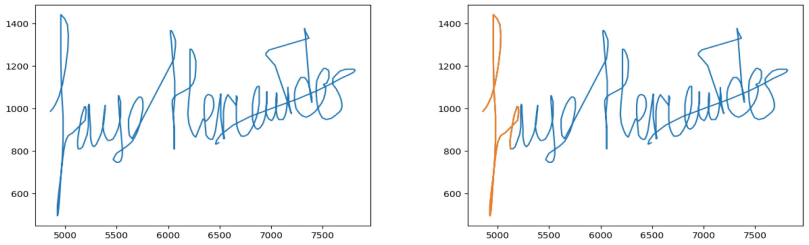


Fig. 7. Comparison between original and generated signatures - Sample 1.

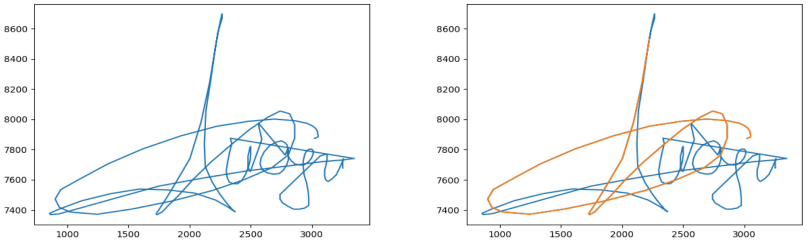


Fig. 8. Comparison between original and generated signatures - Sample 2.

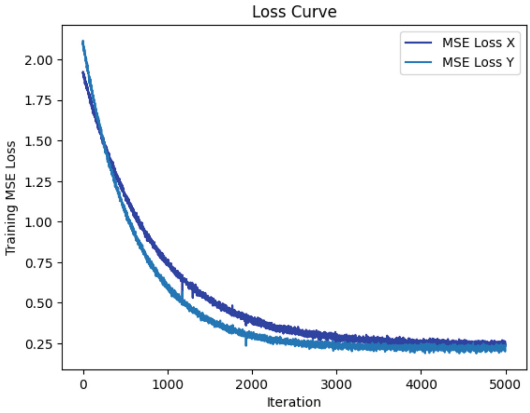


Fig. 9. Convergence of training loss of Sign-Diffusion Network while training on MCYT dataset

generated output. orange stroke is generated using the initial prompt value i.e., 100 points given to the model, and blue strokes are generated from the generated output of the model.

Figure 9, represents the loss convergence while training on the MCYT dataset. The loss curve represents the convergence separately for x and y coordinates while training the model. In Table 3, the calculated metrics MSE, Euclidean, and cosine similarity for X and Y coordinates are presented for 100 initial points to generate the signature. Table 4, represents the MSE, ED, and Cosine values between the original and generated signature as well as original and manual forgery samples already present in datasets. It has been observed that oscillation across the y axis is more than x axis, as behavioral traits of

Table 4. Comparison of Original Signature w.r.t System Generated Signature and Manual Forgeries.

Dataset	Original v/s Generated			Original v/s Manual Forgery		
	MSE	ED	Cosine	MSE	ED	Cosine
MCYT	X:2.768 Y : 4.773	1.1760	0.9934	X:70.432 Y:75.956	4.6707	0.9757
EBDS1	X:1.884 Y:4.546	1.9432	0.9956	X:83.433 Y:84.909	3.9680	0.9681
EBDS2	X:1.772 Y:4.833	1.7890	0.9914	X:84.915 Y:82.178	4.8206	0.9747
Biosecure-ID	X:2.575 Y:5.906	0.9901	0.9978	X:75.991 Y:70.123	4.9218	0.9601

Table 5. FID Comparision for existing SOTA in handwriting generation w.r.t SOTA values achieved for signature generation.

Approach	Dataset	Mode	Data Type	FID
Luhman et al. [17]	IAM	Offline	Handwriting	7.10
Sign-Diffusion	MCYT	Online	Signatures	6.65
	EBDS1	Online	Signatures	6.45
	EBDS2	Online	Signatures	6.52
	Biosecure-ID	Online	Signatures	6.48

signature, hence values observed are high against y axis. The results of online signature generation demonstrate a close match with the original online signature data. Additionally, the online signature data generated by Sign-Diffusion shows a higher degree of similarity to the original data compared to manually forged data present in the datasets.

As stated earlier Sign-Diffusion presents the foundational approach in online signature generation although we have compared it with existing offline handwriting generation techniques and represented it in Table 5. This comparison is done between original and generated online signature data versus original and generated offline handwriting data. This doesn't provide an exact SOTA comparison but gives a good view of the presented approach being capable of generating near-equal results. As Sign-Diffusion is a foundational model for online signature generation and no other existing approach is present in the literature, given comparison seems to be the best fit to evaluate the robustness of the proposed approach.

4.3 Ablation Study

To see the effect of different components of the proposed approach Sign-Diffusion, we also perform the following ablation studies by removing certain components. We have experimented on multiple combinations of network architectures as shown in Table 6. Ablation study includes following versions:

- **Sign-Diffusion** proposed approach with two SSM blocks.
- **Sign-Diffusion-1** corresponds to Sign-Diffusion with one transformer and one SSM block.
- **Sign-Diffusion-2** corresponds to Sign-Diffusion with two transformer blocks and no SSM block.

It is evident from the results, that having two SSM blocks as present in Sign-Diffusion, gives the best results because the network learns spatial and long spatio-temporal features at the same time. As signature length is highly variable in nature, possesses long temporal information to be catered during generation. Time series imputation generation model using convolution or attention mechanism lacks in longer time series forecasting. SSM approach overcomes this problem and presents a robust approach to handle longer time series sequence generation such as signatures and handwriting.

Table 6. Ablation study across different Sign-Diffusion architectures with 50 initial coordinates as points

Approach	Average MSE of coordinate x and y			
	MCYT	EBDS1	EBDS2	Biosecure-ID
Sign-Diffusion	3.637	3.240	3.483	3.460
Sign-Diffusion-1	5.890	5.587	5.665	5.560
Sign-Diffusion-2	8.833	8.562	8.692	8.681

Sign-Diffusion's superior performance over the ablation models Sign-Diffusion-1 and Sign-Diffusion-2 can be attributed to several technical factors. Firstly, the incorporation of conditional diffusion enables Sign-Diffusion to iteratively refine its predictions based on the context of the entire signature sequence, facilitating the generation of realistic and coherent signatures, especially with complex patterns. Additionally, the use of two-state space models in Sign-Diffusion allows for more effective modeling of temporal dependencies in signature data, which is crucial for handling signatures of varied lengths. Unlike bidirectional dilated convolution, which may struggle with capturing long-range dependencies, the state space models in Sign-Diffusion provide better temporal dependency handling for longer sequences. This capability enables Sign-Diffusion to capture intricate signature details, leading to more realistic and coherent generated signatures. Furthermore, Sign-Diffusion's probabilistic framework allows for uncertainty estimation in generated samples, ensuring the generation of diverse and natural-looking signatures while preserving the individuality of the signer's style.

5 Conclusion

This paper introduces the foundational model for generating online signature data, known as Sign-Diffusion, which is based on the conditional diffusion model. It demonstrates the efficacy of using conditional diffusion and SSM for generating signatures of varying lengths without requiring additional modalities. The Sign-Diffusion architecture is shown to have applications beyond time series imputation generation. This work also opens up avenues for research in deep fake identification within online signatures and handwriting.

References

1. Aksan, E., Pece, F., Hilliges, O.: DeepWriting: making digital ink editable via deep generative modeling. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2018)
2. Fierrez, J., et al.: BiosecurID: a multimodal biometric database. Pattern Anal. Appl. **13**, 235–246 (2010)
3. Fierrez, J., Ortega-Garcia, J.: On-line signature verification. In: Jain, A.K., Flynn, P., Ross, A.A. (eds.) Handbook of biometrics, pp. 189–209. Springer, Boston (2008). https://doi.org/10.1007/978-0-387-71041-9_10

4. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
5. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
6. Gu, A., Johnson, I., Timalsina, A., Rudra, A., Ré, C.: How to train your hippo: state space models with generalized orthogonal basis projections. arXiv preprint arXiv:2206.12037 (2022)
7. Gurvich, I.: Diffusion models and steady-state approximations for exponentially ergodic Markovian queues (2014)
8. Hodson, T.O.: Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci. Model Dev. Discussions **2022**, 1–10 (2022)
9. Jia, Z., Yang, L., Zhang, Z., Liu, H., Kong, F.: Sequence to point learning based on bidirectional dilated residual network for non-intrusive load monitoring. Int. J. Electr. Power Energy Syst. **129**, 106837 (2021)
10. Kang, L., Riba, P., Wang, Y., Rusiñol, M., Fornés, A., Villegas, M.: GANwriting: content-conditioned generation of styled handwritten word images. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12368, pp. 273–289. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58592-1_17
11. Kholmatov, A., Yanikoglu, B.: Identity authentication using improved online signature verification method. Pattern Recogn. Lett. **26**(15), 2400–2408 (2005)
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
13. Kiranyaz, S., Ince, T., Gabbouj, M.: Real-time patient-specific ECG classification by 1-D convolutional neural networks. IEEE Trans. Biomed. Eng. **63**(3), 664–675 (2015)
14. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: DiffWave: a versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761 (2020)
15. Kotani, A., Tellex, S., Tompkin, J.: Generating handwriting via decoupled style descriptors. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 764–780. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_45
16. Lian, Z., Zhao, B., Xiao, J.: Automatic generation of large-scale handwriting fonts via style learning. In: SIGGRAPH Asia 2016 Technical Briefs, pp. 1–4 (2016)
17. Luhman, T., Luhman, E.: Diffusion models for handwriting generation. arXiv preprint arXiv:2011.06704 (2020)
18. Meijer, C., Chen, L.Y.: The rise of diffusion models in time-series forecasting. arXiv preprint arXiv:2401.03006 (2024)
19. Ortega-Garcia, J., et al.: MCYT baseline corpus: a bimodal biometric database. IEE Proc.-Vision Image Signal Process. **150**(6), 395–401 (2003)
20. Poh, N., Bourlai, T., Kittler, J.: Biosecure DS2: a score-level quality-dependent and cost-sensitive multimodal biometric test bed. Pattern Recogn. J. (2010)
21. Sherstinsky, A.: Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys. D **404**, 132306 (2020)
22. Vera-Rodriguez, R., Tolosana, R., Ortega-Garcia, J., Fierrez, J.: e-BioSign: stylus- and finger-input multi-device database for dynamic signature recognition. In: 3rd International Workshop on Biometrics and Forensics (IWBF 2015), pp. 1–6. IEEE (2015)
23. Wijmans, J.G., Baker, R.W.: The solution-diffusion model: a review. J. Membr. Sci. **107**(1–2), 1–21 (1995)

C-PPT: A Channel-Wise Prototypical Part Transformer for Interpretable Perioperative Complication Prediction with Blood Pressure

Jingwei Zhang[1,2], Xiaodong Yang[2], Yiqiang Chen[2,3(✉)], and Ruizhe Sun[2]

[1] Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou, China
z1209673281@gs.zzu.edu.cn
[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{yangxiaodong,yqchen,sunruizhe18s}@ict.ac.cn
[3] Peng Cheng Laboratory, Shenzhen, China

Abstract. Continuous monitoring of multi-channel blood pressure during the perioperative period is crucial for predicting complications. ProtoPNet has garnered attention as powerful tools for providing interpretable support for complication prediction, enabling clinicians to better understand the impact of blood pressure trajectories on complications. However, existing prototype-based models for complication prediction tasks based on channel-wise blood pressure fail to recognize the unique characteristics of individual channels such as systolic pressure, diastolic pressure, and mean arterial pressure. Instead, they treat them as a unified entity for prediction and interpretation, resulting in performance degradation. To address this issue, we proposed the Channel-wise Prototypical Part Transformer (C-PPT). Firstly, we match the encoded data independently for each channel using a ProtoPNet, allowing for the extraction of unique features. Secondly, we enhance the prototypes by incorporating distance information between prototypes in the loss function. Finally, we optimize the influence of different channels on the results using pre-set weights. Experimental results conducted on a real dataset of perioperative blood pressure and cardiovascular adverse events classification tasks in a hospital setting demonstrate that our proposed method effectively interprets the relationship between blood pressure trajectories in different channels and the occurrence of cardiovascular adverse events, outperforming other benchmark models on relevant metrics.

Keywords: blood pressure · interpretable · prototype · multivariate time series

1 Introduction

Cardiovascular disease is the leading cause of death worldwide, claiming more lives each year than any other cause. Blood pressure is a key indicator of cardiovascular health [7]. It consists of multiple channels, including systolic pressure, diastolic pressure, and mean arterial pressure, each of which has distinct

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 46–60, 2025.
https://doi.org/10.1007/978-3-031-78341-8_4

physiological significance and clinical relevance, varying in its importance for different types of cardiovascular diseases [14]. As such, continuous monitoring of multi-channel blood pressure during surgical procedures is crucial. Research has demonstrated that significant fluctuations in blood pressure during surgery can trigger adverse cardiovascular events, such as stroke, which in turn increase the risk of postoperative mortality [11]. By carefully analyzing blood pressure data from these different channels during surgery, clinicians are better positioned to intervene in a timely manner, potentially preventing serious complications.

During clinical surgical procedures, clinicians typically predict potential complications by monitoring blood pressure levels that deviate from normal thresholds [15]. For instance, when blood pressure values consistently exceed or fall below a certain threshold, clinicians may perceive health risks and anticipate impending complications [9]. However, this approach struggles to comprehensively capture the underlying connections between the trajectory of blood pressure changes and the onset of complications.

To provide clinicians with a more comprehensive understanding of the relationship between blood pressure trajectories and complications, ProtoPNet [1], an interpretable tool, has been introduced for this task. Originally designed for interpretable fine-grained image recognition, ProtoPNet has since been extended to other domains, demonstrating strong performance.

However, existing prototype-based models for predicting complications using multi-channel time-series blood pressure data fail to fully account for the unique information inherent in each blood pressure channel. Instead, they treat all channels as a unified entity for both prediction and interpretation, which ultimately leads to suboptimal performance. This highlights the urgent need for methods that emphasize the independence of different channels to enhance prediction accuracy and interpretability.

To address these challenges, we proposed C-PPT, a novel interpretable prototype-based network designed for end-to-end fine-grained complication detection. C-PPT leverages a transformer as the backbone network and introduces a channel-wise prototype network that accounts for the unique characteristics of multi-channel blood pressure data, providing interpretable insights. This approach allows for the precise identification of representative prototypes specific to each channel. Furthermore, we conduct targeted optimizations on the channel-wise prototype network.

In summary, the main contributions of this paper are as follows:

(1) We proposed the Channel-wise Prototypical Part Transformer (C-PPT), an end-to-end model designed for perioperative multi-channel time-series blood pressure data, which enhances the accuracy and interpretability of perioperative complication prediction tasks.
(2) To address the characteristics of multi-channel time-series data, we developed a channel-wise prototype network. This network identifies key blood pressure trajectory regions influencing model predictions, enabling detailed monitoring of different channel data in complication manifestation. By analyzing prototype features from both global and local perspectives, the model demonstrates comprehensive interpretability.

(3) To optimize the interpretability of the model, we introduce a novel loss mechanism that enhances prototype distinctiveness by computing distances between different prototypes. Additionally, considering the varying contributions of blood pressure channels to complication prediction, we adjust the influence of each channel in the model through preset weights, optimizing the accuracy and relevance of predictions.
(4) Experimental validation on real datasets demonstrates that the C-PPT model not only exhibits excellent performance but also possesses high accuracy and interpretability. Visual analysis of global and local prototype visualization provides transparent and intuitive explanations for the reasoning process between different channels and classifications.

2 Related Work

2.1 Complication Prediction Task Based on Blood Pressure Data

Early machine learning-based methods often extract relevant features through some metric relationships and then input these features into a model for training. For example, Wijnberge et al. [19] utilized 23 variables extracted from the arterial pressure waveform to calculate an alert value, predicting the likelihood of hypotension occurring in the next few minutes. Song et al. [16] addressed the problem of historical information loss by proposing an improved Long Short-Term Memory - Convolutional Neural Network (LSM-CNN) combined with attention mechanisms, thereby improving the accuracy of hypertension prediction. With the widespread application of Transformer [17] models in time series classification tasks and their impressive performance, further exploration in this field has been prompted. Liu et al. [12] proposed Gated Transformer Networks (GTN), which combine two Transformer towers with gating mechanisms-one modeling channel-wise correlations and the other modeling step-wise correlations. This approach has been effectively applied to multivariate time series classification tasks. This approach has been effectively applied to multivariate time series classification tasks. These characteristics of GTN make it well-suited for tasks such as predicting complications from multi-channel blood pressure data. Therefore, we chose GTN as the backbone network for C-PPT.

2.2 Interpretability of the Time Series Classification Methods

To enhance physicians' understanding of predictive models for complications, recent research has made significant strides in enhancing model interpretability [13]. Current interpretable methods often employ bottom-up approaches, such as using saliency methods to ascertain the importance of input features. However, due to the time-dependency of feature importance, such methods typically perform poorly on data with strong temporal dependencies [6,10]. Consequently, these methods struggle to effectively capture the dynamic importance of data like blood pressure over time, limiting their practicality in clinical applications.

Hence, ProtoPNet [1] has been introduced to address interpretability in this task. ProtoPNet was initially developed primarily for interpretable methods in fine-grained image recognition, achieved by combining the model with sample-based methods. Xue et al. [20] were the first to combine Transformer with ProtoPNet, improving interpretive recognition for images. In subsequent research, ProtoPNet was attempted to be applied to relevant tasks in time-series classification. Despite improvements by Gao et al. [3] to adapt ProtoPNet for time series data, this approach still fails to fully consider the independence between different channels of data, instead treating them collectively for prediction and interpretation. When handling data such as multi-channel blood pressure data, where each channel holds distinct physiological significance, this approach may impact model performance. For instance, if the current channel's prototype fails to consider the distinctiveness of other channels and is directly used for matching, it may result in reduced prototype similarity, thus diminishing overall model interpretability.

Therefore, by specifically enhancing the prototype network, we have improved the interpretability of our model for predicting complications using multi-channel perioperative blood pressure data. This enhancement enables accurate matching of representative prototypes for each channel while also improving model performance.

3 Method

Our proposed method primarily consists of three components: 1) The Gated Transformer Multivariate Time-Series Encoder Network, which is responsible for extracting features from multi-channel blood pressure data. 2) The channel-wise Prototype Network, which learns representative blood pressure trajectories for different channels and classes from the training data. Subsequently, it utilizes the similarity between unknown samples and corresponding channel prototypes as evidence for decision-making in interpreting blood pressure trajectories during complication occurrences. 3) We have improved the loss mechanism to facilitate the acquisition of more distinctive prototypes for different classifications of samples and to better capture the impact of each channel on complication outcomes, thereby enhancing model performance. Our model structure is illustrated in Fig. 1.

3.1 Gated Transformer Encoder Network

Encoder-decoder is a common model framework in deep learning, utilized to transform input sequences into a fixed-length vector or convert previously generated fixed vectors into output sequences. In our model, the input sequence consists of a set of multivariate, multi-channel time-series physiological signal data. Through this encoder, our input sequence undergoes feature extraction, transforming into a fixed-length vector used for fine-grained complication detection and subsequent interpretability tasks.

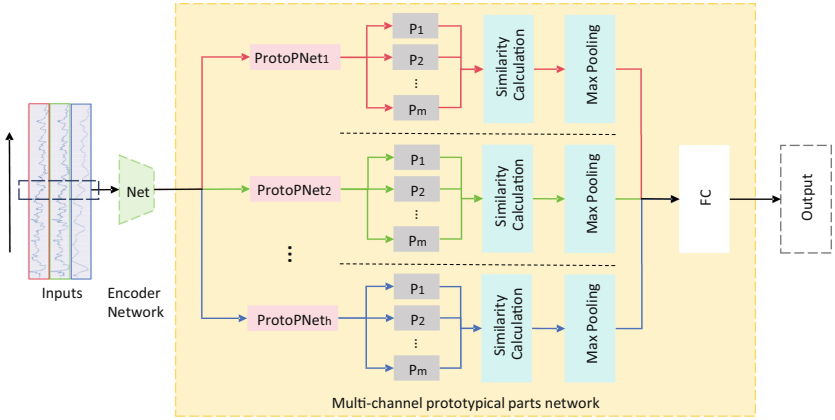


Fig. 1. Our C-PPT model structure consists of (1) The Encoder part of the model is the feature selection of the input sequence using the time series model, (2) The channel-wise ProtoPNet corresponds to the characteristics of the respective channels and contributes to the final prediction.

Due to the presence of multiple channels in blood pressure sequences, our encoder network needs to effectively extract features from multi-channel time-series data. Gated Transformer Network [12], as a variant of the Transformer, demonstrates superior performance in handling multi-channel time-series data. It can simultaneously capture temporal features of blood pressure data and inter-channel information, effectively addressing the issue of long-term data dependencies. This paper utilizes an existing Gated Transformer network as an encoder to achieve improved feature extraction for multivariate time-series data.

3.2 Channel-Wise Prototypical Part Network

Due to the inadequate consideration of the unique information among different channels in existing ProtoPNet-based models for predicting complications from multi-channel time-series blood pressure data, we optimized the network structure to better match multi-channel time-series data while preserving the distinctive features of each channel. As illustrated in Fig. 1, after encoding by the Encoder Network, multi-channel blood pressure data is matched with their respective channel-specific prototypes, ensuring that each prototype adequately captures the representative segments of its respective channel.

For example, in this blood pressure task, for the blood pressure sequence, we input 3×700 time-series data into the network. Here, 3 represents the different channels of blood pressure, and 700 indicates a total duration of 350 min, with data sampled at intervals of 30 s. Let $f(x)$ represent the feature map of input data x after passing through the encoder network, where $f(x) = Z \in \mathbb{R}^{D \times H \times W}$. To enable the prototypes to fully capture the blood pressure trajectories, the dimensions of the feature map for the blood pressure data

are set to $D = 1406, H = 3, W = 23$, where D represents the feature dimension, H denotes the number of channels, and W corresponds to the encoded sequence length.

In other words, the original blood pressure sequence of length 700 is compressed into a feature sequence of length 23, with each feature representing approximately 30 time points of the original data. The network will learn $M = H \times m$ pre-determined prototypes $P = \{\{p_{h,j}\}_{j=1}^{m}\}_{h=1}^{H}$, where the m denotes the number of prototypes per channel, the k denotes the number of classes. In this task, to adequately capture prototypes for different classifications of samples, the quantity of m is set as $6 \times k$ (six prototypes are assigned for each of the k classifications). The shape of each prototype is $D \times 1 \times 1$.

By computing the reciprocal of the squared L2 distance between the feature mapping matrix of the test data and each prototype within the corresponding channel, we obtain the similarity scores between the features and prototypes for each respective channel. After performing max-pooling on these similarity scores, we obtain the similarity score matrix. Upsampling the similarity score matrix enables us to obtain a corresponding matrix of the same length as the input sample, with each data point corresponding one-to-one with the data points in the input sample.

By identifying the regions with the highest similarity in the corresponding matrix, we can obtain the prototypes that best match the current sample channel. Further visualization processing of these captured prototype regions can be used to support interpretive analysis.

To visualize the prototypes, during the training process, we map each prototype $p_{h,j}$ to the corresponding channel of the training sample with the highest similarity and of the same classification as $p_{h,j}$, thereby obtaining an equivalent prototype projection. Mathematically, we need to perform the following updates:

$$p_{h,j} \leftarrow \arg\min_{z \in \mathcal{Z}_{h,j}} \|z - p_{h,j}\|_2, \tag{1}$$

where $\mathcal{Z}_{h,j} = \{\tilde{z_h} : \tilde{z_h} \in patches(f(x_{h,i}))\forall i s.t. y_i = k\}$, $p_{h.j} \in P_{h,k}$.

After passing through the channel-wise prototypical parts network, the data is processed by fully connected layers to produce the final sample classification results.

3.3 Training and Inference

During the process of prototype matching, it is possible for prototypes from different classifications to have distances that are too close, or even overlap. This can result in prototypes being matched to the same regions, leading to a decrease in the quality of interpretability. Additionally, the varying contributions of different channels of multi-channel blood pressure to complications can also hinder the model from fully realizing its potential. In this paper, considering the characteristics of multi-channel time-series data and prototypes, we proposed a new loss mechanism to mitigate the occurrence of this issue. The first part of the loss

function is a cross-entropy loss, which penalizes misclassifications in the training data:

$$L_{pred}(y_{true}, y_{pred}) = -\frac{1}{n}\sum_{i=1}^{n} y_{true}^{i} \log\left(y_{pred}^{i}\right), \tag{2}$$

where n represents the total number of data batches.

In addition, we have integrated distance information between channel prototypes and different targets into the loss function, denoted as $L_{clus}(P)$:

$$L_{clus(P)} = \frac{1}{H}\sum_{h=1}^{H}(\lambda_h \cdot L_{clst}(P_h) + \lambda_h \cdot L_{sep}(P_h) + L_{dist}(P_h)), \tag{3}$$

where $L_{clst}(P_h)$, $L_{sep}(P_h)$ and $L_{dist}(P_h)$ are defined by

$$L_{clst}(P_h) = \frac{1}{n}\sum_{i=1}^{n} \min_{j:p_{h,j}\in P_{h,y_i}} \min_{z_h \in patches(f(x_i)))} \|z - p_{h,j}\|_2^2; \tag{4}$$

$$L_{sep}(P_h) = -\frac{1}{n}\sum_{i=1}^{n} \min_{j:p_{h,j}\notin P_{h,y_i}} \min_{z_h \in patches(f(x_i))} \|z - p_{h,j}\|_2^2, \tag{5}$$

$$L_{dist}(P_h) = -\frac{1}{n}\sum_{i=1}^{n} \|p_{h,j} - p_{h,l}\|_2^2, \tag{6}$$

where L_{clst} and L_{sep} represent the distances between the sample and the nearest prototype of the same class, p_1, and the nearest prototype of a different class, p_2, respectively, which were originally existing information. We multiplied L_{clst} and L_{sep} for different channels by λ_h as weights to adjust their contributions to predicting complications. The value of λ_h can be set based on the experience of medical professionals. Additionally, we introduced L_{dist} to capture the distance information between p_1 and p_2.

In this process, within the same channel, the distance information between the sample and prototype with the same classification is utilized, along with the distance information of different classifications, as a penalty. This encourages each training sample's temporal feature sequence to be close to its assigned category and simultaneously distant from categories to which it does not belong. Meanwhile, the distance information between the prototype of the same classification nearest to the current sample and the prototype of different classification also serves as a penalty, enhancing the interpretability of the prototypes. By adjusting the weights assigned to different channels, we optimize the influence of each channel on the results, thereby further improving the performance of our network in predicting complications. This helps to address the issue of prototype overlap and further promotes the accuracy and interpretability of our network in classification results.

In the initial stages of the model, our loss function is updated using only L_{pred}. After the model has been running for some time, the prototype matching mechanism of the prototype network comes into play. At this point, the loss function is modified to $L_{pred} + L_{clus}$ for updates. The complete training and inference procedure of C-PPT is shown in Algorithm 1.

Algorithm 1: Training and Inference Procedure of C-PPT

Input: The perioperative multi-channel time-series blood pressure dataset X
Output: The output of classification and prototype P.

Transform the training data to the shape (N, 3, 700), where N means the number of samples, 3 means the number of channels, and 700 means the length of the time-series blood pressure.
for *epoch from 1 to 100* **do**
  For each sample, learn the feature matrix Z through the encoder network
  for *h from 1 to H* **do**
    Extract the current channel features : $z = Z_h$.
    Calculate similarity scores with prototypes using the squared L^2 distance and invert it.
    Maxpolling the similarity scores to get the activation map and softmax.
    Obtain the final similarity matrix $Scores_h$.
  Concatenate the similarity matrices of each channel to obtain Scores.
  $y_{pred} = FC(Scores)$
  $Loss_{all} = L_pred.$
  if *epoch* $>= 10$ *and epoch*$\%5 == 0$ **then**
    for *h from 1 to H* **do**
      Iterate over all prototypes P_h to project with the nearest training features of the same channel.
  if *epoch* $>= 10$ **then**
    $L_{clus} = \lambda_h \cdot L_{clst}(P_h) + \lambda_h \cdot L_{sep}(P_h) + L_{dist}(P_h)$
    $Loss_{all} += L_{clus}$
  Update the $Loss_{all}$ loss function
Return y_{pred}

4 Experiments

4.1 Datasets

To demonstrate the feasibility of the proposed method in real-world scenarios, we utilized authentic intraoperative blood pressure data. The dataset was obtained from the Chinese People's Liberation Army (PLA) General Hospital and includes time-series blood pressure data from patients during clinical surgeries, along with the classification of major adverse cardiac events (MACE) occurring after surgery. The blood pressure data consists of systolic pressure, diastolic pressure, mean arterial pressure, etc. All three types of blood pressure data were collected at a frequency of 0.5 min and have a duration of 350 min each, resulting in a data length of 700 for each channel. The dataset comprises a total of 7489 data instances, with a positive-to-negative ratio of 1:9. We partitioned the dataset into training and testing sets in an 8:2 ratio. Before the model was executed, we normalized the data. The weight λ for the loss functions mentioned in the text was pre-set to $[0.1, 0.1, 1]$.

4.2 Technical Detail

We constructed our model based on the PyTorch library. In the experiments, all models were trained using the Adam optimizer with a batch size of 32 and scheduled for training over 100 epochs using the ReduceLROnPlateau scheduler. Through the experiments, we ultimately established 6 prototypes of size $1406 * 1 * 1$ for each class in every channel of the data. Due to the blood pressure dataset containing a total of 3 channels and having 2 categories for the outcome variable, there are a total of 36 prototypes here. Our experiments were conducted on Intel(R) Xeon(R) Bronze 3204 CPU @1.90 GHz and 2 NVIDIA RTX 3090 graphics cards on Ubuntu 64-bit.

4.3 Comparison Methods

We evaluated the C-PPT model, simultaneously listing several baseline models and conducting comparisons:

- **LSTM** [5]**:** Introducing three gates (input gate, forget gate, output gate) and a cell state, this deep learning model is designed to better handle long-term dependencies in sequential data.
- **FCN** [18]**:** A deep learning model utilizes CNN to automatically discover and extract the internal structures of input time-series data, generating deep features for classification.
- **ResNet** [4,18]**:** A deep convolutional neural network addressing the performance degradation caused by increasing CNN depth through the utilization of residual modules.
- **ProtoPLSTM** [3]**:** A model that combines CNN-LSTM with ProtoPNet to achieve interpretable results for time series data.
- **TapNet** [21]**:** A deep learning model for multi-dimensional time-series classification learns low-dimensional feature representations and interactive features through an attentional prototype network and multi-layer convolutional networks.
- **LSTMFCN** [8]**:** A multi-dimensional time-series classification model is realized by extending a fully convolutional block with a squeeze-excitation block.
- **InceptionTime** [2]**:** A deep learning model for time-series classification inspired by the Inception network structure.
- **Gated Transformer Network(GTN)** [12]**:** A multivariate time series classification model that captures channel-specific and step-wise dependencies by merging two Transformer towers with gating mechanisms.

4.4 Performance Comparison

In Table 1, we present the comprehensive results of C-PPT and related time-series classification methods on the blood pressure dataset. It is evident that our proposed C-PPT demonstrates outstanding performance, showing significant improvements in key metrics such as prediction accuracy, precision, F1 score, and

AUC compared to baseline models. These results not only validate the effectiveness of C-PPT in analyzing perioperative blood pressure data but also demonstrate its advantage in maintaining efficient discriminative power when dealing with highly imbalanced datasets. Furthermore, this underscores the potential wide applicability and excellent generalization ability of our model in clinical settings.

Table 1. Prediction scores of all models on blood pressure data.

Methods	Acc (%)	Pre (%)	F1 (%)	Spe (%)	AUC (%)
LSTM	84.03	29.01	38.87	86.40	77.62
FCN	84.24	30.55	41.58	86.04	80.44
ResNet	84.17	28.74	38.12	86.77	77.22
ProtoPLSTM	76.29	22.44	34.14	76.75	80.44
TapNet	77.10	22.80	34.40	77.80	79.20
LSTMFCN	88.58	36.88	40.83	92.62	71.58
InceptionTime	69.94	17.18	27.18	70.39	70.92
GTN	79.96	24.32	35.06	81.58	77.89
C-PPT	89.65	40.71	42.38	92.84	80.96

In Table 2, we conducted ablation experiments on the proposed channel-wise prototype network using the following symbols: a) C-PPT$_{one-channel}$ employs a single-channel prototype network structure; b) C-PPT$_{no-channel}$ represents the basic network when the prototype network module is removed; c) C-PPT represents the original model using a channel-wise prototype network. It can be observed that the AUC with a single-channel prototype network is not significantly different from the case without the prototype module. However, our proposed C-PPT outperforms other methods in terms of precision and AUC. This indicates that adding the prototype network enhances the interpretability of the model without adversely affecting its performance. In fact, it may even improve the model's performance. The optimization of the C-PPT model for the multi-channel characteristics of blood pressure data results in particularly noticeable improvements in model performance.

Table 2. The C-PPT model on blood pressure data underwent ablation experiments with different channel prototype networks.

Methods	Acc (%)	Pre (%)	F1 (%)	Spe (%)	AUC (%)
C-PPT$_{one-channel}$	77.22	24.52	37.43	77.05	77.92
C-PPT$_{no-channel}$	79.96	24.32	35.06	81.58	77.89
C-PPT	89.65	40.71	42.38	92.84	80.96

4.5 Comparison of Different Loss Mechanisms

We evaluated the impact of the optimized loss functions on model performance through ablation experiments. With other parameters fixed, we compared the following scenarios in Table 3: a) using only the cross-entropy loss function L_{pred}; b) using both $L_{pred} + L_{clst} + L_{sep}$; c) C-PPT, where the loss function is $L_{pred} + L_{clst} + L_{sep} + L_{dist}$. Furthermore, to assess the influence of different channel weights on model performance, we compared the following scenarios in Table 4: a) Not setting the λ value; b) setting the λ. The results from Table 3 and Table 4 indicate that by optimizing the loss mechanism, the model's metrics such as accuracy, precision, F1 score, and AUC have been improved to varying degrees. This enhancement contributes to an overall improvement in model performance and aids in achieving more balanced results.

Table 3. The C-PPT model on blood pressure data underwent ablation experiments with different channel prototype networks.

Methods	Acc (%)	Pre (%)	F1 (%)	Spe (%)	AUC (%)
C-PPT$_{L_{pred}}$	82.97	28.42	39.43	84.72	80.37
C-PPT$_{L_{pred}+L_{clst}+L_{sep}}$	87.64	35.86	43.43	90.72	79.34
C-PPT	89.65	40.71	42.38	92.84	80.96

Table 4. The C-PPT model on blood pressure data underwent ablation experiments with different channel loss weights.

Methods	Acc (%)	Pre (%)	F1 (%)	Spe (%)	AUC (%)
C-PPT (no λ)	83.77	29.50	40.29	85.67	80.42
C-PPT	89.65	40.71	42.38	92.84	80.96

4.6 Visualization Analysis

In order to assess the interpretability of perioperative complication prediction based on continuous blood pressure, we conducted a visual analysis by post-training and marking, with bounding boxes, the parts of each test sample that were most similar to the corresponding prototypes. For ease of analysis, the marked region width was set to 15 min.

The C-PPT model utilizes representative prototypes trained from the entire training set as global prototypes and references them when verifying local interpretability on test samples. To demonstrate the model's global interpretability in perioperative complication prediction, we conducted an evaluative analysis in Fig. 2. As shown in the figure, by comparing the global prototypes across three

channels and their corresponding test samples, the proposed C-PPT model intuitively allows doctors to identify key regions contributing to complication decisions. This region represents common features found by the model when similar blood pressure waveforms occur in the corresponding results from a large number of samples in the training set. It indicates that this region has a more significant association with complication decisions compared to other regions. From the graph, it is evident that the corresponding region in the test sample also matches a similar area and maintains high similarity with the global prototype, demonstrating the effectiveness of the global prototype.

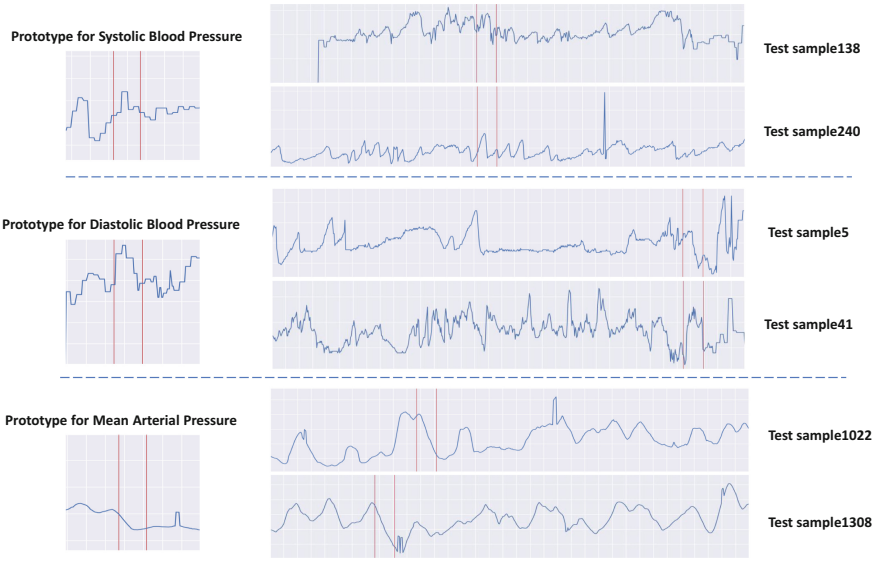


Fig. 2. The global prototypes constructed by C-PPT, encompassing perioperative systolic pressure, diastolic pressure, and mean arterial pressure, serve as the basis for decision-making in test samples.

In Figure 3, we demonstrate the locally interpretable effects of the C-PPT model. We verify interpretability by matching the local prototypes, obtained by selecting the prototype with the highest similarity in each respective channel of the test samples to the corresponding channel, with the global prototypes trained from the training samples. In the test samples, each channel of each sample will have a set of local prototypes for the respective channel. We take one sample from this to evaluate. We can see that for this test sample, systolic blood pressure and mean arterial pressure play a more important role in the decision of this cardiovascular adverse event. Furthermore, we can observe that the prototype region in the test sample corresponds to the corresponding region of the global prototype with the same semantic concept. This indicates that the C-PPT model can successfully capture representative prototypes from

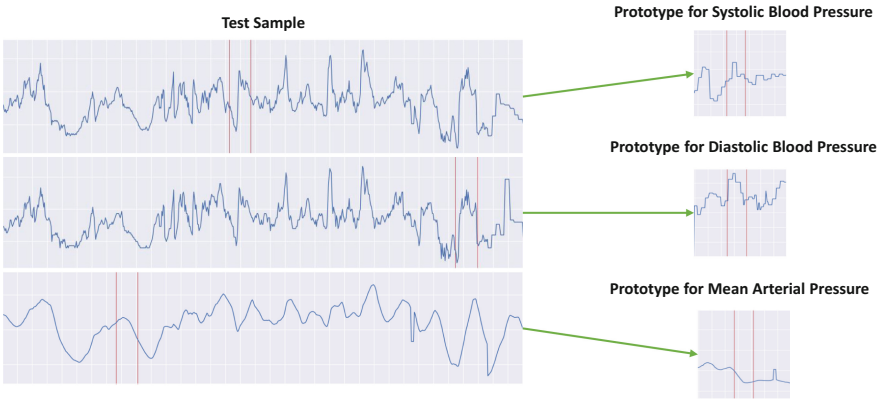


Fig. 3. The three rows in the test sample represent systolic pressure, diastolic pressure, and mean arterial pressure during the perioperative period. These are matched with the corresponding global prototypes obtained from the training set to achieve the interpretability effect.

the test samples. By comparing the local prototypes with the global prototypes, doctors can better capture the detailed features in blood pressure. Providing this as a reference to doctors can effectively improve the efficiency of clinical decision-making for them.

5 Conclusion

In this study, we proposed the Channel-wise Prototypical Part Transformer (C-PPT), a novel fine-grained interpretable method specifically tailored for predicting complications based on multi-channel time-series blood pressure data during the perioperative period. We systematically address the characteristics of multi-channel blood pressure by developing a multi-channel prototype network. This network is designed to capture the unique manifestations of each channel in complication prediction tasks while effectively reducing unnecessary interference between channels. Furthermore, we introduce a novel loss optimization mechanism that significantly enhances the interpretability of prototypes by considering the distances between different prototypes and incorporating them into the loss function. We also pay particular attention to the contributions of different channels to the prediction results by weighting the distance information of different channels and integrating it into the loss function to optimize model performance. Also conduct experiments on a real perioperative dataset, where C-PPT outperforms all baseline models in all relevant metrics. The experimental results not only demonstrate the superiority of our approach but also showcase its potential in practical clinical applications. In future work, we plan to explore the application potential of C-PPT in other key areas, such as emotion recognition and behavior recognition, to further validate its wide-ranging applicability.

Acknowledgement. This work is supported by the National Natural Science Foundation of China No. 62202455, Beijing Natural Science Foundation No. L222100 and Beijing Municipal Science & Technology Commission No. Z221100002722009.

References

1. Chen, C., Li, O., Barnett, A.J., et al.: This looks like that: deep learning for interpretable image recognition. In: Neural Information Processing Systems (2018). https://api.semanticscholar.org/CorpusID:49482223
2. Fawaz, H.I., Lucas, B., Forestier, G., et al.: InceptionTime: finding AlexNet for time series classification. Data Min. Knowl. Discov. **34**, 1936–1962 (2019). https://api.semanticscholar.org/CorpusID:202572652
3. Gao, C., Zhang, T., Jiang, X., et al.: ProtoPLSTM: an interpretable deep learning approach for wearable fine-grained fall detection. In: 2022 IEEE Smartworld, Ubiquitous Intelligence I& Computing, Scalable Computing I& Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous I& Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta), pp. 516–524 (2022). https://doi.org/10.1109/SmartWorld-UIC-ATC-ScalCom-DigitalTwin-PriComp-Metaverse56740.2022.00091
4. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2015). https://api.semanticscholar.org/CorpusID:206594692
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
6. Kacprzyk, K., Liu, T., van der Schaar, M.: Towards transparent time series forecasting. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=TYXtXLYHpR
7. Kannel, W.B.: Role of blood pressure in cardiovascular morbidity and mortality. Prog. Cardiovasc. Dis. **17**(1), 5–24 (1974). https://doi.org/10.1016/0033-0620(74)90034-6, https://www.sciencedirect.com/science/article/pii/0033062074900346
8. Karim, F., Majumdar, S., Darabi, H., et al.: Multivariate LSTM-FCNs for time series classification. Neural Netw. Off. J. Int. Neural Netw. Soc. **116**, 237–245 (2018). https://api.semanticscholar.org/CorpusID:22665671
9. Le, T., Ellington, F., Lee, T.Y., et al.: Continuous non-invasive blood pressure monitoring: a methodological review on measurement techniques. IEEE Access **8**, 212478–212498 (2020). https://doi.org/10.1109/ACCESS.2020.3040257
10. Leung, K.K., Rooke, C., Smith, J., et al.: Temporal dependencies in feature importance for time series predictions (2023)
11. Liang, J.: Clinical management of perioperative blood pressure and complications in patients with hypertension. Prevention and Treatment of Cardiovascular Disease (16-18) (2018)
12. Liu, M., Ren, S., Ma, S., et al.: Gated transformer networks for multivariate time series classification. ArXiv abs/2103.14438 (2021)
13. Patrício, C., Neves, J.C., Teixeira, L.F.: Explainable deep learning methods in medical image classification: a survey. ACM Comput. Surv. **56**, 1 – 41 (2022). https://api.semanticscholar.org/CorpusID:261934517
14. Rapsomaniki, E., Timmis, A., George, J., et al.: Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1.25 million people. Lancet **383**(9932), 1899–1911 (2014).

https://doi.org/10.1016/S0140-6736(14)60685-1, https://www.sciencedirect.com/science/article/pii/S0140673614606851
15. Sessler, D.I., Bloomstone, J.A., Aronson, S., et al.: Perioperative quality initiative consensus statement on intraoperative blood pressure, risk and outcomes for elective surgery. Br. J. Anaesth. **122**(5), 563–574 (2019). https://doi.org/10.1016/j.bja.2019.01.013, https://www.sciencedirect.com/science/article/pii/S0007091219300509
16. Song, X., Zhu, L., Feng, X., et al.: Combined forecast model of LSTM-CNN hypertension based on EEMD. In: Proceedings of the 2021 4th International Conference on Signal Processing and Machine Learning, SPML 2021, pp. 117–122. Association for Computing Machinery, New York (2021).https://doi.org/10.1145/3483207.3483227
17. Vaswani, A., Shazeer, N.M., Parmar, N., et al.: Attention is all you need. In: Neural Information Processing Systems (2017). https://api.semanticscholar.org/CorpusID:13756489
18. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: a strong baseline, pp. 1578–1585 (2017). https://doi.org/10.1109/IJCNN.2017.7966039
19. Wijnberge, M., Geerts, B.F., Hol, L., et al.: Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the hype randomized clinical trial. JAMA **323**(11), 1052–1060 (2020). https://doi.org/10.1001/jama.2020.0592
20. Xue, M., Huang, Q., Zhang, H., et al.: ProtoPFormer: concentrating on prototypical parts in vision transformers for interpretable image recognition. ArXiv abs/2208.10431 (2022). https://api.semanticscholar.org/CorpusID:251718906
21. Zhang, X., Gao, Y., Lin, J., et al.: TapNet: multivariate time series classification with attentional prototypical network. In: AAAI Conference on Artificial Intelligence (2020). https://api.semanticscholar.org/CorpusID:210703726

Visible-Infrared Person Search: A Novel Benchmark and Solution

Hongxu Chen[1], Jianghao Xiong[1], Yuheng Huang[2], Xiaohua Xie[1,3], and Jian-Huang Lai[1,3(✉)]

[1] School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510006, China
{chenhx87,xiongjh7}@mail2.sysu.edu.cn,
{xiexiaoh6,stsljh}@mail.sysu.edu.cn
[2] Guangzhou GRGBanking Equipment Co., Ltd., Guangzhou, China
[3] Guangdong Key Laboratory of Information Security Technology, Guangzhou 510006, China

Abstract. Person search aims to simultaneously localize and identify a query person from realistic and uncropped images, which consists of person detection and re-identification (Re-ID). Existing person search methods and datasets predominantly focus on the visible light domain, and have difficulty in alleviating modality discrepancies. Furthermore, existing visible-infrared person Re-ID methods struggle to adequately address occlusions and handle background interference effectively. To address the above issues simultaneously, we first construct a new large-scale dataset, Multi-Modality Person Search (MMPS), which tackles the lack of suitable benchmarks for person search in the visible-infrared domain. Encompassing challenges of complex background interferences and occlusions under modality discrepancies, MMPS includes 21,470 images and 1,012 identities across six different cameras. Furthermore, we propose a novel visible-infrared person search method that integrates detection and Re-ID into a progressive process. Specifically, Progressive Inclusion (PI) is proposed to explore backgrounds and provide adaptive proposals. To better tackle the complex occlusions under significant modality discrepancies, we present Discriminative Mix (DM) to synthesize more diverse samples, leveraging specific pattern map embedding. This strategy ensures that our model is not overfitted to specific patterns and is capable of identifying diverse and distinctive human parts. Extensive experiments demonstrate that our method (PI-DM) achieves state-of-the-art performance on the task of visible-infrared person search. Our dataset has been released on https://github.com/sysuchx/MMPS.

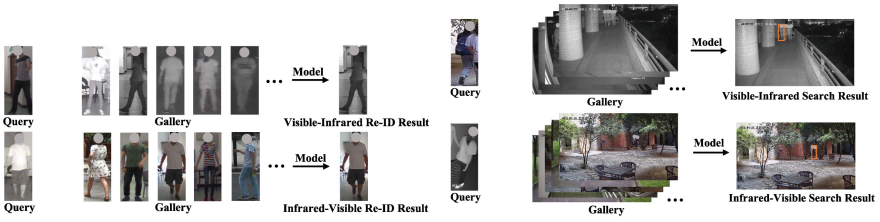
Keywords: Person search · Visible-Infrared · Novel benchmark

1 Introduction

Person search [22] aims to find specific pedestrians in scene images or video frames obtained under different cameras. This task needs to simultaneously

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 61–76, 2025.
https://doi.org/10.1007/978-3-031-78341-8_5

address the tasks of person detection [26] and re-identification (Re-ID) [5,29,31], and has emerged as an important task with real-world applications [8]. Existing methods [11,25] primarily focus on person search in visible light scenes. Moreover, the mainstream datasets CUHK-SYSU [19] and PRW [32] contain images from a single modality. This presents environmental limitations, as in dark conditions we may only acquire infrared images. On the other hand, in visible-infrared scenarios, existing work [13,24,30] concentrates on person Re-ID (as shown in Fig. 1a). In real life, infrared images usually encompass a substantial amount of unrelated backgrounds and occlusions, as shown in Fig. 1b. This presents significant challenges to researchers, which we summarize as follows. First, existing person search methods struggle to adequately alleviate modality discrepancies, resulting in unsatisfactory performance in visible-infrared applications. Moreover, existing methods for visible-infrared person Re-ID can not handle occlusions and background interference well. Even other strategies specifically for occlusions struggle to effectively extract discriminative pedestrian features in infrared scenarios. Besides, there is a lack of large-scale benchmark datasets in the field of visible-infrared person search.



(a) Cross-modality person Re-ID. (b) Cross-modality person search.

Fig. 1. Introduction of the task of (a) visible-infrared person Re-ID and (b) visible-infrared person search in real-life cross-modality scenarios.

To address above issues, we construct a new large-scale dataset Multi-Modality Person Search (MMPS), which effectively promotes the progress of research and application in the cross-modality domain. Specifically, MMPS includes challenges such as complex occlusions, scale variations, background noise, and perspective distortion. This dataset comprises 21,470 images, which surpasses that of existing person search datasets. These images contain 1,012 different pedestrian identities captured by six cameras. Our constructed MMPS dataset can be used as an important benchmark for visible-infrared person search.

Furthermore, we propose a novel method for visible-infrared person search. Our method incorporates Progressive Inclusion (PI) and Discriminative Mix (DM), which effectively handles the complex occlusions and background interferences under significant modality discrepancies. We first observe that the background information present in raw frames serves as a rich source of potential

variations within the scene. By selectively integrating the information from the foreground (person) and the background, we are able to construct more diverse samples, simulating occlusion problems. These enhanced samples assist the model in capturing a wide range of distinct human body parts, thereby improving its robustness. PI incorporates an exponentially decaying threshold, which applies heightened restrictions on background samples during the early stage of training. This strategy effectively curtails the excessive introduction of these samples, safeguarding the stability of early-stage training. As training progresses, PI dynamically expands the search scope. Upon acquiring these progressive samples, we adopt a feature-tiered augmentation strategy DM, merging the pattern maps of pedestrian samples with those of the backgrounds. Based on the more diversified sample repository, DM constructs a dual-task harmonization loss that tackles modality discrepancies under complex occlusions. In comparison with existing data augmentation methods based on random strategies [24,27], DM employs targeted augmentation, selectively mixing the specific pattern map from the chosen samples. This empowers the extraction of finer and more diverse discriminative pedestrian features under different modalities.

The main contributions can be summarized as follows:

- We first construct a new visible-infrared person search dataset Multi-Modality Person Search (MMPS). This dataset encompasses challenges such as complex background interferences and occlusions under modality discrepancies, serving as a new benchmark.
- We propose a novel visible-infrared person search method that utilizes Progressive Inclusion (PI) strategy to adaptively filter proposals for our subsequent Discriminative Mix (DM) processing.
- Our DM mixes the specific pattern map of particular samples and employs a dual-task harmonization loss, thereby effectively addressing complex occlusions under modality variations and focusing on more diverse human parts.
- Results of extensive experiments demonstrate that our method achieves state-of-the-art performance on visible-infrared person search.

2 Related Work

2.1 Person Search

Since two large-scale datasets CUHK-SYSU [19] and PRW [32] are introduced, person search has garnered significant interest within the computer vision community. Most existing methods for this task can be roughly grouped into two-stage and end-to-end approaches. Two-stage methods [3,7,32] separately train the detection and Re-ID models. For end-to-end methods, OIM [19] first employs the Faster R-CNN [15] as the detector, and shares base layers with the person Re-ID network. Based on OIM, NAE [4] segregates the person embedding into norm and angle components, which are respectively employed for detection and Re-ID. Based on NAE, SeqNet [11] considers detection and Re-ID as a progressive process and tackles two sub-networks sequentially. Some approaches

(COAT [25], PSTR [1]) employ an end-to-end architecture based on transformer [16]. COAT emulates the influence of external objects occluding a person of interest at the token level. PSTR introduces a specialized module for person search, which includes a detection encoder-decoder for detection and a distinctive Re-ID decoder for person Re-ID. In the weakly supervised domain, CGPS [22] utilizes two sub-networks to investigate different levels of context clues. Other research efforts, such as AlignPS [23], first utilize an anchor-free basis. However, most person search studies primarily focus on visible images, which limits the applicability of these methods. In this paper, we expand the scope by incorporating the use of infrared scenarios.

2.2 Visible-Infrared Person Re-identification

Infrared imagery is a critical component in a majority of surveillance systems. A large-scale visible-infrared person Re-ID dataset SYSU-MM01 [17] is proposed as an important benchmark. And the authors who constructed this dataset present a deep zero-padding strategy, which trains a single-stream network toward automatically evolving domain-specific nodes. XIV-ReID [10] introduces an auxiliary X modality, serving as an assistant, and reconceptualizes infrared-visible dual-mode cross-modality learning as a tri-modal learning problem. Using a single-stream network structure, MPANet [18] identifies cross-modality nuances in various patterns and jointly extracts discriminative features. To take the gray-scale images as an auxiliary modality, a progressive modality-shared transformer PMT [13] is proposed. Besides the single-stream methods, some methods employ a dual-stream network. FMCNet [28] compensates for missing modality-specific information at the feature level. A transformer-based visible-infrared network [2] captures the long-range dependencies of person images and modality-specific representations. CMTR [12] explicitly mines the information from various modalities to deal with the insufficient perception of modality information. Other methods, such as CAJL [24], employ data augmentation strategies through random color channel exchanging and random erasing. These methods typically work on pre-cropped images, whereas our method processes raw frames directly and is more attuned to real-world applications with complex occlusions.

3 Multi-modality Person Search Dataset

3.1 Motivation

At present, person search datasets mainly focus on retrievals in visible light images, with relatively less exploration in the infrared domain. In the visible-infrared field, existing datasets focus on cropped pedestrian images, excluding full video frames, which somewhat hampers the advancement of person search technology. Thus, there is a pressing need for a dataset that aligns closer with practical application scenarios, enabling the establishment of relevant benchmarks. Moreover, the task of visible-infrared person search needs to confront

challenges such as background interference and member occlusion in cross-modality scenarios, which presents greater complexity. Considering these emerging challenges, we propose Multi-Modality Person Search (MMPS) dataset to promote the development of cross-modality person search research and applications.

Table 1. Statistical comparison of datasets on person search.

Dataset	#Images	#Cameras	#Person ID	#B-Box	Modality
CUHK-SYSU [19]	18184	Vary	8432	96143	Visible
PRW [32]	11816	6	932	43110	Visible
MMPS	21470	6	1012	50142	Visible, Infrared

(a) Occlusion/scale variation. (b) Clothes change/ occlusion. (c) Background noise. (d) Perspective distortion.

Fig. 2. Illustration of diverse challenges in our MMPS dataset.

3.2 Dataset Description

Before releasing the dataset, we ensure that every individual appearing in the dataset has **signed a consent letter** for video recording and data collection for **academic use**. The video frames in our dataset are derived from the original videos of another dataset CM-Group [20]. Differing from CM-Group, we increase the time intervals between adjacent frames within the same video, allowing for larger relative position changes of pedestrians across different images and thus increasing the complexity. We also select video frames with more complex scenarios for person search tasks. Our MMPS dataset contains more images than the existing mainstream person search datasets, with detailed statistics and comparisons shown in Table 1. Moreover, compared to other visible-infrared person Re-ID datasets, MMPS is more challenging. First, visible-infrared person search

needs to handle issues of occlusion. These occlusions may arise from the background (Fig. 2a) or from other pedestrians (Fig. 2b). Second, the interference of irrelevant backgrounds is also a challenge, as some backgrounds are quite similar to real pedestrians, and we need to eliminate such disturbances (Fig. 2c). Third, our dataset also contains some perspective distortion challenges (Fig. 2d). Besides these challenges, visible-infrared person search also faces similar challenges to person Re-ID, such as scale variations (Fig. 2a) and clothes changes (Fig. 2b). For the selected video frames, we employ PP-YOLOE [21] to obtain preliminary pedestrian bounding boxes. Subsequently, we manually correct any erroneous or missed bounding boxes. Finally, we perform cross-camera associations for pedestrians and provide labels.

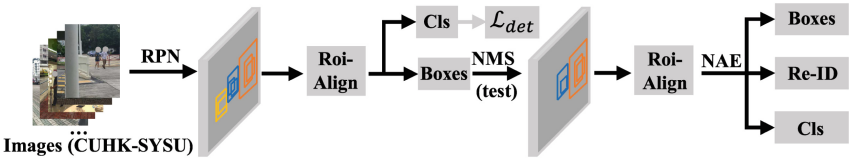
3.3 Evaluation Protocol

Our MMPS dataset is randomly divided into a training set and a test set, with a near 1:1 ratio. There is no overlap in identities between the training and test sets, and the ratio of the number of identities in the training set to that in the testing set is approximately 1:1. The test set is further divided into two subsets: one composed of infrared images, and the other of visible images. MMPS has two different evaluation settings: the Infrared-Visible mode and the Visible-Infrared mode. In the Infrared-Visible mode, all cropped pedestrian images from the infrared subset serve as the query, while all original images from the visible subset form the gallery for search operations. In the Visible-Infrared mode, the roles of the subsets are swapped. To evaluate the performance of the models, we utilize the Cumulative Matching Characteristics (CMC) curve at Rank-1 (R1), Rank-5 (R5), Rank-10 (R10), and the mean Average Precision (mAP) metric. For each pedestrian query, a match is deemed successful only when the identity of the pedestrian located by the algorithm in the retrieved frame corresponds with the query.

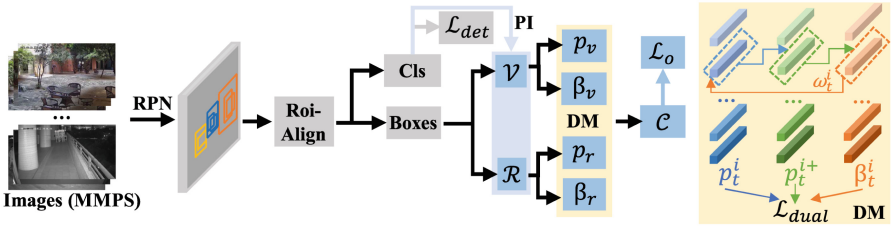
4 Method

4.1 Preliminary

In this work, we adopt the typical person search framework SeqNet [11], which achieves sequential person search with two sub-networks. Specifically, this framework is developed upon the Faster R-CNN architecture [15], as shown in Fig. 3a. The first sub-network, a standard Faster R-CNN head, is used as a Region Proposal Network (RPN). This can generate a set of proposals, including positive samples for each pedestrian and some negative samples corresponding to the background. The second head NAE [4] is utilized to further refine these bounding boxes and extract their distinctive features.



(a) SeqNet.



(b) Our proposed method PI-DM.

Fig. 3. Introduction of the (a) SeqNet and (b) our method PI-DM. Our method preserves part of the SeqNet (gray part), then incorporates Progressive Inclusion (PI) and Discriminative Mix (DM). (Color figure online)

4.2 Progressive Inclusion

Our motivation is that during training, the model's comprehension of pedestrian and background features evolves progressively. As shown in Fig. 3b, we keep part of the SeqNet (gray part). This part generates a substantial number of proposals that could potentially contain pedestrians or backgrounds. For each proposal, the network provides a prediction score (Cls) with respect to person detection. During the initial stages of training, the model might not be adequately robust. To lay the groundwork for a stable feature space, we meticulously filter the proposals. This prevents the excessive inclusion of background samples from compromising the learning of pedestrian features. As the training advances to the later stages, our model develops a refined understanding of pedestrian features. To enhance the feature discrimination capabilities within complex backgrounds and to emulate the genuine environmental feature distribution, we progressively introduce more background proposals for subsequent mixing. Our method initiates by mapping the predicted values within the $[0, 1]$ range, with higher values suggesting a higher likelihood that the proposal contains a real pedestrian. To facilitate the progressive inclusion of background samples, we then set a threshold τ that gradually adjusts, allowing for background samples exceeding this threshold to be forwarded for subsequent network processing:

$$\tau = \eta_0 e^{-\gamma/\Gamma}, \tag{1}$$

where γ/Γ represents the current/total training epoch and η_0 is the hyperparameter. We feed both the background proposals that meet the condition and

other positive proposals into the subsequent network for processing. While maintaining a comprehensive understanding of pedestrian features, our method also enhances its capability to handle backgrounds, thereby improving the overall performance of the model. For the sake of clarity, we denote the visible and infrared modalities as $t \in \{v, r\}$, except when specified otherwise. Suppose our preceding network generates two sets of proposals $\mathcal{V}$ and $\mathcal{R}$:

$$\mathcal{V} = \{X_v^i\}_{i=1}^{N_v} = \{P_v^i\}_{i=1}^{N_{v(p)}} \cup \{B_v^i\}_{i=1}^{N_{v(b)}}, \tag{2}$$

$$\mathcal{R} = \{X_r^i\}_{i=1}^{N_r} = \{P_r^i\}_{i=1}^{N_{r(p)}} \cup \{B_r^i\}_{i=1}^{N_{r(b)}}. \tag{3}$$

In Eq. 2 and Eq. 3, X represents all obtained proposals. $\mathcal{V}$ contains $N_{v(p)}$ visible pedestrian proposals (P_v^i) and $N_{v(b)}$ visible background proposals (B_v^i). $\mathcal{R}$ contains $N_{r(p)}$ infrared pedestrian proposals (P_r^i) and $N_{r(b)}$ infrared background proposals (B_r^i). The features $\mathcal{F}$ corresponding to these proposals are:

$$\mathcal{F} = \{f_v^i\}_{i=1}^{N_{v(p)}+N_{v(b)}} \cup \{f_r^i\}_{i=1}^{N_{r(p)}+N_{r(b)}}. \tag{4}$$

4.3 Discriminative Mix

We utilize the proposals from our Progressive Inclusion as samples for Discriminative Mix. We first split the proposal feature maps into l patterns. Different patterns focus on the distinctive features of various parts of pedestrians. Specifically, we use $\mathcal{P}(\cdot)$ to generate initial human pattern maps. These are subsequently passed through a sigmoid function $\sigma(\cdot)$, which yields the corresponding pattern map probabilities $F_t^i(k)$. This process can be mathematically represented as:

$$\{F_t^i(k)\}_{k=1}^{l} = \sigma(\mathcal{P}(f_t^i)), \tag{5}$$

where $\mathcal{P}$ is a convolution with kernel size 1. Following this, the pattern map embeddings $\mathbf{x}_t^i$ are computed as below:

$$\mathbf{x}_t^i = [\mathbf{x}_t^i(k)]_{k=1}^{l} = [\mathcal{G}(F_t^i(k) \odot f_t^i)]_{k=1}^{l}. \tag{6}$$

Herein, $\odot$ denotes element-wise multiplication, $\mathcal{G}(\cdot)$ denotes Global Average Pooling, and $[\cdot]$ is concatenate operation. Note that $\mathbf{x}$ in Eq. 6 is not the same as X in Eq. 2 and Eq. 3. The pattern map embeddings $\mathbf{x}_t^i$ refer to features, whereas X in Eq. 2 and Eq. 3 represent proposals. During training, within the current batch, we define the multi-pattern distance descriptor $\tilde{\Theta}(\mathbf{x}_t^i, \mathbf{x}_t^j)$ between two embeddings $\mathbf{x}_t^i$ and $\mathbf{x}_t^j$:

$$\tilde{\Theta}(\mathbf{x}_t^i, \mathbf{x}_t^j) = [\Theta(\mathbf{x}_t^i(1), \mathbf{x}_t^j(1)), \cdots, \Theta(\mathbf{x}_t^i(l), \mathbf{x}_t^j(l))], \tag{7}$$

where $\Theta(\cdot, \cdot)$ represents the Euclidean distance between two features. Then our method creates new positive and negative samples by exchanging the specified pattern map between particular samples. This not only simulates occlusions, but also enables the model to capture discriminative features of diverse human

body parts. For the given pedestrian pattern map embeddings $\mathbf{x}_t^i = \mathbf{p}_t^i$, we first identify the most similar one $\boldsymbol{\beta}_t^{i'}$ within the background samples (similarity is determined by the Euclidean distance between the entire embeddings). Upon identifying the corresponding $\boldsymbol{\beta}_t^{i'}$ for $\mathbf{p}_t^i$, we delve into exploring the specific pattern map between them:

$$\omega_t^i = \arg \max_{k \in \{1,2,\ldots,l\}} \tilde{\Theta}(\mathbf{p}_t^i, \boldsymbol{\beta}_t^{i'})[k], \tag{8}$$

where ω_t^i is the index of the pattern map that we aim to mix. Mixing the most dissimilar pattern enables the model in focusing on more diverse human parts. Then we seek out the hardest positive sample $\mathbf{p}_t^{i+}$ (the most dissimilar pedestrian sample with the same ID as $\mathbf{p}_t^i$) within the current batch. We conduct following exchanges to construct new samples:

$$\begin{aligned} \mathbf{p}_t^{i\prime} &= [\mathbf{p}_t^i(1), \cdots, \boldsymbol{\beta}_t^{i'}(\omega_t^i), \cdots, \mathbf{p}_t^i(l)], \\ \mathbf{p}_t^{i+\prime} &= [\mathbf{p}_t^{i+}(1), \cdots, \mathbf{p}_t^i(\omega_t^i), \cdots, \mathbf{p}_t^{i+}(l)], \\ \boldsymbol{\beta}_t^{i'\prime} &= [\boldsymbol{\beta}_t^{i'}(1), \cdots, \mathbf{p}_t^{i+}(\omega_t^i), \cdots, \boldsymbol{\beta}_t^{i'}(l)]. \end{aligned} \tag{9}$$

To distinguish different pedestrians while simultaneously differentiating between foreground and background, we employ a dual-task harmonization loss $\mathcal{L}_{dual}$, which is designed to better address the challenge of person search.

$$\mathcal{L}_{dual} = \alpha \mathcal{L}_1 + (1 - \alpha)\mathcal{L}_2, \tag{10}$$

$$\mathcal{L}_1 = \left[\left\| \mathbf{p}_t^{i\prime} - \mathbf{p}_t^{\star\prime} \right\| - \left\| \mathbf{p}_t^{i\prime} - \boldsymbol{\beta}_t^{i'\prime} \right\| + m \right]_+, \tag{11}$$

$$\mathcal{L}_2 = \left[\left\| \mathbf{p}_t^{i\prime} - \mathbf{p}_t^{i+\prime} \right\| - \left\| \mathbf{p}_t^{i\prime} - \mathbf{p}_t^{i-\prime} \right\| + m \right]_+, \tag{12}$$

where m controls the margin and α is the hyperparameter to balance two terms. $[\cdot]_+$ means $max(\cdot, 0)$. $\|\cdot\|$ denotes the Euclidean norm of the features. $\mathbf{p}_t^{i-\prime}$ is the pedestrian sample from other pedestrian identities that is most similar to $\mathbf{p}_t^{i\prime}$, while $\mathbf{p}_t^{\star\prime}$ is the most dissimilar pedestrian sample to $\mathbf{p}_t^{i\prime}$ among all pedestrian samples. This can effectively leverage a diverse range of complex positive and negative samples, thereby mitigating overfitting to specific parts and modalities. Our Discriminative Mix can be categorized into two types: intra-modality and inter-modality. With $\mathbf{p}_v^i$ as an example, for inter-modality mix, we identify $\boldsymbol{\beta}_r^{i'}$ and $\mathbf{p}_r^{i+}$ in another modality and then execute the mix. Similarly, for $\mathbf{p}_v^{i\prime}$ in Eq. 11 and Eq. 12, we search in another modality to find corresponding $\mathbf{p}_r^{\star\prime}$ and $\mathbf{p}_r^{i-\prime}$.

Besides the above optimizations, in $\mathcal{C}$ of Fig. 3b, we also apply center cluster loss $\mathcal{L}_{cc}$, identity classification loss $\mathcal{L}_{id}$, separation loss $\mathcal{L}_{sep}$, modality-specific ID loss $\mathcal{L}_{sid}$, and modality learning loss $\mathcal{L}_{MM}$ in MPANet [18] with coefficients of 1, 1, 0.5, 0.5, 2.5. These five losses are combined into $\mathcal{L}_o$, and together with our specially designed $\mathcal{L}_{dual}$, they jointly optimize the model.

Table 2. Comparison of CMC (%) and mAP (%) performance with state-of-the-art person search methods on our MMPS dataset.

Methods	Publication	Visible-Infrared				Infrared-Visible			
		mAP	R1	R5	R10	mAP	R1	R5	R10
SeqNet [11]	AAAI 2021	17.2	28.7	41.6	49.4	17.4	31.2	43.6	50.8
COAT [25]	CVPR 2022	18.0	32.3	44.3	50.7	18.2	33.0	48.1	52.3
PI-DM	–	36.1	51.2	70.6	77.1	36.7	51.9	71.6	77.9

5 Experiment

5.1 Implementation and Performance

Our model implementation utilizes PyTorch [14], with all experiments carried out on a single NVIDIA RTX 8000 GPU. The backbone network comprises a ResNet50 architecture [9] pre-trained on the ImageNet dataset [6]. During training, we follow SeqNet [11] and train for 25 epochs. As shown in Table 2, we have conducted a comparison of the performance with current mainstream person search methods on the MMPS dataset. It is evident that our method PI-DM achieves substantial improvement compared to state-of-the-art techniques. Methods such as SeqNet largely limit themselves to comparing Re-ID features corresponding to the input pedestrian proposals with those stored in the memory bank, without sufficiently exploring the relationship with Re-ID features within the same batch. However, in the cross-modality context, we need to deeply associate instances of different modalities within the same batch, effectively minimizing inter-modality discrepancies. Compared to the current occlusion-handling method COAT [25], our method demonstrates significant improvements in both Visible-Infrared mode and Infrared-Visible mode, as observed in the R1/mAP metrics. Infrared images exhibit lower contrast, making occlusions and background interference more challenging to address; as a result, method COAT underperforms. In contrast, our proposed PI-DM method progressively explores background information, mixing specific pattern map of particular proposals. Through our designed dual-task harmonization loss, PI-DM can focus on a broader range of human parts and capture more precise discriminative features under modality discrepancies.

5.2 Ablation Study

To further demonstrate the effectiveness of our method, we conduct ablation experiments in Table 3. In the first row of Table 3, S* indicates that we retain the gray part of SeqNet in Fig. 3 and apply the loss $\mathcal{L}_o$. S* does not utilize our Progressive Inclusion strategy (PI) and Discriminative Mix strategy (DM). The effectiveness of each component is revealed. The comparison between the third and fifth rows indicates that our PI strategy respectively improves the mAP/R1 accuracy by 0.6%/1.2% on Visible-Infrared mode, 0.4%/0.4% on Infrared-Visible

Table 3. Ablation experiments for different components on our MMPS dataset.

Methods	MMPS			
	Visible-Infrared		Infrared-Visible	
	mAP	R1	mAP	R1
S*	33.6	48.0	34.3	49.2
PI-DM w/o DM	34.9	49.3	35.2	50.1
PI-DM w/o PI	35.5	50.0	36.3	51.5
PI-DM (intra)	35.4	50.0	36.1	51.4
PI-DM	36.1	51.2	36.7	51.9

mode. The comparison between the second and fifth rows demonstrates that our DM strategy also enhances the model's performance. Specifically, on Visible-Infrared mode, DM improves the mAP/R1 accuracy by 1.2%/1.9%; on Infrared-Visible mode, DM improves the mAP/R1 accuracy by 1.5%/1.8%. To compare the effectiveness of intra-modality and inter-modality strategies within our DM approach, we conduct an experiment on the intra-modality strategy, as shown in the fourth row (whereas other rows utilized the inter-modality strategy if using DM). The results from the fourth and fifth rows reveal that employing the inter-modality strategy leads to a more substantial improvement in performance. This is attributable to the fact that inter-modality exploration is more effective at generating diverse samples for visible-infrared person search, enabling the model to better learn distinctive features from different human parts.

5.3 Discriminative Mix

We also compare the performance of our Discriminative Mix (DM) strategy on the visible-infrared person Re-ID task. We conduct experiments on the mainstream visible-infrared dataset SYSU-MM01 [17]. This dataset has two different evaluation settings: the All-search(A) mode and the Indoor-search(I) mode.

Since it is difficult for us to obtain appropriate background samples on this dataset, we set α in Eq. 10 to 0. Moreover, during the mix operation, we only consider exchanges between pedestrian pattern maps, that is, we replace $\boldsymbol{\beta}_t^{i'}$ in Eq. 8 with hard negative pedestrian samples $\mathbf{p}_t^{i-}$. In Eq. 9, we only consider the mix between $\mathbf{p}_t^i$ and $\mathbf{p}_t^{i-}$, after which we proceed with subsequent operations. As shown in Table 4, our strategy achieves satisfactory performance on both two modes, surpassing methods such as CAJL [24], MPANet [18], CMTR [12], and PMT [13]. This demonstrates that our DM strategy can also effectively extend to gain richer samples on the visible-infrared person Re-ID task, enabling the model to better learn discriminative features in different pattern maps. In contrast, the features learned by other methods lack robustness. For instance, CAJL may introduce potential information loss and significant uncertainty with its random strategy on data augmentation. This hinders the model from stably learning modality-specific features.

Table 4. Comparison with other person Re-ID methods on SYSU-MM01 dataset. This dataset has two testing modes: A (All-search) mode and I (Indoor-search) mode.

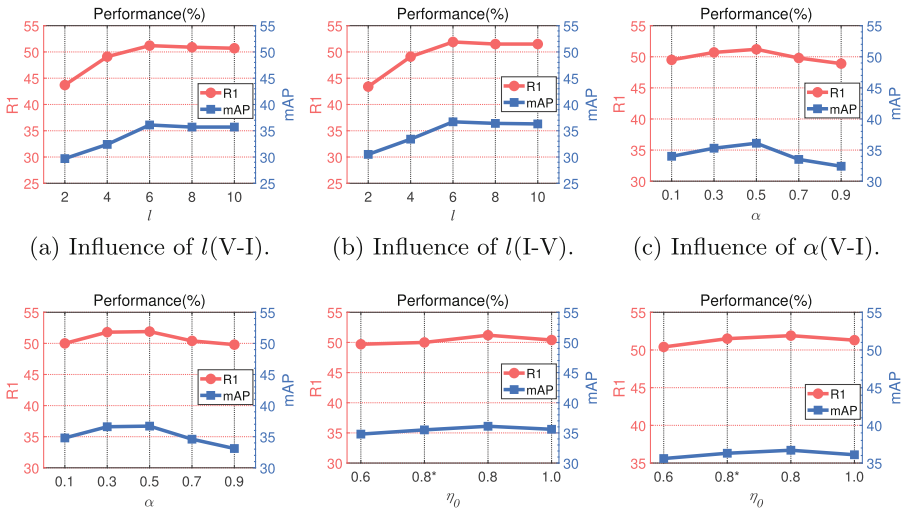
	CAJL [24]	MPANet [18]	FMCNet [28]	CMTR [12]	TransVI [2]	PMT [13]	**Ours**
mAP(A)	66.9	68.2	62.5	62.9	68.6	65.0	**69.9**
R1(A)	69.9	70.6	66.3	65.5	71.4	67.5	**73.1**
mAP(I)	80.4	81.0	74.1	76.7	81.3	76.5	**81.6**
R1(I)	76.3	76.7	68.2	71.5	77.4	71.7	**77.9**

5.4 Hyperparameter Analysis

Influence of l. l represents the total number of pattern maps. We test our method using different number of pattern maps, and the results of mAP and R1 are shown in Fig. 4a and Fig. 4b. Our evaluations are also divided into two modes: Visible-Infrared (V-I) mode and Infrared-Visible (I-V) mode. When l is relatively small, the performance of the model tends to be unsatisfactory. As l increases gradually, there is a corresponding enhancement in the accuracy of the model. The model reaches optimal performance when $l = 6$, after which point there is a slight decrease. Hence, we set l to 6 for all remaining experiments in our paper.

Influence of α. We investigate the impact of the hyperparameter α in Eq. 10, which is tasked with striking a balance between the $\mathcal{L}_1$ and $\mathcal{L}_2$ losses. We experiment with a series of candidate values, namely $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, in order to scrutinize their effects on the metric of R1 and mAP, as presented in Fig. 4c and Fig. 4d. We conduct tests on both two modes: V-I mode and I-V mode. Our observations reveal that both losses contribute positively to the learning of Re-ID features, with $\alpha = 0.5$ culminating in the most effective performance.

Influence of η_0. In Fig. 4e and Fig. 4f, we compare different initial threshold values of η_0, set to $0.6, 0.8, 1.0$, respectively. Experiment results indicate that when η_0 is set to 0.8, the model performance is optimal on V-I mode and I-V mode. Excessively high thresholds may result in insufficient background samples being introduced, lacking adequate background information to construct a richer sample set. When the threshold is set too low, too much noise may be introduced in the early stages of training, undermining the model's stability. Additionally, we compare the effects of setting the threshold to a fixed value of 0.8, as represented by the horizontal coordinate marked with 0.8* in Fig. 4e and Fig. 4f. Setting a constant threshold might lead to an insufficient diversity of samples in the later stages. When the threshold is set to a fixed value, the model is unable to progressively explore the background samples. In the latter stages of training, once the model has stabilized, the background samples cannot be properly expanded, preventing the model from exploring a richer set of samples to enhance its robustness.



(a) Influence of l(V-I). (b) Influence of l(I-V). (c) Influence of α(V-I).

(d) Influence of α(I-V). (e) Influence of η_0(V-I). (f) Influence of η_0(I-V).

Fig. 4. Influence of hyperparameter.

(a) Visible-Infrared mode. (b) Infrared-Visible mode.

Fig. 5. Visualizing of the top five retrieval results in two modes. Each row represents a different method. Green/red bounding boxes represent correct/wrong matchings. (Color figure online)

5.5 Visualization

In Fig. 5, we enumerate two visual retrieval examples of different methods (SeqNet [11] and COAT [25]) on person search. It is evident that our approach performs better in terms of search compared to existing methods. The SeqNet and COAT methods struggle to effectively address the differences between modalities and are unable to capture more precise discriminative information for search purposes. In Fig. 5b, these methods fail to properly deal with occlusions under different modalities. Likewise, in Fig. 5a, these methods show limitations

when different pedestrian features are in close spatial proximity. In contrast, our approach is capable of overcoming these challenges. Furthermore, the comparative results in the third and fourth rows of Fig. 5a and Fig. 5b validate the efficacy of our DM strategy.

6 Conclusion

In this paper, we propose a novel method for person search, integrating detection and Re-ID into a progressive process. Specifically, we first construct a new large-scale visible-infrared person search dataset MMPS. Furthermore, we propose Progressive Inclusion to adaptively provide proposals. The Discriminative Mix strategy is proposed to mix specific feature pattern maps, simulating occlusions and providing a more diverse sample pool. This enables our model to effectively identify a variety of human parts and learn discriminative features better. Extensive experiments demonstrate that our method achieves state-of-the-art performance on visible-infrared person search.

Acknowledgements. This project was supported in part by the National Natural Science Foundation of China (62076258, U22A2095) and the Key-Area Research and Development Program of Guangzhou (202206030003).

References

1. Cao, J., et al.: PSTR: end-to-end one-step person search with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9458–9467 (2022)
2. Chai, Z., Ling, Y., Luo, Z., Lin, D., Jiang, M., Li, S.: Dual-stream transformer with distribution alignment for visible-infrared person re-identification. IEEE Trans. Circuits Syst. Video Technol. (2023)
3. Chen, D., Zhang, S., Ouyang, W., Yang, J., Tai, Y.: Person search via a mask-guided two-stream CNN model. In: Proceedings of the European Conference on Computer Vision, pp. 734–750 (2018)
4. Chen, D., Zhang, S., Yang, J., Schiele, B.: Norm-aware embedding for efficient person search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12615–12624 (2020)
5. Chen, H., Zhang, Q., Lai, J., Xie, X.: Unsupervised group re-identification via adaptive clustering-driven progressive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 1054–1062 (2024)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
7. Dong, W., Zhang, Z., Song, C., Tan, T.: Instance guided proposal network for person search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2585–2594 (2020)
8. Han, C., et al.: DMRNet++: learning discriminative features with decoupled networks and enriched pairs for one-step person search. IEEE Trans. Pattern Anal. Mach. Intell. (2022)

9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an X modality. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 4610–4617 (2020)
11. Li, Z., Miao, D.: Sequential end-to-end network for efficient person search. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2011–2019 (2021)
12. Liang, T., Jin, Y., Liu, W., Li, Y.: Cross-modality transformer with modality mining for visible-infrared person re-identification. IEEE Trans. Multimed. (2023)
13. Lu, H., Zou, X., Zhang, P.: Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1835–1843 (2023)
14. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
16. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
17. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: RGB-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380–5389 (2017)
18. Wu, Q., et al.: Discover cross-modality nuances for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4330–4339 (2021)
19. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3415–3424 (2017)
20. Xiong, J., Lai, J.: Similarity metric learning for RGB-infrared group re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13662–13671 (2023)
21. Xu, S., et al.: PP-YOLOE: an evolved version of YOLO. arXiv preprint arXiv:2203.16250 (2022)
22. Yan, Y., et al.: Exploring visual context for weakly supervised person search. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 3027–3035 (2022)
23. Yan, Y., et al.: Anchor-free person search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7690–7699 (2021)
24. Ye, M., Ruan, W., Du, B., Shou, M.Z.: Channel augmented joint learning for visible-infrared recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13567–13576 (2021)
25. Yu, R., et al.: Cascade transformers for end-to-end person search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7267–7276 (2022)
26. Yu, X., Gong, Y., Jiang, N., Ye, Q., Han, Z.: Scale match for tiny person detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1257–1265 (2020)
27. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: CutMix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)

28. Zhang, Q., Lai, C., Liu, J., Huang, N., Han, J.: FMCnet: feature-level modality compensation for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7349–7358 (2022)
29. Zhang, Q., Lai, J., Feng, Z., Xie, X.: Seeing like a human: asynchronous learning with dynamic progressive refinement for person re-identification. IEEE Trans. Image Process. **31**, 352–365 (2021)
30. Zhang, Q., Lai, J., Xie, X.: Learning modal-invariant angular metric by cyclic projection network for VIS-NIR person re-identification. IEEE Trans. Image Process. **30**, 8019–8033 (2021)
31. Zhang, Q., Wang, L., Patel, V.M., Xie, X., Lai, J.: View-decoupled transformer for person re-identification under aerial-ground camera network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22000–22009 (2024)
32. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3346–3355 (2017)

Merging Multiple Datasets for Improved Appearance-Based Gaze Estimation

Liang Wu(✉) and Bertram E. Shi

Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
liang.wu@connect.ust.hk, eebert@ust.hk

Abstract. Multiple datasets have been created for training and testing appearance-based gaze estimators. Intuitively, more data should lead to better performance. However, combining datasets to train a single estimator rarely improves gaze estimation performance. One reason may be differences in the experimental protocols used to obtain the gaze samples, resulting in differences in the distributions of head poses, gaze angles, illumination, etc. Another reason may be the inconsistency between methods used to define gaze angles (label mismatch). We propose two innovations to improve the performance of gaze estimation by leveraging multiple datasets, a change in the estimator architecture and the introduction of a gaze adaptation module. Most state-of-the-art estimators merge information extracted from images of the two eyes and the entire face either in parallel or combine information from the eyes first then with the face. Our proposed *Two-stage Transformer-based Gaze-feature Fusion* (TTGF) method uses transformers to merge information from each eye and the face separately and then merge across the two eyes. We argue that this improves head pose invariance since changes in head pose affect left and right eye images in different ways. Our proposed *Gaze Adaptation Module* (GAM) method handles annotation inconsistency by applying a Gaze Adaption Module for each dataset to correct gaze estimates from a single shared estimator. This enables us to combine information across datasets despite differences in labeling. Our experiments show that these innovations improve gaze estimation performance over the SOTA both individually and collectively (by 10%–20%). Our code is available at https://github.com/HKUST-NISL/GazeSetMerge.

Keywords: gaze estimation · transformers · feature fusion · multi-dataset training

1 Introduction

Estimation of human gaze plays important roles in many applications, such as human-computer interaction [2,3], virtual reality [1], attention analysis [4,5] and psychological studies [6].

Conventional methods, such as those based on pupil center corneal reflections (PCCR), use 3D eye models to compute the gaze direction [11]. These

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 77–90, 2025.
https://doi.org/10.1007/978-3-031-78341-8_6

require special measurement setups, such as active infrared illumination, to estimate model geometry. In contrast, appearance-based gaze estimators use input from commonly available RGB web cameras, which are more convenient and less expensive. Unfortunately, estimates from them are less accurate than those from PCCR-based systems. The current lowest reported within-person error of gaze estimation is 4.04° [42] on MPIIFaceGaze. In contrast, manufacturers of PCCR-based systems typically report accuracies of less than one degree.

However, the gap between the two continues to shrink, most recently due to the use of Convolutional Neural Networks (CNN) [7–9] and transformers. Many CNN architectures have been proposed for appearance-based gaze estimation. Zhang et al. employed a multi-modal model that used eye images and an estimated head pose vector as inputs to estimate gaze direction [7]. Later, they applied spatial weighting to feature maps from the face image to enhance information from eye regions [8]. Other studies used three separate pipelines to extract features from images of the head and the two eyes and then fused them to predict the gaze [9,12]. Merging information from the eyes and the face improves estimation accuracy.

Since appearance-based gaze estimators rely heavily on training data, many datasets have been proposed to train gaze estimators. Initial datasets were collected under fairly well-controlled and limited conditions (e.g., ranges of head poses and gaze angles). More recent datasets have been collected on conditions of increased diversity. The availability of more data can potentially increase the performance of appearance-based gaze estimators, but can also introduce new challenges. This paper seeks to address two of these challenges.

First, increases in the head pose range have spurred the development of new architectures that combine information from images of the two eye regions (which primarily indicate gaze direction in head-centric coordinates) and an image of the entire face (which primarily indicates head pose). Many SOTA (state-of-the-art) methods combine this information in parallel [9], or combine information from the eyes first followed by the face image [12].

To improve upon these approaches, we propose a Two-stage Transformer-based Gaze-feature Fusion (TTGF) architecture, which combines information from each eye image with the face image separately and then integrates information across the two eyes. This approach is motivated by the fact that the head-centric gaze directions of the two eyes differ and should thus each be merged with the face image. This may also compensate for situations where the reliability of information from the two eyes may differ, e.g., due to occlusion.

Second, although intuitively increasing the amount of data by combining datasets should improve performance, inconsistencies in annotation among datasets make it difficult to improve accuracy by simply combining multiple gaze datasets. To provide a normalized gaze annotation, a common scheme is to rotate the gaze vector from the gaze origin to the target point by a rotation matrix that depends upon the head pose [31]. Differences between the methods for head pose estimation and target point estimation lead to inconsistency among different datasets. Even when the subject's head is constrained by a chin rest [29], head pose estimation error can still exist due to the placement of the subject's head in the chin rest.

To address this, we propose the use of a Gaze Adaption Module (GAMs) for each dataset, which adjusts the gaze label from a shared estimator so it is consistent with the dataset of the source image. This enables multi-dataset training by simply adding GAM to the model's gaze regression head.

Our experimental results demonstrate that these two innovations lead to state-of-the-art performance on multiple datasets, under training with both single datasets and mixed datasets.

2 Related Work

Gaze Estimation Methods. Gaze estimation methods can typically be categorized as either model-based or appearance-based. Model-based methods usually construct the 3D model of the head and eyes. The gaze direction is calculated by utilizing geometric information [11,13–15]. Model-based methods usually require time-consuming personal calibration to fit the subject-specific parameters, such as cornea radius and kappa angles.

In contrast, appearance-based methods directly learn mapping functions from a large number of image-gaze sample pairs. Early approaches used conventional regression to perform the mapping [16–18]. More recently, CNNs have significantly improved the performance of appearance-based gaze estimation. Zhang et al. proposed the first CNN-based network to regress the gaze direction from a cropped eye image, and a head pose vector [7]. They later proposed to use the learnable spatial weights to enhance the information from the eye regions in the face image [8]. Krafka et al. proposed iTracker, a multi-region CNN model, which takes both the head and eye images as input. To further improve the accuracy, Chen et al. investigated the dilated convolution layers to efficiently increase the receptive field sizes of the features [9]. Researchers have now started to use transformer-based networks, which can further improve gaze estimation accuracy [19–21].

Transformers. The Transformer architecture was first introduced by Vaswani et al. for natural language processing [38]. It consists of self-attention layers, layer normalization, and multi-layer perceptron layers. Compared with recurrent networks, the global computations and efficient memory of self-attention layers make transformers more suitable for long sequences.

The Vision Transformer (ViT) was proposed by Dosovitskiy et al. for image classification tasks [37]. ViT divides one image into non-overlapping patches. A transformer encoder is applied to the features extracted from the patches. Transformers have achieved state-of-the-art in large-scale image classification tasks, leading to their application to many other vision tasks [39–41].

Recently, a few researchers have explored the capability of transformers in gaze estimation. Cheng et al. proposed GazeTR-Hybrid where they used convolutional neural networks to extract the feature map of an input head image, then treated the features at different positions as a sequence of features input to a transformer encoder [19]. Cai et al. proposed iTracker-MHSH [21]. Inspired by iTracker, it uses a transformer to integrate the features of the head and eye images.

Mixed Dataset Training. There are two main advantages to mixed dataset training. First, it provides a single model applicable to multiple datasets. Second, model training may benefit from the increased amount of data. Mixed dataset training has been applied to many computer vision tasks, such as person reidentification [22,23], monocular depth estimation [24], semantic image segmentation [25,26], video quality assessment [27,30] and 3D object detection [28]. Addressing the challenges of mixed dataset training is task-specific. For example, to mix image segmentation datasets, category merging needs were conducted before training [25,26]. For video quality assessment [27], the challenge was to resolve inconsistent ranges of subjective quality scores across datasets.

To the best of our knowledge, we are the first to propose mixed dataset training for gaze estimation. There are two challenges that must be addressed. First, the distribution of gaze vectors and head poses varies between different gaze datasets. Second, there exists annotation inconsistency in gaze vectors from different gaze datasets.

3 Annotation Inconsistency

The gaze vector is defined as the vector starting from the gaze origin to the gaze target. Gaze dataset collection requires an experimental setup to capture three types of information in camera coordinates: 1) the position of the visual target P_t, 2) the position of gaze origin P_o, and 3) the head pose R [32]. However, different datasets utilize different methods to get these values, leading to different annotations.

Inconsistency in Gaze Target Estimation. Usually, the visual target is indicated by a moving dot on a screens. To determine the position of the dot target, the intrinsic parameters of the camera must be obtained beforehand. MPIIGaze uses a mirror-based calibration method [32] to estimate the 3D positions of each screen plane. Finally, the position of the moving dot is computed based on the screen size and resolution. In addition to a moving dot on the screen, EYEDIAP has an additional floating ball visual target. Its position is estimated first in an RGB-D sensor coordinate system and then transformed to the camera coordinate system. Imprecision in the RGB-D sensor, errors in the screen-to-camera calibration and RGB-D to-camera calibration will all contribute to the inconsistency of the gaze target position p_t.

Inconsistency in Gaze Origin and Head Pose Estimation. There are inconsistencies between datasets in the selection of gaze origin and the estimation of head pose. In early work, gaze was estimated eye images, where the eye center defined the gaze origin [7,35,36]. More recently, people estimate gaze from the whole head image, where the gaze origin is usually set at the center of the head [10,33,34]. To get the 3D head pose, MPIIGaze and ETH-XGaze detect landmarks from the 2D head image and fit a 3D morphable model of the head to the detected landmarks. EYEDIAP directly uses the depth data from the RGB-D sensor to fit a 3D Morphable Model.

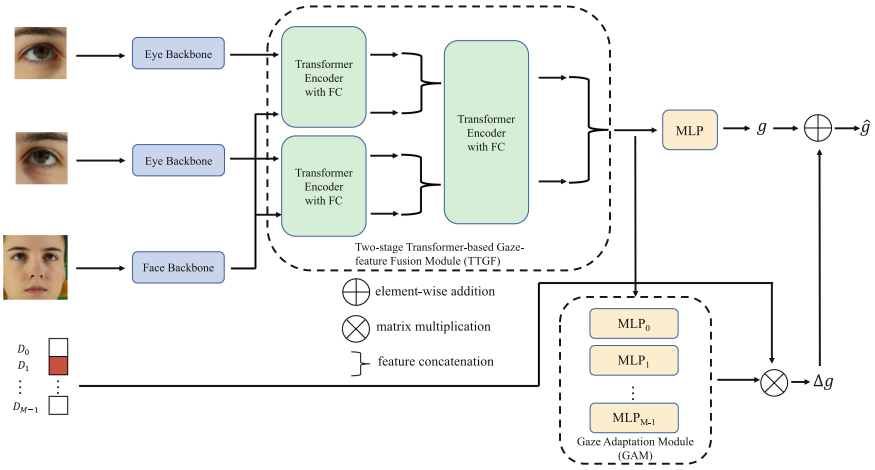


Fig. 1. The proposed framework contains two modules: 1) TTGF and 2) GAM. The TTGF applies two-stage feature fusion to the features of the head and eyes with transformers, and the GAM produces a gaze offset to adjust the predicted gaze for mixed datasets training.

4 Method

Figure 1 shows our framework, which consists of an eye-head transformer-based feature fusion module for gaze estimation followed by a set of gaze adaptation modules. We described these in more detail below.

4.1 Feature Fusion with Transformers

A typical transformer encoder contains L transformer blocks, each containing multi-head self-attention (MHSA) layers, layer normalization (LN), and multi-layer perceptron layers (MLP). To process an input feature matrix $\boldsymbol{Z} \in \mathbb{R}^{n \times d}$, MHSA projects $\boldsymbol{Z}$ into $\boldsymbol{Q} \in \mathbb{R}^{n \times d_k}$, keys $\boldsymbol{K} \in \mathbb{R}^{n \times d_k}$ and values $\boldsymbol{V} \in \mathbb{R}^{n \times d_v}$ where n is the number of tokens and d, d_k, d_v are the dimension of the feature, key/query and value.

The attention is computed through the following equation:

$$\text{Attention}(\boldsymbol{Q}; \boldsymbol{K}; \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\boldsymbol{V}). \tag{1}$$

Combined with LN and MLP, the overall equations for the transformer encoder with L transformer blocks are

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \qquad l = 1...L, \tag{2}$$

$$z_l = \text{MLP}(\text{LN}(z'_l)) + z'_l, \qquad l = 1...L, \tag{3}$$

$$y = \text{LN}(z_L). \tag{4}$$

Krafka et al. proposed iTracker [12] to estimate gaze by integrating the features of the head and eyes using several fully connected layers. To better fuse features, we propose the two-stage transformer-based gaze-feature fusion (TTGF) architecture shown in Fig. 1. This architecture applies three transformer encoders to fuse the features from the head and eye images in two fusion steps, 1) head-eye fusion and 2) left-right fusion. The idea of using two-step fusion is based on the intuition combining information of the head and one eye enable rough inference of the person's gaze direction. The second step combines the two rough estimates into a single more precise estimate.

In our design, the architectures of all three fusion modules are identical. One TGF module accepts two gaze-related features and produces a fused feature. We describe the computation in ta TGF formally with the following equation:

$$\mathrm{TGF}(f^*, f^\dagger) = \mathrm{CAT}(\mathrm{FC}(\mathrm{Trans}([f^*; f^\dagger]))), \tag{5}$$

where $\mathrm{Trans}([f^*; f^\dagger])$ is the transformer used for fusing the head-eye features or eye-eye features, FC is a linear layer used to project the features to a specific size, and CAT concatenates the outputs of the transformer to generate fused features. In the head-eye fusion stage, each eye feature f^{le} or f^{re} is fused with the head feature f^h:

$$f^{lh} = \mathrm{TGF}^{lh}(f^{le}, f^h) \tag{6}$$

$$f^{rh} = \mathrm{TGF}^{rh}(f^{re}, f^h) \tag{7}$$

In the second stage, the two fused eye-head features are fed into a third TGF module to fuse features from left and right:

$$f^{lr} = \mathrm{TGF}^{lr}(f^{lh}, f^{rh}) \tag{8}$$

Finally, the fused feature f^{lr} is fed to an MLP to get the predicted gaze g:

$$g = \mathrm{MLP}(f^{lr}). \tag{9}$$

4.2 Gaze Adaptation Module

Suppose we have M gaze datasets, $D = \{D_0, D_2, ..., D_{M-1}\}$. Typically, we need to train M models: one for each dataset to get good performance. A model trained on D_i typically performs poorly on D_j where $i \neq j$.

Instead, our approach trains only one model and $M-1$ Gaze Adaptation Modules (GAMs). The GAM is a module consisting of a M MLPs, one for each dataset $i \in \{0, \ldots, M-1\}$. Each MLP, $\mathrm{MLP}_i(\cdot)$, accepts the extracted feature f^{lr} and produces a gaze offset assuming the sample comes from dataset i. D_0 is regarded as the anchor dataset, so its offset is always zero, i.e., $\mathrm{MLP}_0(\cdot) = \mathbf{0}$ and does not need to be trained. For the others, the MLP has two layers with GELU nonlinearities. If the sample comes from dataset i, the corrected gaze vector is given by $\hat{g} = g + \Delta g$, where $\Delta g = \mathrm{MLP}_i(f^{lr})$.

4.3 Architecture Details

The whole architecture contains three pipelines for the face and two eye images. All the backbones are ResNet18 networks, which are initialized from the model trained on ImageNet. The input face image size is $224 \times 224 \times 3$. We crop the eye patches according to the landmarks and use RoI align to resize the cropped patches to $128 \times 128 \times 3$. The estimated gaze contains the yaw and pitch representing the 3D gaze direction in the camera coordinate system. We chose L1 loss as the loss function for gaze estimation.

For TTGF, we set the number of heads of all MSAs as 8 and the hidden size of the MLP is 2048. We use 8 repeated blocks in each transformer encoder. After each transformer encoder, the features are projected with a linear layer whose output size is 128. For the MLPs for both gaze regression and the GAMs, the sizes of the hidden layers are identically set to 128.

5 Experiments

In this section, we introduce the experimental settings and the evaluation datasets we selected and evaluate our proposed TTGF and GAM in two types of experiments. We first compare our method with the state-of-the-art methods for gaze estimation performance. Then we perform ablation studies to determine the effects due to TTGF and GAM respectively and study the effect of multiple dataset training.

Dataset for Evaluation. For evaluating gaze estimation performance, we used three gaze datasets to evaluate the gaze estimation performance as shown in Table 1: MPIIFaceGaze [8], RT-GENE [10], and EYEDIAP [34]. MPIIFaceGaze dataset is based on MPIIGaze, but includes face and eye images. It contains 45K images collected from 15 subjects. We used leave-one-person-out cross-validation with this dataset. The RT-GENE dataset consists of 123K samples from 15 participants. We used three-fold cross-validation with this dataset. The raw data of the EYEDIAP dataset has 94 videos collected from 16 subjects. We used the sampling scheme from [33] to extract face images and four-fold cross-validation. For our experiments on multi-dataset training, we trained 15 models (one for

Table 1. Overview of the datasets used for evaluation and anchor dataset in our experiments. We show the number of subjects, the range of gaze, and the head pose in both horizontal and vertical directions in the camera coordinate systems.

Dataset	# Subjects	Gaze	Head Pose	# Data
MPIIFaceGaze [8]	15	$\pm 20°$, $\pm 20°$	$\pm 15°$, $30°$	45K images
RT-GENE [10]	15	$\pm 40°$, $-40°$	$\pm 40°$, $\pm 40°$	123K images
EYEDIAP [34]	16	$\pm 25°$, $20°$	$\pm 15°$, $30°$	94 videos
ETH-XGaze [33]	110	$\pm 120°$, $\pm 70°$	$\pm 80°$, $\pm 80°$	1.1M images

each subject left out from MPIIFaceGaze), where each person was assigned to one of the folds in the other two datasets. Performance for each fold in the other two datasets was computed by averaging the performance of the models from the MPIIFaceGaze subjects assigned to that fold.

Anchor Dataset and Pre-training. ETH-XGaze [33] is a large-scale gaze dataset that consists of 1,083,492 image samples from 110 participants (47 female and 63 male). It has the largest range of head poses compared to the evaluation dataset and the gaze direction is evenly sampled both horizontally and vertically as shown in Table 1. The large variation and scale make it a suitable dataset as the anchor dataset D_0 and for pre-training. The whole dataset contains three parts: the training set, the within-dataset, and the person-specific evaluation set. The training set has 765K images of 80 subjects. We use this part as the anchor set and also for pre-training. The person-specific evaluation consists of 15 subjects but is not related to this task. The within-dataset which includes 15 subjects is used for validation of multiple datasets training and the pre-training model.

Experimental Settings. The optimizer applied for model training is AdamW with a linear scheduled warm-up strategy. The initial learning rate is set to 0.0001 for all the training and uses the exponential schedule to update it. For multiple-set training, in each iteration, we randomly sample the same number of samples from each set to form a batch fed to the model. The batch size is set to 64. The number of iterations in one training epoch is determined by the size of the dataset with the smallest number of samples. The number of epochs is 50 and gamma is 0.96. For single-set training for the TTGF-only model, the batch size is also set to 64. For ETH-XGaze, we train the model for 50 epochs with the exponential gamma setting to 0.95. For MPIIFaceGaze and RT-GENe, the total number of epochs is 30 epochs with the exponential gamma setting to 0.95. For EYEDIAP, the number of epochs is 50 and gamma is 0.096. Our experiments are all conducted on a single GeForce RTX 3090 GPU.

5.1 Comparison with State-of-the-Art Methods

In this part, we compare the gaze estimation performance of our proposed model with state-of-the-art methods. Our model is a single model trained on multiple datasets: one anchor dataset and three evaluation datasets, while the existing methods were tested with separated models for different evaluation datasets. We trained our TTGF-only model on ETH-XGaze and got a testing error of 3.58° and the proposed TTGF+GAM trained on multiple datasets achieved a slightly better error of 3.54°.

Table 2 shows the angular errors of each method on the evaluation datasets: MPIIFaceGaze, RT-GENE, and EYEDIAP. As iTracker and iTracker-MHSA did not provide the performance on the evaluation datasets, we re-implemented them by replacing their backbones with ResNet18 for fair comparison. In the table, among existing models, FullFace [8], GazeTR [19], and GazeCADSE [42] only use the full face image as the input for gaze estimation. RT-GENE [10] feeds two

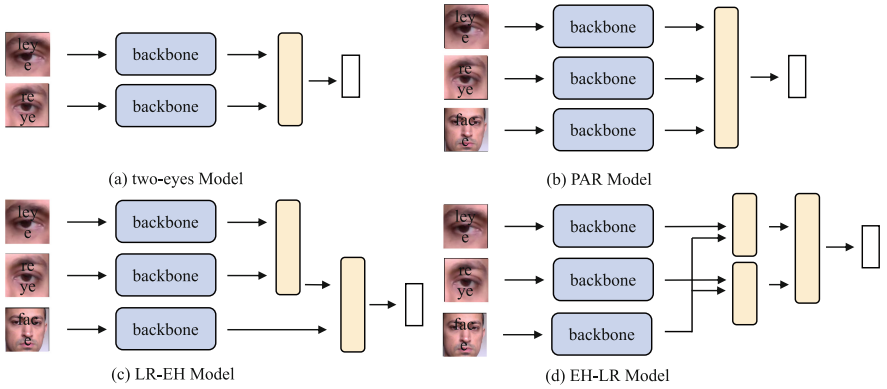


Fig. 2. Four types of feature fusion for gaze estimation models: (a) two-eyes model uses the cropped eye patches as inputs. (b) PAR indicates left eye, right eye, and head features are combined in parallel. (c) LR-EH indicates that left and right eye features are combined first then combined head features. (d) EH-LR indicates that single eye and head features are combined first followed by a combination across the left and right.

Table 2. Comparison with the state-of-the-art methods. The proposed method outperforms state-of-the-art results in estimation error.

Model	Transformer	Feature Fusion	MPIIFaceGaze	RT-GENE	EYEDIAP
FullFace [8]	NO	Face Only	4.93°	10.00°	6.53°
RT-GENE [10]	NO	Two Eyes	4.66°	8.00°	6.02°
DilatedNet [9]	NO	PAR	4.42°	8.38°	6.19°
iTracker [12]	NO	LR-EH	4.33°	7.12°	5.28°
iTracker-MHSA [21]	YES	LR-EH	4.05°	7.06°	5.17°
GazeTR-Hybrid [19]	YES	Face Only	4.18°	7.12°	5.33°
GazeCADSE [42]	YES	Face Only	4.04°	7.00°	5.25°
Proposed	YES	EH-LR	**3.88°**	**6.46°**	**4.89°**

cropped eyes to a VGG16 model. DilatedNet [9] fuses the features of the left eye, right eye, and head directly. iTracker [12], iTracker-MHSA [21] fuse the features of the left and right eyes first then with the head features. Our proposed method also uses both the face and the eye images as inputs but has different ways of feature fusion we fuse the features of each eye and head in the first stage and then fuse the left and right features in the second stage. In addition, GazeTR, GazeCADSE, and our proposed methods utilize the transformers in the model. We show different types of gaze estimation models in Fig. 2.

As shown in Table 2, our proposed methods TTGF with GAM achieved the state-of-the-art performance of gaze estimation on all the selected evaluation datasets. Among the methods using the feature fusing, our eye-head first then left-right combination shows the best performance. Overall the transformer-

based methods show advantages in the performance of gaze estimation compared with non-transformer methods. Among the transformer-based methods, our model uses both the face and eye images, we used RoI alignment to resize the eye region to 128×128, which enables the model to extract features directly from the eye patches.

Table 3. Comparison of Computational Costs.

Model	Params	FLOPs
RT-GENE [10]	82.0M	30.81G
GazeTR-Hybrid [19]	11.4M	1.82G
GazeCADSE [42]	74.8M	12.78G
proposed method	65.3M	3.03G

By using GAM, our proposed model achieves better performance on multiple datasets using only a single main model. This results in a smaller number of parameters compared with other methods. Suppose the number of parameters of the feature extractor is N and that of each gaze regressor is K. For M datasets, without GAM we need to train M models for each dataset resulting in total MN parameters. On the contrary, by applying GAM to train on multiple datasets, we only need one single model with one feature extractor, one gaze regressor and $M-1$ MLPs as the gaze offset for the anchor set is always $\mathbf{0}$. So the total number of parameters for our proposed model is $N+MK$. As K is much smaller than N, our method needs fewer parameters to achieve better performance.

Table 3 shows the number of parameters and the flops for each model. We can see that our proposed method has a fairly low computational cost which we believe is related to two reasons: 1) a relatively smaller model ResNet18 is applied as the backbone, and 2) a smaller size for the two eye patches as inputs.

5.2 Ablation Study

To study the individual contributions of the TTGF and GAM modules, we conducted ablation experiments by removing one of them from the entire framework.

Effect of TTGF. To study the TTGF, we trained a TTGF-only model on each evaluation dataset and compared the results with itracker-MHSA. We compare with itracker-MHSA because it also uses a transformer encoder to combine eye and head features in a different order. The itracker-MHSA fuses features first from the left and right eyes and then with the head feature. TTGF fuses features from each eye with head features and then across the two eyesAs we mentioned before, we re-implemented itracker-MHSA with the same backbone as our model for a fair comparison. As we mentioned before, we re-implemented itracker-MHSA with the same backbone as our model for a fair comparison.

Table 4. Ablation study.

Model	Multiple Sets	MPIIFaceGaze	RT-GENE	EYEDIAP
itracker-MHSA [21]	NO	4.05°	7.06°	5.17°
TTGF-Only	NO	3.98°	6.89°	5.11°
TTGF-Only	YES	4.12°	7.14°	5.20°
proposed method (TTGF+GAM)	YES	**3.88°**	**6.46°**	**4.89°**

Table 4 shows the angular errors of each method on the evaluation datasets. The TTGF-only model outperforms the itracker-MHSA on all evaluation datasets.

Effect of GAM. We compared our proposed model with GAM with the TTGF-only model trained on multiple datasets. Table 4 shows that with GAM the accuracy of the TTGF-only model without multiple sets of training is improved on all three datasets from 0.1° to 0.43° respectively.

To confirm the performance gain in multiple dataset training is due to the use of GAM, we trained the TTGF-only model with the combination of the ETH-XGaze and the evaluation datasets. The TTGF-only model trained on mixed datasets performed even worse than the TTGF-only model trained on each single evaluation set. This supports our claim that GAM can address the inconsistency in annotation across different datasets.

6 Conclusion

We proposed a Two-stage Transformer-based Gaze-future Fusion (TTGF) and the use of Gaze Adaption Modules (GAMs) for improving gaze estimation accuracy. The TTGF uses two-stage fusion for the features of the head and eye images through three transformer encoders. The proposed GAM generates gaze corrections to gaze estimates for one dataset (chosen here to be ETH-Gaze) to create estimates for images from other datasets. Our experiments show that our method surpasses the state-of-the-art by a significant margin. Ablation studies show that both innovations result in improvements when applied in isolation and that improvements compound when they are applied together. However, our proposed model still has some limitations. For example, the proposed TTGF needs cropped eye patches as input. The GAM does not address all issues arising from annotation inconsistency among gaze datasets.

References

1. Patney, A., et al.: Towards foveated rendering for gaze-tracked virtual reality. ACM Trans. Graph. (TOG) **35**, 1–12 (2016)
2. Chen, Z., Shi, B.: Using variable dwell time to accelerate gaze-based web browsing with two-step selection. Int. J. Hum.-Comput. Interact. **35**, 240–255 (2019)

3. Pi, J., Shi, B.: Probabilistic adjustment of dwell time for eye typing. In: 2017 10th International Conference on Human System Interactions (HSI), pp. 251–257 (2017)
4. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: Advances in Neural Information Processing Systems, vol. 28 (2015)
5. Chong, E., Wang, Y., Ruiz, N., Rehg, J.: Detecting attended visual targets in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5396–5406 (2020)
6. Gehrer, N., Schönenberg, M., Duchowski, A., Krejtz, K.: Implementing innovative gaze analytic mods in clinical psychology: a study on eye movements in antisocial violent offenders. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, pp. 1–9 (2018)
7. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4511–4520 (2015)
8. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: full-face appearance-based gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 51–60 (2017)
9. Chen, Z., Shi, B.: Appearance-based gaze estimation using dilated-convolutions. In: Asian Conference on Computer Vision, pp. 309–324 (2018)
10. Fischer, T., Chang, H., Demiris, Y.: RT-gene: real-time eye gaze estimation in natural environments. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 334–352 (2018)
11. Guestrin, E., Eizenman, M.: General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Trans. Biomed. Eng. **53**, 1124–1133 (2006)
12. Krafka, K., et al.: Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2176–2184 (2016)
13. Chen, J., Ji, Q.: 3D gaze estimation with a single camera without IR illumination. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4 (2008)
14. Valenti, R., Sebe, N., Gevers, T.: Combining head pose and eye location information for gaze estimation. IEEE Trans. Image Process. **21**, 802–815 (2011)
15. Wood, E., Bulling, A.: EyeTab: model-based gaze estimation on unmodified tablet computers. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 207–210 (2014)
16. Tan, K., Kriegman, D., Ahuja, N.: Appearance-based eye gaze estimation. In: Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision, WACV 2002, pp. 191–195 (2002)
17. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Adaptive linear regression for appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. **36**, 2033–2046 (2014)
18. Williams, O., Blake, A., Cipolla, R.: Sparse and semi-supervised visual mapping with the sundefined 3GP. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 230–237 (2006)
19. Cheng, Y., Lu, F.: Gaze estimation using transformer. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 3341–3347 (2022)
20. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: End-to-end human-gaze-target detection with transformers. arXiv Preprint arXiv:2203.10433 (2022)
21. Cai, X., et al.: Gaze estimation with an ensemble of four architectures. arXiv Preprint arXiv:2107.01980 (2021)

22. Lv, J., Chen, W., Li, Q., Yang, C.: Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7948–7956 (2018)
23. Li, Y., Lin, C., Lin, Y., Wang, Y.: Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7919–7929 (2019)
24. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 1623–1637 (2020)
25. He, H., Zhang, J., Zhang, Q., Tao, D.: Grapy-ML: graph pyramid mutual learning for cross-dataset human parsing. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020, pp. 10949–10956 (2020). https://ojs.aaai.org/index.php/AAAI/article/view/6728
26. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V. MSeg: a composite dataset for multi-domain semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2879–2888 (2020)
27. Li, D., Jiang, T., Jiang, M.: Unified quality assessment of in-the-wild videos with mixed datasets training. Int. J. Comput. Vis. **129**, 1238–1257 (2021)
28. Zhang, W., Li, W., Xu, D.: SRDAN: scale-aware and range-aware domain adaptation network for cross-dataset 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6769–6779 (2021)
29. Smith, B., Yin, Q., Feiner, S., Nayar, S.: Gaze locking: passive eye contact detection for human-object interaction. In: Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, pp. 271–280 (2013)
30. Korhonen, J.: Two-level approach for no-reference consumer video quality assessment. IEEE Trans. Image Process. **28**, 5923–5938 (2019)
31. Zhang, X., Sugano, Y., Bulling, A.: Revisiting data normalization for appearance-based gaze estimation. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, pp. 1–9 (2018)
32. Rodrigues, R., Barreto, J., Nunes, U.: Camera pose estimation using images of planar mirror reflections. In: European Conference on Computer Vision, pp. 382–395 (2010)
33. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: ETH-XGaze: a large scale dataset for gaze estimation under extreme head pose and gaze variation. In: European Conference on Computer Vision, pp. 365–381 (2020)
34. Funes Mora, K., Monay, F., Odobez, J.: Eyediap: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 255–258 (2014)
35. Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 721–738 (2018)
36. Liu, G., Yu, Y., Mora, K., Odobez, J.: A differential approach for gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 1092–1099 (2019)
37. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv Preprint arXiv:2010.11929 (2020)

38. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
39. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems, vol. 34, pp. 12077–12090 (2021)
40. Kim, B., Lee, J., Kang, J., Kim, E., Kim, H.: HOTR: end-to-end human-object interaction detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 74–83 (2021)
41. Zhang, W., et al.: Transformer-based multimodal information fusion for facial expression analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2428–2437 (2022)
42. Oh, J.O., Chang, H., Choi, S.: Self-attention with convolution and deconvolution for efficient eye gaze estimation from a full face image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4992–5000 (2022)

A Dual Adaptation Approach for EEG-Based Biometric Authentication Using the Ensemble of Riemannian Geometry and NSGA-II

Aashish Khilnani[1], Jyoti Singh Kirar[2(✉)], and Ganga Ram Gautam[1]

[1] Banaras Hindu University, Varanasi, India
{aashishkhilnani,gangacims}@bhu.ac.in
[2] Jawaharlal Nehru University, New Delhi, India
kirarjyoti@gmail.com

Abstract. Recently, it has been discovered that EEG signals have enormous potential to be used as biometric authentication. Although, its practical implementation is limited due to the intricate and dynamic nature of EEG signals. To overcome these challenges, we need to simplify the analysis and preserve the spatial attributes of the EEG signals. In this work, a methodology using an ensemble of Riemannian geometry and a genetic algorithm for EEG-based biometric authentication is devised. Here, the symmetric positive definite covariance matrices of the EEG signals are calculated and classified using the Minimum distance to the Riemannian Mean (MDRM) and the Tangent space LDA (TSLDA) classifier. Furthermore, NSGA-II is used to optimize the number of channels and to reduce the computational complexity. This study achieved an accuracy of 99.9% on average with all the datasets used. Multiple publicly available datasets are used to compare the proposed approach with existing methods. Results obtained show the efficacy of the proposed method. Friedman's statistical test also supports the statistical significance difference between the proposed and existing methods.

Keywords: Riemannian Geometry · NSGA - II · Covariance Matrices · Biometric · EEG

1 Introduction

In this era, characterized by the rapid convergence of technology with human lives, safeguarding data privacy, ensuring security, and verifying individual identities are paramount. Thus, biometrics has emerged as an essential tool. Biometrics involves the statistical analysis of unique physical, physiological and biological attributes.

Although conventional authentication systems such as fingerprints, retinal scans, and facial recognition are well established, research is being conducted to

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 91–109, 2025.
https://doi.org/10.1007/978-3-031-78341-8_7

use brain activity for biometric identification. This emerging field has gained popularity with advancements in the brain-computer interface (BCI). Brain activity as biometrics can benefit border protection, security services, financial safety, and access control for consumer electronics. [29] The fundamental concept is to scan and match an individual's physiological or electrical brainwave patterns with those stored in a database. Several methods are available for tracking brain function, such as functional magnetic resonance imaging (fMRI), near-infrared spectroscopy (NIRS), positron emission tomography (PET), magnetoencephalography (MEG), and electroencephalography (EEG) [21]. This study specifically highlights EEG signals for biometrics due to their affordability and ease of use.

Utilizing EEG signals for biometrics presents several advantages. [5] 1) Non-Duplicable: EEG signals are unique to individuals and cannot be duplicated. 2) Real-Time Monitoring: They allow continuous, real-time biometric identification. 3) Inconspicuous: EEG-based biometrics are invisible and imperceptible to the human eye, enhancing security. 4) Emphasis on Universal Traits: EEG-based biometrics emphasize universality, distinctiveness, permanence, and circumvention.

Despite these advantages, EEG signals remain underutilized due to their dynamic nature and sensitivity to the subject's mental and physical state. The computational complexity of using EEG as biometric identification may be immense.

In this paper, we have implemented a classification model to identify subjects while optimizing the number of channels using a combination of Riemannian geometry and a non-dominated sorting genetic algorithm (NSGA-II). Covariance matrix features are extracted using the Riemannian framework, and the optimal number of channels is estimated using NSGA-II for EEG-based Biometric authentication. We harnessed two Riemannian Geometry methods: the Minimum Distance to Mean (MDM) classifier and the Tangent Space LDA (TSLDA) models for EEG-based Biometric classification. Friedman statistical test was performed to estimate the significance of the proposed model.

The major contributions of this research work are: 1. Using the Riemannian framework for the classification of Biometric EEG data. 2. Optimizing the number of channels using the NSGA-II algorithm for biometric authentication. 3. comparison of proposed work with existing studies on publicly available datasets.

The paper is organised as follows. The relevant studies are discussed in Sect. 2. The proposed method is described in detail in Sect. 3. Analysis of dataset and experimental setup is given in Sect. 4 and finally conclusion and future work are discussed in Sect. 5.

2 Related Works

EEG as biometrics has attracted attention in past few years. [30] delved into the authentication of individuals using EEG signals obtained during periods of rest. They achieved an impressive 99.7% genuine accept rate (GAR) with

EO and 98.6% with EC, specifically in the beta frequency band. Similarly, [24] adopted frequency-weighted power (FWP) as an alternative approach to representing power within a specific frequency band. Their system underwent evaluation using two EEG datasets recorded during periods of rest. This system achieved a remarkable 0.0039 equal error rate (EER), mainly when EEG signals were collected during EC resting states. These studies give high accuracies while using EEG-based biometrics, even in different experimental conditions.

An EEG emotion recognition system tailored to individual subjects, as introduced by [2], utilized an MDRM classifier to differentiate between high and low levels of valence and arousal by analyzing multidimensional EEG signals. In [32], an innovative transfer learning algorithm grounded in Riemannian geometry was presented for code-modulated visual evoked potential (c-VEP) based Brain-Computer Interface (BCI). [1] concentrated on the classification of four human emotions by using EEG signals. The approach employed the Log-Euclidean Riemannian Metric (LERM) distance metric on a symmetric positive definite manifold (SPD) while considering various channel combinations (2, 7, 10, and 18-channels) across the various frequency bands.

[23] compared the performance of the greedy backward elimination algorithm with two versions of the non-dominated sorting genetic algorithm (NSGA), specifically NSGA-II and NSGA-III. Their findings indicated an accuracy range of 0.98 to 1 using only one to two channels, which was comparable to the accuracy achieved with the full set of electrodes, where accuracy ranged from 1 to 0.97. Numerous other research works assert flawless recognition accuracy when employing EEG signals as biometric identifiers [9]. Nonetheless, the reliability of a substantial portion of these studies is frequently questioned due to a prevailing practice. Such practice involves collecting the data on a single day or under the same circumstances [28]. In such instances, the system's performance evaluation often hinges on the conditions specific to the recording session rather than the distinct characteristics of the individuals [25]. To address this limitation, it is essential to properly test EEG-based biometric recognition systems

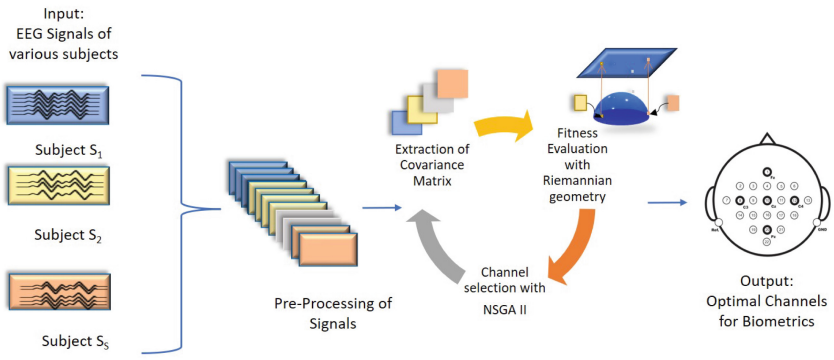


Fig. 1. Proposed framework of Methodology

using multi-session datasets that are recorded on different days. When considering such conditions, the attainable recognition rates often exhibit a notable decrease compared to scenarios where data from the same session is employed for enrollment and recognition [22].

3 Proposed Framework

The flow diagram of our proposed model is described in Fig. 1. First, the EEG signals are pre-processed by the Butterworth low pass filter and scaling. Then, covariance matrices are extracted from each epoch as a feature. For classification, we used two methods based on the framework of Riemannian geometry - Minimum Distance to Mean (MDM) and Tangent Space LDA (TSLDA) [6]. As the name says, the MDM classifier classifies the test epoch based on the minimum distance between the covariance matrix of the test epoch and the mean covariance matrix of the classes. TSLDA projects the covariance matrices to the tangent plane and then with help of LDA, the classification is performed. The Mathematical framework behind the Riemannian geometry is given in Sect. 3.3. Then, the number of channels is optimised with NSGA=II, and the fitness of the generation of NSGA-II is calculated using the MDM and TSLDA classifiers. The best accuracy with the least number of channels is obtained. The Algorithm 1 briefly explains the proposed method.

3.1 Pre-processing

Data Segmentation. The EEG signals were segmented into shapes (epochs, channels, time points) to ensure consistent analysis across datasets. The number

Algorithm 1: Proposed Framework for Classification through MDM and TSLDA

Input: Training Data: a set of EEG epochs and corresponding subject labels

Steps:
1. Pre-process and filter the data.
2. Extract the covariance matrix from each epoch as a feature.
3. Compute the mean covariance matrix for each class.
4. For MDM classification:
   - To classify a new test EEG epoch, calculate the distance from the covariance matrix of the test EEG epoch to each class mean matrix.
   - Assign the epoch to the class with the minimum distance to its mean covariance matrix.
5. For TSLDA classification:
   - Map each covariance matrix to the tangent space using the matrix logarithm.
   - Perform Linear Discriminant Analysis (LDA) on the tangent space representations.

Output: Classification of EEG epochs to the class labels based on MDM and TSLDA.

of time points is taken to be the same as the sampling frequency, such that each epoch represents one second. All relevant EEG Epochs from subjects performing the same task were concatenated. This built a dataset specific to the chosen task. Each epoch was labelled according to its subject to facilitate supervised machine learning.

Data Filtering and Normalization. A Butterworth low-pass filter of order 25 was applied to each epoch, with a 50 Hz cutoff frequency. This filter reduced the noise and maintained the quality of EEG data. Then, MinMax scaling was performed to ensure consistent amplitude scaling across all datasets. This process helped to align the amplitude of EEG data while preserving their relative differences. Then, a variance threshold of 0.03 was employed to ensure data quality. Epochs with variance below this threshold were considered insufficient and excluded from the dataset.

3.2 Extraction of Covariance Matrix as a Feature

Since EEG signals are non-stationary in both the frequency and time domain, computerized analysis emerges as a crucial tool [17]. To address this challenge, spatial covariance matrix for each epoch of the processed EEG signals is calculated. This reduces the dimensionality of EEG signals and extracts salient features, ultimately enhancing classification accuracy. Then, the mean of all the epochs of each subject is computed. These means of SPD matrices and their subjects' labels are now ready to be classified.

3.3 Classification Using Riemannian Framework

Riemannian geometry is used to analyse the properties of smooth and curved spaces. These smooth curved spaces hold local characteristics similar to Euclidean spaces [31]. We used the principles of Riemannian geometry to analyse the space of covariance matrices extracted from EEG signals. These covariance matrices of shape (channels, channels) preserve the spatial properties of EEG Signals [12].

Since Covariance matrices are symmetric positive definite (SPD) matrices, they are suitable for analysis with the help of Riemannian Geometry. These covariance matrices can be treated as points on the Riemannian manifold. Various metrics have been defined in the domain of Riemannian geometry to compare the SPD matrices on the manifold. Each metric has its properties and implications for brain signal classification.

[10] reviews the properties of these distance metrics, exploring their applications in EEG-based brain-computer interfaces. It provides extensive literature on various formulations of different metrics and their computational sensitivity. The formulae to find distance and mean among SPD matrices is provided in Table 1.

Table 1. Distance Metrics defined between SPD matrices

Metric	Distance $\delta(\mathrm{C}_1, \mathrm{C}_2)$	Mean of $\mathrm{C}_1, \mathrm{C}_2, ..., \mathrm{C}_\mathrm{m}$
Harmonic [20]	$\|\mathbf{C_1}^{-1} - \mathbf{C_2}^{-1}\|_F$ where $\|.\|_F$ is Frobenius norm.	$\left(\frac{1}{m}\sum_i \mathbf{C}_i{}^{-1}\right)^{-1}$
Log-Det [8]	$\left(\log(\det\left(\frac{\mathbf{C_1}+\mathbf{C_2}}{2}\right)) - \frac{1}{2}\log(\det(\mathbf{C_1}\mathbf{C_2}))\right)^{1/2}$	Algorithm 1 in [8]
Log-Euclidean [4]	$\|\log(\mathbf{C_1}) - \log(\mathbf{C_2})\|_F$	$\exp\left(\frac{1}{m}\sum_i \log \mathbf{C}_i\right)$
Riemannian [26]	$\delta_R(\mathbf{C}_1, \mathbf{C}_2) = \left(\sum_i \log(\lambda_i)^2\right)^{1/2}$ where λ_i are the joint eigenvalues of C_1 & C_2.	$\arg\min_{\mathbf{C}} \sum_i \delta_R(\mathbf{C}, \mathbf{C}_i)^2$
Wasserstein [3]	$\left(\mathrm{tr}\left(\mathbf{C_1} + \mathbf{C_2} - 2(\mathbf{C_2}{}^{1/2}\mathbf{C_1}\mathbf{C_2}{}^{1/2})^{1/2}\right)\right)^{1/2}$	Eq. (19) in [3]

Riemannian Manifolds: The concept of surface in higher dimensions can be understood better by topological manifolds. This is because every point in the topological manifold has a neighbourhood homoeomorphic (one-to-one, onto, and continuous mapping in both directions) to R_n. In other words, it looks like a flat space locally. Possessing a differential structure facilitates the transformation of a topological manifold into a differential manifold. Smooth manifolds, which are a type of differential manifold, exhibit smooth transitions between maps. These smooth transitions establish rules for locally translating a point on the manifold to its linear approximation.

For simplicity, we will first see how it works with two signals. Consider $e_1(t)$ and $e_2(t)$ be two EEG signals recorded as a function of time. Let e_{1k} and e_{2k} be the k^{th} epoch under analysis. Then the covariance matrices C_k generated will be

$$\mathbf{C}_k(e_1(t), e_2(t)) = \begin{bmatrix} \mathrm{Var}(e_{1k}) & \mathrm{Cov}(e_{1k}, e_{2k}) \\ \mathrm{Cov}(e_{2k}, e_{1k}) & \mathrm{Var}(e_{2k}) \end{bmatrix} \tag{1}$$

where $\mathrm{Cov}(e_{1k}, e_{2k}) = \mathrm{Cov}(e_{2k}, e_{1k})$, making the covariance matrix SPD.

Thus, we are concerned with the variance and the covariance of EEG signals. Here, we have three elements since we are demonstrating 2 EEG signals. C_k is generally determined by $(\mathcal{N}+1)/2$ elements where $\mathcal{N}$ is the number of signals. C_k, with two signals, can be represented in 3D space along the axes $\mathrm{Var}(e_{1k}), \mathrm{Cov}(e_{2k}, e_{1k})$ and $\mathrm{Var}(e_{2k})$. Since the Covariance matrix is SPD, the Cauchy-Schwartz inequality states that $\mathrm{Cov}(e_{1k}, e_{2k}) \leq \mathrm{Var}(e_{1k})\mathrm{Var}(e_{2k})$

Geometrically, this signifies that our covariance matrices are constrained within the symmetric cone. From an electrophysiological perspective, the C_k (the covariance matrices) behaviour varies along the variance axis when changes in electrode energy occur. It also varies along the covariance axis during phase synchronization and amplitude co-modulation between the signals. On a similar conceptual note, in higher dimensions, the manifold takes the shape of a hyper-cone.

A suitable distance metric for this positive matrix cone is defined to facilitate meaningful analysis, similar to the geometric distance [12]. The linear space of $\mathcal{N}$x$\mathcal{N}$ symmetric matrices possesses a natural inner product $\langle \mathcal{X}, \mathcal{Y} \rangle \geq \mathrm{tr}(\mathcal{X}\mathcal{Y})$, along with the associated Euclidean norm $\|\mathcal{X}\|_2$. This Euclidean norm is defined as

$$\|\mathcal{X}\|_2^2 = \mathrm{tr}(\mathcal{X})^2 = \sum_{i=1}^{\mathcal{N}} \lambda_i(\mathcal{X})^2 \tag{2}$$

where λ_i represents the $\mathcal{N}$ eigenvalues of matrix $\mathcal{X}$, and $\mathrm{tr}(\mathcal{X})$ is the sum of the diagonal values of $\mathcal{X}$. As our interest primarily lies in Symmetric Positive Definite (SPD) matrices, they naturally inherit this Euclidean norm $\|\mathcal{X}\|_2$.

Let $S_{++}(\mathcal{N})$ be a set of $\mathcal{N} \times \mathcal{N}$ positive matrices. Since it can be regarded as a differentiable manifold, every small neighbourhood around point C exhibits properties similar to Euclidean space of symmetric matrices, an open set. In Riemannian geometry, the first step is to equip each tangent space with an inner product, ensuring that the resulting metric varies smoothly from one point to another. In this context, Inner Product at a point C is given by [12]

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \mathrm{tr}(C^{-1}\mathcal{X}C^{-1}\mathcal{Y}) \tag{3}$$

The corresponding norm $\|\mathcal{X}\|_{2,C}$ is given as

$$\|\mathcal{X}\|_{2,C}^2 = \|C^{-1}\mathcal{X}\|_2^2 = \|C^{-\frac{1}{2}}\mathcal{X}C^{\frac{1}{2}}\|_2^2 \tag{4}$$

Geodesic Riemannian Distance: Suppose $\Psi(\mathtt{t}) : [0,1] \rightarrow C(\mathtt{n})$ is a differential path from $\Psi(0) = C_1$ to $\Psi(1) = C_2$. Then length $\mathtt{L}$ of $\Psi(\mathtt{t})$ is stated as [6]

$$\mathtt{L}(\Psi(\mathtt{t})) = \int_0^1 ||\Psi(\mathtt{t})||_{\Psi(\mathtt{t})} d\mathtt{t} \tag{5}$$

The geodesic in the space $S_{++}(\mathcal{N})$ connects two points $\mathcal{X}$ and $\mathcal{Y}$ with the shortest curve, while the Riemannian distance between these points is defined as the length of this curve.

$$\delta_R(C_1, C_2) = \|Log(C_1^{-1/2}C_2C_1^{-1/2})\|_F = \sqrt{\sum_{n=1}^{\mathcal{N}} \log^2 \lambda_n} \tag{6}$$

Here, λ_n represents the eigenvalues of the matrix $C_1^{-1/2}C_2C_1^{-1/2}$, which can also be expressed as $C_1^{-1}C_2$. Additionally, "Log(.)" denotes the matrix logarithm, while "log(.)" refers to the scalar logarithm.

Exponential Map: At each point C belonging to $C(\mathtt{n})$, it's feasible to define a tangent space by aggregating tangent vectors at the point C [6]. Each of these tangent vectors denoted as T_i, represents the derivative at $\mathtt{t} = 0$ of the geodesic $\Psi_i(\mathtt{t})$ linking point C and the exponential mapping $C_i = Exp_c(T_i)$. This relationship can be expressed mathematically as follows:

$$\mathrm{Exp}_C(T_i) = C_i = C^{1/2}\exp(C^{-1/2}T_iC^{-1/2})C^{1/2} \tag{7}$$

Algorithm 2: Pseudo code of NSGA II algorithm

Input: Random initial population P of channels, Number of generations `g`

```
Evaluate accuracy through MDM and TSLDA classifier as given in algorithm 1 ;
Rank the individuals using Pareto sorting.;
Create offspring population for P.;
for i = 1 : g do
    Evaluate accuracy of through MDM and TSLA classifier as given in
    algorithm 1;
    for each parent and child in P do
        Reassign ranks of individuals using Pareto sorting. ;
        Create collections of solutions that are not dominated by others. ;
        Calculate crowding distance ;
        Apply crossover and mutation ;
        Update population based on the evolutionary strategy
    Choose points on the lower frontier with significant crowding distance. ;
    Create offspring population
```

Output: Child Generation

Consequently, the inverse mapping can be defined using the logarithmic mapping, given by:

$$\mathrm{Log}_C(C_i) = T_i = C^{1/2} \log(C^{-1/2} C_i C^{-1/2}) C^{1/2} \tag{8}$$

The corresponding definitions of Riemannian distance are:

$$\begin{aligned} \delta_R(C, C_i) &= \|\mathrm{Log}_C(C_i)\|_C = \|T_i\|_P \\ &= \|\mathrm{upper}(C^{-1/2} \mathrm{Log}_C(C_i) C^{-1/2})\|_2 \\ &= \|t_i\|_2 \end{aligned} \tag{9}$$

Here, the operator "upper(.)" vectorises the upper triangular symmetric matrix.

Under specific conditions satisfied by matrices C and C_i, the above expression can be interpreted as an approximation of the distance between the Riemannian manifold and the tangent space. This approximation adheres to the following relationship for all i and j:

$$\forall i, j, \delta_R(C_i, C_j) \sim \|t_i - t_j\|_2$$

Mean of Covariance Matrices. Let $C_{ii=1,\ldots,N}$ be a set of SPD matrices, then using the Riemannian geodesic distance, the Riemannian mean is defined as

$$\Delta(C_1, C_2, \ldots C_N) = \arg \min_{C \in C(n)} \sum_{i=1}^{N} \delta^2(C, C_i) \tag{10}$$

The notation $\arg\min g(x)$ designates the point at which the function g reaches its minimum value. In simpler terms, the Riemannian mean represents the precise location at which the dispersion or variance of the Symmetric Positive Definite (SPD) matrices is minimized.

3.4 Channel Selection

To optimize the informative channels that contribute to better classification, a Non-dominated Sorting Genetic Algorithm - II (NSGA II) [13] is used. NSGA - II chooses parent chromosomes for reproduction based on the classifier's accuracy and the number of channels selected [11]. The algorithm for NSGA II is given in Algorithm 2.

4 Experimental Setup and Results

This section introduces the dataset on which the proposed method was executed. The datasets chosen are characterised by various task types, signal frequencies, participant numbers, and multiple sessions, enhancing the findings' robustness. Various experiments are conducted thereafter to validate the proposed model. Thus, this section discusses the step-by-step observations that led us to the desired result.

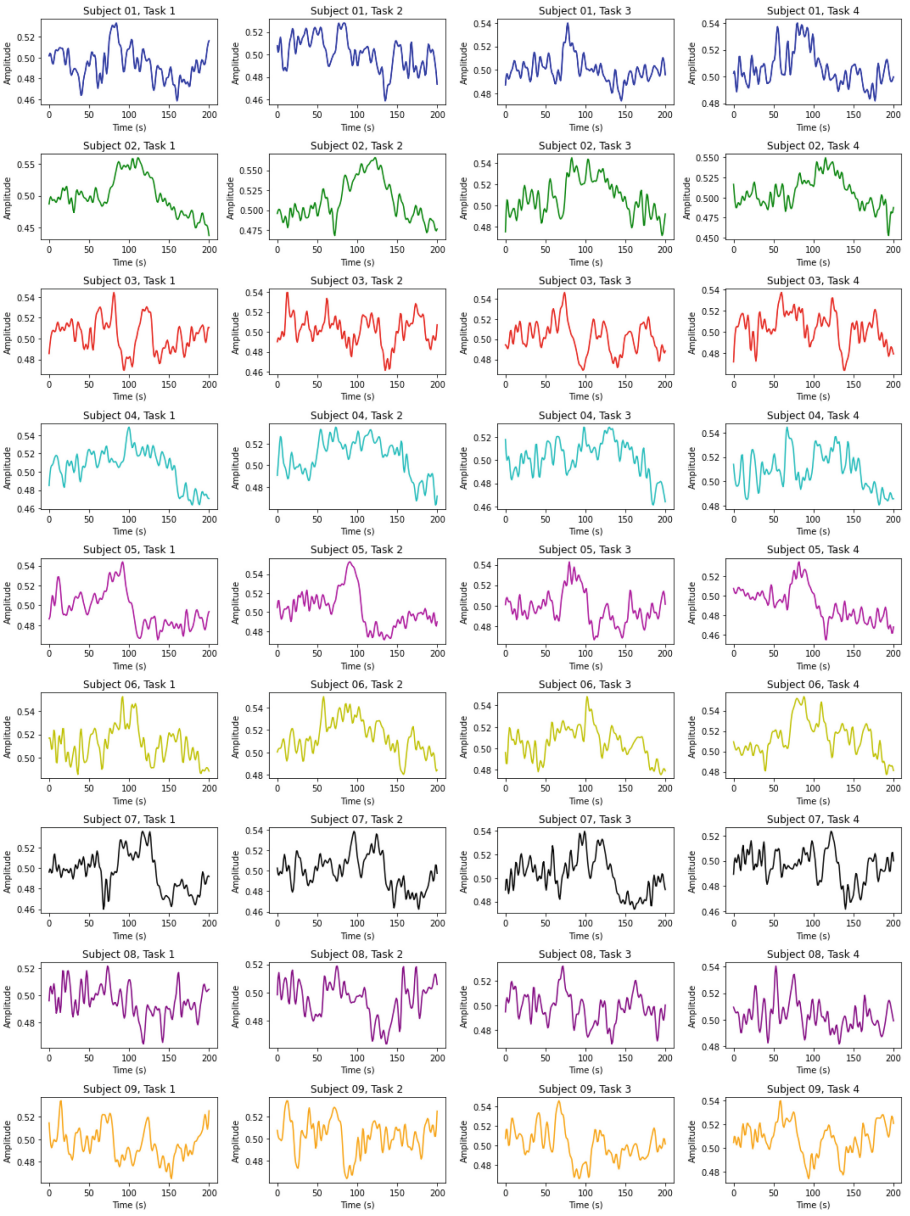
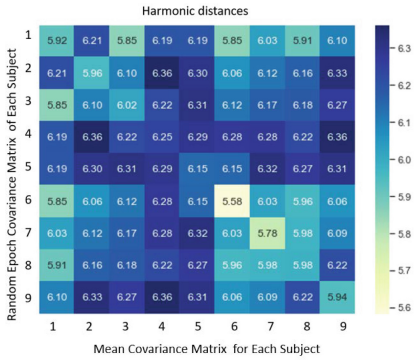
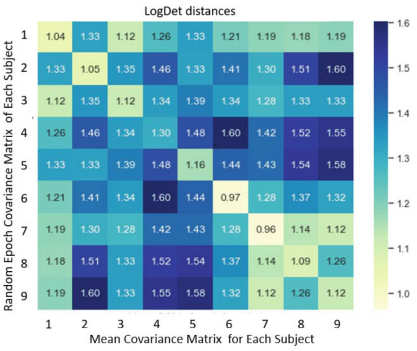


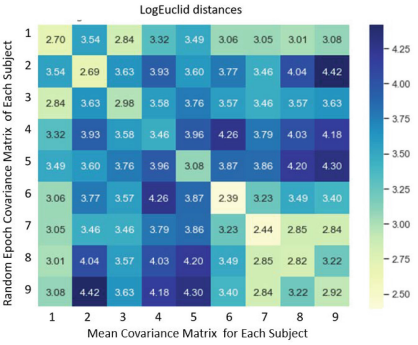
Fig. 2. Mean of all epochs with given the subject and task within dataset 2a of BCI Comp IV



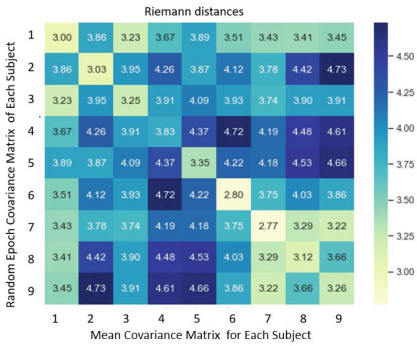
(a) with Harmonic metric



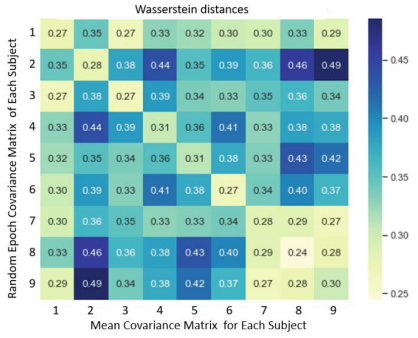
(b) with LogDet metric



(c) with LogEuclid metric



(d) with Riemannian metric



(e) with Wasserstein metric

Method	Average Ranking
Harmonic	5.5
LogDet	3
LogEuclid	3.5
Riemann	2.5
Wasserstein	5.5
TSLDA	1

(f) Average Ranking of various classifiers

Fig. 3. (a)–(e) Heat maps comparing the distance between covariance matrix of 9 test epoch and mean covariance matrix of each Subject; (f) Friedman ranking of various methods

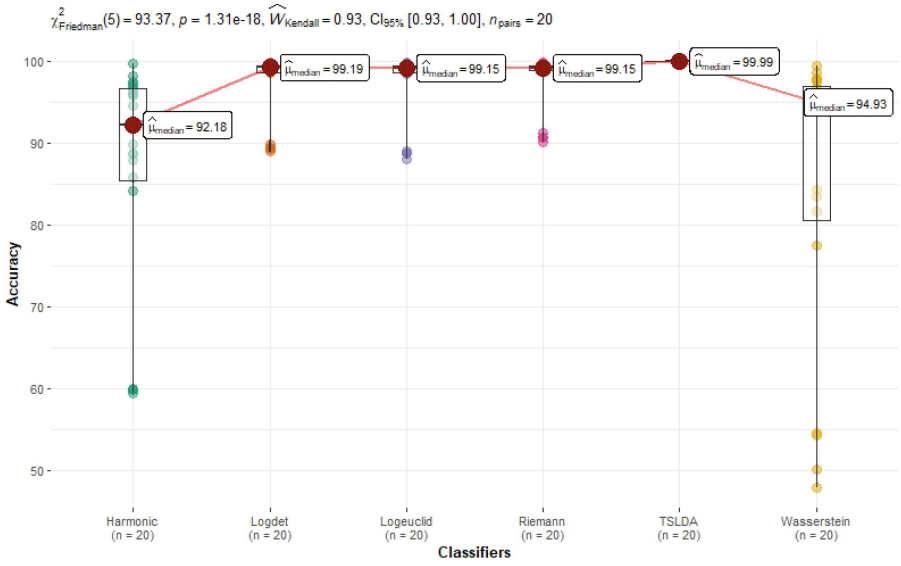


Fig. 4. Box plot representing the accuracies of the methods on all 20 tasks

4.1 Datasets

All the experiments are executed on three datasets. The respective numbers of tasks within these datasets are 12,4 and 4. In total, we have 20 tasks performed by a certain number of subjects. The experiments were performed for each dataset with 10 cross-fold validations and a testing ratio of 80:20. There was no overlapping between training and testing epochs in each run. A brief description of each dataset is given below:

Dataset 1. As presented in [19], the dataset consists of EEG recordings from 30 subjects performing 12 distinct tasks. Each subject engaged in multiple sessions, ranging from 2 to 5, and executed 3 to 6 distinct tasks. The data was collected with 128 channels with sampling frequency of 250 Hz. The details of the data are given in [19]. This segmentation procedure was performed without bias toward the experimental protocol. Consequently, these meticulously crafted epochs remain task-independent. This approach encompasses the resting state, passive observation or listening to various stimuli or instructions, and active engagement in specific tasks.

M3CV Competition. The EEG Biometric Competition on Kaggle, M3CV [16], is a Multi-subject, Multi-session, and Multi-task Database for Investigation of EEG Commonality and Variability. This dataset originally included 106 participants in specific tasks across six distinct paradigms. However, for this research, we concentrated on trials where both the subject and task labels were available. This narrowed our focus to four paradigms: Motor Execution (ME),

Table 2. Optimized channels and their performance obtained through GA

Task	MDRM				TS+LDA			
	Channels used	Accuracy (%)	F1 Score	Recall	Channels used	Accuracy (%)	F1 Score	Recall
Dataset 1								
1	12	100	1	1	13	99.83	0.9983	0.9983
2	14	99.70	0.9970	0.9970	11	100	1	1
3	14	99.91	0.9892	0.9891	12	99.84	0.9984	0.9984
4	15	98.35	0.9836	0.9835	14	100	1	1
5	11	99.38	0.9938	0.9938	5	100	1	1
6	15	99.27	0.9928	0.9927	14	99.83	0.9983	0.9983
7	14	99.52	0.9952	0.9952	16	100	1	1
8	10	99.70	0.9969	0.997	11	100	1	1
9	14	98.37	0.984	0.9837	18	100	1	1
10	22	99.50	0.9950	0.9950	14	100	1	1
11	13	98.98	0.9899	0.9899	11	99.87	0.9987	0.9987
12	16	98.27	0.9828	0.9827	13	99.46	0.9946	0.9946
M3CV dataset								
13	35	97.11	0.9706	0.9711	23	99.93	0.9999	0.9993
14	37	98.09	0.9812	0.9809	23	100	1	1
15	33	98.37	0.984	0.9837	21	99.84	0.9983	0.9983
16	34	98.98	0.9899	0.9899	22	100	1	1
Dataset 2a of BCI Competition IV								
17–20	12	100	1	1	13	100	1	1

Transient-State Sensory (TSS), Resting-State (RS), and Steady-State Sensory (SSS). Consequently, our dataset now includes 95 subjects. These participants are part of the enrolment and calibration subsets of the initial dataset. The data was collected with 64 EEG channels with a sampling rate of 1000 Hz.

Dataset 2a of BCI Competition IV. Dataset 2a of BCI competition IV [7] is compiled for 4-class motor imagery tasks from 9 healthy individuals. Each experimental session consists of 72 trials per class. The experimental timeline is structured as a 2-s fixation period, a 1.25-s cue presentation, and a subsequent 4-s MI process. The data was collected with 22 channels on a frequency of 250 Hz.

4.2 Analysis

Visual Signature. Initially, the mean of all epochs associated with a specific task and a particular subject is evaluated for comparison. We took the dataset 2a of BCI competition IV. Figure 2 displays the plot of mean values of the epochs associated with tasks and subjects of dataset 2a BCI Comp IV. We can see that the pattern of EEG signals is distinct for all the individuals across a task. So this motivated us to investigate the metrics available to compare the mean of these signals.

Performance of Various Metrics. Five different metrics are considered to calculate the mean of covariance metrics and the distance between covariance

metrics. These are Harmonic, LogDet, LogEuclid, Riemann, and Wasserstein. To understand the difference among the metrics used with the MDM classifier, we took the first task, 'Imagining Left Hand Movement' of dataset 2a of the BCI competition, performed by nine subjects. Then, nine random epochs were taken from different subjects, and their covariance matrices were computed. The mean covariance matrics of all classes were also computed. Figure 3 shows the pairwise distance between covariance matrices of randomly selected epochs and the mean of covariance matrices of all nine subjects, with all five metrics. The harmonic distances came out to be comparatively larger, so a log of all the values is taken. The heatmap of these metrics in Fig. 3 exhibited a diverse range of values. The variance of the given values in the heatmap comes out to be 0.025, 0.024, 0.23, 0.23 and 0.003, respectively. We can observe that Riemannian and Logeuclid metrics have the most variance, highlighting their sensitivity to differences. Moreover, the diagonal values are minimum in all the heatmaps, indicating that if an epoch belongs to subject k, then its distance from the mean of covariance matrices of subject k is minimum.

All five metrics are used to evaluate the mean of covariance matrices under the MDM method. The mean of covariance metrics is then projected in Tangent space and classified with the help of LDA. All the channels available were initially used with a 10-fold cross-validation scheme on all 20 tasks. The accuracies with these configurations in MDM and TSLDA are given in Table 3. For a better visualization of performance by these metrics, Fig. 4 shows the box plot of the accuracies achieved by these methods. The median of the accuracies is also mentioned along side the plot of each method. We can see that the Riemann metric performed better across all the tasks using the MDM method. Moreover, TSLDA performed best overall, giving an accuracy of more than 99.8% for all tasks.

Optimization of Channels with NSGA-II. A non-dominated sorting genetic algorithm - II [13] was employed on both classifiers—MDRM (Minimum Distance to Riemannian Mean) and TSLDA—to reduce the number of channels across all tasks. Given that the Riemann metric demonstrated superior performance within the MDM classifier, the Riemann metric with MDM was used to evaluate the fitness of generations in NSGA-II. The NSGA evaluated 30 generations with a population of 20 binary chromosomes, where the length of the chromosome is equal to the total number of channels present, and each binary bit tells the presence or absence of a channel. Initially, a random population is chosen for all the tasks. This random population was kept consistent for all the tasks within each dataset. In each generation, pre-processing is performed with selected EEG signals as given in Sect. 3.1. Then the fitness of each chromosome in a generation is evaluated with MDRM and TSLDA. To generate the next generation, a single-point crossover between the parent chromosomes is performed. The mutation rate is set at 0.05 to introduce diversity in the population. Also, 4 best-performing chromosomes were preserved for the next generation to introduce elitism. This formulation ensures that the NSGA-II seeks high classification accuracy and prioritizes the parsimonious selection of EEG channels [18]. The

number of channels selected after 30 iterations and the accuracies are given in Table 2. In the dataset provided by [19], the MDRM classifier utilized an average of 14 channels, achieving an accuracy of 99.4%, while TSLDA employed 12 channels, resulting in an accuracy of 99.94%. For the M3CV dataset, MDRM utilized 34 channels, giving an accuracy of 98.13%, whereas TSLDA utilized 22 channels to achieve an accuracy of 99.94%. Similarly, in the case of dataset 2a from BCI Competition IV, MDRM utilized 12 channels, achieving an accuracy of 100%, and TSLDA used 13 channels, resulting in an accuracy of 100%.

Friedman Statistical Test. To assess the statistical differences among the 6 metrics employed on 20 tasks in our study, Friedman statistical tests [14] is executed. The null hypothesis assumes that all six metrics have equal classifica-

Table 3. Accuracies achieved by metrics when utilized in the MDM classifier and by TSLDA

Task	MDM					TS+LDA
	harmonic	logdet	logeuclid	riemann	wasserstein	
Dataset 1						
1	96.951	99.098	99.098	99.098	95.209	99.969
2	89.804	98.601	98.601	98.684	93.996	100
3	84.085	99.714	99.714	99.733	96.264	100
4	98.109	99.7	99.7	99.7	99.43	100
5	99.668	99.785	99.785	99.785	94.65	100
6	97.112	99.392	99.341	99.417	97.872	100
7	97.508	99.225	99.225	99.225	97.785	100
8	95.782	99.208	99.208	99.208	96.509	99.935
9	94.563	98.854	98.821	98.854	96.757	99.967
10	96.421	99.199	99.199	99.199	98.581	99.877
11	95.729	99.181	99.267	99.288	97.52	100
12	96.057	99.54	99.54	99.54	96.679	99.988
M3CV dataset						
13	59.43	89.85	89.015	91.158	50.085	99.987
14	60.025	89.321	88.664	90.61	54.577	99.984
15	59.813	89.379	88.68	90.715	47.969	99.981
16	59.814	88.98	88.077	90.104	54.299	99.968
BCI Competition IV Dataset 2a						
17	87.815	99.536	99.382	99.536	84.276	100
18	85.784	98.606	98.45	98.916	81.635	99.846
19	88.567	99.228	98.767	99.075	77.457	100
20	88.731	99.075	99.075	99.075	83.471	99.846
Mean	86.5884	97.2736	97.08045	97.546	84.75105	99.9674

tion accuracy. Under this test, the methods are assigned rank values based on their classification accuracy, where the best method and worst method receive rank values 1 and 6, respectively. The P value is calculated as 1.31E−18, which rejects the null hypothesis. Thus, there is a statistically significant difference in the accuracy of the methods with X^2 Friedman (df = 5) = 93.37. The effect size $W_{\text{Kendall}} = 0.93$ with 95% confidence interval [0.93,1] turned out to be substantial. To assess the performance of all other methods in comparison to the top-ranked method (Control method, namely TSLDA), p-values are calculated through prescribed post hoc analyses, including Nemenyi, Bonferroni, Holm, Hochberg, and Hommel methods. The p values are given in Fig. 3f. All the p-values derived from post hoc procedures were less than 0.001. This means that the results are statistically significant.

Comparision with Existing Studies. Studies that have performed Biometric authentication using EEG signals are mentioned in Table 4. [19] used subspace system techniques to identify individuals with accuracies of 86.4% using just nine EEG channels. On the same dataset, we got accuracy more than 99% with MDRM and TSLDA but used 14 and 12 channels respectively. M3CV dataset is comparatively new, and no studies have been published on this dataset yet. [27] proposed adaptive transfer learning-based multiscale feature fused deep convolutional neural network model that achieved average classification accuracy of

Table 4. Performance comparison of existing methods on the datasets used in this study.

Method	Subjects	Channels	Accuracy
Dataset 1			
IX-VECTOR [19]	30	9	86.40%
Proposed:			
MDRM	30	14	99.40%
TSLDA	30	12	99.94%
M3CV dataset			
Proposed:			
MDRM	95	34	98.13%
TSLDA	95	22	99.94%
Dataset 2a of BCI Competition IV			
TST-ICA [15]	9	22	97.77% (With ICA) 96.11% (Without ICA)
MSFFCNN-TL [27]	9	22	94.06%
Proposed:			
MDRM	9	12	100%
TSLDA	9	13	100%

94.06% on BCI competition dataset 2a. On this dataset, we achieved accuracy of 100% with 12 and 13 Channels using MDRM and TSLDA, respectively.

5 Conclusion

This study aims to identify the subjects based on EEG signals. Spatial covariance matrices are calculated from each EEG epoch. Since covariance matrices are symmetric positive definite (SPD) matrices, they can be represented on the Riemannian manifold. Two classification methods grounded in Riemannian geometry: Minimum Distance to Mean (MDM) and Tangent Space LDA (TSLDA) are employed. The MDM classifier utilized the minimum distance between test epoch covariance matrices, and class mean covariance matrices for classification. Meanwhile, TSLDA projected these covariance matrices onto a tangent plane and used Linear Discriminant Analysis (LDA) for classification.

While calculating the mean and minimum distance among covariance matrices for the MDM classifier, 5 metrics, namely Harmonic, LogDet, LogEuclid, Riemann, and Wasserstein and then TSLDA are used. The respective accuracies of all classifiers were 86.58, 97.27, 97.08, 97.55, 84.75 and 99.97. Friedman statistical test concluded that there is a significant difference between all the methods. Moreover, TSLDA performed remarkably well with 99.9% accuracy.

These methods' classification accuracy is high, but so is the computational cost. To resolve this, the number of channels is optimized using NSGA-II, where the fitness of each generation is evaluated by both MDM and TSLDA classifiers. Table 2 shows the optimal number of channels for various tasks. On average, TSLDA used 11,25 and 14 channels for our three datasets giving accuracies of 99.3%, 99.78% and 99.61%, respectively. Our approach successfully identified the optimal channel configuration, achieving the best accuracy with the least number of channels.

One limitation of our study is that it doesn't account for the participants' geological, physiological, environmental, and societal conditions. In the future, we intend to integrate these factors into our research for a more comprehensive analysis.

In conclusion, this research successfully investigated the complexities of EEG-based person identification. By adopting innovative strategies, such as Riemannian geometry and Genetic Algorithms, we've achieved high accuracy and reduced the number of signals. The insights of this work contribute significantly to the field of biometric identification and have the potential to be applied to a wide range of real-world applications.

Acknowledgements. This work has been funded by IOE Seed grant BHU, BIC JNU Grant and UGC research fellowship. All the datasets used are publicly available. The authors declare that they do not have any competing interests.

References

1. Abdel-Ghaffar, E.A., Daoudi, M.: Emotion recognition from multidimensional electroencephalographic signals on the manifold of symmetric positive definite matrices. In: 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 354–359. IEEE (2020)
2. Abdel-Ghaffar, E.A., Wu, Y., Daoudi, M.: Subject-dependent emotion recognition system based on multidimensional electroencephalographic signals: a Riemannian geometry approach. IEEE Access **10**, 14993–15006 (2022)
3. Álvarez-Esteban, P.C., Del Barrio, E., Cuesta-Albertos, J., Matrán, C.: A fixed-point approach to barycenters in Wasserstein space. J. Math. Anal. Appl. **441**(2), 744–762 (2016)
4. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J. Matrix Anal. Appl. **29**(1), 328–347 (2007)
5. Ashenaei, R., Beheshti, A.A., Rezaii, T.Y.: Stable EEG-based biometric system using functional connectivity based on time-frequency features with optimal channels. Biomed. Signal Process. Control **77**, 103790 (2022)
6. Barachant, A., Bonnet, S., Congedo, M., Jutten, C.: Multiclass brain-computer interface classification by Riemannian geometry. IEEE Trans. Biomed. Eng. **59**(4), 920–928 (2011)
7. Brunner, C., Leeb, R., Müller-Putz, G., Schlögl, A., Pfurtscheller, G.: BCI competition 2008–Graz data set a. Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology, vol. 16, pp. 1–6 (2008)
8. Chebbi, Z., Moakher, M.: Means of Hermitian positive-definite matrices based on the log-determinant α-divergence function. Linear Algebra Appl. **436**(7), 1872–1889 (2012)
9. Chen, Y., et al.: A high-security EEG-based login system with rsvp stimuli and dry electrodes. IEEE Trans. Inf. Forensics Secur. **11**(12), 2635–2647 (2016)
10. Chevallier, S., Kalunga, E.K., Barthélemy, Q., Monacelli, E.: Review of Riemannian distances and divergences, applied to SSVEP-based BCI. Neuroinformatics **19**(1), 93–106 (2021)
11. Cîmpanu, C., Ferariu, L., Dumitriu, T., Ungureanu, F.: Multi-objective optimization of feature selection procedure for EEG signals classification. In: 2017 E-Health and Bioengineering Conference (EHB), pp. 434–437. IEEE (2017)
12. Congedo, M., Barachant, A., Bhatia, R.: Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. Brain-Comput. Interfaces **4**(3), 155–174 (2017)
13. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)
14. Derrac, J., García, S., Molina, D., Herrera, F.: A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. Swarm Evol. Comput. **1**(1), 3–18 (2011)
15. Hameed, A., et al.: Temporal-spatial transformer based motor imagery classification for BCI using independent component analysis. Biomed. Signal Process. Control **87**, 105359 (2024)
16. Huang, G., et al.: M3CV: a multi-subject, multi-session, and multi-task database for EEG-based biometrics challenge. Neuroimage **264**, 119666 (2022)
17. Kirar, J.S., Agrawal, R.: Relevant feature selection from a combination of spectral-temporal and spatial features for classification of motor imagery eeg. J. Med. Syst. **42**, 1–15 (2018)

18. Kirar, J.S., Agrawal, R.: A combination of spectral graph theory and quantum genetic algorithm to find relevant set of electrodes for motor imagery classification. Appl. Soft Comput. **97**, 105519 (2020)
19. Kumar, M.G., Narayanan, S., Sur, M., Murthy, H.A.: Evidence of task-independent person-specific signatures in EEG using subspace techniques. IEEE Trans. Inf. Forensics Secur. **16**, 2856–2871 (2021)
20. Lim, Y., Pálfia, M.: Matrix power means and the karcher mean. J. Funct. Anal. **262**(4), 1498–1514 (2012)
21. Maiorana, E.: Learning deep features for task-independent EEG-based biometric verification. Pattern Recogn. Lett. **143**, 122–129 (2021)
22. Maiorana, E., Campisi, P.: Longitudinal evaluation of EEG-based biometric recognition. IEEE Trans. Inf. Forensics Secur. **13**(5), 1123–1138 (2017)
23. Moctezuma, L.A., Molinas, M.: Towards a minimal EEG channel array for a biometric system using resting-state and a genetic algorithm for channel selection. Sci. Rep. **10**(1), 14917 (2020)
24. Monsy, J.C., Vinod, A.P.: EEG-based biometric identification using frequency-weighted power feature. IET Biometrics **9**(6), 251–258 (2020)
25. Özdenizci, O., Wang, Y., Koike-Akino, T., Erdoğmuş, D.: Adversarial deep learning in EEG biometrics. IEEE Signal Process. Lett. **26**(5), 710–714 (2019)
26. Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. Int. J. Comput. Vis. **66**, 41–66 (2006)
27. Roy, A.M.: Adaptive transfer learning-based multiscale feature fused deep convolutional neural network for EEG MI multiclassification in brain-computer interface. Eng. Appl. Artif. Intell. **116**, 105347 (2022)
28. Ruiz-Blondet, M.V., Jin, Z., Laszlo, S.: Cerebre: a novel method for very high accuracy event-related potential biometric identification. IEEE Trans. Inf. Forensics Secur. **11**(7), 1618–1629 (2016)
29. Tatar, A.B.: Biometric identification system using EEG signals. Neural Comput. Appl. **35**(1), 1009–1023 (2023)
30. Thomas, K.P., Vinod, A.P.: Biometric identification of persons using sample entropy features of EEG during rest state. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 003487–003492. IEEE (2016)
31. Yger, F., Berar, M., Lotte, F.: Riemannian approaches in brain-computer interfaces: a review. IEEE Trans. Neural Syst. Rehabil. Eng. **25**(10), 1753–1762 (2016)
32. Ying, J., Wei, Q., Zhou, X.: Riemannian geometry-based transfer learning for reducing training time in C-VEP BCIS. Sci. Rep. **12**(1), 9818 (2022)

Comprehensive Equity Index (CEI): Definition and Application to Bias Evaluation in Biometrics

Imanol Solano[1,2(✉)], Alejandro Peña[1], Aythami Morales[1], Julian Fierrez[1], Ruben Tolosana[1], Francisco Zamora-Martinez[2], and Javier San Agustin[2]

[1] BiDA-Lab, Universidad Autónoma de Madrid, 28049 Madrid, Spain
{alejandro.penna,aythami.morales,julian.fierrez,ruben.tolosana}@uam.es
[2] Veridas, 31192 Navarre, Spain
{isolano,pzamora,jsanagustin}@veridas.com

Abstract. We present a novel metric designed, among other applications, to quantify biased behaviors of machine learning models. As its core, the metric consists of a new similarity metric between score distributions that balances both their general shapes and tails' probabilities. In that sense, our proposed metric may be useful in many application areas. Here we focus on and apply it to the operational evaluation of face recognition systems, with special attention to quantifying demographic biases; an application where our metric is especially useful. The topic of demographic bias and fairness in biometric recognition systems has gained major attention in recent years. The usage of these systems has spread in society, raising concerns about the extent to which these systems treat different population groups. A relevant step to prevent and mitigate demographic biases is first to detect and quantify them. Traditionally, two approaches have been studied to quantify differences between population groups in machine learning literature: 1) measuring differences in error rates, and 2) measuring differences in recognition score distributions. Our proposed Comprehensive Equity Index (CEI) trade-offs both approaches combining both errors from distribution tails and general distribution shapes. This new metric is well suited to real-world scenarios, as measured on NIST FRVT evaluations, involving high-performance systems and realistic face databases including a wide range of covariates and demographic groups. We first show the limitations of existing metrics to correctly assess the presence of biases in realistic setups and then propose our new metric to tackle these limitations. We tested the proposed metric with two state-of-the-art models and four widely used databases, showing its capacity to overcome the main flaws of previous bias metrics.

Keywords: Biometrics · Face Recognition · Fairness · Bias

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 110–126, 2025.
https://doi.org/10.1007/978-3-031-78341-8_8

1 Introduction

In the past decade, we have experienced a revolution in the field of Artificial Intelligence (AI). The surprising capabilities of data-driven automatic systems have made possible the development of AI-based solutions in a variety of domains, such as health [15], education [5], or recruitment [35]. Among these application areas, biometric recognition technology, or biometrics, is growing significantly due to its advantages over traditional security/authentication approaches. Compared to the latter, biometric recognition systems are capable of authenticating the identity of a person using features extracted from biometric data of the individual, usually known as biometric traits [26]. Some traits that have been studied in the field include iris [1], fingerprint [31], or human-computer interaction signals [8]. Of the different traits traditionally addressed in biometrics, Face Recognition (FR) is probably the one that has benefited the most from the emergence of Deep Learning. The success of novel architectures [22] and learning strategies [9,45], has notably raised the performance of these systems compared to traditional handcrafted approaches [17]. Consequently, during the last decade a lot of attention has been paid to face recognition [47], while its use in real systems has become more widespread, with applications ranging from border control [4] to mobile phone authentication [2,33].

Although several AI-based systems may appear to be ready for large-scale deployment in a vast array of domains and applications attending solely to the performance, some unsolved issues around their use must be first addressed for successful and trustworthy applications [11,35]. Several scholars have raised concerns about aspects such as vulnerabilities to attacks [14,23,34,36], or potential algorithmic discrimination effects [29,39]. Attending to the latter, in recent years we have observed a significant amount of systems exhibiting biased behaviors, leading to unfair treatment towards certain individuals based on their membership to demographic groups [3]. In addition to these concerns, biometrics has been singled out in both past and future legislation on data privacy and AI, due to the sensitive nature of the data it deals with. The European General Data Protection Regulation (GDPR)[1] already imposed several restrictions on how to store and process personal data [18]. Furthermore, the recently approved European AI Act[2] includes new requirements that biometric systems shall meet, including the prevention of the aforementioned problems. Biometrics has been particularly fruitful in this scenario, with the study of demographic biases, including their prevention and mitigation [10,13]. Methods tackling bias in biometrics span the evaluation of trained models in different populations [6,38,40], learning strategies including fairness constraints in their optimization objectives [16,39], or the development of new databases with a broad and fair representation of the demographic groups [32,46,48]. However, a key point for analyzing and mitigating demographic biases is to be able to measure them properly.

[1] https://gdpr-info.eu/.
[2] https://www.euaiact.com/.

Traditionally, the measurement of demographic bias in biometric systems has been based on the differences among the error rates for each demographic group. This approach has received the name of the differential outcome [25]. In [13] the authors propose to measure fairness using the maximum difference (i.e., worst-case differentials) of both False Match Rate (FMR) and False Non-Match Rate (FNMR) between demographic groups at a given operating threshold τ. They combine both measures in a single metric, known as Fairness Discrepancy Rate (FDR). In [24] the authors introduced the Gini Aggregation Rate for Biometric Equitability (GARBE) metric, a fairness measure inspired by the Gini coefficient, computed for both FMR and FNMR. The National Institute of Standards and Technology (NIST) has also highlighted the use of FMR/FNMR discrepancies as a quantitative measure of system fairness [20]. They proposed the Inequity metric (IN), an alternative measure to the FDR [19]. In this case, instead of computing the maximum difference of FMR/FNMR across groups, the maximum ratios of these between demographic groups at a given threshold τ are obtained, which are combined to obtain the Inequity metric. The metric has a direct operational sense since it directly represents the number of times that the individuals from the disfavoured group are more likely to be confused, compared to the favored group (i.e., the one with the best performance). A common drawback of all the previous proposals is the need to select a concrete operational point to measure fairness, which may hinder a complete assessment of the model performance. An exception to the aforementioned threshold-focused approach that still is based on differential outcomes is the work proposed by Gong *et al.* [16]. They proposed to measure bias as the standard deviation of performance across demographic groups, reported in terms of the Area Under the Curve (AUC). Whilst it does not require explicitly fixing a threshold, the metric cannot measure fairness at the distribution level.

Contrary to the aforementioned differential outcome approaches, Kotwal and Marcel focused on a differential performance approach [27]. Instead of measuring the fairness at a specific operating point τ, they introduced a metric directly working with score distributions of demographic groups. Concretely, the Distribution Fairness Index (DFI) measures the difference in score distributions between demographic groups by leveraging the Kullback-Leibler (KL) divergence. While measuring fairness at the score distribution level poses some benefits compared to other methods, the DFI exhibits some limitations. By considering the whole distribution equally, biases appearing in the tail of those distributions, which ultimately condition the recognition performance, are not always properly represented. This nuance is especially relevant for high-performance systems, such as the ones leading the NIST's Face Recognition Technology Evaluation (FRTE). These FR systems, mostly commercial from the industry, usually present low error rates, and hence their differences are determined by extremely competitive operational points. In other words, it is mostly in the matching score distributions tails where the main differences across systems arise. The differential outcome approach can be useful to measure bias in these scenarios. Still, it should be noted that performance differences in the distribution tails may

not be always related to demographic biases. As Therhörst *et al.* showed, non-demographic attributes such as head-pose, illuminations, brightness, resolution, or even black and white images can affect the performance of FR systems [43]. Therefore, it would be desirable to use a metric that strikes a balance between the two extremes: based on decision thresholds or full score distributions.

In this work, we propose a metric that can capture differences in the distributions tails without fixing specific operational points. We were inspired by recent fairness metrics [19,27] and we have addressed some of the flaws observed when using these metrics in real and synthetic scenarios [30,41]. We present a modification of the DFI metric [27] that solves the previously mentioned drawbacks, which we use to evaluate a high-performance algorithm presented to the NIST FRTE challenge in several state-of-the-art face recognition databases. The main contributions of this work are:

1. We present an evaluation of six fairness metrics applied to face recognition in synthetic and real-world scenarios. We characterize its performance when measuring the fairness of commercial algorithms in state-of-the-art datasets. Our experiments demonstrate the difficulties in detecting biased behaviors in high-performance algorithms characterized by long-tail score distributions.
2. We propose a new metric, the Comprehensive Equity Index (CEI) that addresses the drawbacks of the existing DFI metric when evaluating high-performance systems. The proposed metric aims to detect demographic biases in the distributions tails and in the center part of the distributions.

2 Measuring Fairness in Biometric Systems

2.1 Problem Formulation

Let us consider any 2-class classification problem (n-class can be developed as multiple 2-class problems), in our case exemplified using Face Recognition (FR). Other AI setups apart from classification in which output probabilities for different data populations can be obtained are also easily covered by our methods. Other AI setups in which class probabilities are not straightforward, e.g., regression, will need further work for our ideas to be properly developed. Let's now focus for concreteness and without loss of generality in Face Recognition.

Traditionally, a FR system operates in one of the following setups: *i)* Identification or *ii)* Verification. Our interest here is in the latter, where the task is to determine whether two samples belong to the same identity or not with a 1:1 similarity comparison, or *match*. If both samples belong to the same identity, the samples are said to form a *genuine* pair, otherwise, we refer to it as *impostor*, i.e., a 2-class classification problem. In a real scenario, it is common that one of the biometric samples in the pair is pre-enrolled in the system (reference sample), whose identity is known. Thus, the system is presented with a second sample (probe sample) that claims to belong to the same identity as the reference.

Formally, to measure the performance of the system let us consider a dataset of biometric samples, which contains N samples, i.e., face images $\mathbf{I}$ in the case

of FR systems. Each of the images was captured from a subject, who is part of a demographic group d_i according to its demographic traits (e.g., gender, ethnicity, age). We assume here a set $\mathcal{D}$ of K demographic groups, which are disjoint (i.e., a subject can only be a member of one group). A FR model $\mathbf{w}^F$ is trained to extract face representations $\mathbf{x} = f(\mathbf{I}|\mathbf{w}^F)$ discriminant for the task of identity recognition from the images. The dissimilarity of a pair of face representations $(\mathbf{x}_n, \mathbf{x}_m)$ is usually computed using a vector distance metric, such as the Euclidean or the cosine distances, which can then be normalized [12] to a score $s(\mathbf{x}_n, \mathbf{x}_m)$. In this context, different considerations of what is meant for the system to be demographically fair have been proposed in the literature [13,21,27,39], but they all follow a similar hypothesis: the "performance" of the system should be equal across groups. The nuance here is how to define performance. If we abstractly formulate the performance of a model for a demographic group d_i as Performance$(\mathbf{w}^F|d_i)$, the previous idea would be satisfied if Performance$(\mathbf{w}^F|d_i) \approx$ Performance$(\mathbf{w}^F|d_j)\ \forall d_i, d_j \in \mathcal{D}$ (where $\approx$ should be carefully defined). In traditional Machine Learning [21], these performances have been measured as the probability of a certain decision of the model (i.e., Demographic Parity), or even as the True Positive Rate (i.e., Equal of Opportunity). Particularly in biometrics, differential outcome approaches have considered error rates, i.e., FMR/FNMR, as the basis for these performance measurements [13]. On the other hand, differential performance approaches such as [27] consider the entire distributions of scores $z = p(s|\mathbf{w}^F)$ (where p denotes probability) to represent the performance of the model.

2.2 Fairness Metrics: Existing Methods

Recently, Kotwal and Marcel have addressed the problem of measuring demographic fairness in biometric systems [27]. They argued how the community has paid mostly attention to differential outcome metrics, i.e., those which measure fairness as gaps in classification rates across groups [13,16,19,24]. Compared to these, they proposed a differential performance metric based on the distances between score distributions z. This approach presents the main advantage of being agnostic to the operational point selected, thus measuring the fairness of the overall system.

Concretely, the metric proposed in [27] leverages the Kullback-Leibler (KL) divergence as the basic distance measure among the distributions of each demographic group. The metric, known as the Distribution Fairness Index (DFI), spans values between 0 and 1, where a value close to the latter represents a fairer model. Formally, DFI is defined as follows (using the same notation as in [27]):

$$z_{D_{\text{mean}}} = \frac{1}{K} \sum_{i=1}^{K} z_{D_i} \tag{1}$$

$$\text{DFI}_{\text{N}} = 1 - \frac{1}{K \log_2 K} \sum_{i=1}^{K} S_i \tag{2}$$

where z_{D_i} is the combined (genuine + impostor) distribution for the demographic group d_i, normalized so that the curve area sums one, and S_i is the KL divergence among z_{D_i} and the mean distribution $z_{D_{\text{mean}}}$. The formulation in Eq. 2 corresponds to a baseline definition of DFI (Normal, therefore N), which is based on the average of the dissimilarities of all the demographic distributions from the mean distribution. Additionally, it is a common approach to measure the fairness of a system considering the group that is disfavoured the most, as it represents the worst-performing case. Thus, another formulation of DFI can be made, which only considers the distribution of the demographic group that diverges the most from the mean:

$$\text{DFI}_{\text{E}} = 1 - \frac{1}{\log_2 K} \max(S_i) \tag{3}$$

Apart from DFI, two interesting differential outcome metrics are highlighted in the latest NIST FRVT report on demographic differentials measurement [19]. The first of them is an updated version of the Inequity metric, which computes the ratio between maximum and minimum FMR/FNMR values across demographic groups. As noted in [19], using the minimum values is not robust in general setups, thus a better measurement can be obtained by including in the ratio the geometric mean of FMR/FNMR across groups, instead of the minimum. Thus, the modified Inequity metric is formulated as follows:

$$\text{IN}_{\text{FMR}} = \frac{\max_{d_i} \text{FMR}(\tau)}{\text{FMR}_{\text{geom}}} \tag{4}$$

$$\text{IN}_{\text{FNMR}} = \frac{\max_{d_i} \text{FNMR}(\tau)}{\text{FNMR}_{\text{geom}}} \tag{5}$$

In addition to the previous metric, the NIST's report proposes as well the use of the GARBE metric to measure fairness in terms of both FMR and FNMR [25]. This metric is inspired by the Gini coefficient, a commonly used measurement of income disparity, and is formulated as follows:

$$\text{GARBE}_{\text{FMR}} = \frac{\sum_i \sum_j |\text{FMR}_{d_i}(\tau) - \text{FMR}_{d_j}(\tau)|}{2K^2 \, \text{FMR}_{\text{arith}}} \tag{6}$$

$$\text{GARBE}_{\text{FNMR}} = \frac{\sum_i \sum_j |\text{FNMR}_{d_i}(\tau) - \text{FNMR}_{d_j}(\tau)|}{2K^2 \, \text{FNMR}_{\text{arith}}} \tag{7}$$

where $\text{FMR}_{\text{arith}}$ and $\text{FNMR}_{\text{arith}}$ represent arithmetic means of each demographic group considered. The Gini metric yields values on the interval $[0, 1]$, with high values being associated with unfair systems. Higher values are a sign of unfairness as well for the Inequity metric. For both Inequity and GARBE metrics, an operational point needs to be selected. In its evaluations, NIST fixes the operational point as that for which the systems give an overall FMR of 0.0003. Then, FMR and FNMR values for each demographic group can be computed and aggregated using any of the previous metrics. While both of them can be

further aggregated into a single value, having a separate value for FMR and FNMR allows us to analyze the fairness with regard to different kinds of errors, i.e., whether the model exhibits more bias in the genuine distribution (FNMR) or in the impostor distribution (FMR).

2.3 Proposed Metric: Comprehensive Equity Index (CEI)

In this section, we present an extension of the metric of [27] to measure fairness. Our proposal tries to keep the benefits of performance-based metrics while integrating the error-based perspective of differential outcome metrics. With this balance, we are not only aiming to measure the model's bias but also to consider how competitive the recognition system is, a relevant aspect in systems with very small error rates.

By examining the evaluation of high-performance models (e.g., those presented to NIST FRTE) with the DFI metric on state-of-the-art datasets, such as RFW [48] or BUPT-B [46], we noticed that error rates associated to demographic biases are not captured with the cited metric. We hypothesize that, since DFI uses the entire distribution to measure fairness *i)* the tail has a little relevance in the computation and *ii)* genuine and impostor distributions are treated as a whole, hence hindering the assessment of any bias present in either of them. In comparison, differential outcome metrics such as GARBE [25] or Inequity [19] can capture these biases, since the selection of an operational point directly focuses the evaluation on the tails of the distributions. However, measuring fairness for a concrete operational point presents some drawbacks. First, the demographic bias underlying the core of the biometric system is not captured at all, so information about the rest of the distribution is lost. Second, by considering only the tail of the distribution, the performance is measured in a lower percentage of samples than when using the entire curve. Thus, outcome differences could be due to reasons beyond demographic attributes, for instance, image resolution, brightness, or pose covariates.

We aim to overcome the aforementioned shortcoming by presenting a new fairness measure built on the proposal of Kotwal and Marcel [27]. Specifically, our objective is to have a metric that is both threshold-agnostic and able to measure bias in genuine and impostor distributions independently while properly accounting for the tails, i.e., where errors occur. We introduce here the Comprehensive Equity Index (CEI). For every demographic group, the CEI first splits each distribution (i.e., genuine or impostor) into two groups based on a given percentile P_s (i.e., score threshold s corresponding to certain accumulated probability P), dividing the tail from the rest of the distribution (referred to as center distribution from now on). The intuition here is to have independent components so we can assign them different weights when computing fairness. Once the distribution is split, we can compute a score S'_i (dissimilarity score as we are using distance measures) between a demographic distribution and the mean distribution as follows:

$$S'_i(P_s) = w_t \cdot D_{\mathrm{KL}}(z^t_{D_i} || z^t_{D_{\mathrm{mean}}}) + w_c \cdot D_{\mathrm{KL}}(z^c_{D_i} || z^c_{D_{\mathrm{mean}}}) \tag{8}$$

where $z^t_{D_i}$ and $z^c_{D_i}$ are respectively the tail and center distributions from z_{D_i}, $z^t_{D_{\text{mean}}}$ and $z^c_{D_{\text{mean}}}$ refer to mean distributions as defined in Eq. 1, and w_t and w_c are manually-tuned weights controlling the trade-off between the relevance of each part in the similarity score. The term in Eq. 8 is computed for each demographic group, then the CEI is calculated in a similar way as the DFI, having Normal and Extreme variants:

$$\text{CEI}_{\text{N}}(P_s) = 1 - \frac{1}{K \log_2 K} \sum_{i=1}^{K} S'_i \tag{9}$$

$$\text{CEI}_{\text{E}}(P_s) = 1 - \frac{1}{\log_2 K} \max(S'_i) \tag{10}$$

Both proposed metrics CEI_{N} and CEI_{E} are on the interval $[0, 1]$, with a higher value being associated with a fairer model.

3 Material and Methods

3.1 Models and Databases

For the present work, we have trained two face recognition models from scratch for face recognition. The models were trained with a margin-based loss, i.e., CosFace [45], on the WebFace database [49], which contains 260M images from 4M identities. The database includes images from 7 different race groups, with more than half of the identities being Caucasian. Similarly to the models evaluated in [49], we assessed the performance of the trained models on IJB-C [28]. These models will be used later in our experiments:

- **ResNet-100** [22]. The ResNet architecture is one of the most famous convolutional models of the last decade. Here, we have used the architecture with 101 convolutional layers. The trained ResNet-100 model exhibits a FNMR@FMR = 1e−5 of 0.0407.
- **Proprietary Model**. A commercial model submitted to the NIST FRTE 1:1 with a FNMR@FMR = 0.0003 of 0.0058. When evaluating this model on the IJB-C [28] dataset, we obtained FNMR@FMR = 1e−5 of 0.037.

Throughout the experiments carried out in the present work, we will use the following publicly available databases: MORPH [37,44], RFW [48], and BUPT-B [46]. All three databases include demographic labels with the gender and ethnicity of each subject. In addition to the aforementioned databases, we have used in this work a synthetic database recently released for the FRCSyn Challenge [30] with realistic conditions and controllable demographics.

4 Experiments

In this section, we present different experimental scenarios in which we show the usefulness of the proposed metric to measure the (un)fairness of high-performing

face recognition models. In Sect. 4.1 we present a toy scenario to elaborate on and numerically assess the advantages of the presented metric in comparison to existing metrics. Finally, experiments on real images are conducted in Sect. 4.2, where we evaluate a high-performance industry model and compare our proposed metric with existing methods.

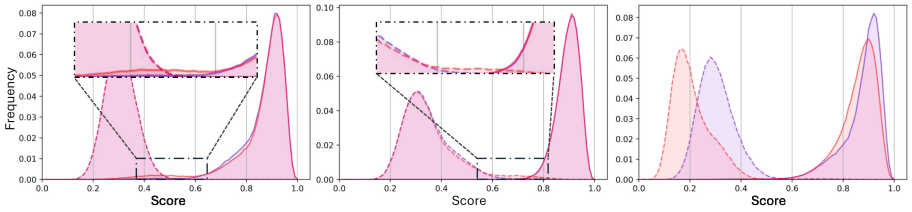


Fig. 1. Genuine and impostor synthetically-generated similarity score distributions, in different scenarios: (Left) Biased Genuine distribution tail (BG); (Center) Biased Impostor distribution tail (BI); and (Right) Biased genuine-impostor distribution Center (BC).

4.1 Synthetically-Generated Distributions

In the following, we present experiments on synthetically-generated similarity score distributions, simulating the performance of a competitive model. Three scenarios are considered (see Fig. 1). First, we manipulated the left tail of the genuine distribution, i.e., the right distribution in Fig. 1 (left), to increase the false rejections in that region. We have called this scenario Biased Genuine distribution tail (BG). This name is given because in the overlapping region between the two distributions (genuine and impostor), the genuine tail is forced (biased) to have an atypically high probability (considering as typical a rapid decrease similar in nature to a normal distribution tail, e.g., as shown in the impostor distribution).[3] The second scenario is similarly created for a Biased Impostor distribution tail (BI). Finally, in the third scenario, both distributions (genuine and impostor) have similar probabilities in their tails, but their centers are shifted. We have called this scenario Biased genuine-impostor distribution Centers (BC). The first two scenarios are expected to be well captured by the IN_{FMR} (IN_{FNMR}) and $\text{GARBE}_{\text{FMR}}$ ($\text{GARBE}_{\text{FNMR}}$) metrics, as changes in the tail are more relevant here, whereas the distribution changes introduced in the third scenario will, in principle, be better captured using both DFI_{N} and DFI_{E} metrics, as the distribution tails in that case are similar and hence present an identical error rate.

[3] In a general sense, bias in machine learning can be considered a systematic error that occurs in a model due to incorrect assumptions in the machine learning process. Technically, we can define bias as the error between the model behavior and the ground truth. In practical terms, measuring that error will normally mean measuring differences between score distributions, as done in the present paper for the particular case of systematic demographic differences in biometric systems [6].

The evaluation presented in Table 1 confirms the initial intuition. On one hand, both variants of the DFI metric are not able to detect any bias in those cases where the differences are found in the distribution tails (BG and BI), but the GARBE and IN metrics seem to capture those differences. We hypothesize that DFI_N and DFI_E are not being able to capture the generated bias because of: *i*) the minor impact that differences in the distribution tail have compared to the center of the distribution, and *ii*) the metric using the genuine + impostor distribution as a whole, ignoring particular differences found in each one. On the other hand, in the third scenario, it can be observed that the GARBE and IN metrics are not able to capture any differences, whereas both variants of

Table 1. Values of the DFI and NIST-related metrics and the proposed CEI_N and CEI_E on the simulated scenario.

	BG	BI	BC
DFI_N [27]	0.9983	0.9974	0.8361
DFI_E [27]	0.9982	0.9970	0.8112
$GARBE_{FMR}$ [25]	0.0050	0.2950	0.0433
$GARBE_{FNMR}$ [25]	0.3326	0.0025	0.0208
IN_{FMR} [19]	1.1249	2.0989	1.0697
IN_{FNMR} [19]	2.2331	1.0037	1.0416
$CEI_{N_{Genuine}}$ [**ours**]	0.5678	0.9991	0.9919
$CEI_{N_{Impostor}}$ [**ours**]	0.9992	0.6223	0.3767
$CEI_{E_{Genuine}}$ [**ours**]	0.4714	0.9990	0.9916
$CEI_{E_{Imporstor}}$ [**ours**]	0.9992	0.5372	0.2986

Table 2. Values of CEI_N on three synthetically generated cases: *i)* BG, *ii)* BI, and *iii)* BC. We evaluate each case using three different percentiles (75, 90, 95) and three different weight sets (i.e., $\mathbf{w}_1 = (0.2, 0.8)$, $\mathbf{w}_2 = (0.5, 0.5)$, and $\mathbf{w}_3 = (0.8, 0.2)$) for the tail and center of the distributions, respectively.

		Genuine			Impostor		
		BG	BI	BC	BG	BI	BC
P_{75}	$\mathbf{w}_1$	0.9897	1.0000	0.9110	0.9999	0.9908	0.4008
	$\mathbf{w}_2$	0.9757	0.9999	0.8990	0.9999	0.9803	0.4083
	$\mathbf{w}_3$	0.9617	0.9998	0.8869	0.9998	0.9698	0.4158
P_{90}	$\mathbf{w}_1$	0.9438	0.9997	0.9726	0.9995	0.9481	0.5319
	$\mathbf{w}_2$	0.8787	0.9994	0.9757	0.9992	0.8901	0.4580
	$\mathbf{w}_3$	0.8136	0.9991	0.9787	0.9989	0.8320	0.3841
P_{95}	$\mathbf{w}_1$	0.8791	0.9998	0.9804	0.9997	0.8936	0.5703
	$\mathbf{w}_2$	0.7235	0.9995	0.9862	0.9995	0.7580	0.4735
	$\mathbf{w}_3$	0.5678	0.9991	0.9919	0.9992	0.6223	0.3767

the DFI seem to be more sensitive to distribution displacements. As the NIST-related metrics need the differences to be related to the performance instead of to the shape of the curve, this does not manifest in this case.

To assess our proposed metric, we evaluated those three scenarios using different configurations of the proposed CEI metric with the normal variant, CEI_N. We have conducted experiments combining percentile values of 0.75, 0.90, and 0.95 and weight values of $(w_\text{tail}, w_\text{center}) = \{(0.2, 0.8)\}, (0.5, 0.5), (0.8, 0.2)\}$. The results are shown in Table 2. For the first two scenarios (BG and BI) of Fig. 1, it is observed that when configuring our metric to give more importance to the distribution tail (both using high percentile values and high w_tail proportions), our metric is able to detect the introduced bias (i.e., the CEI_N value decreases) in each the genuine (for the BG scenario) and impostor distribution (for the BI scenario). (Note that the metric diverting from 1 means that the bias introduced between the two evaluated scenarios with/without bias is properly noticed.) For the last scenario, BC, we observe the biggest decrease in CEI_N (i.e., largest bias detected) for the impostor distribution, as expected given the BC setup considered (see Fig. 1 right, where we can see that the bias introduced makes more different the impostor distributions in comparison to the genuine ones). Therefore, the proposed metric is able to detect the bias in all three presented cases, regardless of the weight parameters used. This is a desired behavior not observed with any of the other metrics in the literature. Thus we conclude that our proposed CEI has the potential to overcome some of the weaknesses of the original DFI. However, this needs to be assessed in real-world scenarios.

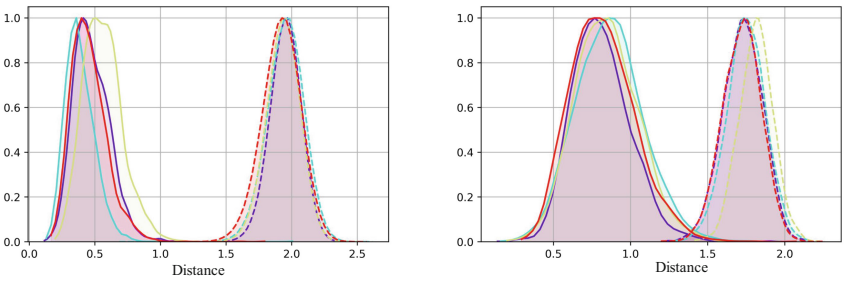


Fig. 2. Genuine (continuous line) and impostor (dashed line) distributions for ResNet-100 [22] model in MORPH [37,44] (Left) and RFW [48] (Right) datasets. The x-axis shows the Euclidean distance between two images. Thus the genuine distributions are on the left and the impostor on the right. Each demographic group is represented by a different color.

4.2 Evaluation in Real Scenarios

In Fig. 2 the genuine and impostor distributions for a ResNet-100 [22] model trained over the WebFace database [49] for the MORPH [37,44] and RFW [48]

Table 3. Values of CEI_N metric using the ResNet-100 [22] model on MORPH [37,44], RFW [48], BUPT-B [46], and the FRCSyn database [30]. We evaluate using three different percentiles (75, 90, 95) and two weight sets (i.e., $\mathbf{w}_2 = (0.5, 0.5)$, and $\mathbf{w}_3 = (0.8, 0.2)$) for the tail and center of the distributions, respectively.

		Genuine				Impostor			
		MORPH	RFW	BUPT-B	FRCSyn	MORPH	RFW	BUPT-B	FRCSyn
p_{75}	$\mathbf{w_2}$	0.9759	0.9318	0.9192	0.9515	0.9492	0.8151	0.9670	0.9316
	$\mathbf{w_3}$	0.9614	0.9073	0.8951	0.9343	0.9368	0.7556	0.9532	0.9138
p_{90}	$\mathbf{w_2}$	0.9202	0.8911	0.8944	0.9317	0.9129	0.7779	0.9272	0.9157
	$\mathbf{w_3}$	0.8723	0.8365	0.847	0.8988	0.8741	0.6804	0.8907	0.8798
p_{95}	$\mathbf{w_2}$	0.8696	0.8461	0.8753	0.9206	0.8871	0.7581	0.9097	0.9046
	$\mathbf{w_3}$	0.7959	0.7622	0.8129	0.8787	0.8298	0.6435	0.8611	0.8578

Table 4. Values of the DFI and NIST-related metrics, and the proposed CEI_N and CEI_E on the simulated scenario obtained on MORPH [37,44], RFW [48], BUPT-B [46], and the FRCSyn database [30]. The two CEI variants use a percentile of 95% and weights $(w_{\text{tail}}, w_{\text{center}}) = (0.8, 0.2)$.

	ResNet-100 [22]			
	MORPH	RFW	BUPT-B	FRCSyn
DFI_N [27]	0.9932	0.9785	0.9965	0.9927
DFI_E [27]	0.9885	0.9529	0.9927	0.9768
$GARBE_{FMR}$ [25]	0.3762	0.2885	0.3289	0.4631
$GARBE_{FNMR}$ [25]	0.1500	0.1377	0.2719	0.0654
IN_{FMR} [19]	2.9418	1.8803	2.2661	4.3461
IN_{FNMR} [19]	1.6818	1.3723	2.0182	1.1754
$CEI_{N_{Genuine}}$ **[ours]**	0.7959	0.7622	0.8129	0.8787
$CEI_{N_{Impostor}}$ **[ours]**	0.8298	0.6435	0.8611	0.8578
$CEI_{E_{Genuine}}$ **[ours]**	0.5425	0.6724	0.6725	0.6717
$CEI_{E_{Imporstor}}$ **[ours]**	0.6797	0.394	0.7973	0.6989
	Proprietary Model			
	MORPH	RFW	BUPT-B	FRCSyn
DFI_N [27]	0.9933	0.9818	0.9983	0.9906
DFI_E [27]	0.9873	0.9647	0.9818	0.9662
$GARBE_{FMR}$ [25]	0.2439	0.2500	0.3075	0.4616
$GARBE_{FNMR}$ [25]	0.1500	0.1941	0.2873	0.0693
IN_{FMR} [19]	2.9410	1.7965	2.1286	4.1876
IN_{FNMR} [19]	1.6818	1.5635	2.0803	1.2038
$CEI_{N_{Genuine}}$ **[ours]**	0.9056	0.7624	0.7831	0.8686
$CEI_{N_{Impostor}}$ **[ours]**	0.9135	0.6744	0.9001	0.8467
$CEI_{E_{Genuine}}$ **[ours]**	0.7953	0.6408	0.6431	0.6704
$CEI_{E_{Imporstor}}$ **[ours]**	0.8585	0.4560	0.8492	0.6867

datasets is depicted. For each of the datasets, each curve represents a demographic group based on the ethnicity. We have evaluated the normal variant of the CEI metric described in Eq. 9 with different configurations. More concretely, we analyze its behavior using percentile values of 75%, 90%, 95%, and, based on the observations from Sect. 4.1, weight values of $(w_{\text{tail}}, w_{\text{center}}) = \{(0.5, 0.5), (0.8, 0.2)\}$. Table 3 it can be observed that, as the distribution tail receives more importance (i.e., using high percentile like P_{95} and a weight combination that prioritizes the tail like $\mathbf{w}_3$). the metric value decreases, indicating that differences among demographic groups exist on those parts of the curves. In Table 3 in that configuration (P_{95} and $\mathbf{w}_3$) we also observe differences in the behavior of genuine and impostor distributions, e.g., MORPH [37,44] and RFW [48] have a larger difference in their CEI_N score between the genuine and impostor distribution, meaning that the bias is different for each one. We have used the distance score distributions over those datasets and the ResNet-100 [22] model in Fig. 2 to confirm the existence of the differences captured by the CEI_N.

The configuration using a percentile of 95% and weights (textsubscript) = (0.8, 0.2) has been used for the two variants of the proposed CEI metric (CEI_N) and the extreme variant described in Eq. 10, (CEI_E) to compare them with other existing metrics (see Sect. 2.2). The results are represented in Table 4. It is shown that the DFI-related metrics (DFI_N and DFI_E) are not able to capture any of the existing differences. As hypothesized before, this may be related to the fact that differences are mainly found in the distribution tails. Moreover, the DFI_N and DFI_E metrics do not separate the genuine and impostor distributions. It uses an aggregation of both distributions to compute the "fairness", provoking bias related to specific distributions not to be captured. That behavior is not observed with the NIST-related metrics, which is especially relevant in this scenario because the metric can detect potential differences between the demographic groups while providing more detailed information about the distribution (genuine or impostor) in which the difference is found. If we analyze the results obtained for both variants of the CEI metric, we find improvements w.r.t. the existing performance-based DFI metric, as it is able to better detect differences between demographic groups. Moreover, we observe that the proposed variant CEI_E is more sensitive when measuring those demographic differences. Thus, the validity of both variants of the proposed metric is confirmed. It is shown to be able to capture existing differences while maintaining the strengths of the performance-based approach.

5 Conclusions

In this work, we follow up on previous efforts to measure "fairness" in biometric recognition systems by using a differential performance-based approach, dependent on the system score function. We have introduced a modification of a previous metric by adapting it into its application to real-world scenarios where the differences are found in the score distribution tails. The proposed metric, called Comprehensive Equity Index (CEI), has been shown to capture existing

differences in the score distributions for different demographic groups when evaluating a high-performance Face Recognition (FR) system presented in the NIST FRVTE 1:1 (with excellent results) in several state-of-the-art datasets.

Our proposal addresses previous weaknesses of differential performance metrics by parameterizing the relevance of the tail distribution differences for diverse demographic groups with a percentile selecting the tail and weights that give more or less importance to differences in that area of the distribution. Our proposed metric CEI also provides information on the bias encountered in each of the genuine and impostor distributions. This way, the metric can adapt to the distribution area where bias is desired to be studied. The proposed metric therefore overcomes observed deficiencies of previous metrics in real-world scenarios while preserving the benefits of the differential performance approach: it does not depend on concrete operational points and knowledge of the intrinsic behavior of the system, i.e., how the model represents biometric samples depending on its demographic attributes.

The introduced metric should be understood as a complement to other performance outcome-based metrics. Ours can detect differences in distributions, but this may not always be enough to determine whether a system is fair (or has bias), as that statement is dependent on the definition of fairness (or bias) and the concrete use case [39]. We propose to use the CEI as an index to detect differences in high-performance model distributions together with other performance metrics such as FMR, FNMR, and outcome differential-based indexes to have a wider view of the biases of the system in terms of the demographic group.

Future work includes continued investigation on data-efficient and cost-effective bias detection and evaluation methods looking both at models internals [38,40] and outputs [6], symbolic methods to analyze biases [42], and exploitation of LLMs to better assess biometric systems [7] including bias evaluation.

Acknowledgements. This paper has been financed by the Government of Navarre within Industrial Doctorates 2022, the company Veridas (https://veridas.com/en/), Cátedra ENIA UAM-VERIDAS en IA Responsible (NextGenerationEU PRTR TSI-100927-2023-2), and project BBforTAI (PID2021-127641OB-I00 MICINN/FEDER).

References

1. Alonso-Fernandez, F., Bigun, J., Fierrez, J., et al.: Periocular biometrics: a modality for unconstrained scenarios. IEEE Comput. **57**(6), 40–49 (2024)
2. Alonso-Fernandez, F., Farrugia, R.A., Fierrez, J., Bigun, J.: Super-resolution for selfie biometrics: introduction and application to face and iris. In: Rattani, A., Derakhshani, R., Ross, A. (eds.) Selfie Biometrics. ACVPR, pp. 105–128. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-26972-2_5
3. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of FAccT, pp. 77–91 (2018)
4. Busch, C., et al.: Facilitating free travel in the schengen area: a position paper by the European Association for Biometrics. IET Biometrics **12**(2), 112–128 (2023)

5. Daza, R., Morales, A., Tolosana, R., Gomez, L.F., et al.: edBB-demo: biometrics and behavior analysis for educational platforms. In: Proceedings of AAAI (2023)
6. DeAlcala, D., et al.: Measuring bias in AI models: an statistical approach introducing N-sigma. In: Proceedings of COMPSAC, pp. 1167–1172 (2023)
7. Deandres-Tame, I., Tolosana, R., Vera-Rodriguez, R., Morales, A., Fierrez, J., Ortega-Garcia, J.: How good is ChatGPT at face biometrics? A first look into recognition, soft biometrics, and explainability. IEEE Access **12**, 34390–34401 (2024)
8. Delgado-Santos, P., et al.: Swipeformer: transformers for mobile touchscreen biometrics. Expert Syst. Appl. **237**, 121537 (2024)
9. Deng, J., Guo, J., Xue, N., et al.: ArcFace: additive angular margin loss for deep face recognition. In: Proceedings of CVPR, pp. 4690–4699 (2019)
10. Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., et al.: Demographic bias in biometrics: a survey on an emerging challenge. IEEE Trans. Technol. Soc. **1**(2), 89–103 (2020)
11. Díaz-Rodríguez, N., et al.: Connecting the dots in trustworthy artificial intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation. Inf. Fusion **99**, 101896 (2023)
12. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Target dependent score normalization techniques and their application to signature verification. IEEE Trans. Syst. Man Cybern. Part C **35**(3), 418–425 (2005)
13. de Freitas Pereira, T., Marcel, S.: Fairness in biometrics: a figure of merit to assess biometric verification systems. IEEE Trans. Biometrics Behav. Identity Sci. **4**(1), 19–29 (2022)
14. Ghafourian, M., Fierrez, J., Vera-Rodriguez, R., Tolosana, R., Morales, A.: SaFL: sybil-aware federated learning with application to face recognition. In: Proceedings of ICIP (2024)
15. Gomez, L.F., Morales, A., Orozco-Arroyave, J.R., Daza, R., Fierrez, J.: Improving Parkinson detection using dynamic features from evoked expressions in video. In: Proceedings of CVPR, pp. 1562–1570 (2021)
16. Gong, S., Liu, X., Jain, A.K.: Jointly de-biasing face recognition and demographic attribute estimation. In: Proceedings of ECCV, pp. 330–347 (2020)
17. Gonzalez-Sosa, E., et al.: Exploring facial regions in unconstrained scenarios: experience on ICB-RW. IEEE Intell. Syst. **33**(3), 60–63 (2018)
18. Goodman, B.W.: A step towards accountable algorithms?: algorithmic discrimination and the European Union general data protection. In: Proceedings of NIPS (2016)
19. Grother, P.: FRVT Part 8: Summarizing Demographic Differentials (2022). National Institute of Standards and Technology (NIST). https://pages.nist.gov/frvt/reports/demographics/nistir_8429.pdf
20. Grother, P., Ngan, M., Hanaoka, K.: FRVT Part 3: Demographic Effects (2019). National Institute of Standards and Technology (NIST). https://doi.org/10.6028/NIST.IR.8280
21. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Proceedings of NIPS, vol. 29, pp. 3323–3331 (2016)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of CVPR, pp. 770–778 (2016)
23. Hernandez-Ortega, J., Fierrez, J., Morales, A., Galbally, J.: Introduction to presentation attack detection in face biometrics and recent advances. In: Marcel, S., Fierrez, J., Evans, N. (eds.) Handbook of Biometric Anti-Spoofing. ACVPR, pp. 203–230. Springer, Singapore (2023). https://doi.org/10.1007/978-981-19-5288-3_9

24. Howard, J.J., Laird, E.J., Sirotin, Y.B., Rubin, R.E., Tipton, J.L., et al.: Evaluating proposed fairness models for face recognition algorithms. In: Proceedings of ICPR Workshops, pp. 431–447 (2023)
25. Howard, J.J., Sirotin, Y.B., Vemury, A.R.: The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In: Proceedings of BTAS, pp. 1–8 (2019)
26. Jain, A.K., Ross, A., et al.: Introduction to Biometrics. Springer, Cham (2011)
27. Kotwal, K., Marcel, S.: Fairness index measures to evaluate bias in biometric recognition. In: Proceedings of ICPR, pp. 479–493 (2023)
28. Maze, B., et al.: IARPA Janus Benchmark-C: face dataset and protocol. In: Proceedings of ICB, pp. 158–165 (2018)
29. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., et al.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 1–35 (2021)
30. Melzi, P., et al.: FRCSyn challenge at WACV 2024: face recognition challenge in the era of synthetic data. In: Proceedings of WACV, pp. 892–901 (2024)
31. Micheletto, M., et al.: LivDet2023 - fingerprint liveness detection competition: advancing generalization. In: IEEE/IAPR International Joint Conference on Biometrics (2023)
32. Morales, A., Fierrez, J., Vera-Rodriguez, R., Tolosana, R.: SensitiveNets: learning agnostic representations with application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **43**(6), 2158–2164 (2021)
33. Perera, P., Fierrez, J., Patel, V.M.: Quickest multiple user active authentication. In: Bourlai, T., Karampelas, P., Patel, V.M. (eds.) Securing Social Identity in Mobile Platforms. ASTSA, pp. 179–196. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39489-9_10
34. Peña, A., Serna, I., Morales, A., Fierrez, J., Lapedriza, A.: Facial expressions as a vulnerability in face recognition. In: Proceedings of ICIP, pp. 2988–2992 (2021)
35. Peña, A., et al.: Human-centric multimodal machine learning: recent advances and testbed on AI-based recruitment. SN Comput. Sci. **4**(5), 434 (2023)
36. Ramachandra, R., et al.: Presentation attack detection methods for face recognition systems: a comprehensive survey. ACM Comput. Surv. **50**(1) (2017)
37. Ricanek, K., Jr., Tesafaye, T.: MORPH: a longitudinal image database of normal adult age-progression. In: Proceedings of FGR, pp. 341–345 (2006)
38. Serna, I., DeAlcala, D., Morales, A., Fierrez, J., Ortega-Garcia, J.: IFBiD: inference-free bias detection. In: AAAI Workshop on Artificial Intelligence Safety (SafeAI). CEUR-WS, vol. 3087 (2022)
39. Serna, I., Morales, A., Fierrez, J., Obradovich, N.: Sensitive Loss: improving accuracy and fairness of face representations with discrimination-aware deep learning. Artif. Intell. **305**, 103682 (2022)
40. Serna, I., Pena, A., Morales, A., Fierrez, J.: InsideBias: measuring bias in deep networks and application to face gender biometrics. In: Proceedings of ICPR, pp. 3720–3727 (2021)
41. Shahreza, H.O., et al.: SDFR: synthetic data for face recognition competition. In: 2024 IEEE International Conference on Automatic Face and Gesture Recognition (FG) (2024)
42. Tello, J., et al.: Symbolic AI (LFIT) for XAI to handle biases. In: European Conference on Artificial Intelligence Workshops (ECAIw). CEUR-WS, vol. 3523 (2023)
43. Terhörst, P., Kolf, J.N., Huber, M., Kirchbuchner, F., Damer, N., et al.: A comprehensive study on face recognition biases beyond demographics. IEEE Trans. Technol. Soc. **3**(1), 16–30 (2022)

44. Tome, P., Fierrez, J., Vera-Rodriguez, R., Ramos, D.: Identification using face regions: application and assessment in forensic scenarios. Forensic Sci. Int. **233**, 75–83 (2013)
45. Wang, H., Wang, Y., Zhou, Z., Ji, X., et al.: CosFace: large margin cosine loss for deep face recognition. In: Proceedings of CVPR, pp. 5265–5274 (2018)
46. Wang, M., Deng, W.: Mitigating bias in face recognition using skewness-aware reinforcement learning. In: Proceedings of CVPR, pp. 9139–9328 (2020)
47. Wang, M., Deng, W.: Deep face recognition: a survey. Neurocomputing **429**, 215–244 (2021)
48. Wang, M., et al.: Racial faces in the wild: reducing racial bias by information maximization adaptation network. In: Proceedings of ICCV, pp. 692–702 (2019)
49. Zhu, Z., et al.: WebFace260M: a benchmark unveiling the power of million-scale deep face recognition. In: Proceedings of CVPR, pp. 10492–10502 (2021)

An Unconstrained Dataset for Face Recognition Across Distance, Pose, and Resolution

Udaybhan Rathore and Akshay Agarwal(✉)

Data Science and Engineering Department, IISER Bhopal, Bhopal, India
{udaybhan19,agarwal}@iiserb.ac.in

Abstract. Face, a non-intrusive recognition modality, is an ideal candidate for identifying criminals or performing general-purpose person identification. On top of that, faces are not only related to identity but other essential attributes such as age and gender can also be extracted. Due to such potential, face recognition has received tremendous attention, yet face recognition from a distance remains challenging. To empower the face recognition research, we have collected a novel unconstrained *video face dataset* namely *MobileFaces* across various distances using mobile phones. Utilizing the proposed dataset, we have performed extensive experiments on face recognition, including verification and identification using state-of-the-art (SOTA) deep face recognition networks. Further, we have evaluated the robustness of current SOTA deep face attributes prediction networks to demonstrate whether the challenge lies in face recognition only or whether the existing algorithms are vulnerable in predicting facial attributes such as age and gender. The results suggest that the existing algorithms are ineffective not only in identifying the identity of the subjects but also fail to detect face attributes when the images are captured in unconstrained environments. For example, deep face networks yield the best macro average accuracy of **65**% and an F-1 score of **0.48** when asked to predict gender on the collected dataset at a distance of 10 m. Based on the comparison with existing unconstrained face datasets and analysis of the effectiveness of image super-resolution techniques, it is showcased that the proposed dataset is significantly challenging compared to them, and hence, we believe that the presence of our dataset can advance the development of unconstrained face recognition algorithms.

Keywords: Face Recognition · MobileFaces · Unconstrained Videos Dataset

1 Introduction

Acquiring faces in the real world and an unconstrained environment requires installing surveillance cameras [9,18,42]. The prime challenge of installing

The dataset can be accessed using the following link.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 127–144, 2025.
https://doi.org/10.1007/978-3-031-78341-8_9

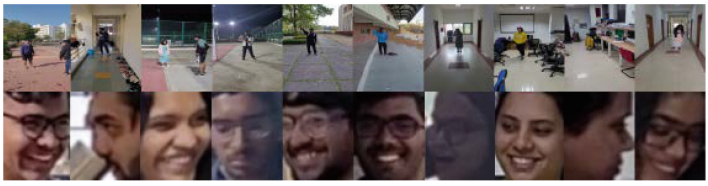
Fig. 1. Challenges present in unconstrained face recognition where the images are captured using a mobile camera. The first row of images is taken at a 5 m distance, and the second row is the low-quality faces when captured from a 10 m distance.

surveillance cameras is the setup cost, along with its infeasibility of covering every possible corner, especially in a country that has a large geographical area such as India. That is where the boom in the mobile camera industry can be a boon as "every pocket" has a camera that can be used to capture images [35]. We assert that these "**mobile surveillance**" images can better act as evidence in a court of law as compared to the testimony of eyewitnesses or sketches developed [33,43]. Therefore, we required an unconstrained dataset captured using mobile phones to develop a robust face recognition technology across varying distances. However, capturing faces from videos recorded on mobile phones presents unique challenges, especially when the distance between the camera and a person is significant, say 10 m. Due to low resolution, pose variation, and poor image quality, recognizing individuals from a long distance poses many challenges. Apart from that, unconstrained acquisition conditions suffer from several other limitations due to a high degree of facial freedom, which results in expression variation, occlusion, and motion blur. Along with these, the environment also plays a critical role in the form of illumination.

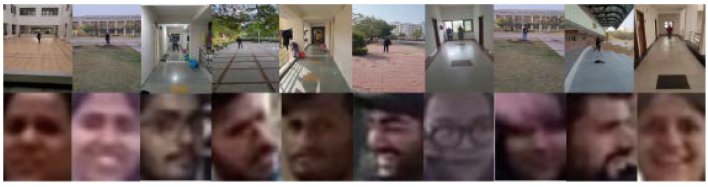
Looking at the potential of face recognition (FR) in person identification and to boost FR algorithms, over the years, substantial progress has been made in collecting the face recognition datasets, such as FERET [34], CMU FIA [16], and CMU PIE [41]. However, these databases are mainly collected from close distances with low and constant camera zoom, making them suitable for short-distance applications. The datasets, such as VGGFace2, MS-Celeb-1M, and MegaFace, are developed over a longer range and captured using surveillance or professional cameras. However, most of these datasets' studies are conducted on European, Caucasian, and American ethnicities, and very limited studies have been conducted on *Indian ethnicity* face images. Therefore, in this research, we have conducted a first-ever large-scale face analysis study using the proposed MobileFaces dataset captured in the unconstrained real world at various distances using mobile phones. In brief, the contributions of this research are:

– We have collected an unconstrained MobileFaces video dataset reflecting several real-world artifacts such as pose, distances, illumination, and resolution. The proposed dataset can be seen as mobile surveillance but with proper ethical guidelines;
– A detailed face analysis study has been conducted to extract facial characteristics related to soft biometrics, such as age and gender, along with face verification. The study reveals several interesting insights that can pave the way for robust, unconstrained FR.

(a) Sample images from videos captured at 2 meters distance.



(b) Sample images from videos captured at 5 meters distance.



(c) Sample images from videos captured at 10 meters distance.

Fig. 2. Full images and cropped face images from our collected dataset reflect the challenges developed due to not only varying distances but also the subject being unconstrained in having natural actions.

Figure 1 shows the challenges of FR when the images in the proposed dataset are captured using mobile phones in completely unconstrained environments at varying distances. The images suffer from quality, pose, illumination, occlusion, and blur. Therefore, analysis of the current deep FR algorithms can help in further improving the FR algorithms in unconstrained mobile settings.

2 Related Work

In the literature, a tremendous amount of work has been done for face recognition at the architecture and optimizer level [6,13], and several datasets have

Table 1. Statistics for face datasets. The proposed Mobilefaces contain images at three different distances, such as 2 m, 5 m, and 10 m, which result in varying poses and resolutions as well. *A few of the large-scale datasets (*) are not publicly available, and most of them have achieved at least 95% accuracy* [4,53], *reflecting the demand for a novel unconstrained face recognition dataset.* The description of existing datasets is taken from VGG-Face2 [5].

Dataset	Total Images	Minimum Per Subject Images	Varying Distance	Varying Pose	Varying Resolution	Mobile Acquisition	Indian Ethnicity	Year
LFW [19]	13,233	1	–	–	–	–	–	2007
CASIA-WebFace [50]	494,414	2	–	–	–	–	–	2014
IJB-A [24]	5,712	11.4	–	–	–	–	–	2015
VGG-Face [31]	2.6 M	1,000	–	–	–	–	–	2015
*MegaFace [22]	4.7 M	3	–	–	–	–	–	2016
*MSCeleb-1M [17]	10 M	100	–	–	–	–	–	2016
UMDFaces [3]	367,920	43.3	–	Yes	–	–	–	2016
CSCRV [42]	40,950	210	Yes	Yes	Yes	–	Yes	2016
IJB-B [49]	11,754	36.2	–	Yes	–	–	–	2017
IJB-C [29]	31,334	36.3	–	Yes	Yes	–	–	2018
VGG-Face2 [5]	3.31 M	80	–	Yes	–	–	–	2018
TinyFace [8]	169,403	1	–	Yes	Yes	–	–	2019
D-Lord [28]	1.2M	60	Yes	Yes	Yes	–	Yes	2023
MobileFaces (Ours)	**52,913**	**224**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**2024**

been collected to feed and test this model. However, few efforts have been made to understand the challenges when face images are captured in an unconstrained environment. Li et al. [15] explored deep learning techniques for face recognition from a distance using mobile devices and have achieved promising results. The UMD Faces [3] challenge dataset has been designed for face detection and recognition under unconstrained and realistic conditions, such as in surveillance and security applications. IJB-C [29] is a benchmark dataset for face recognition under unconstrained conditions. Table 1 shows the statistics of the existing face recognition datasets along with their comparison with the proposed dataset. As can be seen, most existing datasets are not acquired using mobile cameras and do not have explicit labeling concerning varying distances used for acquisition. Further, many of these datasets have a large number of subjects, which is because they might be unethically scrapped from the web[1], whereas the proposed dataset is captured by following ethical guidelines with proper consent from the subjects and hence might look small in terms of the number of subjects. Although we want to highlight the number of subjects is comparable to other ethically collected datasets in the real world, such as FPV (#140) [36] NDPSID (#100) [10], and KaspAROV (#108) [9]. The closest dataset to our dataset is CSCRV [42] and D-Lord [28], which contains face images at varying distances; however, acquired using a sophisticated five mega-pixel camera and in a semi-controlled environment.

[1] It can also be seen from multiple recent incidents including data regulation policy, retrieval of such datasets, and ban of web scrapping APIs.

The prime limitations so far in the literature are (i) the lack of a dataset covering Indian ethnicity, (ii) varying distances for acquisition reflecting in-the-wild real-world settings, (iii) video datasets, and (iv) a mobile camera dataset for unconstrained acquisition. We assert that due to the significant population strength, which is not bound to any geographical boundaries, the lack of datasets covering Indian ethnicities can create biased face recognition systems. Hence, for fair face recognition, it is essential to cover each ethnicity. Therefore, we have developed a dataset from Indian ethnicities in unconstrained real-world settings using mobile phones, where the subjects are free to perform their natural actions. We assert that the presence of such a large-scale dataset of under-represented ethnicity will add to a more balanced representation of various ethnicities in the face research domain and can boost the learning of ethnicity-unbiased face recognition algorithms.

Table 2. Summarizing the characteristics of the proposed MobileFaces dataset.

No. of Subjects	87
Distances from mobile	2, 5, and 10 m
No. of probe videos	87 videos/distance (total 261 videos).
No. of frames in a probe video	200–250
Total no. of probe frames	52,913
Frame width*height, rate	1920 × 1080 pixels, 30.01 frames/second
Gallery	87 (Frontal image of each subject)
Age range	Range 18–34 years, Average 22.6 years
Gender ratio	50 Males, 37 Females
Location	Indoor, Outdoor daytime (natural light), Outdoor nighttime (synthetic light)
Annotation	Age and gender of each subject in CSV file

3 Proposed MobileFaces Dataset for Unconstrained Person Verification and Identification

We have collected a video-based face dataset of Indian individuals in the unconstrained real world. The dataset is a collection of facial images and videos captured using multiple smartphones. The videos in the proposed dataset are collected at varying distances between the camera and the subjects, where no other instructions are provided to the subjects except to ensure the distance is constant for a particular video. In other words, the subjects are free to perform natural tasks to ensure the presence of natural actions, which are highly possible in real-world surveillance scenarios. Table 2 summarizes our proposed dataset's characteristics. In the proposed dataset, the frame's resolution is high, but the cropped face region covers a small portion, especially where the distance is high.

For example, the resolution of face images captured at 10 m distance is 32×32 or lower. The faces extracted from these unconstrained videos are used as probes for face recognition and the analysis of soft biometrics attributes, including age and gender. We collected a high-resolution gallery image of each subject, which has been used for face verification and identification.

Figure 2 shows the unconstrained nature of the proposed dataset and the challenges it brings for face recognition. As mentioned, the images are captured across distances, resulting in a high distribution shift among the images. Due to the high richness of features such as variation in pose, distance, and resolution, the proposed dataset is an ideal candidate to benchmark the performance of current DFR networks. In brief, the proposed dataset has several implications for the field of face recognition: (i) the dataset provides a benchmark for evaluating the performance of face recognition algorithms on images and videos captured from a range of distances. Using a benchmark dataset enables researchers to objectively compare the performance of different algorithms and identify improvement areas; (ii) the dataset will address the current lack of balanced representation

Table 3. Comparison of TPR at different FPR rates using macro and micro average aggregation of score vectors for the face verification task performed on raw images across different distances, using VGG-Face and ArcFace networks.

CNN	Distance	FPR: 10^{-3}		FPR: 10^{-2}	
		Micro	Macro	Micro	Macro
VGG-Face	2 m	**0.82**	0.14	**0.82**	0.14
	5 m	**0.79**	0.08	**0.79**	0.13
	10 m	**0.38**	0.01	**0.38**	0.08
ArcFace	2 m	**0.72**	0.11	**0.72**	0.36
	5 m	**0.64**	0.11	**0.64**	0.39
	10 m	**0.09**	0.05	**0.09**	0.08

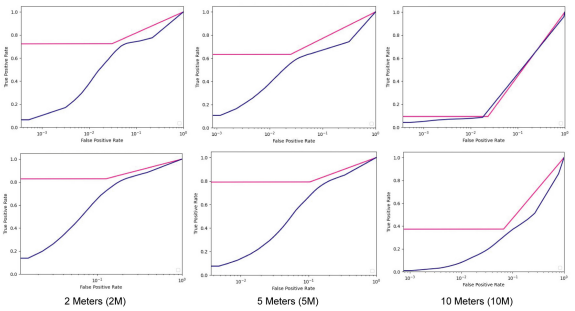


Fig. 3. ROC curves of face verification task using the face images captured at distances 2 m (leftmost), 5 m (middle), and 10 m (rightmost) using two SOTA FR networks, namely ArcFace and VGG-Face. Pink shows the micro average performance, and blue curves the macro average performance of the networks. (Color figure online)

of various ethnicities, such as the Indian, within the computer vision community; (iii) the dataset can aid in the development of new robust face recognition algorithms that are optimized for images and videos captured from different distances and under different conditions even in mobile phones. This is important because current face recognition algorithms are optimized for images captured at close range, and their performance degrades as the distance increases.

4 Experimental Results and Analysis

To utilize the full potential of the proposed MobileFaces dataset, we have performed several experiments, which can be broadly divided into two categories:

Table 4. A comparison between pre-trained and fine-tuned SOTA FR deep networks in terms of TPR at various FPR values, utilizing images acquired at a distance of 10 m. Fine-tuning is done using 8076 raw images captured at 10 m, distinct from the test images.

FR Model	Pre-Trained		Fine-Tuned	
	10^{-3}	10^{-2}	10^{-3}	10^{-2}
ArcFace	**0.056**	**0.161**	**0.099**	**0.165**
AdaFace	0.015	0.041	0.014	0.0663
ElasticFace	0.0049	0.035	0.0088	0.0549
MagFace	0.008	0.04	0.0312	0.0977

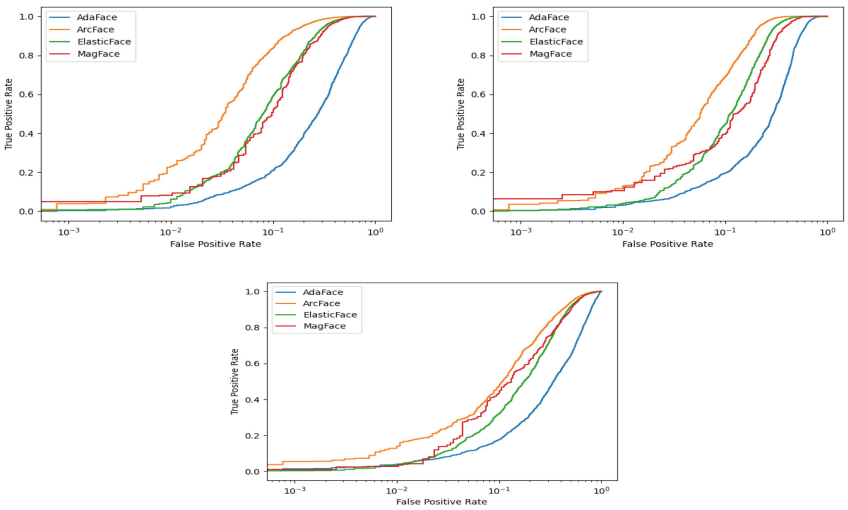
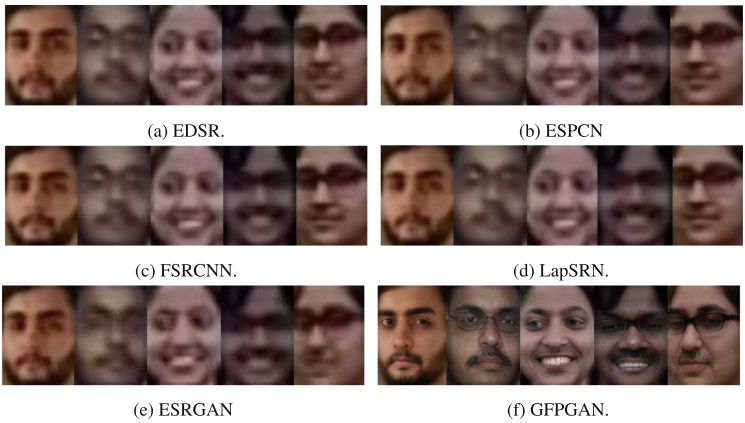


Fig. 4. ROC curves for face verification task using pre-trained AdaFace, ArcFace, ElasticFace, and MagFace FR models at distances 2 m (upper right), 5 m (upper left), and 10 m (lower).

(i) face verification and (ii) soft biometrics analysis, i.e., age and gender classification. In this section, we first describe the experimental results and analysis concerning the face verification algorithms. Later, face attribute classification results are discussed. In the end, we present an interesting study highlighting the need for novel and robust face analysis networks by conducting several experiments using existing gender and ethnicity-balanced datasets such as UTKFace [37] and FairFace [7] along with highly used low-resolution test dataset TinyFace [8].

4.1 Face Verification Results and Analysis

Face verification is a 1:1 matching process where, along with the probe image, an identity is also provided to match the corresponding identity's gallery image. If the matching score exceeds the predefined threshold, the match is termed as genuine else imposter. For face verification, we have used two popular SOTA models, namely VGG-Face [31] and ArcFace [12] available in the deepface library [39]. The ROCs of face verification on unprocessed images are reported in Fig. 3. As expected, as soon as the distance between the subjects and the camera increases, the performance of the face verification networks starts decreasing. The detected faces from the video at 10 m distance are blurred and contain noises with low resolutions, as shown in Fig. 2, which decreases the performance of the face verification networks. Therefore, the analysis of the face verification experiment can be divided into two parts: (i) analysis based on the face verification networks and (ii) impact of distance measure. The ArcFace yields significantly lower performances of the two CNNs used, as shown in Table 3. For example, the face verification accuracy of the ArcFace in terms of micro average [32] is **10**% lower than VGG-face at 10^{-3} false accept rate when the face images captured at 2 m. Further, as the distance increases from 2 m to 10 m, and the gap between the



(a) EDSR. (b) ESPCN

(c) FSRCNN. (d) LapSRN.

(e) ESRGAN (f) GFPGAN.

Fig. 5. Sample images processed using various super-resolution models performed on 10 m raw face images.

Table 5. Comparison of TPR at different FPR rates using macro and micro average aggregation of score vectors for the face verification task on the 10 m processed face images using various SR models.

Models	10^{-3}		10^{-2}	
	Macro	Micro	Macro	Micro
EDSR	0.01	0.41	0.08	0.41
ESPCN	0.01	0.38	0.08	0.38
FSRCNN	0.01	0.38	**0.09**	0.38
LapSRn	0.01	0.40	0.08	0.40
ESRGAN	0.01	0.30	**0.09**	0.30
GFPGAN	**0.02**	**0.52**	0.08	**0.52**

accuracy of both networks increases drastically, i.e., face images captured at 10 m distance are found highly challenging for ArcFace. While the performance of VGG-Face also suffers a drastic reduction, it shows significantly better robustness than ArcFace, even on large stand-off distances.

We have also evaluated other SOTA benchmark DFR networks such as FaceNet [38], OpenFace [2], DeepID [44], ElasticFace [4], AdaFace [23], and MagFace [30]. It is observed that these networks also suffer huge performance drops on the proposed unconstrained dataset and found that their performance is, most of the time, lower than ArcFace for each distance. Figure 4 shows the ROC curves of the face recognition, where the TPR values at different FPR at 10 m are given in Table 4.

Face Super-Resolution. From Figure(s) 1 and 2, we can see the quality of the detected faces at a 10 m distance is poor; therefore, we assert that improving the quality of low-resolution detected faces at a distance of 10 m might be a solution of improved accuracy. For that, we have experimented with the following SOTA super-resolution (SR) models with an upscaling factor of 4: (i) EDSR [26], (ii) FSRCNN [14], (iii) ESPCN [40], (iv) LapSRN [25], (v) ESRGAN [48], and (vi) GFPGAN [47].

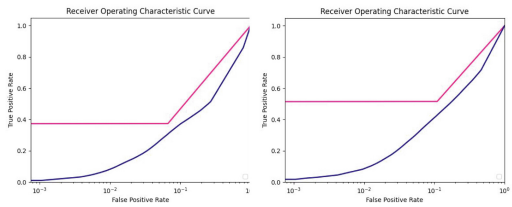


Fig. 6. Comparison of ROC curves for face verification task using VGG-face network as the backbone on the raw face images (left) with enhanced SR face images processed using GFPGAN model (right), at 10 m distance.

Through empirical evaluation, we found that the VGG-Face model performs best; hence, we have utilized that for future face verification experiments on SR face images. It can be seen from Fig. 5 that while the majority of SR models fail drastically in improving the image quality, few models significantly improve the quality. It is also clear from our analysis, presented in Table 5, that EDSR, ESPCN, FSRCNN, and LapSRN models do not show significant performance improvement compared to raw face images (without SR). Further, ESRGAN models yield an 8% lower micro average accuracy than raw images, which might be due to increased image noise in SR images. In contrast, as seen in Fig. 6, GFPGAN increases the image quality significantly, which also shows through a jump of 14% in the verification accuracy.

4.2 Soft Biometrics Analysis

Apart from performing face verification, which aims to find a person's identity, we have also conducted a detailed study to extract several soft biometrics attributes, such as age and gender. In this session, first, we describe the findings related to gender classification, then age identification results are described, and at the end, experimental findings are reported using the balanced datasets, namely UTKFace [52] and FairFace [21].

Gender Classification Results and Analysis. To perform the gender classification, multiple face detectors are used wrapped in the Deepface library, namely (i) OpenCV (Haar-Cascade) [46], (ii) MTCNN [51], (iii) RetinaFace [11], and (iv) SSD [27]. The prime reason for using multiple face detectors is to understand

Table 6. Confusion matrix of gender classification on 1305 random frames at each distance using various face detectors

Predicted / True	RetinaFace		MTCNN		SSD		OpenCV	
	M	F	M	F	M	F	M	F
Acquisiition Distance 2 meters								
M	**749**	1	**748**	2	**733**	17	**732**	18
F	346	**209**	340	**215**	350	**205**	467	**88**
Acquisiition Distance 5 meters								
M	**749**	1	**748**	2	**710**	40	**741**	9
F	402	**153**	444	**111**	540	**15**	450	**105**
Acquisiition Distance 10 meters								
M	**747**	3	**744**	6	**741**	9	**749**	1
F	488	**67**	497	**58**	539	**16**	489	**66**

Table 7. F1-scores for gender classification using various face detectors at 2, 5, and 10 m.

Distance	RetinaFace	MTCNN	SSD	OpenCV
2 m	0.68	**0.69**	0.66	0.51
5 m	**0.61**	0.55	0.38	0.54
10 m	**0.48**	0.47	0.39	**0.48**

the robustness of these SOTA models, which are heavily popular for face detection, and it is shown that the effectiveness of the detection model significantly impacts the recognition performance [1,20]. Table 6 shows the confusion matrix of gender classification concerning face images captured across distance.

Through the evaluation using multiple face detectors, it is found that the gender classification algorithms are highly biased towards the male class. Further, for the male class, the gender classification performance on the majority of the face detectors is either higher or similar on 10 m images as compared to 2 and 5 m images. However, for female images, the detectors show a drastic reduction in gender classification performance. As observed in the literature, Table 7 reveals that the effective selection of face detectors is crucial for gender classification across distances. For example, Retinaface and MTCNN perform similarly in terms of accuracy at 2 m distance images, but in terms of speed and accuracy on large-stand-off, Retinaface outperforms MTCNN for gender classification. In brief, it is observed that RetinaFace performs consistently better or comparable to the other detectors across the acquisition distances. The reason for the poor performance might be the low resolution of the detected faces.

Table 8 shows the confusion matrix of gender classification using various SR models along with various image enchantment techniques. For the male class, the F1 score of gender classification on SR images without filters shows a slight improvement of 2% compared to the raw images. However, when we applied sharpening and denoising filters on these SR images, the performance increased by 9%. For instance, FSRCNN classified 200 female images captured with filters, compared to 82 images captured without filters on SR images, and for males, almost all are correctly classified. This demonstrates the models' biased nature towards male classifiers at a threshold value of 50%. To eliminate bias, we experimented with 26 random validation subjects to determine the best threshold value, which is found to be 20%. When the percentage of female prediction exceeds 20%, the model classifies the image as a female. Therefore, from the Tables 8 and 9, among all the experiments, the FSRCNN super-resolution model with filters at a threshold value of 20% (TH-20) provided the best results, with an F1-score of 0.77. Figure 7 shows a few samples where gender classification fails and works correctly on 10 m distance images. A few reasons for misclassification include occlusion, the presence/absence of a mustache, and glasses.

Apart from utilizing existing pre-trained models, we fine-tuned the VGG16 network using a training set from the proposed dataset consisting of 26 subjects

Table 8. Confusion matrix of gender classification using Retinaface face detector and various SR models with image processing techniques at 10 m distances.

Predicted	EDSR		ESPCN		FSRCNN		LapSRN	
True	M	F	M	F	M	F	M	F
Without Filters								
M	**743**	7	**742**	8	**739**	11	**743**	7
F	484	**71**	485	**70**	473	**82**	473	**82**
With Filters								
M	**730**	20	**718**	32	**710**	40	**726**	24
F	415	**140**	405	**150**	355	**200**	402	**153**
With Filters and Threshold value of 20								
M	**633**	117	**595**	155	**599**	151	**610**	140
F	232	**323**	207	**348**	137	**418**	206	**349**

Table 9. F1-scores for gender classification using SR models without filters and with filters at threshold (TH) values of 50 and with filters at TH of 20 at 10 m.

Models	EDSR	ESPCN	FSRCNN	LapSRN
Without Filters	0.49	0.49	0.50	0.51
With Filters	0.58	0.59	0.64	0.60
TH-20	**0.72**	**0.71**	**0.77**	**0.72**

with an equal gender ratio. Our experiments involved two types of images: raw detected faces and FSRCNN super-resolution images after filter-based enhancement. Once the model is trained for gender classification, we evaluated the model's performance on 61 subjects and found the macro F1-score of 0.50 when raw face images are used. It increases to 0.52 once the enhanced images are used for evaluation. These results obtained using pre-trained and model trained from scratch demand a robust gender classification network capable of handling artifacts present in mobile faces captured at large-standoff.

Age Classification Results and Analysis. We have also studied the performance of deep face attribute classifiers using different face detectors for age classification. In this research, the age classification has been performed based on the following principle: "if the predicted age by the model is in the range of $\pm k$ of the true age, then treated as a correct classified age else labeled as misclassified age sample". Where k is referred to as the tolerance age limit, for example, if the true age is 24 and $k = 2$, then the correct classification will be when the network predicts the age in the range of 22 to 26. Table 10 shows the number of correctly classified samples using different face detectors across distances. As expected,

Table 10. Age identification with different tolerance limits (2, 3, 5, and 10) in terms of the number of images correctly classified samples.

Detectors	2 m				5 m				10 m			
	2	3	5	10	2	3	5	10	2	3	5	10
RetinaFace	72	114	227	573	55	91	221	564	62	135	386	851
MTCNN	66	102	247	557	34	68	197	595	**82**	**154**	**463**	**902**
SSD	58	111	293	763	5	14	283	719	4	12	227	766
OpenFace	**128**	**195**	**399**	**835**	**107**	**181**	**407**	**826**	74	135	402	853

Table 11. Age identification with different tolerance limits in terms of the number of images correctly classified using various SR models without filters (w/o) and with filters (w/) at 10 m.

Super-Resolution	2		3		5		10	
	w/	w/o	w/	w/o	w/	w/o	w/	w/o
EDSR	170	161	303	275	716	670	1140	1122
FSRCNN	167	97	294	210	**741**	504	**1166**	1076
LapSRN	164	158	227	267	713	684	1139	1134
ESPCN	175	**179**	307	**310**	740	709	1161	1162

with the increase of the tolerance limit, the performance of each face detection-based age classification network increases drastically. However, one caveat of such performance improvement is that the higher the tolerance limit, the search space in which we need to find a person will also be significantly high. The age classification performance also shows somewhat improved performance even when the distance of acquisition increases. For example, for the RetinaFace detector, the number of correctly classified samples at 2 m distance increases from 573 to 851 when the images are captured at a 10 m distance. A similar jump can also be observed when the MTCNN face detector-based age classification has been used. However, for tolerance level 2, the number of correctly classified images is very low, which needs to be considered for accurate prediction. Therefore, for the age classification, we have experimented with various super-resolution models with and without filters, as shown in Table 11. The age classification performance on images obtained after super-resolution algorithms shows better results than raw images. For instance, the number of correctly classified samples increases from 82 to 179 for images with ESPCN resolution with a tolerance level of 2 at a 10 m distance. However, adding filters did not significantly affect the age classification task, so we can exclude this step to optimize time.

Comparison with Other Datasets. In this experiment, we employed pre-trained models that are trained on the UTKFace [37] and FairFace [7] datasets to evaluate the robustness of soft biometrics analysis networks when they are

trained on balanced datasets and evaluated on our unconstrained dataset. Both datasets are comprised of face images annotated for age, gender, and ethnicity with variations in pose and resolution. We tested the performance of these models [7,37] on raw face images captured at distances of 2, 5, and 10 m and also with processed images using the best-performing super-resolution algorithm. The experimental results are shown in Table 12. It is worth noting that even in the model trained on one balanced dataset, namely UTKface, the gender classification performance shows a significant decline in performance and even yields lower accuracy than the deepface model. The UTKface model for age prediction performed slightly better than the deepface model for a tolerance limit of 2 but poorly for a limit of 10. On the other hand, Gender classification using a FairFace pre-trained model shows significantly higher performance, but the network shows bias towards males as the recall percentage for men is 97% compared to 63% for women at 10 m distance images. On the other hand, age prediction using the FairFace model performed poorly and only correctly classified 544 images, even at a tolerance limit of 10 for 10 m distance images. Furthermore, We also analyzed the ethnicity results produced using the FairFace model on 10 m images. Out of 1305 Indians, this model could only predict 271, indicating

Fig. 7. Full images and cropped face region samples from our collected dataset. Along with images, we have mentioned the true gender and age of each sample and the predicted age and gender by the deep face model. The SOTA models are highly ineffective on Indian ethnicity subjects in identifying age and gender. Red and blue represent the true and predicted values, respectively. (Color figure online)

Table 12. Macro-F1 Scores for gender classification of our proposed dataset at varying distances, and with processed images obtained using the FSRCNN SR model, at a distance of 10 m.

Distance	Model trained on		
	UTKFace	FairFace	MobileFaces
2 m	0.70	**0.96**	0.77
5 m	0.52	**0.93**	0.68
10 m	0.44	**0.81**	0.62
SR(10 m)	0.48	**0.80**	0.64

Table 13. Top 3 Rank accuracies of TinyFace and MobileFaces dataset for face identification task.

FRModel	TinyFace			MobileFaces		
	R1	R2	R3	R1	R2	R3
ArcFace	45.06	49.49	51.88	30.91	38.88	46.94
AdaFace	67.81	69.37	70.25	25.02	34.54	40.04
MagFace	61.45	63.60	64.86	23.78	32.22	37.49

that the current dataset lacks Indian ethnicity data. Therefore, at last, we only experiment on the Indian ethnicity data extracted from UTKface and FairFace dataset, and the model [45] trained using the VGG16 network as the backbone for the model. Even though this model also does not give good results at distance images, it shows interesting results with oppositely biased nature dominance of female prediction. This shows the lack of Indian ethnicity data, especially the male and long-range variations in the dataset. In the end, we have also experimented with the recently highly used low-quality testing dataset TinyFace [8] for face identification task (1:N matching). Table 13 shows that the rank accuracy of our MobileFaces is significantly lower than the TinyFace, which shows that our dataset can become a new benchmark testing dataset for improving model unbiasedness in a real scenario.

5 Conclusions

In this research, we have presented a unique mobile surveillance dataset, namely *'MobileFaces dataset'*[2]. The images are captured at multiple standoff distances between subjects and the camera, multiple times a day, at multiple locations, and in natural, unconstrained conditions. Extensive experiments performed using the proposed dataset demonstrate that current DFR models are not robust enough to identify individuals or perform soft biometrics prediction, especially when the images are captured at large standoff (say 10 m). We have also evaluated several image super-resolution models to check whether image enhancement is effective in improving performance; however, based on the performance, it observed that these models are not only ineffective, but also the selection of an accurate super-resolution model is also necessary. Based on the limitations of the existing models, we aim to develop robust face recognition networks. Apart from that, we aim to increase the number of subjects drastically to ensure coverage of a wide range of populations and ethnicities.

References

1. Amato, G., Falchi, F., Gennaro, C., Massoli, F.V., Vairo, C.: Multi-resolution face recognition with drones. In: ICSSIP, pp. 13–18 (2020)

[2] The dataset has been captured by following the proper user consent and is released to the research community on the following link.

2. Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al.: OpenFace: a general-purpose face recognition library with mobile applications. CMU Sch. Comput. Sci. **6**(2), 20 (2016)
3. Bansal, A., Nanduri, A., Castillo, C.D., Ranjan, R., Chellappa, R.: UMDFaces: an annotated face dataset for training deep networks. In: IEEE IJCB, pp. 464–473 (2017)
4. Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A.: ElasticFace: elastic margin loss for deep face recognition. In: IEEE/CVF CVPR, pp. 1578–1587 (2022)
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. In: IEEE FG (FG 2018), pp. 67–74 (2018)
6. Chalabi, N.E., Attia, A., Bouziane, A., Hassaballah, M., Akhtar, Z.: Recent trends in face recognition using metaheuristic optimization. In: Mohamed, A.W., Oliva, D., Suganthan, P.N. (eds.) Handbook of Nature-Inspired Optimization Algorithms: The State of the Art. SSDC, vol. 213, pp. 85–112. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-07516-2_5
7. Chen, D.: Fairface: face attribute dataset for balanced race, gender, and age (2021). https://github.com/dchen236/FairFace
8. Cheng, Z., Zhu, X., Gong, S.: Low-resolution face recognition. In: ACCV, pp. 605–621 (2019)
9. Chhokra, P., Chowdhury, A., Goswami, G., Vatsa, M., Singh, R.: Unconstrained kinect video face database. Inf. Fusion **44**, 113–125 (2018)
10. Czajka, A., Fang, Z., Bowyer, K.: Iris presentation attack detection based on photometric stereo features. In: IEEE WACV, pp. 877–885 (2019)
11. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: single-shot multi-level face localisation in the wild. In: IEEE/CVF CVPR, pp. 5203–5212 (2020)
12. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: IEEE/CVF CVPR, pp. 4690–4699 (2019)
13. Djouamai, Z., Attia, A., Chalabi, N.E., Hassaballah, M.: A face recognition system based-almmo-0 classifier. Evol. Syst. 1–18 (2023)
14. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 391–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_25
15. Duong, C.N., Quach, K.G., Jalata, I., Le, N., Luu, K.: Mobiface: a lightweight deep learning face recognition on mobile devices. In: IEEE BTAS, pp. 1–6 (2019)
16. Goh, R., Liu, L., Liu, X., Chen, T.: The CMU face in action (FIA) database. In: Zhao, W., Gong, S., Tang, X. (eds.) AMFG 2005. LNCS, vol. 3723, pp. 255–263. Springer, Heidelberg (2005). https://doi.org/10.1007/11564386_20
17. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 87–102. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_6
18. Gupta, S., et al.: Facesurv: a benchmark video dataset for face detection and recognition across spectra and resolutions. In: IEEE FG, pp. 1–7 (2019)
19. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition (2008)
20. Kalra, I., Singh, M., Nagpal, S., Singh, R., Vatsa, M., Sujit, P.: DroneSURF: benchmark dataset for drone-based face recognition. In: IEEE FG, pp. 1–7 (2019)

21. Karkkainen, K., Joo, J.: Fairface: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: IEEE/CVF WACV, pp. 1548–1558 (2021)
22. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: IEEE CVPR, pp. 4873–4882 (2016)
23. Kim, M., Jain, A.K., Liu, X.: Adaface: quality adaptive margin for face recognition. In: IEEE/CVF CVPR, pp. 18750–18759 (2022)
24. Klare, B.F., et al.: Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A. In: IEEE CVPR, pp. 1931–1939 (2015)
25. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep Laplacian pyramid networks for fast and accurate super-resolution. In: IEEE CVPR, pp. 624–632 (2017)
26. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: IEEE CVPRW, pp. 136–144 (2017)
27. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
28. Manchanda, S., et al.: D-LORD: DYSL-AI database for low-resolution disguised face recognition. IEEE TBIOM (2023)
29. Maze, B., et al.: IARPA Janus benchmark-C: face dataset and protocol. In: IEEE ICB, pp. 158–165 (2018)
30. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: a universal representation for face recognition and quality assessment. In: IEEE/CVF CVPR, pp. 14225–14234 (2021)
31. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
32. Pedregosa, F., et al.: Scikit-learn: machine learning in Python (2021). https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html
33. Pezdek, K., Lerer, T.: Let's go to the tape: science-based standards for non-eyewitness identifications in a surveillance world. **59** (2023)
34. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. Image Vis. Comput. **16**(5), 295–306 (1998)
35. Rathore, U., Agarwal, A.: Is DFR for soft biometrics prediction in unconstrained images fair and effective? In: The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, 5 May 2023 (2023)
36. Ren, H., Sun, L., Guo, J., Han, C.: A dataset and benchmark for multimodal biometric recognition based on fingerprint and finger vein. IEEE TIFS **17**, 2030–2043 (2022)
37. Roy, A.: Face data based deeplearning (2020). https://github.com/abr-98/Face_data_based_deeplearning
38. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: IEEE CVPR, pp. 815–823 (2015)
39. Serengil, S.I., Ozpinar, A.: Lightface: a hybrid deep face recognition framework. In: IEEE ASYU, pp. 23–27 (2020). https://doi.org/10.1109/ASYU50717.2020.9259802
40. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: IEEE CVPR, pp. 1874–1883 (2016)
41. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: IEEE FG, pp. 53–58 (2002)
42. Singh, M., et al.: Cross-spectral cross-resolution video database for face recognition. In: IEEE BTAS, pp. 1–7 (2016)

43. Sujitha, S.: The admissibility & reliability of electronic evidence in Indian courts & its legal framework
44. Sun, Y., Liang, D., Wang, X., Tang, X.: Deepid3: face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
45. Tyagi, U.: Age and gender prediction model (trained on Indian faces) (2021). https://github.com/ujjwaltyagi355/Age-Gender-Classfication
46. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: 2IEEE CVPR, vol. 1, p. I (2001)
47. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: IEEE/CVF CVPR, pp. 9168–9178 (2021)
48. Wang, X., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. In: ECCVW (2018)
49. Whitelam, C., et al.: IARPA Janus benchmark-B face dataset. In: IEEE CVPRW, pp. 90–98 (2017)
50. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
51. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE SPL **23**(10), 1499–1503 (2016)
52. Zhang, Zhifei, S.Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: IEEE CVPR (2017)
53. Zhong, Y., Deng, W., Hu, J., Zhao, D., Li, X., Wen, D.: SFace: sigmoid-constrained hypersphere loss for robust face recognition. IEEE TIP **30**, 2587–2598 (2021)

UAV-Enhanced Combination to Application: Comprehensive Analysis and Benchmarking of a Human Detection Dataset for Disaster Scenarios

Ragib Amin Nihal[1(✉)], Benjamin Yen[1], Katsutoshi Itoyama[2], and Kazuhiro Nakadai[1]

[1] Institute of Science Tokyo, Tokyo, Japan
{ragib,benjamin,nakadai}@ra.sc.e.titech.ac.jp
[2] Honda Research Institute Japan Co., Ltd., Wako, Japan
itoyama@ra.sc.e.titech.ac.jp

Abstract. Unmanned aerial vehicles (UAVs) have revolutionized search and rescue (SAR) operations, but the lack of specialized human detection datasets for training machine learning models poses a significant challenge. To address this gap, this paper introduces the Combination to Application (C2A) dataset, synthesized by overlaying human poses onto UAV-captured disaster scenes. Through extensive experimentation with state-of-the-art detection models, we demonstrate that models fine-tuned on the C2A dataset exhibit substantial performance improvements compared to those pre-trained on generic aerial datasets. Furthermore, we highlight the importance of combining the C2A dataset with general human datasets to achieve optimal performance and generalization across various scenarios. This points out the crucial need for a tailored dataset to enhance the effectiveness of SAR operations. Our contributions also include developing dataset creation pipeline and integrating diverse human poses and disaster scenes information to assess the severity of disaster scenarios. Our findings advocate for future developments, to ensure that SAR operations benefit from the most realistic and effective AI-assisted interventions possible. The dataset, code, and model are publicly available at: https://github.com/Ragib-Amin-Nihal/C2A.

Keywords: Aerial Object Detection · UAV (Unmanned Aerial Vehicle) · Human Detection · Disaster Response · Search and Rescue (SAR) · Artificial Intelligence in Disaster Relief · Emergency Management · Benchmark Dataset

1 Introduction

The advancement of UAVs, colloquially known as drones, has signaled a new era in the field of emergency response and disaster management. With their unparalleled agility and ability to provide an aerial perspective, drones have rapidly become indispensable assets in the arsenal of SAR operations worldwide.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 145–162, 2025.
https://doi.org/10.1007/978-3-031-78341-8_10

These technological marvels significantly improve the efficiency and effectiveness of missions aimed at locating and aiding people in disaster-hit areas [13]. Drones can have a significant impact on minimizing the aftermath of disasters through time efficiency, making a difference in survival and fatality rates.

Despite these advancements, a major shortcoming exists in the deployment of drone technologies—particularly in the area of object detection via drone vision. Existing computer vision or drone vision systems significantly depend on datasets to train detection algorithms. However, these datasets are primarily designed for general situations and do not adequately address the specialized and intricate requirements of disaster contexts. The shortage of disaster detection datasets is mostly owing to logistical and ethical obstacles in capturing and annotating real events, which need substantial resources and often involve sensitive circumstances. The ethical dilemmas of capturing vulnerable people during actual catastrophes aggravate the issue. This lack of specialized human detection datasets for SAR operations hinders the capability of drones to effectively identify human figures disaster scenarios. This deficiency is particularly acute, as our findings indicate current pre-trained detection models fall short of effectively identifying humans (more discussed in Sect. 6.1) amidst the multifaceted chaos of disaster scenes—where the stakes are very high.

In critical scenarios, the lack of datasets with the necessary detail for training advanced machine learning algorithms hinders the optimization of drone capabilities. Recognizing the critical need for a specialized dataset, our research introduces a novel dataset explicitly designed to improve human detection capabilities in disaster scenarios through drone vision. This dataset is created by combining human posture images with disaster scene backgrounds, resulting in an intricate collection of images that simulate the diverse settings faced during real-world SAR operations. By integrating human figures into a variety of disaster scenes, our approach aims to mirror the array of challenges that SAR drones are likely to face, facilitating the development of more robust and effective machine learning models tailored for disaster response applications [21].

Our endeavor is inspired by the growing field of disaster management technology, where recent breakthroughs highlight the revolutionary potential of machine learning and computer vision in improving SAR operations [4]. Our research focuses on creating a challenging dataset to train AI models for detecting partially occluded humans, a common scenario in disaster-stricken environments. Such occlusions, resulting from individuals being trapped under debris or obscured by various objects, represent a significant hurdle for human detection systems. Existing datasets scarcely address this challenge, primarily due to the inherent difficulty in replicating these complex scenarios accurately. Our dataset, therefore, signifies a pioneering step towards addressing this gap, offering a resource that simulates the conditions under which SAR operations unfold in reality.

We aim to enhance the precision and reliability of drone-operated human detection in emergency scenarios, underscoring the nuanced requirements of effective disaster response strategies. The construction of this dataset involved

Fig. 1. This collection of images presents a selection from our proposed "Combination to Application" (C2A) dataset, a specialized compilation designed to refine machine learning algorithms for SAR operations in diverse disaster scenarios. Within the bounding boxes, human figures are superimposed onto various disaster scenes, demonstrating the intricate process of overlaying accurately segmented human poses onto different disaster backdrops such as rubble, traffic incidents, flood, and fire. This synthetic approach is crucial for creating more challenging training conditions that AI models may encounter in actual SAR missions. Furthermore, the dataset is enriched with detailed pose information-such as bent, kneeling, lying, sitting, and upright-providing comprehensive data for AI to learn and recognize human forms even when partially occluded by environmental obstacles.

a process of incorporating elements of obstruction to emulate the visibility challenges frequently encountered in real-world disaster settings. This approach facilitates the training of AI models that are not only adept at recognizing human forms in clear view, but are also capable of inferring the presence of individuals in less than optimal visibility conditions. Such abilities are necessary for quickly finding emergency areas and finding survivors who need help right away, which makes SAR missions much more effective [2,8].

Through this comprehensive paper, we delineate the process undertaken to create our novel synthesized dataset (Fig. 1), emphasizing its designed complexity and the specific challenges it poses to AI models. We delve into the selection criteria for the images, the sophisticated image processing techniques employed, and the rationale underlying the dataset's structure. We also present our findings from using this dataset to train deep learning models, which show notable improvements in detection accuracy and the feasibility of operational deployment. In essence, our research essentially provides the field of disaster management with a substantial C2A dataset. Researchers and professionals can use this dataset to extend the potential of drone technology, which will make emergency responses and life-saving operations more effective in chaotic situations. This paper presents our contributions to the field as follows:

- We introduce a novel synthesized dataset that mitigates the gap in current SAR operations by providing imagery capturing human figures in disaster contexts, designed to train machine learning models for complex human detection tasks.
- We present a comprehensive dataset creation pipeline that combines advanced image processing techniques and domain-specific knowledge, resulting in a dataset that represents the complexity and unpredictability of disaster scenarios.
- Our dataset includes a variety of human poses and disaster scene information, allowing researchers to develop models that can assess the severity of a disaster scene and prioritize rescue efforts.
- Our preliminary results demonstrate that the dataset significantly improves the detection performance and operational feasibility of deep learning models, indicating the dataset's potential to transform disaster response efforts.

2 Literature Review

Robust datasets are fundamental to developing and training precise machine learning models. The importance of datasets tailored to specific application domains is well established in the literature. Studies such as those by [1] and [27] emphasize the need for datasets that encompass the intricacies of various disaster scenarios. These resources are pivotal for calibrating UAV-operated detection systems to recognize human subjects under a multitude of conditions. Yet, there remains a scarcity of datasets that accurately mirror the complexities of disaster-hit environments.

Table 1. Comparison of Datasets for Human Detection in Disaster Scenarios

Feature	C2A (proposed)	LoveNAS [30]	Smoke Scene [31]	M4SFWD [29]	SARD [25]
Focus	Human detection in disaster scenarios with partial occlusion	Land-cover mapping for multiple scenes including disaster scenarios	Detection of smoke scenes from satellite imagery for early disaster response	Synthetic dataset for remote sensing forest wildfires detection	Human detection in search and rescue operations using drone imagery
Scenarios Covered	Various disaster scenarios (earthquakes, flood, fire) with partial occlusion	Urban, rural, and disaster scenes	Areas prone to fire outbreaks	Forest and wildfire scenarios	Search and rescue operations in non-urban areas (no disaster scene)
Partial Occlusion	Yes	No	No	No	Yes
Image Diversity	High, with images from multiple disaster types including occluded humans	High, including three normal and two disaster scenes	Moderate, focused on smoke detection	High, developed through post-processing and synthesis	Moderate, focused on search and rescue scenarios
Realism	Moderate, designed to mimic real disaster conditions with individuals	Moderate, lacks specific focus on human detection	High for smoke detection, moderate for overall disaster realism	High for wildfires, moderate for human detection relevance	High, includes realistic search and rescue scenes captured by drones
Human Poses	Bent, Kneeling, Lying, Sitting, Upright	Not applicable	Not applicable	Not applicable	Diverse, not included in annotation

Debris and destruction frequently obscure human subjects in disaster zones, biasing UAV detection. According to the literature, existing datasets and models are progressing, but they fall short of providing the granular detail required for reliable detection in such complex circumstances [18,19]. A significant body of research, including work by [11], emphasizes the pressing need for advancing UAV technology to navigate these obstacles adeptly. Yet, the development of datasets that reflect the reality of partial occlusions in disaster contexts is still in its early stages, indicating a pivotal area for future research.

The evolution of machine learning and computer vision has been enhancing to UAV capabilities. These advancements have paved the way for more nuanced data analysis, crucial for discerning human presence within complex terrains. Even with these improvements, the research shows that algorithms and models still need to be improved, especially to make human detection more reliable [12].

One notable effort in this direction is the Search and Rescue Drone (SARD) dataset [25], which focuses on human detection in search and rescue operations using drone imagery. The SARD dataset includes images of people in various poses simulating exhausted or injured individuals, captured in non-urban environments. While it provides a valuable resource for developing detection models for SAR scenarios, the SARD dataset does not explicitly include disaster scenes, which pose additional challenges such as debris, occlusions, and clutter. Moreover, although the dataset contains diverse human poses, this information is not directly incorporated into the annotations, limiting its utility for pose-aware detection.

The main gap is the absence of comprehensive datasets capturing the full spectrum of disaster scenarios, specifically focusing on the aspect of partial human occlusion. Moreover, there is an evident need for further exploration into advanced machine learning and computer vision applications tailored for SAR operations. The new dataset is compared to existing datasets about finding people in disaster situations in Table 1.

This comparison indicates the distinctive contributions of the newly developed dataset, particularly its emphasis on effective human detection in disaster scenarios, a topic not explicitly addressed by the existing datasets. The newly developed dataset sets itself apart from existing ones, which narrowly focus on specific types of disaster scenes or aspects like smoke and fire detection.

The new dataset mitigates in a gap in specifically designed datasets for finding partially occluded individuals in disaster scenarios. Its goal is to greatly enhance the abilities of machine learning models for disaster response and SAR operations, ultimately leading to more effective and timely humanitarian efforts.

3 Dataset Creation Pipeline

We developed a systematic pipeline to produce a comprehensive set of images for training machine learning models to detect humans in disaster scenarios. The dataset combines parts of the Aerial Image Dataset for Emergency Response Applications (AIDER) and the LSP/MPII-MPHB dataset. It shows a variety of human poses on a range of disaster backgrounds.

3.1 Data Sources and Composition

AIDER (Aerial Image Dataset for Emergency Response Applications): The AIDER dataset [14] serves as the foundation for the disaster scene backgrounds. It comprises images from four major disaster types: Fire/Smoke (320 images), Flood (370 images), Collapsed Building/Rubble (320 images), and Traffic Accidents (335 images). These authentic disaster images offer a realistic portrayal of the chaotic and unpredictable conditions typical in emergency scenarios. We did not utilize the 1,200 normal case images to keep the focus on emergency situations. This dataset offers a glimpse into the chaotic and unpredictable environments that characterize disaster scenes, making it an ideal choice for our purposes.

LSP/MPII-MPHB (Multiple Poses Human Body): For the human subjects, we sourced images from the LSP/MPII-MPHB dataset [3,10], which contains 26,675 images featuring 29,732 instances of human bodies in various poses. This dataset is specifically designed to capture a wide range of human body positions, including bent, kneeling, sitting, upright, and lying, providing the necessary diversity to train models for detecting humans under different conditions. The detailed annotations of human poses in this dataset are critical for training models to recognize human figures in complex disaster environments.

3.2 Pipeline Steps

1. Background Removal and Image Preparation: Using the U2Net segmentation model [23], we isolated human figures from the LSP/MPII-MPHB dataset by removing the background. The U2-Net, short for "U-Squared Net", is a deep neural network known for its powerful performance in salient object detection and image segmentation tasks. It employs a nested U-structure that enhances the learning of local and global features within images, enabling precise segmentation of objects, including human figures, from their backgrounds. This process involved saving each figure with its respective pose in a separate folder, ensuring that the focus remained on the human subject without any background distractions.

2. Image Cropping and Cleaning: In the next step, the isolated images were then cropped to highlight the human figures, removing unnecessary peripheral content. This step involved calculating the minimum and maximum indices of non-zero pixel elements to determine the bounding box for each figure. Images where non-zero indices constituted less than 2% of the total image area were excluded to minimize noise and inaccuracies.

3. Overlay Process: For each disaster background from the AIDER dataset, human figures from the LSP/MPII-MPHB dataset were overlaid at random positions. The scaling of human figures in the dataset was randomized within specified lower and upper bounds, taking into account the dimensions of the disaster scene backgrounds. This approach was employed to mimic the diverse scales at which humans may be observed in real disaster scenarios. This process also included checks for collisions and adjustments to the placement of figures to ensure a realistic composition. The final images were annotated with bounding boxes, accurately reflecting the position and scale of each human figure within the disaster scene.

The dataset creation pipeline integrates human poses with disaster scene backgrounds to construct a comprehensive dataset for training machine learning models for human detection within disaster scenarios. The Algorithm 1 outlines the steps involved in the dataset creation process.

Algorithm 1. Dataset Creation Pipeline for Human Detection in Disaster Scenarios. This pipeline integrates human poses with disaster scene backgrounds, involving background removal, cropping, random scaling, and overlaying to simulate realistic disaster environments for training AI models.

Input:
$AIDER$: Set of disaster scene images
$MPHB$: Set of human pose images
Output: Combined dataset D with annotated human poses in disaster scenes
procedure U2NetRemoveBackground($image$)
    Apply U2-Net model to $image$ for segmentation
    Extract foreground (human) based on segmentation result
    return foreground
end procedure
procedure CropFocusedObject($image$)
    Compute bounding box around non-zero pixels in $image$
    Crop $image$ to the bounding box
    return cropped image
end procedure
procedure RandomScale($image$)
    $scale \leftarrow$ random value between predefined min and max
    Resize $image$ by $scale$
    return resized image
end procedure
procedure RandomPosition($background, object$)
    $bgWidth, bgHeight \leftarrow$ dimensions of $background$
    $objWidth, objHeight \leftarrow$ dimensions of $object$
    $x \leftarrow$ random integer from 0 to $bgWidth - objWidth$
    $y \leftarrow$ random integer from 0 to $bgHeight - objHeight$
    return (x, y)
end procedure
procedure CreateDataset($AIDER, MPHB$)
    $D \leftarrow \emptyset$
    for each pose p in $\{bent, kneeling, sitting, upright, lying\}$ **do**
        for each image i in $MPHB$ corresponding to pose p **do**
            $i_{bg\_removed} \leftarrow$ U2NetRemoveBackground(i)
            $i_{cropped} \leftarrow$ CropFocusedObject($i_{bg\_removed}$)
            if Size of $i_{cropped} \geq 0.02\times$ Size of $i_{bg\_removed}$ **then**
                Add $i_{cropped}$ to $MPHB_p$
            end if
        end for
    end for
    for each image a in $AIDER$ **do**
        $H \leftarrow$ Random selection of human poses from $MPHB_p$
        for each human pose h in H **do**
            $h_{scaled} \leftarrow$ RandomScale(h)
            $pos \leftarrow$ RandomPosition(a, h_{scaled})
            Overlay h_{scaled} on a at position pos
            Compute bounding box $bbox$ for h_{scaled} at pos
            Add $(a, bbox)$ to D
        end for
    end for
    return D
end procedure

Table 2. Comparison of various datasets [33] including the proposed C2A dataset

Dataset	Annotation way	# main categories	# Instances	# Images	Image width
SARD	horizontal BB	1	6,532	1,981	1920
M4SFWD	oriented BB	2	17,613	3,946	776–1480
Smoke Scene	oriented BB	2	18,849	8,735	95–6000
Tiny Persons	horizontal BB	1	70,702	1,570	765–2048
Crowd Human	horizontal BB	2	456,098	19,370	400–10800
PASCAL VOC	horizontal BB	20	27,450	11,530	640
MS COCO	horizontal BB	80	∼2.5M	∼328,000	640
NWPU VHR-10	horizontal BB	10	3,651	800	∼1000
3K Vehicle Detection	oriented BB	2	14,235	20	5616
DOTA	oriented BB	14	188,282	2,806	800–4000
C2A (proposed)	**horizontal**	**1 (with 5 poses)**	**>360,000**	**10,215**	**150–3400**

4 Properties of C2A Dataset

The C2A (Combination to Application) dataset[1] is a curated collection specifically designed for advancing human detection disaster scenarios by combining AIDER dataset images (disaster scene backgrounds) and diverse human poses from the LSP/MPII-MPHB dataset. Some of the samples of the dataset are presented in Fig. 1. Comparison of various datasets is shown on Table 2. In this section, we present a comprehensive analysis of the dataset's properties.

4.1 Number of Images and Image Size

In the C2A dataset, the total number of images is $10,215$, encompassing over $360,000$ objects for human detection within disaster scenarios. The original size of the images spans a wide range from approximately 123×152 pixels to high-resolution images of 5184×3456 pixels. This range is significantly broader than what is commonly found in standard datasets like PASCAL VOC or MSCOCO, where the image dimensions generally do not exceed 1000×1000 pixels. The wide range of resolution in the C2A dataset ensures the inclusion of various granular details necessary for the precise detection of humans in diverse and challenging disaster environments. Furthermore, the most common image width range within the C2A dataset is between 322 and 600 pixels, with over 50.32% of images falling within this range. The median image width is noted at 428 pixels, indicative of the dataset's central tendency toward mid-range resolutions. The dataset preserves the integrity of the scenes and avoids potential complications that may arise from segmenting an instance across multiple image pieces.

[1] Dataset available at: https://github.com/Ragib-Amin-Nihal/C2A.

Table 3. Comparison of instance size (in terms of width) distribution of some datasets in aerial images and natural images. Some statistics collected from [33]

Dataset	<10 pixel	10–50 pixel	50–300 pixel	>300 pixel
SARD	0.01	0.66	0.32	0.01
M4SFWD	0.02	0.38	0.51	0.09
Smoke Scene	0	0.25	0.61	0.14
Tiny Persons	0.56	0.42	0.02	0
Crowd Human	0.01	0.42	0.50	0.07
PASCAL VOC	0	0.14	0.61	0.25
MSCOCO	0	0.43	0.49	0.08
NWPU VHR-10	0	0.15	0.83	0.02
3K Munich Vehicle	0	0.93	0.07	0
DOTA	0	0.57	0.41	0.02
C2A (proposed)	**0.47**	**0.52**	**0.01**	**0**

4.2 Objects Size

In our C2A dataset, the pixel size of objects is distributed across a broad spectrum, accommodating the real-world variability in human sizes from an aerial perspective. Specifically, we observe that a substantial 47% of instances are under 10 pixels, indicative of individuals who appear extremely small due to the altitude of the imagery. This reflects realistic scenarios where people are often tiny and challenging to detect. The dataset also contains 52% of instances in the range of 10–50 pixels and a minimal 1% within the 50–300 pixel bracket. There are no instances above 300 pixels, reinforcing the dataset's focus on detecting smaller objects. In Table 3, when compared to datasets like PASCAL VOC and DOTA, the C2A dataset demonstrates a more balanced distribution between small and middle-sized instances. It is challenging for the models to detect the objects that are in tiny size.

4.3 Aspect Ratio of Objects

The aspect ratio (AR) is a critical parameter in anchor-based detection models, influencing the design and effectiveness of detectors like Faster R-CNN and YOLO series. In the C2A dataset, we analyze the AR of the minimally circumscribed horizontal bounding boxes encompassing each object. The histogram in Fig. 2a displays the distribution of these aspect ratios. The distribution is skewed towards smaller ARs, with the majority of objects having an AR less than 1. This suggests that most bounding boxes are wider than they are tall, a likely scenario when dealing with collapsed individuals or those in horizontal positions in disaster scenarios. A noticeable amount of instances have ARs between 1 and 2, aligning with natural human proportions when standing or sitting. Very few

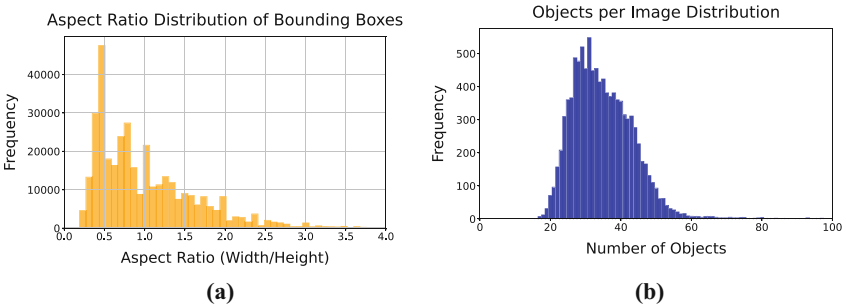


Fig. 2. (a) Aspect Ratio of C2A Dataset (b) Object Density

instances possess a high AR, which is expected as elongated bounding boxes would be less common unless representing individuals in highly unusual orientations or in motion.

4.4 Object Density of Images

Aerial image datasets often exhibit a far greater number of objects per image when compared to datasets composed of natural images. Typical datasets, such as ImageNet, have an average of 2 objects per image, while MSCOCO averages 7.7. In stark contrast, our C2A dataset showcases a higher object density, reflective of the real-world complexity found in disaster-stricken environments.

The histogram depicted in Fig. 2b outlines the frequency of object instances per image within our dataset. The distribution peaks significantly around 20 to 40 objects, with a notable extension towards images containing up to 100 instances. This dense distribution is a testament to the C2A dataset's capacity for providing a challenging and enriched learning context for object detection algorithms, pushing the envelope of their detection and discrimination capabilities.

4.5 Human Pose and Disaster Scene Information

In the pursuit of advancing SAR operations through machine learning, our C2A dataset offers more than object detection; it integrates critical contextual data by providing detailed annotations for both human poses and disaster scene types. The dataset categorizes human figures into five distinct poses: 'Bent', 'Kneeling', 'Lying', 'Sitting', and 'Upright'. These annotations are crucial as they potentially correlate with the urgency and type of assistance required; for example, individuals found 'Lying' or 'Bent' in a disaster scene could indicate severe injury, necessitating immediate medical attention.

Alongside pose information, the C2A dataset is annotated with disaster scene context, such as 'traffic incident', 'fire', 'flood', and 'collapsed building'. This level of detail allows for a nuanced understanding of the environment, providing vital

clues about the challenges present in each unique scenario. Such information is instrumental for developing machine learning models that can not only detect humans in aerial images but also assess the severity and nature of the disaster context. The addition of these rich contextual layer opens new avenues for machine learning applications, potentially transforming the landscape of disaster response and emergency aid.

5 Evaluation

5.1 Evaluation Metrics

The evaluation of object detection models was conducted using the mean Average Precision (mAP) [22], a prevalent metric that integrates both precision and recall aspects of the predictions. Precision, defined as Precision $= \frac{TP}{TP+FP}$, measures the correctness of the predictions, while recall quantifies the model's ability to identify all relevant instances. The mAP is the mean of Average Precision (AP) across all classes, computed for varying Intersection over Union (IoU) thresholds, typically ranging from 0.5 to 0.95. The AP at a specific IoU threshold is the area under the precision-recall curve. The mAP at IoU threshold of 0.5, denoted as mAP@.50, is represented as $\text{mAP@.50} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \big|_{IoU=0.5}$, highlighting a model's proficiency in detecting objects with a moderate overlap with the ground truth.

5.2 Training Options

The evaluation of the models on the C2A dataset was conducted using NVIDIA A100 GPUs, with a uniform batch size of 24 and an image resolution of 640×640 pixels across 50 epochs. The ADAM optimizer was chosen for its efficiency in handling large datasets and complex image structures. Basic data augmentation techniques, such as flipping and resizing, were employed to enhance model robustness and prevent overfitting. The experiments were facilitated by popular deep learning frameworks, specifically mmDetection [6], Detectron2 [32], and Ultralytics [9], known for their high performance in object detection tasks. These frameworks provide extensive support for custom dataset training, enabling the effective application of state-of-the-art detection models to our specialized dataset.

5.3 Benchmarking

The C2A dataset was subjected to a rigorous evaluation process using a suite of state-of-the-art object detection models. These evaluations aimed to benchmark the dataset's performance in training machine learning algorithms for the task of human detection in various disaster scenarios. The models were chosen for their relevance and proven accuracy in similar tasks, with an emphasis on assessing their capability to handle the complexities introduced by varied disaster backgrounds within the dataset. The Table 4 illustrates the results found.

Table 4. Performance Evaluation of state-of-the-art Models on the C2A Dataset

Model	mAP	mAP@.50	Model	mAP	mAP@.50
Faster R-CNN [24]	0.3656	0.6340	Dino [16]	0.4710	0.7890
RetinaNet [15]	0.3834	0.6933	Rtmdet [17]	0.4420	0.7080
Cascade R-CNN [5]	0.4860	0.7350	YOLOv5 [9]	0.4920	0.8080
YOLOv9-e [28]	**0.6883**	**0.8927**	YOLOv9-c [28]	0.5562	0.7996

5.4 Result Analysis

The evaluation results demonstrate a range of performance metrics across different models, reflecting the diverse strengths of each approach. YOLOv9-e outperformed other models with the highest mAP (mean Average Precision) score, indicating its superior ability to detect objects with a high degree of accuracy across varying Intersection over Union (IoU) thresholds. This suggests that the architectural improvements in YOLOv9, particularly for detecting small and partially occluded objects, are beneficial in the context of disaster scenarios.

On the other hand, Faster R-CNN and RetinaNet, while offering competitive performance, particularly at the AP50 metric, fell short of the YOLO models. Dino and Cascade R-CNN showed substantial performance, with Cascade R-CNN achieving the second-highest mAP score, indicating its effectiveness in handling complex object relationships, likely due to its multi-stage detection process.

The analysis of AP50 scores, which are based on a lower IoU threshold, reveals that most models perform significantly better when the requirement for the overlap between predicted and ground truth bounding boxes is relaxed. This discrepancy suggests that while the models are capable of identifying the presence of objects, refining the accuracy of bounding box predictions remains a challenge and an area for potential improvement in future research iterations.

6 Discussion

6.1 Model Optimization for Complex Disaster Scenarios

To investigate the impact of domain-specific training on model performance in complex disaster scenarios, we conducted a comparative analysis using several datasets and a model [20]: C2A (synthetic disaster scenes), SARD (real-world search and rescue images), and "General Human Detection" (a combination of crowd human [26], tiny person [34], and VisDrone [7] datasets). By training models on these datasets and evaluating their performance across different validation sets, we aimed to identify the most effective approach for detecting humans in challenging disaster environments. Table 5 presents the results of this experiment, showcasing the mAP scores achieved by models trained on various datasets and validated on different test sets.

Table 5. Comparative Performance Across Different Training and Validation Datasets

Trained on\Validated on	General Human	SARD	C2A	General Human+C2A
General Human	0.77	0.347	0.168	0.159
SARD	0.036	0.931	0.168	0.071
C2A	0.168	0.259	0.784	0.462
General Human+C2A	0.855	0.66	0.874	0.862

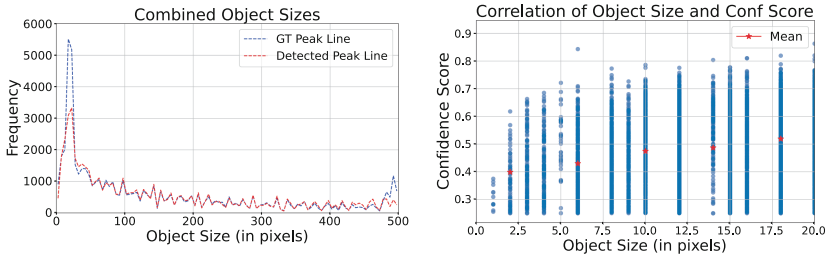
The model trained exclusively on the C2A dataset demonstrates a significant improvement in performance (0.784 mAP) when validated on the C2A test set compared to models trained on other datasets. This substantial increase in performance highlights the importance of domain-specific training using a dataset tailored to the task at hand, such as C2A, for developing models that can effectively detect humans in complex disaster scenarios.

One notable observation from the results is that although models trained on general human datasets perform poorly on search and rescue (SARD) and disaster (C2A) scenarios, the addition of the C2A dataset to the training process leads to a considerable improvement in performance. For instance, the model trained on the combined "General-Human + C2A" dataset achieves an mAP of 0.660 on the SARD validation set and 0.874 on the C2A validation set, surpassing the performance of models trained on either dataset alone. This finding suggests that incorporating disaster-specific data, such as the C2A dataset, can significantly enhance the model's ability to generalize to various challenging environments.

Furthermore, the results demonstrate that the combination of the C2A dataset and general human datasets yields better generalization performance across all validation sets. The model trained on the "General-Human + C2A" dataset achieves the highest mAP scores on the "General-Human" (0.855), SARD (0.660), and C2A (0.874) validation sets, indicating its robustness and versatility in handling diverse scenarios.

While the C2A dataset primarily consists of synthetic images, relying solely on synthetic data for training may undermine confidence in the model's real-world efficacy. Therefore, it is advisable to combine general human datasets with the C2A dataset to improve the model's ability to detect people in real-world disaster situations. The strong performance of the model trained on the "General-Human + C2A" dataset on the SARD validation set, which contains real-world search and rescue images, supports this recommendation.

To further validate the model's performance and increase confidence in its real-world applicability, future work should focus on evaluating the model on a more extensive set of real-world disaster images. This evaluation will help identify any potential gaps between the model's performance on synthetic and real-world data, guiding efforts to refine the dataset and training process.



(a) (b)

Fig. 3. Comparative Analysis of Object Detection (a) The frequency distribution of ground truth object sizes (blue) showcases a clear decline in detection rates for smaller objects (red), highlighting the challenges current detection algorithms face with objects less than 20 pixels in size. (b) The detection confidence scores across varying object sizes, with mean confidence indicated by red points, emphasize the higher reliability of detecting larger objects. These visualizations underscore the need for refining detection algorithms to better recognize small objects, which are critical for comprehensive disaster scene analysis. (Color figure online)

6.2 Object Size and Detection Confidence

In-depth analysis of detection performance reveals a notable size bias where smaller objects (less than 20 pixels) are detected with less frequency and lower confidence scores. This trend, observable in Fig. 3(a), points to a potential size-dependent limitation inherent in current detection algorithms. Conversely, larger objects demonstrate higher detection confidence, as seen in Fig. 3(b), where the mean confidence score, represented by red points, scales with object size. This size-detection relationship suggests an avenue for model improvement-specifically, enhancing the sensitivity of detection algorithms to smaller objects could significantly improve performance in complex disaster environments, where small-scale features can be critical.

6.3 Dataset Limitations and Prospects for Improvement

The C2A dataset, while effective, encounters limitations due to its synthetic nature. The overlay of human figures from the LSP/MPII-MPHB dataset onto disaster scenes can sometimes result in unrealistic scaling and positioning, potentially compromising the model's ability to generalize to real-world scenarios. Interestingly, this element of unrealism could also serve as a form of data augmentation, introducing variability that may help in training more robust and generalized models. Despite this, it is better to have context-aware adaptive scaling and improved spatial algorithms to enhance the realism of the training images. Moreover, transitioning to dynamic 3D models could more accurately depict human movement, overcoming the static nature of 2D images. Another limitation of the C2A dataset is that it consists of single images, whereas in most actual disaster scenarios, the input data could be in the form of video footage.

This discrepancy between the training data and real-world application data may impact the model's performance in practical settings. Future work should focus on expanding the dataset to include video sequences of disaster scenes, enabling the development of models that can effectively process and analyze real-time video feeds from UAVs during SAR operations. The dataset's variety in human poses and disaster scenarios is designed to aid in assessing disaster severity, enhancing its utility for SAR operations. Future enhancements should include real disaster footage to further validate and refine the dataset, optimizing AI model performance for real-world applications.

7 Conclusion

In the rapidly changing field of disaster response, our research introduces the C2A dataset as a crucial resource, connecting AI with humanitarian efforts. This work advances the technical capabilities of UAV-assisted search and rescue operations and represents a significant shift in how we integrate machine learning into crisis management.

The C2A dataset fills a critical need in disaster response, offering a comprehensive, synthetic environment that represents the complexities of real-world catastrophes. This dataset forms a foundation for training more robust and adaptable AI models. Our comparative analysis across various datasets highlights the importance of combining domain-specific data (like C2A) with general human datasets, resulting in models that are both specialized and widely applicable.

These advancements are initial steps in an ongoing process. As we expand the capabilities of AI-assisted disaster response, we must consider the ethical implications and real-world applicability of our work. The limitations we've identified, particularly in synthetic data generation and real-world validation, serve as guides for future research.

We envision a future where AI becomes an essential tool in crisis management, working alongside human expertise to save lives and reduce suffering. To achieve this vision, we encourage the research community to:

1. Test the C2A dataset in real-world pilot studies, linking synthetic training with practical application.
2. Collaborate to expand and refine the dataset, including diverse disaster scenarios and cultural contexts.
3. Pursue research that combines computer vision, disaster management, and ethics to ensure responsible and effective use of AI in humanitarian efforts.

As the field advances, our ultimate goal goes beyond technological progress; we aim to create tools and methods that are reliable in critical situations. By continually improving our approach, incorporating real-world feedback, and fostering collaboration across disciplines, we move towards a future where technology and human compassion work together, addressing the complexities of disaster response with increased precision and reliability.

In conclusion, the C2A dataset and our findings represent not only a technical achievement but also progress towards a more resilient and responsive global community. As we confront increasingly complex global challenges, the combination of AI and human ingenuity offers hope for more effective, efficient, and compassionate disaster response strategies in the years to come.

Acknowledgements. This work was supported by JSPS KAKENHI Grant No. JP22F22769 and JP22KF0141. Also, this work was performed the commissioned research fund provided by F-REI (JPFR23010102).

References

1. AlAli, Z.T., Alabady, S.A.: Techniques and methods for managing disasters and critical situations. Nat. Hazards 1–47 (2024)
2. Alsamhi, S.H., et al.: UAV computing-assisted search and rescue mission framework for disaster and harsh environment mitigation. Drones **6**(7), 154 (2022)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686–3693 (2014)
4. Aryal, R., Karki, M., Thapa, P.: Smoke scene detection from satellite imagery using deep learning. Retrieved from ResearchGate (2024)
5. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)
6. Chen, K., et al.: MMDetection: open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
7. Du, D., et al.: VisDrone-DET2019: the vision meets drone object detection in image challenge results. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
8. Hu, D., Chen, L., Du, J., Cai, J., Li, S.: Seeing through disaster rubble in 3D with ground-penetrating radar and interactive augmented reality for urban search and rescue. J. Comput. Civ. Eng. **36**(5), 04022021 (2022)
9. Jocher, G., et al.: ultralytics/YOLOv5: v7. 0-YOLOv5 SOTA realtime instance segmentation. Zenodo (2022)
10. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC, vol. 2, p. 5. Aberystwyth, UK (2010)
11. Kang, D.K., Olsen, M.J., Fischer, E.: Residential wildfire structural damage detection using deep learning to analyze uncrewed aerial system (UAS) imagery. In: Computing in Civil Engineering 2023, pp. 849–856. ASCE (2023)
12. Khial, N., Mhaisen, N., Mabrok, M., Mohamed, A.: An online learning framework for UAV search mission in adversarial environments. Available at SSRN 4725375 (2024)
13. Kucukayan, G., Karacan, H.: YOLO-IHD: improved real-time human detection system for indoor drones. Sensors **24**(3), 922 (2024)
14. Kyrkou, C., Theocharides, T.: Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In: CVPR Workshops, pp. 517–525 (2019)
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)

16. Liu, S., et al.: Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
17. Lyu, C., et al.: RTMDet: an empirical study of designing real-time object detectors. arXiv preprint arXiv:2212.07784 (2022)
18. Malandrino, F., Chiasserini, C.F., Casetti, C., Chiaraviglio, L., Senacheribbe, A.: Planning UAV activities for efficient user coverage in disaster areas. Ad Hoc Netw. **89**, 177–185 (2019)
19. Munawar, H.S., Ullah, F., Qayyum, S., Khan, S.I., Mojtahedi, M.: UAVs in disaster management: application of integrated aerial imagery and convolutional neural network for flood detection. Sustainability **13**(14), 7547 (2021)
20. Nihal, R.A., Yen, B., Itoyama, K., Nakadai, K.: From blurry to brilliant detection: YOLOv5-based aerial object detection with super resolution. arXiv preprint arXiv:2401.14661 (2024)
21. Núñez-Marcos, A., Arganda-Carreras, I.: Transformer-based fall detection in videos. Eng. Appl. Artif. Intell. **132**, 107937 (2024)
22. Padilla, R., Netto, S.L., Da Silva, E.A.: A survey on performance metrics for object-detection algorithms. In: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 237–242. IEEE (2020)
23. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recogn. **106**, 107404 (2020)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
25. Sambolek, S., Ivasic-Kos, M.: Automatic person detection in search and rescue operations using deep CNN detectors. IEEE Access **9**, 37905–37922 (2021)
26. Shao, S., et al.: Crowdhuman: a benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)
27. Tang, P., Li, J., Sun, H.: A review of electric UAV visual detection and navigation technologies for emergency rescue missions. Sustainability **16**(5), 2105 (2024)
28. Wang, C.Y., Yeh, I.H., Liao, H.Y.M.: YOLOv9: learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616 (2024)
29. Wang, G., et al.: M4SFWD: a multi-faceted synthetic dataset for remote sensing forest wildfires detection. Expert Syst. Appl. 123489 (2024)
30. Wang, J., Zhong, Y., Ma, A., Zheng, Z., Wan, Y., Zhang, L.: LoveNAS: towards multi-scene land-cover mapping via hierarchical searching adaptive network. ISPRS J. Photogramm. Remote. Sens. **209**, 265–278 (2024)
31. Wu, S., Zhang, X., Liu, R., Li, B.: A dataset for fire and smoke object detection. Multimed. Tools Appl. **82**(5), 6707–6726 (2023)
32. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2 (2019). https://github.com/facebookresearch/detectron2
33. Xia, G.S., et al.: DOTA: a large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3974–3983 (2018)
34. Yu, X., Gong, Y., Jiang, N., Ye, Q., Han, Z.: Scale match for tiny person detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1257–1265 (2020)

DREAMS: Diverse Reactions of Engagement and Attention Mind States Dataset

Monisha Singh[1(✉)], Gulshan Sharma[2,1], Ximi Hoque[2], and Abhinav Dhall[3,1]

[1] Indian Institute of Technology Ropar, Rupnagar, India
monisha.21csz0023@iitrpr.ac.in
[2] Kroop AI, Gandhinagar, India
[3] Monash University, Melbourne, Australia

Abstract. Active attention and engagement are important in improving users' learning experiences. Engagement refers to the level of involvement and interest individuals show towards a particular task. Attention, on the other hand, refers to a state where someone is entirely focused on a particular task with conscious awareness. Engagement and attention are different but closely linked concepts and can influence each other bidirectionally [23]. To explore the relationship between user engagement and attention, we introduce the *Diverse Reactions of Engagement and Attention Mind States (DREAMS)* dataset. The dataset includes facial video recordings of 32 users in naturalistic settings watching various stimuli to evoke diverse emotions. We then analyze user engagement and attention states in these videos by framing it as a classification problem, exploring single-task, transfer learning task, and multi-task settings. In single and transfer learning task settings, separate networks are applied to predict engagement and attention states. Whereas in multi-task settings a shared network is applied, which jointly learns to predict both engagement and attention states. Moreover, we examine participants' performance on video-based questionnaires and evaluate their perceived cognitive workload. In our findings, we observe (a) better classification performance in predicting engagement states in both transfer and multi-task learning compared to single-task learning and (b) higher engagement and attention states correlate with lower cognitive load and improved task performance. The dataset and the code are publicly available and can be accessed through https://sites.google.com/view/dreams-dataset/dataset.

Keywords: Engagement · Attention · Single-Task Learning · Transfer Learning · Multi-Task Learning · Transformers

1 Introduction

Human-Computer Interaction (HCI) has witnessed remarkable advancements in recent years, transforming how users engage with interactive systems. Within

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 163–179, 2025.
https://doi.org/10.1007/978-3-031-78341-8_11

this context, *Engagement* and *Attention* are two intertwined concepts which have emerged as key elements in the design and evaluation of interactive systems. *Engagement* mainly includes three aspects: behavioral, affective, and cognitive [24]. Behavioral engagement refers to observable actions, behaviors, and interactions displayed by individuals while performing a task. Affective engagement refers to the emotions and feelings that users experience during their interaction. Cognitive engagement involves mental investment exerted by users while engaging with a particular task. On the other hand, *Attention* refers to a state of focused cognitive engagement and conscious awareness directed toward a specific task.

Understanding the interplay between engagement and attention is useful for the design of interactive systems, as it significantly influences user experience. Positive engagement in interactive systems refers to users' active participation, while positive attention reflects the higher degree of focus users dedicate to the system. Conversely, negative engagement, often driven by confusing interfaces or uninspiring content, may lead to user disinterest and a shift of attention away from the intended interaction. The preceding statements suggest a positive correlation between engagement and attention, implying that higher engagement levels result in heightened attention. However, it is noteworthy that individuals who behaviorally appear engaged may not necessarily be attentive as it is plausible for one's mind to wander despite outward signs of engagement. This underscores the importance of refraining from accepting the conventional notion of a correlation between engagement and attention, but rather considering an orthogonal relationship between them. Consequently, understanding the intricate dynamics of this relationship becomes crucial for enhancing various facets of user interactions.

We aim to understand how engagement and attention are related to help content creators develop more immersive content. The main contributions of the paper are summarized below:

- We introduce *DREAMS* (Diverse Reactions of Engagement and Attention Mind States), a self-annotated engagement and attention state dataset, collected in an in-the-wild setting.
- We designed single, transfer learning, and multi task experiments to evaluate engagement and attention, assuming that the performance of related tasks would exhibit improvement in transfer learning and multi task setup compared to single task setup.
- We applied NASA Task Load Index(NASA-TLX) workload [11] assessment questions to identify the variations in cognitive load levels experienced by individuals while watching the diverse set of stimuli. Additionally, we explored the causal relationship between engagement, attention, and cognitive load levels.
- We study the impact of engagement and attention on task performance by analyzing the percentage of correct responses for various engagement and attention levels.

2 Background and Related Work

In this section, we present an overview of contributions and recent advancements in the fields of engagement and attention research.

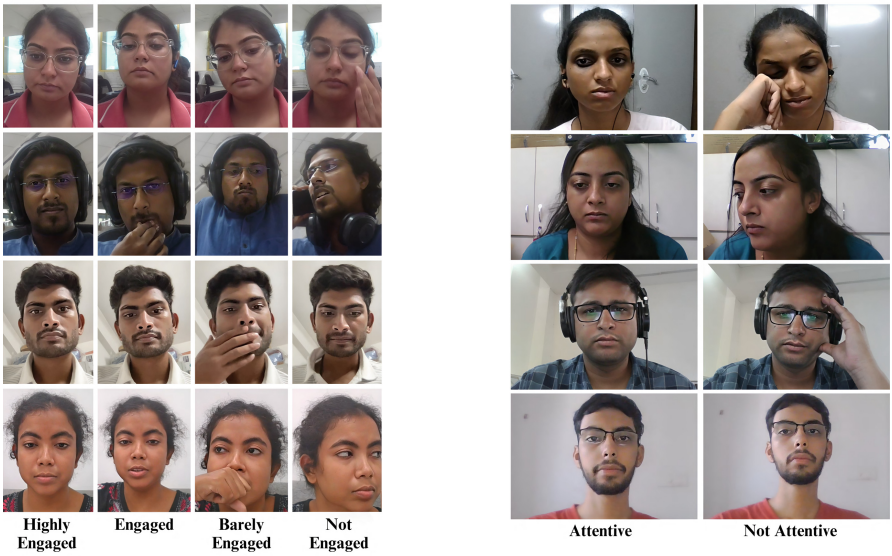


Fig. 1. Engagement(left) and Attention(right) states from the *DREAMS* dataset. On the left, notice that *Highly Engaged* participants are glued to the screen. Subtle changes in expressions and gestures can be observed as engagement decreases gradually. On the right, notice that outward signs of attention are displayed by the first-three-row participants. However, the last-row participant appears to be visually attentive in both states, but his mind may have been wandering in the second instance.

2.1 User Engagement

Engagement encompasses users' observable actions (Behavioral), emotional connection (Affective), and cognitive investment [28]. Behavioral engagement can be assessed through qualitative observations of user behavior, such as gaze, head pose, facial expressions, etc. Various methods, including surveys, physiological measures, behavioral analysis, and neuroimaging techniques, can measure affective engagement. Multiple approaches, such as evaluating task performance, self-reporting, analyzing response times, monitoring physiological indicators, and employing neuroimaging techniques, can be applied to gauge cognitive engagement. User engagement plays a crucial role in achieving various goals, including educational success, productivity, customer satisfaction, and business growth. User engagement has been extensively studied in the context of HCI [10,14,17].

One of the earliest efforts in this direction is by D'Mello et al. [8]. They investigated the relationship between facial expressions, linguistic cues, and engagement detection and explained the importance of non-verbal cues in comprehending user engagement. Facial Action Coding System was used by authors in [9,21] to assess distinct emotions associated with various facial muscle movements to emphasize the relationship between particular engagement labels and facial action units. In [21], Whitehill et al. showed that automated engagement detectors work as accurately as people using non-verbal features. To assess engagement, Booth et al. [2] evaluated the performance of an LSTM-based method. In the EmotiW 2018 [6] sub-challenge, *Engagement Prediction in the Wild*, authors proposed using TCN [19] to predict user engagement in an in-the-wild setting. TCN network was applied here to show improvements over LSTM in improving baseline models. In the recent works, authors tried to understand patterns that relate to User Engagement by demonstrating and focusing on interpretability and having a simpler SVM model to train for predicting engagement [18].

2.2 User Attention

Attention can be described as the mental state in which an individual directs their entire focus and conscious awareness toward a particular task or stimulus. Attention is vital for selectively processing information, facilitating cognitive functions, and enabling efficient task performance. Attention can be measured through various methods such as reaction time tasks, eye-tracking technology, neuroimaging techniques like Electroencephalography or Functional Magnetic Resonance Imaging, and behavioral observation. Attention, a crucial process, is one of the main focus of HCI research. Eye-tracking experiments have greatly added to the study of visual attention patterns. A significant contribution was made by Buscher et al. [3] in which they investigated the dynamics of visual attention while browsing the web. Their research provided an understanding of the hierarchy of visual saliency and user preferences by demonstrating how consumers distribute their attention across various web page elements. The work by D'Mello et al. [7] stands out in the field of attention-aware, intelligent tutoring systems. In their study, they introduce a pioneering approach that harnesses gaze data to enhance the effectiveness of tutoring interactions. The Gaze Tutor system employs eye-tracking technology to track students' gaze patterns and alter instructional content based on their visual attention. Their work demonstrates the potential of attention-aware technology in building more responsive and personalized educational experiences by smoothly incorporating gaze tracking into the tutoring process.

In their study, Hutt et al. [13] offer a compelling investigation into the use of gaze-based attention-aware technologies to combat students' tendency to daydream in classroom environments. The authors suggest a novel strategy that uses gaze monitoring to identify instances of mind wandering and then implements targeted interventions to draw students' attention back. This work provides important insights into the practical implications and efficacy of such interventions inside classroom situations by expanding the deployment of gaze-based

attention-aware technologies outside controlled laboratory settings. Recent work in attention(mind wandering) detection by Lee et al. [15] involved the use of facial action units extracted from videos of students participating in online learning. The authors emphasized that their proposed feature-based model outperformed random and gaze-only baselines and suggested using landmark-based features over gaze features.

Researchers from various fields have been intrigued by the complex interplay between engagement and attention, which has provided light on the dynamic nature of human cognition and interaction. In his study, Heath [12] investigates how attention and engagement mutually influence each other, contributing to the overall impact of advertisements on viewers. His research reveals a strange pattern. A higher level of emotional content in the advertising correlates with lower levels of attentiveness. This discovery casts doubt that engagement and attention have a simple causal relationship. Instead, the experiment demonstrates that engagement and attention can function separately and are not necessarily associated. The study by Leiker et al. [16] used neurophysiological measures of attention to investigate the relationship between engagement and attention in motion-controlled video games. This study found that engagement elicits increased information processing, which reduces attentional reserve. These initiatives underscores the importance of considering both engagement and attention in designing effective educational interventions and interactive systems.

3 DREAMS Dataset

3.1 Ethics

We obtain user consent for the collection of their data and meta-information, ensuring transparency and respect for individual autonomy. Also, all stimuli used in the study are sourced either from publicly available datasets or videos having Creative Commons licenses. By doing this, we protect user's rights and promote transparency in our research processes.

3.2 Stimuli

We use three stimuli from the EngageNet [17] dataset, (i) *Schrödinger's cat: A thought experiment in quantum mechanics* (ii) *What is cryptocurrency?* (iii) *Where did English come from?*. The first and second stimuli feature a digital teacher avatar (refer Fig. 2). These stimuli provide instruction on quantum mechanics and cryptocurrency, respectively. While the third stimulus offers insights into the evolution of the English language. The avatar based videos are generated using the Artiste platform from Kroop AI. Additionally, we include two humorous videos with Creative Commons licenses to diversify the stimuli set. The educational stimuli typically have an average duration of 5 min, whereas the

humorous stimuli are shorter, lasting approximately for 2 min. The background sources of the stimuli can be found here[1].

3.3 Data Collection Protocol

We develop a web-based interface to collect data on user's engagement and attention states. The data is collected in an in-the-wild settings, where participants have the flexibility to record data at their preferred time and location. Participants can use a computer or laptop with a reliable internet connection and a good webcam. The experiment starts with collecting participants' demographic data, including biological sex and age. Afterwards, a pre-study questionnaire is given to gauge user personality traits.

Following the completion of the questionnaire, the video stimuli are presented in a randomized order. As the stimuli play, automatic prompts are integrated into the web interface, asking participants to self-assess their attention and engagement states based on certain guidelines provided to them.

For attention state, the participants could choose from the following options: *Focused/Attentive* (thinking about the stimulus), *Not Focused/Not Attentive* (mind-wandered, doing or thinking something unrelated to the lecture), or *Skip* (participant is indecisive about his/her state).

For engagement, participants could choose from the following engagement levels: *Highly Engaged* (participant is attentive and glued to the screen), *Engaged* (participant is interested in the content, and appeared to like it), *Barely Engaged* (participant is minimally attentive, fidgeted restlessly in the chair or hardly opened his or her eyes), and *Not Engaged* (participant are disengaged, frequently glanced away from the screen and are disinterested).

After completion of a stimuli, participants are required to fill out a brief questionnaire, which includes stimulus content based questions, engagement-related questions and cognitive load assessments using NASA-TLX. The typical duration of a user session varies from 25 to 30 min, depending on the questionnaires' response time of participants.

3.4 Participants

The study involved 32 college students, comprising 13 females, with age ranging from 21 to 38 years. All participants are proficient in English with educational backgrounds in science and engineering.

3.5 Data Collection and Annotations

During the experiment, we collect the following information from each participant: (i) responses to the personality questionnaire, (ii) video recordings of participants as they watch the stimuli, (iii) self-reported engagement and attention

[1] Video Source: Quantum Mechanics, Cryptocurrency, English Language, Humorous Stimulus 1, Humorous Stimulus 2.

states provided by the participants, (iv) responses to questionnaires regarding the content of the viewed stimuli, and (v) responses to questions assessing behavioral, cognitive, and affective engagement, followed by NASA-TLX responses.

Automatic prompts are generated every 40 s, asking participants to self-identify their attention and engagement states. We chose a 40-second interval based on the typical time it takes for attention to shift [22] and the findings of an in-lab pilot study conducted by Lee et al. [15]. We collected 832 videos of 40 s duration, out of which 781 videos were successfully processed. The resolution of the recorded videos were at least 640×480 pixels.

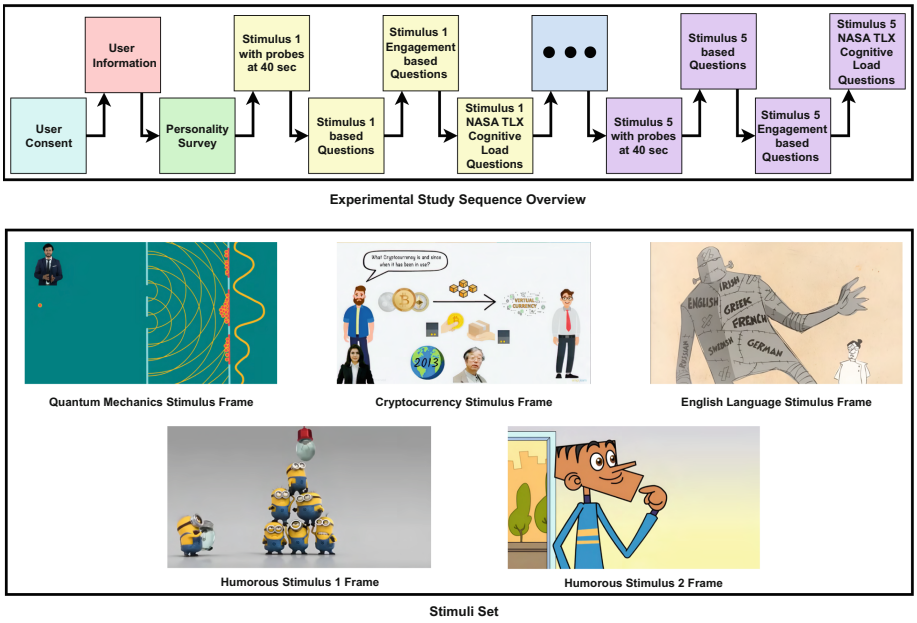


Fig. 2. (Top) Sequence Flow of Experimental Study, and (Bottom) Frames from Stimuli Set used in the study. Notice the digital avatars in upper left corner of the Quantum Mechanics Stimulus Frame, and in lower left corner of the Cryptocurrency Stimulus Frame. For details, refer to Sect. 3.

As a pre-processing step, we combine the *Not Focused/Not Attentive* and *Skip* labels of attention into a single category called *Not Attentive*. The rationale behind this merging was the participants' lack of attention to the task, which prevented them from making a decision.

Table 1 illustrates the distribution of self-labeled instances of Engagement and Attention. To quantify the relationship between these self-labels, we applied *Cramér's V* correlation coefficient. The coefficient for the self-labeled data is 0.7289, signifying a strong association between engagement and attention.

3.6 Data Split

We split the data into subject independent sets, with 25 subjects in *Train* and 7 subjects in *Test* set. There are 600 videos in the *Train* set whereas 181 videos in the *Test* set, amounting to a total of approximately 9 h of data.

4 Identifying Relationship Between Attention and Engagement

In psychology studies, attention and engagement are closely connected concepts. Multiple studies [25–27] have identified attention as an important factor in determining behavioral engagement. However, authors in [16] observe that higher engagement is associated with low *eP3a* levels. *eP3a* is a brain wave component typically observed in response to stimuli, where higher levels are usually associated with increased attention. These studies suggests that the relationship between engagement and attention is complex and sometimes contradictory. Hence it is important to understand this relationship prior to devising effective learning strategies.

Table 1. Distribution of self labelled instances under various Engagement and Attention states.

	Attentive	Not-Attentive
Highly Engaged	260	2
Engaged	220	59
Barely Engaged	32	133
Not Engaged	8	67

To understand this relationship, we apply supervised learning based approach, where we explore single-task learning, transfer learning, and multi-task learning.

- In single-task learning, we apply separate neural networks to independently learn about attention and engagement states. The idea behind this approach is to benchmark the performance of each network in identifying attention and engagement states.
- In transfer learning, we first pretrain the network on engagement or attention task, and later fine-tune on the related counterpart task. The idea behind this approach is to transfer the learned feature representations from one task to another to improve performance.
- In multi-task learning, we train a unified network which jointly learns to identify attention and engagement states. The idea behind this approach is to learn shared representations, allowing the network to capture underlying correlations and dependencies between attention and engagement.

4.1 Feature Extraction

We extract facial features such as eye-gaze, facial landmarks, head pose, facial action units, and Point Distribution Model (PDM) by applying OpenFace [1] framework on user videos. Additionally, we extract spatiotemporal representation of the face by applying MARLIN [4] framework.

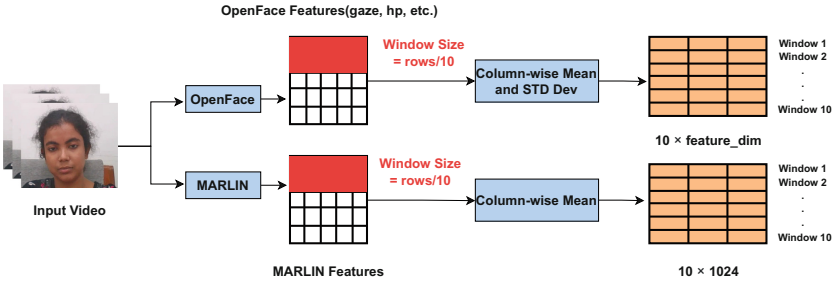


Fig. 3. Feature extraction and statistical feature aggregation. For details, refer to Sect. 4.1.

After extracting facial features and spatiotemporal representations using OpenFace and MARLIN, respectively, we partition the extracted row-wise data into 10 evenly distributed segments. Then, we compute the mean and standard deviation for each segment and combine them into a single vector. This vector statistically captures the changes in facial features over time in the video. The overall process of feature extraction is depicted in Fig. 3.

4.2 Experiments

To understand the relationship between engagement and attention, we apply classification methods based on Transformer [20] architecture. Transformers can learn temporal and spatial relationships in the data. The attention mechanism within the Transformer allows for identifying important relationships in the data and effectively managing long-range relationships.

This section describes the supervised methods employed to uncover the relationship between engagement and attention. Figure 4 illustrates these classification methods.

Single Task and Transfer Learning Task. Single-task learning is a methodology that optimizes a single objective function using a dedicated feature representation. Transfer learning, on the other hand, leverages knowledge gained from solving one problem and applies it to a different but related problem. We conducted single-task and transfer learning experiments using a Transformer architecture incorporating positional encoding, a transformer encoder, and a

Multi-Layer Perceptron (MLP) for engagement and attention prediction. Positional encoding enhances the input data while maintaining its sequential context. The order of each component inside the sequence is considered, which is crucial for tasks involving temporal or spatial interactions.

Multi-task Learning. Multi-task learning is a way to learn multiple objective functions using a common feature representation. We achieve this by using the same time-series transformer-based model to train on the OpenFace and MARLIN features. In the multi-task network, we have a common positional encoding and a transformer encoder layer. The shared transformer encoder layer serves as a foundational feature extractor. This layer is excellent at capturing intricate dependencies and interactions among input sequences. Two different MLP, one for attention (mind wandering) and the other for engagement, handle the encoded features. The task-specific MLP heads are designed to extract task-specific features from the shared transformer encoder's output. This specialization enables the model to excel at each task's unique challenges. Over the positional encodings, we add task-specific attention layers and establish a residual connection with MLP heads. These task-specific attention layers grant our model the ability to dynamically distribute attention across the input sequence in accordance with the unique demands of each task. The network thus benefits from a common transformer encoder layer and task-specific attention from the

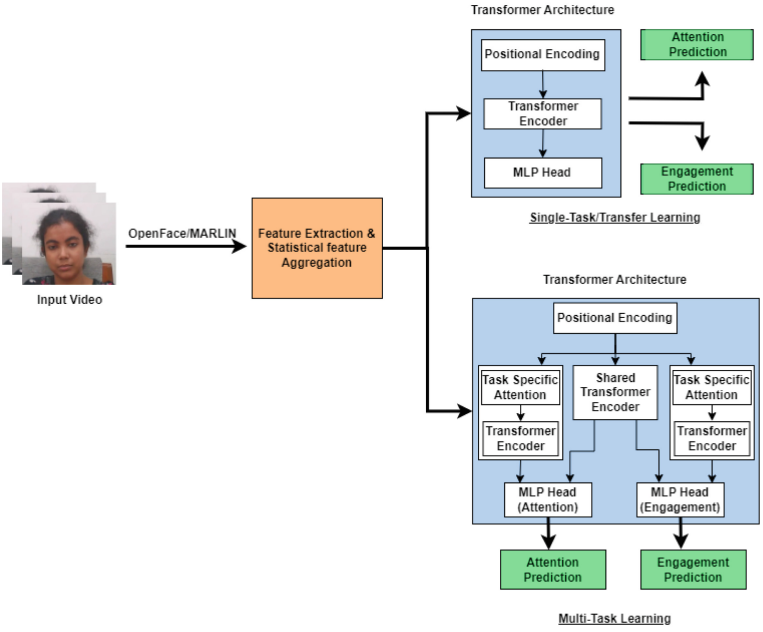


Fig. 4. Single-Task/Transfer Learning and Multi-Task Learning. For details, refer to Sect. 4.2.

positional embeddings. We train the model where we try to minimize engagement and attention losses jointly, using the same encoder. The network splits into two task-specific MLP where they reduce their individual losses.

We use cross entropy loss with assigned class weights for engagement and attention prediction. We choose the best model by maximizing the sum of the weighted F1 score, weighted Precision, and weighted Recall for engagement and attention prediction.

4.3 Results

The experiments were performed in single-task, transfer learning, and multi-task setups using supervised approach. We considered gaze as our base feature, a crucial indicator of attention and behavioral engagement, and incrementally added other OpenFace features to assess their contribution in predicting engagement and attention levels. For performance comparisons we took the summation of weighted precision, weighted F1, and weighted recall scores. Below is a detailed analysis of the results of the various experiments performed as part of this study.

Table 2. Results of Engagement prediction in Single-Task, Transfer Learning and Multi-Task setups. Here HP, AU, LMK, and PDM refers to Head Pose, Action Units, Landmarks, and Point Distribution Model respectively. P, F1, and R refer to Weighted Precision, F1, and Recall scores respectively.

	Engagement								
	Single-Task			Transfer Learning			Multi-Task		
Features	P	F1	R	P	F1	R	P	F1	R
GAZE	**0.125**	**0.185**	**0.354**	**0.125**	**0.185**	**0.354**	**0.125**	**0.185**	**0.354**
GAZE+HP	**0.178**	0.181	**0.332**	0.110	0.165	0.331	0.163	**0.218**	**0.332**
GAZE+HP+AU	0.219	0.256	0.320	0.167	0.216	**0.354**	**0.295**	**0.306**	0.332
GAZE+HP+AU+LMK	0.496	0.226	0.320	0.400	**0.293**	**0.392**	**0.596**	0.208	0.365
GAZE+HP+AU+LMK+PDM	0.260	**0.305**	0.376	**0.606**	0.284	**0.398**	0.311	0.299	0.343
MARLIN	0.267	0.256	0.298	0.310	0.275	0.326	**0.382**	**0.276**	**0.354**

Single-Task vs Transfer Learning. From the engagement results, (refer to Table 2) we can observe that for the feature combination of Gaze + HP + AU + LMK, Gaze + HP + AU + LMK + PDM, and MARLIN, there is an improvement in the performance of transfer learning(attention pre-training and engagement fine-tuning) over the single-task learning. However, in the attention results (refer to Table 3) no improvement is observed in the performance of transfer learning(engagement pre-training and attention fine-tuning) over the single-task learning for any feature combination. Based on these observations, one can infer that attention may serve as a clearer indication of engagement, but engagement may not be a reliable indicator of attention.

Table 3. Results of Attention prediction in Single-Task, Transfer Learning and Multi-Task setups. Here HP, AU, LMK, and PDM refers to Head Pose, Action Units, Landmarks, and Point Distribution Model respectively. P, F1, and R refer to Weighted Precision, F1, and Recall scores respectively.

	Attention								
	Single-Task			Transfer Learning			Multi-Task		
Features	P	F1	R	P	F1	R	P	F1	R
GAZE	**0.763**	**0.465**	**0.608**	0.158	0.226	0.398	0.158	0.226	0.398
GAZE+HP	**0.775**	0.534	**0.641**	0.564	**0.566**	0.580	0.158	0.226	0.398
GAZE+HP+AU	**0.703**	0.574	**0.652**	0.690	**0.579**	**0.652**	0.158	0.226	0.398
GAZE+HP+AU+LMK	**0.675**	0.615	**0.663**	0.646	**0.617**	0.652	0.363	0.453	0.602
GAZE+HP+AU+LMK+PDM	0.668	**0.637**	**0.669**	0.641	0.606	0.646	**0.763**	0.465	0.608
MARLIN	**0.643**	**0.623**	**0.619**	0.630	0.622	**0.619**	0.621	0.546	0.553

Single-Task vs Multi-Task. From the engagement results Table 2, we can observe that for all feature combinations except Gaze, there is an improvement in performance of multi-task over the single-task learning. However, in attention results Table 3 no improvement was observed in performance of multi-task over single-task learning for any feature combination. The results of this experiment also suggest that the presence or absence of attention directly affects engagement, but the reverse relationship may not be as strong or direct. In other words, the model might find it easier to learn features or patterns related to engagement when attention information is available, as attention could be a contributing factor to engagement. However, the absence of engagement information may not have a substantial impact on the model's ability to predict attention.

4.4 Engagement and Attention NASA-TLX Workload Analysis

We applied NASA-TLX assessment questions to assess the perceived workload experienced by the participants while watching the diverse set of stimuli. The NASA-TLX, which originally had six arbitrary subscales, rates the following aspects: (i) Mental Demand, (ii) Physical Demand, (iii) Temporal Demand, (iv) Performance, (v) Effort, and (vi) Frustration Level.

In order to calculate the average task load for this study, we concentrated on three distinct subscales: Mental Demand, Effort, and Frustration. Using a four-point scale, our assessment classified the task load as *very low*, *low*, *high*, or *very high*. This method gave insightful information about the perceived workload related to the tasks under evaluation.

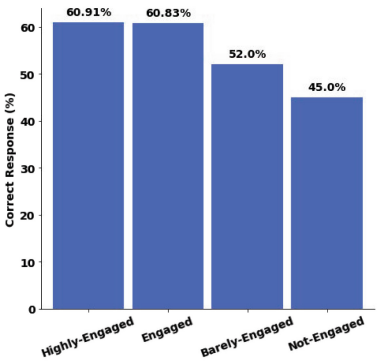
The assessment of workload across the different educational video categories yielded distinct findings. Specifically:

– The average workload associated with the Quantum Mechanics educational video fell within the range between *low* and *high*. This indicates that viewers perceived a moderate level of mental demand and effort while engaging with this educational content.

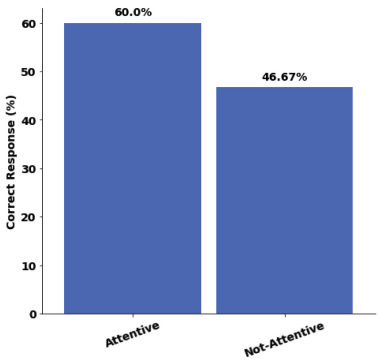
– In contrast, both the Bitcoin and English language educational videos exhibited an average workload categorized as *low*. This suggests that these videos required relatively less mental demand and effort compared to the Quantum Mechanics video.
– Notably, the category of humorous videos was found to have an exceptionally *low* workload. This implies that the act of viewing humorous content involved minimal mental demand, required little effort, and was associated with a *low* level of frustration.

Our observations further revealed intriguing patterns in workload perception among participants with varying levels of engagement and attention. Specifically:

– Participants classified as *Highly Engaged* and *Engaged* consistently reported lower workload scores. This phenomenon can be attributed to their elevated levels of engagement, which likely facilitated memory retention, and content recall. Consequently, these individuals encountered reduced mental demand, spent less effort, and experienced diminished frustration when responding to questions related to the viewed content.
– Similarly, participants categorized as *Attentive* exhibited workload scores that aligned with the trend observed among *Highly Engaged* and *Engaged* subjects. Their attentive demeanor likely contributed to a smoother cognitive process, translating to a lower perceived workload.
– In contrast, participants characterized as *Not Attentive* consistently reported higher workload scores. This outcome can be attributed to their need to exert additional effort to recall and process the presented concepts. This heightened mental demand, coupled with a higher level of frustration, reflects the challenges faced by individuals in this category when responding to questions associated with the viewed stimuli.



(a) % of correct responses of engagement classes.



(b) % of correct responses of attention classes.

Fig. 5. In (a) *Highly Engaged* and *Engaged* classes performed comparatively better than *Barely Engaged* and *Not Engaged* classes, whereas in (b) *Attentive* class performed comparatively better than *Not Attentive* class. For details, refer to Sect. 4.5.

These findings highlight the crucial role of attentiveness and active engagement in shaping users' task experiences and perceived cognitive demands.

4.5 Engagement and Attention Percentage of Correct Responses

We assessed participants' self-labels for engagement and attention and analyzed their performance in video-based questionnaires. Each questionnaire contained two moderately difficult questions related to the stimulus content. By calculating the mode of self-labels, we assigned overall engagement and attention labels to participants. We then computed the percentage of correct responses for each label category. Figures 5a and 5b illustrate the relationship between the percentage of correct responses and engagement and attention labels. We found that the *Highly Engaged*, *Engaged*, and *Attentive* categories gave the highest percentage of correct responses. As engagement and attention levels decrease, the percentage of correct responses also declines. This suggests a direct relationship between task performance and the level of engagement or attention.

5 Discussion

In this study, we investigate the relationship between engagement and attention using the *DREAMS* dataset. Our experimental results indicate that attention may be a more definitive indicator of engagement than the reverse. The underlying cause of this observation can be attributed to the fact that for Engagement, self-labels were assigned based on participants' visible reactions rather than their cognitive involvement. Conversely, attention labels were assigned based on the participants' immediate thoughts, whether related to the stimuli or not. Thus, while an individual may display visible signs of engagement their mind might be elsewhere, as illustrated by the fourth user in the attention sample frames in Fig. 1. On the other hand, attentive individuals tend to unintentionally exhibit positive signs of engagement. When these visible reactions data is fed into our model, it more readily learns that attention can lead to engagement, whereas the reverse relationship is more complex. This complexity arises because a visually engaged individual may or may not be attentive, complicating the model's ability to predict attention based on visual data.

Understanding the interplay between engagement and attention is important for optimizing user interaction with multimedia content. Our findings suggests that attention is an important factor to be considered in educational settings. The educational content should be designed to maintain the user's attention, which can improve learning outcomes. Whereas for recreational content, allowing users the freedom to engage without strict attention is beneficial, as it helps them enjoy the content and relax mentally.

6 Conclusion and Future Work

We present the *DREAMS* dataset, self-labelled with *Engagement* and *Attention* mind states. We perform experiments in single-task, transfer learning, and multi-task setups. The results demonstrate the richness and complexity of the data and

suggests that attention could be considered a clearer or more informative signal of engagement compared to the reverse relationship. Additionally, We analyzed the cognitive workload experienced by the participants of the experimental study and found that the workload was directly correlated to the complexity of the content of the video stimulus. We also observed that Higher Engagement and Attention states resulted in lower workload and vice versa. Furthermore, we analyzed the performance of the subjects belonging to various engagement and attention states and found that the percentage of correct responses was directly related to higher engagement and attentiveness. A limitation of our study is the narrow range of stimuli, primarily focusing on educational and humorous videos. This may not fully represent the broader spectrum of real-world scenarios. In the dataset extension, we will introduce a diverse set of stimuli, including topics outside of computing, and invite participants from non-technical backgrounds. In this work we explored attention and engagement in supervised manner. It will be intriguing to uncover this relationship in self-supervised settings.

References

1. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.-P.: Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66. IEEE Press (2018)
2. Booth, B.M., Ali, A.M., Narayanan, S.S., Bennett, I., Farag, A.A.: Toward active and unobtrusive engagement assessment of distance learners. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 470–476. IEEE (2017)
3. Buscher, G., Cutrell, E., Morris, M. R.: What do you see when you're surfing? using eye tracking to predict salient regions of web pages. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2009, pp. 21–30, New York, NY, USA, 2009. Association for Computing Machinery (2009)
4. Cai, Z., et al.: Marlin: masked autoencoder for facial video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1493–1504 (2023)
5. Copur, O., Nakıp, M., Scardapane, S., Slowack, J.: Engagement detection with multi-task training in e-learning environments. In: Sclaroff, S., Distante, C., Leo, M., Farinella, G.M., Tombari, F. (eds.) ICIAP 2022 Part III. LNCS, vol. 13233, pp. 411–422. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-06433-3_35
6. Dhall, A., Kaur, A., Goecke, R., Gedeon, T.: EmotiW 2018: audio-video, student engagement and group-level affect prediction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI 2018, pp. 653–656. Association for Computing Machinery, New York, NY, USA (2018)
7. D'Mello, S., Olney, A., Williams, C., Hays, P.: Gaze tutor: a gaze-reactive intelligent tutoring system. Int. J. Hum.-Comput. Stud. **70**(5), 377–398 (2012)

8. D'Mello, S.K., Craig, S.D., Graesser, A.C.: Multimethod assessment of affective experience and expression during deep learning. Int. J. Learn. Technol. **4**(3/4), 165–187 (2009)
9. Grafsgaard, J., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.: Automatically recognizing facial expression: predicting engagement and frustration. In: Educational Data Mining 2013 (2013)
10. Gupta, A., D'Cunha, A., Awasthi, K., Balasubramanian, V.: DAiSEE: towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885 (2016)
11. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. **52**, 139–183 (1988)
12. Heath, R.: How do we predict advertising attention and engagement. School of Management University of Bath Working Paper, vol. 9 (2007)
13. Hutt, S., Krasich, K., Brockmole, J.R., D'Mello, S.K.: Breaking out of the lab: mitigating mind wandering with gaze-based attention-aware technology in classrooms. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI 2021. Association for Computing Machinery New York, NY, USA (2021)
14. Kaur, A., Mustafa, A., Mehta, L., Dhall, A.: Prediction and localization of student engagement in the wild. In: 2018 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8. IEEE (2018)
15. Lee, T., Kim, D., Park, S., Kim, D., Lee, S.-J.: Predicting mind-wandering with facial videos in online lectures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2104–2113 (2022)
16. Leiker, A.M., Miller, M., Brewer, L., Nelson, M., Siow, M., Lohse, K.: The relationship between engagement and neurophysiological measures of attention in motion-controlled video games: a randomized controlled trial. JMIR Serious Games **4**(1), e5460 (2016)
17. Singh, M., Hoque, X., Zeng, D., Wang, Y., Ikeda, K., Dhall, A.: Do i have your attention: a large scale engagement prediction dataset and baselines. In: Proceedings of the 25th International Conference on Multimodal Interaction, ICMI 2023, pp. 174–182. Association for Computing Machinery, New York, NY, USA (2023)
18. Stappen, L., Baird, A., Lienhart, M., Bätz, A., Schuller, B.: An estimation of online video user engagement from features of time-and value-continuous, dimensional emotions. Front. Comput. Sci. **4**, 37 (2022)
19. Thomas, C., Nair, N., Jayagopi, D.B.: Predicting engagement intensity in the wild using temporal convolutional network. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 604–610 (2018)
20. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 6000–6010. Curran Associates Inc, Red Hook, NY, USA (2017)
21. Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., Movellan, J.R.: The faces of engagement: automatic recognition of student engagementfrom facial expressions. IEEE Trans. Affect. Comput. **5**(1), 86–98 (2014)
22. Klinger, E.: Modes of normal conscious flow.. the stream of consciousness (1978)
23. Matthews, G., Warm, J.S., Reinerman, L.E., Langheim, L.K., Saxby, D.J.: Task engagement, attention, and executive control. In: Gruszka, A., Matthews, G., Szymura, B. (eds.) Handbook of Individual Differences in Cognition. The Springer Series on Human Exceptionality, pp. 205–230. Springer, New York (2010). https://doi.org/10.1007/978-1-4419-1210-7_13
24. O'Brien, H.L., Toms, E.G.: The development and evaluation of a survey to measure user engagement. J. Am. Soc. Inform. Sci. Technol. **61**(1), 50–69 (2010)

25. Reeve, J., Tseng, C.-M.: Agency as a fourth aspect of students' engagement during learning activities. Contemp. Educ. Psychol. **36**(4), 257–267 (2011)
26. Fredricks, J.A., McColskey, W.: The measurement of student engagement: a comparative analysis of various methods and student self-report instruments. In: Christenson, S., Reschly, A., Wylie, C. (eds.) Handbook of Research on Student Engagement, pp. 763–782. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-2018-7_37
27. Skinner, E.A., Kindermann, T.A., Furrer, C.J.: A motivational perspective on engagement and disaffection: conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. Educ. Psychol. Measur. **69**(3), 493–525 (2009)
28. Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: potential of the concept, state of the evidence. Rev. Educ. Res. **74**(1), 59–109 (2004)

PolyGlotFake: A Novel Multilingual and Multimodal DeepFake Dataset

Yang Hou(✉), Haitao Fu, Chunkai Chen, Zida Li, Haoyu Zhang, and Jianjun Zhao

Kyushu University, Fukuoka, Japan
hou.yang.549@s.kyushu-u.ac.jp

Abstract. With the rapid advancement of generative AI, multimodal deepfakes, which manipulate both audio and visual modalities, have drawn increasing public concern. Currently, deepfake detection has emerged as a crucial strategy in countering these growing threats. However, as a key factor in training and validating deepfake detectors, most existing deepfake datasets primarily focus on the visual modal, and the few that are multimodal employ outdated techniques, and their audio content is limited to a single language, thereby failing to represent the cutting-edge advancements and globalization trends in current deepfake technologies. To address this gap, we propose a novel, multilingual, and multimodal deepfake dataset: PolyGlotFake. It includes content in seven languages, created using a variety of cutting-edge and popular Text-to-Speech, voice cloning, and lip-sync technologies. We conduct comprehensive experiments using state-of-the-art detection methods on PolyGlotFake dataset. These experiments demonstrate the dataset's significant challenges and its practical value in advancing research into multimodal deepfake detection. PolyGlotFake dataset and its associated code are publicly available at: https://github.com/tobuta/PolyGlotFake.

Keywords: Multimodal deepfake · Multilingual deepfake · Deepfake Dataset · Deepfake detection

1 Introduction

In recent years, the emergence of deepfake technology, which leverages advanced deep learning techniques to generate forged content, has captured global attention [18]. A particularly notable significant advancement is the development of multimodal deepfakes [26], which manipulate both visual and audio components in videos. This enhancement substantially increases the realism of the forged content, making it increasingly challenging to differentiate from reality.

Recently, the advancement and popularization of cutting-edge technologies such as Text-to-Speech (TTS), voice cloning, and lip-sync have led to the emergence of a new type of multimodal deepfake on the web. Using Platforms like Heygen [13] and RaskAI [4], producers can easily alter the language spoken by characters in videos. creating convincing fake lip-sync videos. This advancement

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 180–193, 2025.
https://doi.org/10.1007/978-3-031-78341-8_12

in video tampering technology not only overcomes language barriers but also facilitates the rapid global distribution of deepfake content.

The misuse of deepfake technology represents a significant threat to information security. In response, numerous deepfake detection methods have been proposed. These methods [2,9,34,43] are mainly based on deep learning, and their effectiveness is largely dependent on the quality and diversity of the training data. However, the majority of existing deepfake datasets are unimodal [17,21,22,25,32,40,46,47], primarily focusing on visual manipulation and often neglecting the audio aspects. Only a few datasets are multimodal [11,20]. This scarcity of multimodal deepfake datasets leads to the predominance of visual modality focus in current deepfake detection methods.

To the best of our knowledge, DFDC [11] and FakeAVCeleb [20] are among the few publicly accessible multimodal deepfake datasets. While these datasets partially meet the demand for multimodal training data, they employ outdated technologies and are predominantly limited to English content. Consequently, they fail to fully represent the global scope and the cutting-edge status of current deepfake technologies, and these limitations could pose generalization challenges in detecting deepfakes. Furthermore, these datasets usually provide only basic attribute labels, like character attributes (*e.g.*, gender), and lack comprehensive labeling of the techniques used. This deficiency makes it difficult to conduct fine-grained technical traceability analysis of the manipulated videos.

Considering the global trend and technological advancements of deepfake generation technology, we propose PolyGlotFake, a novel multilingual and multimodal deepfake dataset. Specifically, we collected high-quality videos in seven different languages from publicly available video platforms and translate the content of these video into the six other languages. We employ five advanced voice cloning and TTS technologies to generate audio in the target languages. Then, we employ two cutting-edge lip-sync technologies to produce high-quality, realistic, translated videos. Each video is accompanied by detailed technical and attribute labels, which are crucial for analysis and classification in technical traceability. Furthermore, we conduct a comprehensive evaluation of current state-of-the-art deepfake detection methods on our dataset. Experimental results demonstrate the challenges of PolyGlotFake in deepfake detection tasks and its practical value in advancing multimodal deepfake detection research.

Our contributions are summarized as follows:

- We present a novel multimodal, multilingual deepfake dataset comprising seven languages and created using ten multimodal manipulation methods. Notably, no multilingual deepfake dataset has been proposed previously.
- We carefully selected raw videos in seven languages from public platforms and annotated each with fine-grained labels for character features and specific techniques. This deepfake dataset enables more detailed traceability of the technologies used.
- We comprehensively evaluated current state-of-the-art deepfake detection methods on PolyGlotFake and conduct comparative experiments with other datasets. These results demonstrate the challenging nature and the value of PolyGlotFake dataset.

Table 1. Quantitative comparison of PolyGlotFake with existing publicly available video deepfake datasets.

DataSet	Release Data	Manipulated Modality	Mutilingual	Real video	Fake video	Total video	Manipulation Methods	Techniques labeling	attribute labeling
UADFV [46]	2018	V	No	49	49	98	1	No	No
TIMI [21]	2018	V	No	320	640	960	2	No	No
FF++ [40]	2019	V	No	1,000	4,000	5,000	4	No	No
DFD [40]	2019	V	No	360	3,068	3,431	5	No	No
DFDC [11]	2020	A/V	No	23,654	104,500	128,154	8	No	No
DeeperForensics [17]	2020	V	No	50,000	10,000	60,000	1	No	No
Celeb-DF [25]	2020	V	No	590	5,639	6,229	1	No	No
FFIW [47]	2020	V	No	10,000	10,000	20,000	1	No	No
KoDF [22]	2021	V	No	62,166	175,776	237,942	5	No	No
FakeAVCeleb [20]	2021	A/V	No	500	19,500	20,000	4	No	Yes
DF-Platter [32]	2023	V	No	133,260	132,496	265,756	3	No	Yes
PolyGlotFake	2023	A/V	Yes	766	14,472	15,238	10	Yes	Yes

2 Background and Motivation

In this section, we conduct a comprehensive comparison with existing deepfake datasets and detail the limitations of these current datasets. We present a comprehensive list of widely used and publicly available deepfake video datasets for deepfake detection in Table 1. These datasets reflect the gradual evolution of deepfake video generation techniques.

The early deepfake datasets, such as UADFV [46] and TIMIT [21], were created using initial versions of deepfake generation technologies like FakeApp [1] and FaceswapGANs [27]. These early datasets are limited in size, contained a small number of low-quality videos, and suffere from significant visual artifacts. Subsequent studies [25,40] utilized advanced deepfake generation algorithms, targeting creating more diverse and higher-quality deepfake videos with reduced artifacts. Concurrently, several large-scale deepfake datasets [11,17,22,32,47] have been proposed. However, most of these datasets primarily concentrate on visual modalities, focusing on techniques such as face swapping while neglecting the manipulation of audio modalities.

Building on previous work, the DFDC [11] dataset emerged as the first multimodal deepfake dataset, incorporating voice cloning in some videos via TTS Skins [37]. However, DFDC's main emphasis is on visual manipulations, and it does not provide clear labeling for audio manipulations, making it difficult to identify which clips have been audio-manipulated. Subsequently, in 2021, FakeAVCeleb [20] was proposed. This dataset includes four types of multimodal forgeries and provides fine-grained labels for each video. While FakeAVCeleb currently stands as the most prominent multimodal deepfake dataset, it faces limitations, notably in the diversity of manipulation techniques and the linguistic variety of the raw videos. It relies solely on SV2TTS [15] for audio manipulation, a system considered somewhat outdated, resulting in lower-quality voice synthesis compared to cutting-edge TTS technologies. For lip-sync, it uses an older version of Wav2Lip [38], which can produce noticeable artifacts. Another significant limitation is that its real videos are collected from the VoxCeleb2 dataset

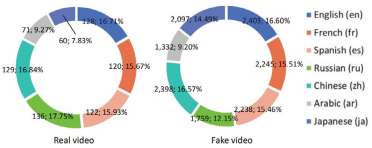


Fig. 1. Language distribution in real and fake videos.

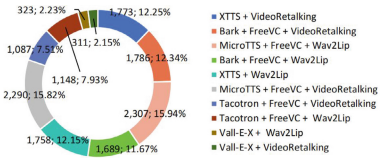


Fig. 2. Synthesis methods distribution in the PolyGlotFake dataset.

Table 2. Attribute distribution by age and sex.

Characteristics		Number	Percentage(%)
Age	0-18	2	0.26
	19-35	366	47.78
	36-55	320	41.78
	56+	78	10.18
Sex	Female	481	62.8
	Male	285	37.2

[8], which is limited to English, thereby restricting the linguistic diversity available for multilingual deepfakes. These constraints diminish the dataset's variety and realism, impacting the generalizability of detectors trained with it.

As a result, current multimodal datasets still exhibit significant limitations in terms of manipulating technical and linguistic diversity. This research gap highlights the urgent need for more technologically advanced, diverse, and globally representative deepfake datasets.

Furthermore, it is worth noting that many current datasets are often promoted based on their large scale. However, for the specialized task of deepfake detection, an excessively large scale can result in longer training periods. This not only reduces experimental efficiency but may also hinder the ability to quickly iterate and test new detection techniques. Additionally, ensuring the quality and consistency of each sample in a very large dataset can be challenging, which in turn affects the performance and reliability of the model. Therefore, in PolyGlotFake, our emphasis is on creating a high-quality, diverse dataset rather than merely focusing on its scale.

3 PolyGlotFake Dataset

The PolyGlotFake dataset comprises a total of 15238 videos, including 766 real videos and 14472 fake videos. The average duration of each video is 11.79 s, with a resolution of 1280*720.

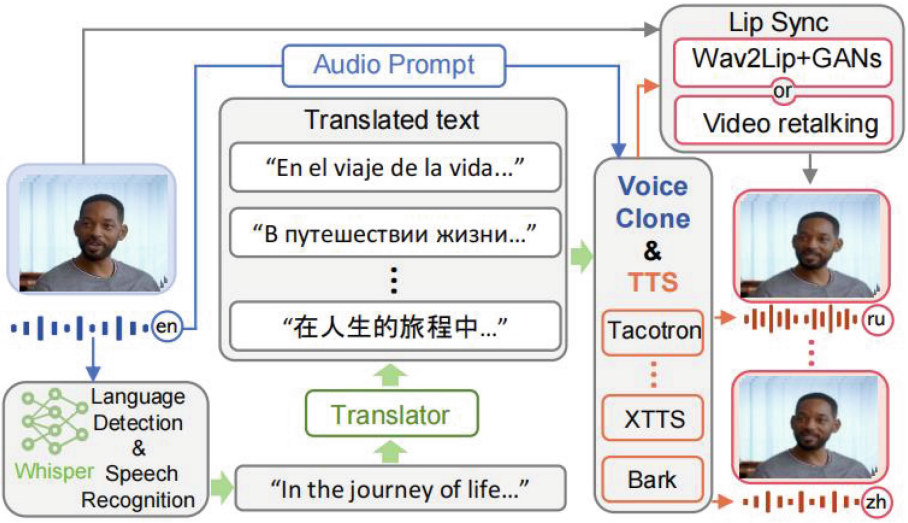


Fig. 3. Generation Pipeline of PolyGlotFake Dataset. Original videos are separated into video and audio. The audio is transcribed into text using Whisper [36] and subsequently translated into multiple languages using a translator. These translated texts are then converted into audio through Text-to-Speech and voice cloning models. Finally, the original video clips are synchronized with the generated audio using a lip-sync model.

3.1 Data Collection

The high-quality raw (*i.e.* real) videos are collected from YouTube, including content in seven different languages. Figure 1 shows the linguistic distribution in collected raw videos and manipulated videos. To ensure the accuracy of subsequent translations, we manually verify that each sentence in the videos is complete. The selection of languages is based on their global popularity and compatibility with existing popular open-source TTS models. These languages include the six official languages of the United Nations (*i.e.*, English, French, Spanish, Russian, Chinese, Arabic) and Japanese. We also conducted detailed labels of the collected videos, encompassing information such as their sources, duration, as well as the gender and age of the characters in videos. The attribute distribution by age and sex is shown in Table 2. Additionally, we preserved the video's background instead of extracting only facial regions, thereby retaining as much of the original video information as possible.

3.2 Synthesized Data

For the generation of fake videos, we employ cutting-edge and popular visual and audio manipulation methods based on realistic deepfake generation cases found in internet media.

For audio modality manipulation, we use the following five methods.

- **XTTS** [3]: A powerful and popular open-source TTS model built on the Tortoise and developed by Coqui AI. XTTS supports 16 languages and enables cross-lingual voice cloning and multilingual speech generation with only three-second audio prompts.
- **Bark** [5] + **FreeVC** [23]: Bark is a Transformer-based multilingual TTS model developed by Suno-AI that supports 13 languages and is capable of generating highly realistic, multilingual speech and other audio content such as music. FreeVC is a high-quality, text-free, one-short voice conversion system. Since Bark does not support cross-language voice clones, we use Bark to generate the corresponding speech first and then FreeVC to realize the voice clone according to the audio prompt.
- **Vall-E-X** [42]: An efficient multilingual text-to-speech synthesis and voice cloning model recently proposed by Microsoft. It can efficiently realize high-quality voice cloning with only three seconds of an audio prompt. It currently supports three languages.
- **Microsoft TTS** [29] + **FreeVC**: Microsoft TTS supports multiple languages and dialects. Given its widespread use on the internet, we design manipulation schemes that combine it with FreeVC.
- **Tacotron** [44] + **FreeVC**: Tacotron is an advanced TTS synthesis system proposed by Google. It is known for its seq2seq architecture and ability to generate highly natural and fluent speech. Similarly, We combine it with FreeVC.

For visual modality manipulation, we employ the following two methods based on the popularity and generation quality:

- **Wav2Lip** [38] + **GANs**: Wav2Lip is a widely used, highly accurate lip-sync model proposed in 2020. This model can accurately match any speech to the lip movements of a character in a video, often utilized in deepfake for face reenactment tasks. The basic Wav2Lip model alone tends to produce videos of low quality. However, by integrating it with Generative Adversarial Networks (GANs), the video quality can be significantly enhanced. In this study, we employ a fine-tuned Wav2Lip plus GANs model to produce high-quality lip-sync videos.
- **VideoRetalking** [7]: VideoRetalking is a audio-driven lip-sync system recently proposed by Cheng *etc.*. This system generates lip-sync videos by processing audio and video in a series of sequential steps. The generated video frames are finally enhanced and repaired using an identity-aware enhancement network.

Additionally, for generated video we label the detailed audio and visual manipulation techniques used, The distribution of the various combinations of techniques is shown in Fig. 2. For instance, in the pie chart, the gray section represents the percentage of videos that use MicroTTS and FreeVC for voice manipulation, and videoRetalking for lip syncing. There are 2,290 such videos, accounting for 15.82% of all fake videos.

Table 3. Visual quality assessment and comparison. The first column shows the different Datasets and the second and third columns show the FID and BRISQUE values measured in that Dataset, respectively. lower values of FID and BRISQUE indicate better quality.

DataSet	FID ↓	BRISQUE ↓
FF++	4.12	52.17
CelebDF	3.72	42.23
DFDC	5.91	74.52
FakeAVCeleb	4.32	69.31
PolyGlotFake	3.25	46.21

Table 4. Audio quality assessment and comparison. The first column shows FakeAVCeleb and the parts of PolyGlotFake that use different sound manipulation techniques. The second column shows the Mos value of the audio in these datasets, where larger indicates higher audio quality.

DataSet	Mos ↑
FakeAVCeleb	3.17
PolyGlotFake(XTTS)	4.12
PolyGlotFake(MicroTTS+FreeVC)	4.51
PolyGlotFake(Vall-E-X)	3.22
PolyGlotFake(Tacotron+FreeVC)	4.57
PolyGlotFake(Bark+FreeVC)	4.30
PolyGlotFake(Overall)	4.12

The fake video generation pipeline is shown in Fig. 3. We first extract the audio from the original video and use Whisper [36] to convert the speech to text while detecting its language. Then, the text output from Whisper [36] is translated into other languages using Microsoft's Translate API. For example, If the output text is in English, the original English text will be translated into Spanish, Russian, Chinese, Japanese, Arabic, and French. We select a suitable TTS model based on the translated text and randomly cut 10 s from the original audio as an audio prompt. The selected TTS model converts the text to audio and performs sound cloning based on the audio prompt. Then, the lip-sync model performs face reenactments of the original video based on the TTS output audios, resulting in a series of high-quality manipulated videos in different languages generated using several techniques.

3.3 Quality Assessment

We perform quality assessments for PolyGlotFake dataset in visual and audio modalities. For the quality assessment of visual modality, we adopt the Frechet Inception Distance (FID) and the no-reference image assessment

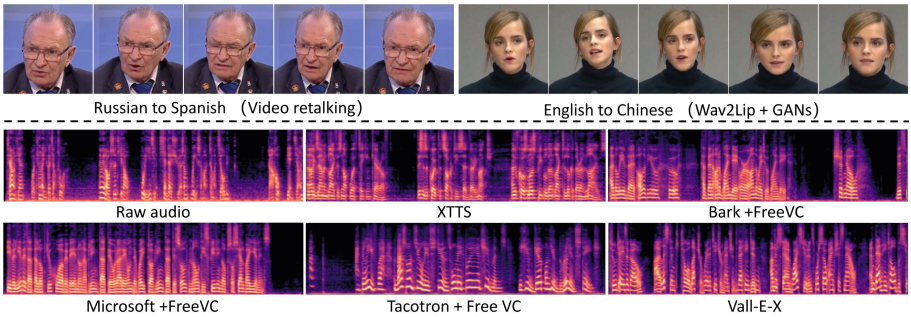


Fig. 4. Visualization of some video frame samples and Mel spectrograms of audio sample clips in the PolyGlotFake dataset.

method BRISQUE [31]. We also compar the quality of the PolyGlotFake dataset with several other well-known datasets, including FF++, Celeb-DF, and FakeAVCeleb, and the related results are presented in Table 3. For the audio modality quality assessment, we employ the non-invasive audio assessment method NISQA [30] to compute the Mean Opinion Score (MOS), and compare the result with FakeAVCeleb. The detailed assessment results for each synthesis method are shown in Table 4.

Based on our quality evaluations, it is clear that the PolyGlotFake dataset exhibits high performance in both visual and audio quality aspects. Additionally, Fig. 4 presents selected video frame samples and Mel spectrograms of audio sample clip from the PolyGlotFake dataset. Both visualization and quantitative quality assessment confirm the superior quality of PolyGlotFake across both visual and audio modalities.

4 DeepFake Detection Benchmark

In this section, we first comprehensively evaluate several existing state-of-the-art deepfake detectors on the PolyGlotFake dataset and compare the performance of these detectors across different datasets. Then, we conduct a qualitative and quantitative experiment on the audio modality of our dataset to explore the impact of different languages on detection performance.

4.1 Selection of Detectors

Current deepfake detection methods can be broadly categorized into three groups: naive detectors, spatial detectors and frequency detectors. ❶ Naive detectors employ CNNs to directly distinguish fake images from real ones. ❷ Spatial detectors examine the spatial domain of images in greater detail using specially designed structures to detect features like fusion boundaries and artifacts. ❸ Frequency detectors analyze the frequency domain of images to identify forgery features such as high-frequency artifacts.

Table 5. Evaluation results and comparisions with other datasets. All detectors were trained on the FakeAVCeleb dataset and tested on FakeAVCeleb, DFDC, and PolyGlotFake. Consequently, the FakeAVCeleb column represents the AUC values obtained from intra-dataset evaluation, while the DFDC and PolyGlotFake columns represent the AUC values from cross-dataset evaluation.

Type	Detector	Backbone	DataSet		
			FakeAVCeleb	DFDC	PolyGlotFake
Naive	MesoNet [2]	Designed	0.7332	0.5906	0.5672
Naive	MesoInception [2]	Designed	0.7945	0.6344	0.5831
Naive	Xception [40]	Xception	0.9169	0.6530	0.6052
Naive	EfficienNet-B4 [41]	EfficienNet	0.9023	0.6020	0.5769
Spatial	Capsule [34]	Capsule	0.8663	0.6146	0.6068
Spatial	FFD [10]	Xception	0.9285	0.6583	0.5960
Spatial	CORE [35]	Xception	0.9345	0.6625	0.6220
Spatial	RECCE [6]	Designed	0.9396	0.6884	0.6596
Spatial	DSP-FWA [24]	Xception	0.9115	0.6929	0.6658
Spatial	TALL [45]	Transformer	0.9387	0.6830	0.6535
Spatial	CADDM [12]	Designed	0.9533	0.6931	0.6737
Frequency	F3Net [39]	Xception	0.9416	0.6452	0.6439
Frequency	SRM [28]	Xception	0.9043	0.6346	0.6143
Ensemble	XRes [19]	Designed	0.9556	0.7042	0.6835

To perform the experiments, we employ a total of 15 state-of-the-art deepfake detectors. This set included four naive detectors, namely MesoNet [2], MesoInception [2], Xception [40], and EfficientNet-B4 [41]; seven spatial detectors, Capsule [34], FFD [10], CORE [35], RECCE [6], and DSP-FWA [24], TALL [45], CADDM [12]; and two frequency detectors, F3Net [39] and SRM [28]. In addition, for multimodal deepfake detection, we use XRex [19], an ensemble model combining Xception and ResNet. In this model, Xception is used for visual modality detection, and ResNet is used for audio modality detection. The selection of these detectors was based on the popularity and public availability of their code.

4.2 Experimental Setting

We divide the dataset into training, validation, and testing sets in the ratio of 8:1:1. To ensure the representativeness of each technique combination in the dataset division; we use a stratified sampling method to ensure that the proportion of each combination is consistent across the datasets. For exisiting detection methods, we follow the respective data preprocessing steps. For the ensemble-based model, we randomly clip three seconds from each audio and convert it into a three-channel Mel Frequency Cepstral Coefficient (MFCC) feature as the input for the audio modality and extract ten frames from each video as input for the visual model.

To ensure fairness, we train all detectors on the FakeAVCeleb dataset and evaluate them on both the DFDC and PolyGlotFake datasets. We use the Area

Table 6. Audio modality analysis. The first column represents the generative models used, while the other columns represent the languages of the audio. "-" indicates that the model does not support generating this language.

Models	en	fr	es	ru	zh	ar	ja
XTTS	0.8552	0.7134	0.6854	0.6764	0.7323	0.6614	-
Bark	0.7336	0.6835	0.7415	0.7632	0.7153	-	0.6824
MicroTTS	0.8364	0.7543	0.7385	0.6978	0.7576	0.6954	0.7143
Tacotron	0.7422	0.6945	0.7223	-	0.8132	-	0.7321
Vall-E-X	0.8064	-	-	-	0.8354	-	0.6974

Under the Curve (AUC), a commonly used evaluation metric for deepfake detection, as our experimental metric.

For the audio modality analysis, we first divid the test set of the dataset according to language, ensuring that the number and ratio of real to fake videos remains the same for each language groups. We use a ResNet trained on the audios (MFCC) of PolyGlotFake dataset to detect the audios of each group and reported the AUC for each language category in Table 6.

4.3 Result and Analysis

Table 5 reports the results of our experiments. The FakeAVCeleb column shows the intra-dataset detection results, which reveal that the spatial detector with a specialized structural design and the frequency detectors outperform the naive detectors. For instance, the detection result of Xception is 0.9169, while CORE, which also utilizes Xception as a backbone, achieves a result of 0.9345.

The DFDC and PolyGlotFake columns present results obtained from cross-dataset detection. Comparing these results with the intra-dataset detection results indicates significant performance degradation for detectors trained on FakeAVCeleb when faced with unseen Deepfake content. Furthermore, the performance of the detectors on the PolyGlotFake dataset is significantly worse than on DFDC. This suggests that PolyGlotFake includes a wider variety of unknown synthesis techniques, making it a more challenging dataset for these detectors.

The result in Table 6 indicates that even when detecting audio with the same semantic content generated by the same model but in different languages, the detection results remain different. This suggests that due to differences in the training data for each language used during the TTS training phase, model adaptability, and the distinct characteristics of each language, the TTS leaves varying degrees of detectable traces when generating audio of different languages. This further underscores the importance and significance of establishing a multilingual, multimodal deepfake dataset.

5 Conclusion

In this study, we propose PolyGlotFake, a multilingual, multimodal deepfake dataset that employs cutting-edge multimodal manipulation techniques. Each technique used in this dataset is meticulously annotated to aid in technical traceability analysis. Furthermore, we comprehensively evaluate various state-of-the-art deepfake detectors on this dataset. The experiment results demonstrate the challenging nature and practical value of our dataset. We comprehensively evaluated various state-of-the-art deepfake detectors using this dataset. The experimental results underscore the challenging nature and the practical value of PolyGlotFake, demonstrating its potential to significantly advance the field of multimodal deepfake detection.

Due to the complexity of data collection and processing, our dataset also faces issues such as an imbalance in the proportions of various attributes and a small scale of real video data. In future research, we aim to address these fairness issues by improving the balance of attributes and expanding the scale of our dataset. We also plan to enhance the linguistic diversity of our dataset. Additionally, in response to recent studies [14,16,33] showing how adversarial perturbations can help evade detection, we plan to explore methods for implementing such perturbations in practical scenarios. This includes incorporating subtle adversarial perturbations into both the audio and video components of our dataset.

Acknowledgements. This work was supported by JST SPRING, Grant Number JPMJSP2136.

Ethics Statement. Access to the dataset is restricted to academic institutions and is intended solely for research use. It complies with YouTube's fair use policy through its transformative, non-commercial use, by including only brief excerpts (approximately 20 s) from each YouTube video, and ensuring that these excerpts do not adversely affect the copyright owners' ability to earn revenue from their original content. Should any copyright owner feel their rights have been infringed, we are committed to promptly removing the contested material from our dataset.

References

1. Deepswap - ai-powered DeepFake technology (2023). https://www.deepswap.net/. Accessed 24 Dec 2023
2. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: MesoNet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7. IEEE (2018)
3. AI, C.: Github repository for coqui AI text-to-speech (2023). https://github.com/coqui-ai/tts. Accessed 29 Dec 2023
4. AI, R.: Rask AI official website (2023). https://zh.rask.ai/. Accessed 29 Dec 2023
5. AI, S.: Github repository for suno ai's bark project (2023). https://github.com/suno-ai/bark. Accessed 29 Dec 2023

6. Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., Yang, X.: End-to-end reconstruction-classification learning for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4113–4122 (2022)
7. Cheng, K., et al.: Videoretalking: audio-based lip synchronization for talking head video editing in the wild. In: SIGGRAPH Asia 2022 Conference Papers, pp. 1–9 (2022)
8. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
9. Coccomini, D.A., Messina, N., Gennaro, C., Falchi, F.: Combining efficientNet and vision transformers for video DeepFake detection. In: Sclaroff, S., Distante, C., Leo, M., Farinella, G.M., Tombari, F. (eds.) ICIAP 2022. LNCS, vol. 13233, pp. 219–229. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-06433-3_19
10. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5781–5790 (2020)
11. Dolhansky, B., et al.: The DeepFake detection challenge (DFDC) dataset. arXiv preprint arXiv:2006.07397 (2020)
12. Dong, S., Wang, J., Ji, R., Liang, J., Fan, H., Ge, Z.: Implicit identity leakage: the stumbling block to improving DeepFake detection generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3994–4004 (2023)
13. Heygen: Heygen official website (2023). https://www.heygen.com/. Accessed 29 Dec 2023
14. Hou, Y., Guo, Q., Huang, Y., Xie, X., Ma, L., Zhao, J.: Evading DeepFake detectors via adversarial statistical consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12271–12280 (2023)
15. Jemine, C.: Real-time-voice-cloning. University of Liége, Liége, Belgium p. 3 (2019)
16. Jia, S., Ma, C., Yao, T., Yin, B., Ding, S., Yang, X.: Exploring frequency adversarial attacks for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4103–4112 (2022)
17. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deeperforensics-1.0: a large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2889–2898 (2020)
18. Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., Liu, Y.: Countering malicious DeepFake: survey, battleground, and horizon. Int. J. Comput. Vision **130**(7), 1678–1734 (2022)
19. Khalid, H., Kim, M., Tariq, S., Woo, S.S.: Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In: Proceedings of the 1st Workshop on Synthetic Multimedia-audiovisual DeepFake Generation and Detection, pp. 7–15 (2021)
20. Khalid, H., Tariq, S., Kim, M., Woo, S.S.: Fakeavceleb: a novel audio-video multimodal deepfake dataset. arXiv preprint arXiv:2108.05080 (2021)
21. Korshunov, P., Marcel, S.: DeepFakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685 (2018)
22. Kwon, P., You, J., Nam, G., Park, S., Chae, G.: Kodf: a large-scale Korean DeepFake detection dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10744–10753 (2021)
23. Li, J., Tu, W., Xiao, L.: Freevc: towards high-quality text-free one-shot voice conversion. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)

24. Li, Y., Lyu, S.: Exposing DeepFake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018)
25. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: a large-scale challenging dataset for DeepFake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3207–3216 (2020)
26. Liz-López, H., Keita, M., Taleb-Ahmed, A., Hadid, A., Huertas-Tato, J., Camacho, D.: Generation and detection of manipulated multimodal audiovisual content: advances, trends and open challenges. Inf. Fusion **103**, 102103 (2024)
27. Lu, S.: faceswap-GAN: A GAN-based faceswap project on github (2023). https://github.com/shaoanlu/faceswap-GAN. Accessed 24 Dec 2023
28. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16317–16326 (2021)
29. Microsoft: Microsoft azure text-to-speech services (2023). https://azure.microsoft.com/en-us/products/ai-services/text-to-speech. Accessed 29 Dec 2023
30. Mittag, G., Naderi, B., Chehadi, A., Möller, S.: Nisqa: a deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. arXiv preprint arXiv:2104.09494 (2021)
31. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. **21**(12), 4695–4708 (2012)
32. Narayan, K., Agarwal, H., Thakral, K., Mittal, S., Vatsa, M., Singh, R.: DF-platter: Multi-face heterogeneous DeepFake dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9739–9748 (2023)
33. Neekhara, P., Dolhansky, B., Bitton, J., Ferrer, C.C.: Adversarial threats to DeepFake detection: a practical perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 923–932 (2021)
34. Nguyen, H.H., Yamagishi, J., Echizen, I.: Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467 (2019)
35. Ni, Y., Meng, D., Yu, C., Quan, C., Ren, D., Zhao, Y.: Core: consistent representation learning for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12–21 (2022)
36. OpenAI: Github repository for openai whisper project (2023). https://github.com/openai/whisper. Accessed 29 Dec 2023
37. Polyak, A., Wolf, L., Taigman, Y.: Tts skins: speaker conversion via asr. arXiv preprint arXiv:1904.08983 (2019)
38. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 484–492 (2020)
39. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: face forgery detection by mining frequency-aware clues. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 86–103. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_6
40. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics: a large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179 (2018)
41. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
42. Wang, C., et al.: Neural codec language models are zero-shot text to speech synthesizers. arXiv preprint arXiv:2301.02111 (2023)

43. Wang, J., et al.: M2tr: multi-modal multi-scale transformers for DeepFake detection. In: Proceedings of the 2022 International Conference on Multimedia Retrieval, pp. 615–623 (2022)
44. Wang, Y., et al.: Tacotron: towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 (2017)
45. Xu, Y., Liang, J., Jia, G., Yang, Z., Zhang, Y., He, R.: Tall: thumbnail layout for DeepFake video detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22658–22668 (2023)
46. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265. IEEE (2019)
47. Zhou, T., Wang, W., Liang, Z., Shen, J.: Face forensics in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5778–5788 (2021)

AnnChor: A Video Dataset for Temporal Action Localization in Classical Ballet Choreography

Margaux Bowditch and Dustin van der Haar(✉)

University of Johannesburg, Auckland Park, Johannesburg 2092, South Africa
dvanderhaar@uj.ac.za
http://www.uj.ac.za

Abstract. *Video action understanding* is a rapidly growing field that has achieved excellent results in various application areas, such as sports and lifestyle applications. However, research that combines computer vision action understanding techniques and the artistic domain of classical ballet choreography is still in its infancy. Publicly available ballet video datasets are limited in number and need more richness to properly explore this specialized field and its extensive collection of actions. Recordings of ballet rehearsals, performances, and competitions have become more readily available on public platforms in recent years, making a substantial amount of data available in this discipline. We propose a novel video dataset, *AnnChor*, for temporal action localization in ballet choreography. The dataset is notable for its quality and the diversity of ballet actions found in the videos of solo ballet performances. The full dataset comprises 1020 videos with over 25 000 temporal annotations for 11 action classes. We evaluate and provide baseline results for temporal action localization using the Coarse-Fine Network and TriDet models. There is much opportunity to advance computer vision technology to aid the classical dance domain. We hope this dataset will benefit the computer vision community and enable researchers to explore the challenges present in action localization, especially in the context of fine-grained ballet movements. The dataset can be found at https://github.com/dvanderhaar/UJAnnChor.

Keywords: Fine-grained Temporal Action Localization · Ballet Dataset · Video Understanding

1 Introduction

The field of video action understanding has continually grown over recent years with the publication of multiple influential and large-scale video datasets. In this

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78341-8_13.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 194–209, 2025.
https://doi.org/10.1007/978-3-031-78341-8_13

paper, we refer to *video action understanding* as an umbrella term that includes the task of temporal action localization. The success of many new machine learning models is driven by state-of-the-art and emerging datasets that explore a variety of domains for the task of temporal action localization [12], such as Charades [28], ActivityNet [3], MultiTHUMOS [37], HACS Segments [39], FineAction [20] and FineGym [26], to name a few. The applications for video understanding are continually branching out into several specialized domains, and temporal action localization remains challenging as each domain presents unique problems.

Classical ballet is an exciting application domain for the temporal action localization task, mainly due to the rich vocabulary of actions present in the art form. Ballet solos in the classical repertoire, also known as ballet *variations*, consist of carefully choreographed step sequences filled with many distinct ballet actions [25] that are performed by a single dancer. However, limited datasets for ballet exist that are sufficiently rich in action classes and appropriate for temporal analysis. There are well-established ballet competitions that take place yearly where multiple dancers perform variations from great classical ballets. With the ever-increasing popularity of online video and streaming platforms, recordings of ballet competitions, rehearsals, and professional performances have become more readily available to the public. There is an opportunity to use the data available in this discipline to explore how the application of computer vision technology may benefit the ballet community and reveal new challenges to researchers.

Creating the AnnChor dataset is motivated by technology's potential to simplify tasks such as notating choreography as well as teaching and correcting ballet movements. The *AnnChor* dataset furthermore demonstrates the difficulty of fine-grained action localization. Automating aspects of the typical tasks found within the fine-grained action setting of ballet, using computer vision requires high-quality data.

This paper proposes a novel video dataset, *AnnChor*, for temporal action localization in ballet choreography. We use the term *temporal action localization*, also known as *temporal action detection* according to [12], to refer to assigning action labels to temporal segments in a video based on the likelihood of an action occurring within a start and end frame window. *AnnChor* is a high-quality dataset consisting of high-resolution videos with a diverse range of annotated ballet actions. Spatial localization is not a part of the dataset as the focus is on solo dance performances where the dancers are mostly isolated. The dataset creation and annotation tasks for *AnnChor* are subject to a carefully defined control process to ensure reliable annotations (Fig. 1).

To take on the challenge of temporal action localization within this relatively unexplored space, we perform a baseline study by using Coarse-Fine Networks [13] and TriDet [27]. The results of our study reveal that there is still much room for progress in applying temporal action localization techniques to the generally complex area of ballet choreography with its many fine-grained movements.

Our work, therefore, contributes to research in video action understanding in the following ways: 1) We create a novel video dataset, *AnnChor* for action understanding in the specialized domain of classical ballet, which is rich in qual-

Fig. 1. An overview of the *AnnChor* dataset showing some of the actions that are present in the selected ballet sequences. We provide fine-grained temporal annotations for 1020 videos.

ity and diversity of actions. 2) We conduct a baseline study on the dataset as a benchmark using Coarse-Fine Networks and TriDet to demonstrate the challenges that arise for this application field's fine-grained temporal action localization task.

Section 2 of this paper presents the background and related work. Section 3 contains information about the dataset presenting details such as dataset construction, annotation, considerations, and the evaluation protocol. Section 4 provides information on the baseline experiments and methods. The results and analysis of the paper are found in Sect. 5. Finally, the paper ends with the conclusion in Sect. 6.

2 Background and Related Work

Classical ballet is a very well-codified art form with precise movements focusing on the correct alignment of the body [5,24]. Interesting research works in the body of literature have combined technological research with ballet. However, the use of action understanding techniques in this context still needs to be improved. Some of the literature has focused on assisting dancers with proper training and correction using various camera and wearable sensors as well as virtual reality technology [1,16,23,30–32]. Other research works have focused on the choreographic side of ballet and dance in general, by considering dance notation systems as well as generative choreography [2,14,17,22,38]. The advancement of work in the combined computer vision and classical ballet area for future research is largely dependent on the availability of substantial, high-quality datasets.

Fig. 2. Frames from the *AnnChor* dataset showing the progressions of the Grand Jeté action from the Black Swan variation

Several existing datasets offer temporal annotations for action localization, primarily in sports or non-artistic domains. For example, the Charades and ActivityNet datasets focus on coarse-grained action understanding, featuring general action instances such as "Watching television" and "Playing hockey" [3,28,37]. In contrast, the FineAction dataset concentrates on fine-grained action instances from multiple domains [20]. However, specialized domains like ballet necessitate domain-specific fine-grained annotations. Other existing specialized datasets include Diving48 and FineDiving for diving activities [18,36], and FineGym for gymnastics [26].

There are a few existing datasets that are mentioned in related work which explore dance as a topic. Examples of general dance datasets include: Let's Dance [4] as well as the AIST++ [18] datasets. The ballet-focused video datasets, however, are limited in number of samples and subjects with little diversity and richness of ballet action annotations. One video ballet dataset that is mentioned in the literature is called Ballet Movements, created by [9], which has 44 video sequences with 8 action classes and 3 subjects [33,34]. The works that make use of this dataset mention the complexity of ballet movement patterns, focusing on action recognition as a task, but not fine-grained temporal action localization [9].

It is also clear from recent literature that performing action localization and recognition on temporally untrimmed videos is a challenging but necessary task for a realistic analysis of human actions [10–12,29]. As such, the *AnnChor* dataset focuses on providing temporal annotations for ballet actions performed in various dance-specific environments. This dataset contributes a relatively significant amount of data together with action classes for the highly nuanced movements within the ballet repertoire. In the next section, we present the details of the *AnnChor* dataset.

3 The AnnChor Dataset

3.1 Dataset Background and Overview

The *AnnChor* video dataset has been carefully constructed by ballet subject matter experts. It presents a new opportunity for researchers to explore action

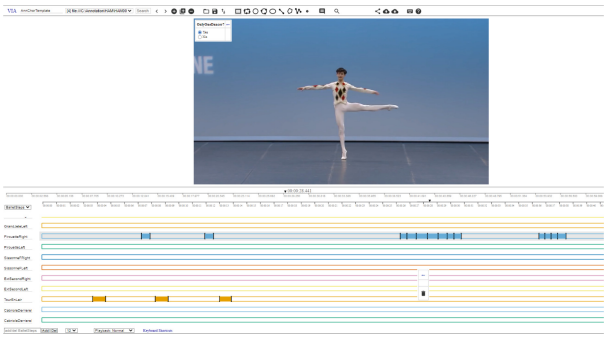


Fig. 3. An example of the *AnnChor* dataset annotation process using the VIA Video annotation tool

understanding problems in a uniquely complex application area. Ballet is a historically rich art form where companies perform masterpiece full-length ballets worldwide. Every full-length ballet contains solo-performance pieces, known as variations, which are performed by a single male or female dancer. Many of these variations from ballet repertoire are well-known within the ballet community, for example, the Black Swan variation from Swan Lake, is a dynamic piece filled with various quick, fine-grained actions as demonstrated by the frames illustrated in Fig. 2. These variations are often performed as stand-alone solo pieces at ballet competitions [21].

A significant number of records of ballet competitions, as well as studio rehearsals and stage performances, have become available on public platforms such as YouTube. Therefore, much data is readily available with the same choreography performed by many dancers. However, little of this data has been utilized for understanding video actions. *AnnChor* addresses the opportunity and need for an adequately annotated high-quality dataset in the classical ballet space with a ballet-specific class vocabulary. The *AnnChor* dataset is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

3.2 Dataset Construction

Video datasets are often collected by crawling the web or capturing the video data by recording study participants. Furthermore, crowd-sourcing labeling services such as Amazon Mechanical Turk are often used to gather annotations [12]. However, due to the required domain knowledge to build a high quality dataset for classical ballet, 3 domain experts with advanced knowledge of classical ballet carefully selected and annotated the data for the *AnnChor* dataset. Fifty-one ballet variations of similar choreography were selected, and 20 high-quality video samples were collected for each of these variations resulting in 1020 videos. In this section, we will unpack various aspects that were considered during the dataset construction phase.

Fig. 4. Sample showing the start and end frames of a Balancing Extension Derriére action

Action Vocabulary. We select the vocabulary by consulting sources such as [25] and observing which basic ballet actions occur most frequently across video footage of well-known ballet variations. Therefore, the vocabulary is based on ballet actions prevalent in classical variation choreography. We also consider which actions are simple to annotate choreographically using a dance notation system such as Benesh Movement Notation [6]. Furthermore, in consultation with the ballet domain experts, the agreed-upon vocabulary was selected based on similar related actions with well-defined starting and ending moments in the ballet context.

Data Preparation and Considerations. We searched for competition, rehearsal, and performance footage on YouTube to gather sufficient high-quality video data. Ethical clearance was obtained prior to the creation of the *AnnChor* dataset which consists of data from publicly available YouTube videos. One constraint in place during the collection phase of the *AnnChor* dataset included that the video background environments had to be consistent with a typical ballet studio or stage environment. We also decided that no mirrors should be present in the background of any video to avoid unnecessary noise that may negatively affect computer vision tasks. The collected video samples had to have one main dancer performing a solo as the focal point and not include any group dance performances. Only high-resolution footage of at least 720P was selected, with clear sections showing the performer's whole body. Some footage in the dataset shows the same dancer performing the same solo on different occasions. Therefore, data collectors carefully inspected the dataset for duplicate videos. Only truly identical duplicates were removed manually to ensure unique samples per solo. Furthermore, the videos were manually trimmed based on each solo's known conventional starting and ending moments to ensure consistency across samples per solo. For example, the Black Swan (extract shown in Fig. 2) variation's choreography typically starts and ends with *Pirouette* actions, and each sample of this solo has, on average, a similar number of actions.

Annotation. In the area of temporal action localization, a common challenge is ensuring consistency across the annotations of the temporal actions in a given dataset [19]. We addressed this challenge by creating an annotation guide that defines the action boundaries of each ballet step in our action vocabulary.

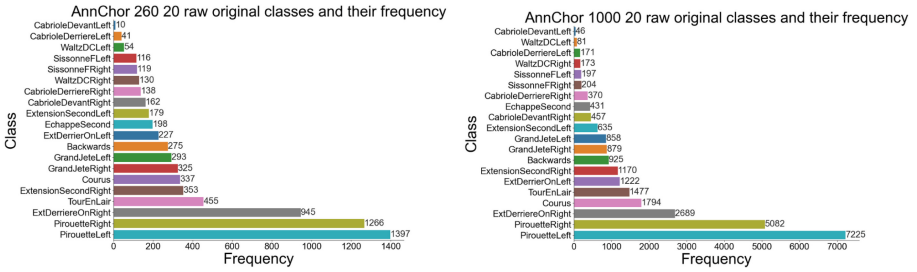


Fig. 5. The number of annotations for the initial annotated 20 classes for the sub-dataset, *AnnChor260* and the full dataset, *AnnChor1000*, sorted in ascending order.

The guide was iteratively improved until consensus among the 3 annotators on action windows and grouping similar actions was reached.

The guide refers to the dancer's *supporting leg* as the leg on which the weight of the body is placed, and the *working leg* as the leg that is lifted in a non-weight-bearing position. An example of a clearly defined action boundary as stipulated in the guide for the *Balancing Extension Derriére* movement, which is presented in Fig. 4, is as follows: The action starts with the dancer's weight transferred onto their straight supporting leg while the working leg is at its lowest point lifted behind the body; The action ends with the dancer's supporting leg still in position, but the working leg is at its highest point lifted behind the body in a balancing pose. The kinds of movements found in the *Balancing Extension Derriére* action class include, in ballet terms, steps such as the *Arabesque* and *Attitude Derriére* actions due to their visual similarity. Therefore, future iterations of the *AnnChor* dataset may contain further fine-grained sub-actions for classes like *Balancing Extension Derriére*.

The annotators were thoroughly briefed, and the pre-defined annotation guide was used. The annotation guide is available in the supplementary material of this paper. Whenever an annotator came across an unclear action instance, all annotators were notified, and the required annotation approach was decided on through collective agreement among the annotators. The average time taken among annotators to complete the annotation of a single video was 12.5 min. Figure 3 shows a screenshot of the annotation process using the open-source VIA Video Annotation tool [7,8]. As seen in Fig. 3 the temporal actions do not overlap, each taking place at distinct time intervals. The only exception in overlapping annotations is for the *backwards* action class which was added to annotate actions that are performed facing away from the audience. The format of the annotations allows for start and end timestamps to be captured along with a class name which is specified based on the provided timelines presented in Fig. 3.

The original raw annotations of the *AnnChor* dataset had 20 different action classes. These classes allow for variation on the sagittal plane, which divides the body into right and left sections. These annotated classes can be seen in Fig. 5.

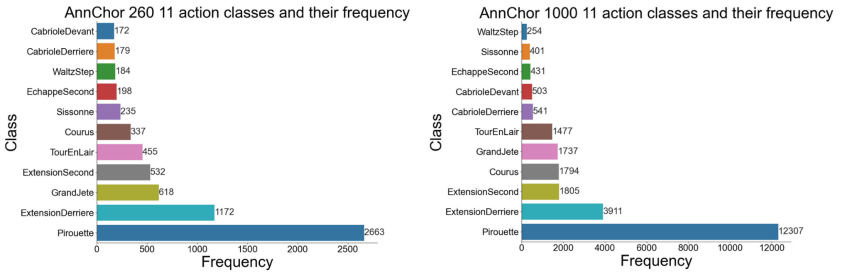


Fig. 6. The number of annotations per class for the final 11 classes for the sub-dataset, *AnnChor260* and the full dataset, *AnnChor1000*, sorted in ascending order.

The reason for initially annotating at this level of detail is due to the fact that the dataset contains identical movements of certain action classes that occur towards the left and the right, respectively. A "backwards" class also allows for annotating actions facing away from the camera. Therefore, there are opportunities to explore various class configurations with AnnChor. A representative sample of the dataset with 260 videos, called *AnnChor260* was created for preliminary benchmark purposes. We refer to the complete dataset with its 1020 videos as *AnnChor1000*. As a starting point for our baseline results, we grouped the relevant classes of the same step for left and right into one, which results in the 11 classes that can be seen in Fig. 6. The distribution of class action annotations are presented in Figs. 5 and 6. The classes in Figs. 5 and 6 have been sorted by the number of annotations for each class. One can clearly see a long-tailed distribution with the pirouettes class being the most frequently occurring action.

Dataset Challenges. There are several difficulties present in the *AnnChor* dataset. Firstly, ballet is a complex performing art containing fast movements with extreme body deformations, which is still challenging for video action understanding models. Secondly, subtle nuances exist in different dancers' approaches to specific ballet movements. For example, the choice of arms and speed at which specific actions are performed vary among dancers. Therefore, another challenge is that the *AnnChor* actions are fine-grained with significant variations in the duration of action instances. The dataset also contains videos with major changes in the field of view and angles from which the dancers are filmed. Another interesting challenge within *AnnChor* includes the symmetry of specific actions as the dancer transitions out of the movement into the next step. An example is the *Balancing Extension Derriére* action presented in Fig. 4 where the transition out of the action is very similar to the transition into the action except for the direction in which the working leg is moving. Lastly, the uneven class distribution, while representative of action frequencies generally found in ballet solos, also contributes to the challenges for action localization models.

Quality Control. It was essential to have ballet dance and teaching experts with extensive knowledge of the classical ballet technique annotate this specialized

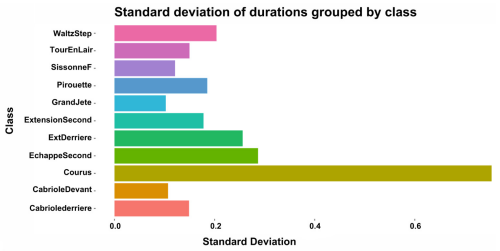


Fig. 7. *AnnChor1000* statistics showing the standard deviation from the average duration per action class

dataset since they could detect necessary details and nuances in the *AnnChor* actions and annotate appropriately. The full dataset has been reviewed to ensure annotation consistency, quality, and consensus among annotators. The review process required annotators to review one another's work by fixing incorrect annotations or adding any missed action instances.

3.3 Dataset Properties

The full *AnnChor* dataset consists of data for 51 ballet solos with 20 high-resolution videos for each solo, which results in an overall total of 1020 videos. For *AnnChor1000*, the overall average temporal length per video is 75.65 s, with the longest video length at 155.56 s. The average duration of the action segments for the entire dataset is 0.7 s. The length of the shortest and longest action segments are 0.02 and 4.98 s, respectively. Figure 7 speaks to the variance of action instance duration by presenting statistics on the standard deviation from the average duration per action class. When one considers the *Courus* action, it is clear that there is a high variance in the durations of this particular action's instances. The reason for this is that the *Courus* action consists of repetitive series of tiny, fast steps on the tips of the toes which, in some solos, are performed for a long duration and in others a concise duration. Further statistics on the complete dataset include an average of 24.13 action annotations per video and an average of 2237.82 action annotations per class. Table 1 compares *AnnChor* with other related datasets. Our dataset is distinguished from other datasets such as MultiSports [19] and Charades [28] with longer untrimmed video clips (75.65 s vs 20.9 s vs 30 s). Notably, the average number of action instances per video for *AnnChor* (24.13) is higher when compared to MultiSports (11.8), Charades (6.8), FineAction (6.17), Thumos14 (15.29), ActivityNet (1.15) and HACS Segment (2.47). Therefore, *AnnChor* is more densely annotated compared to many recent action localization datasets. Furthermore, the average action annotation duration for our dataset of 0.7 s is smaller when compared to MultiSports (1.0 s), FineGym (1.7 s) and FineAction (7.1 s) among others [19,20,26]. It requires tremendous effort to annotate such a large number of highly fine-grained actions. At the same time, it also presents a more significant challenge to the temporal action

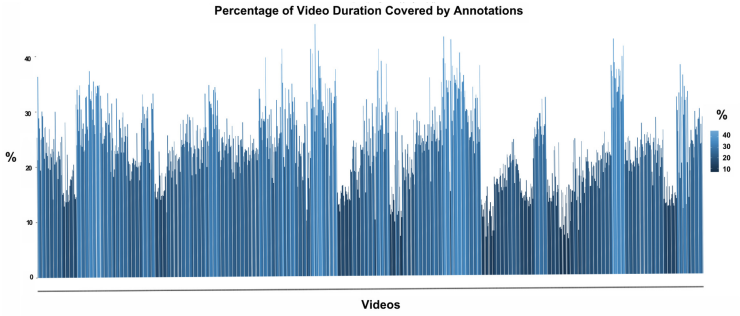


Fig. 8. *AnnChor1000* statistics showing the average percentage of each video which has been covered by annotations

Table 1. *AnnChor* dataset statistics compared to other related datasets: AC (AnnChor), MS (MultiSports), FG (FineGyn), D48 (Diving28), C (Charades), FA (FineAction), TH (Thumos14), AN (ActivityNet), HS (HACSSegment)

Statistic	AC	MS	FG	D48	C	FA	TH	AN	HS
Avg. video duration	75.65 s	20.9 s	-	-	30 s	-	-	-	**156 s**
Avg. action instances per video	**24.13**	11.8	-	-	6.8	6.17	15.29	1.15	2.47
Total action instances	25600	37701	32697	18404	66500	103324	6364	23064	**122304**
Avg. annotation duration	**0.7 s**	1.0 s	1.7 s	-	12.8 s	7.1 s	4.3 s	49.2 s	33.2 s

localization domain. The overall percentage of video durations which are covered by annotations in the *AnnChor1000* dataset is 22.66%. Figure 8 highlights the coverage of annotations as percentages of each video's duration.

3.4 Evaluation Protocol

We report performance as Mean Average Precision (mAP) over all the classes as is prevalent for the temporal action localization task [12,35]. For evaluating the Coarse-Fine model implementation, we use the same evaluation procedure that the authors of [13] used to perform their evaluation on the Charades [28] dataset. The evaluation code provided by [28] evaluates mAP on 25 equally spaced frames throughout the video, where each frame of a video can have zero or more actions. The results are presented as per-class average precision (AP) scores and the overall mAP score.

For the TriDet model, evaluation results are reported as mean average precision (mAP) at different thresholds for the intersection over union (IoU). Following the authors of [27]'s approach with evaluating on the THUMOS14 dataset, results are reported at IoU thresholds of 0.3, 0.4, 0.5, 0.6 and 0.7 along with the average mAP.

4 Baseline Experiments

4.1 Methods

An Nvidia GTX compute system with 4 A100 GPUs was used to conduct the baseline experiments for this study. One of the methods chosen as a baseline for this study is based on an implementation of Coarse-Fine networks [13] which is licensed under an MIT license. This model provides appropriate mechanisms for temporal action detection for video understanding. The Coarse-Fine Networks architecture utilizes two temporal resolutions. The coarse and fine streams are fused during a step the authors call *Multi-stage Fusion*. The Coarse-Fine networks method has achieved excellent results on the Charades dataset. However, the Charades dataset contains a much larger number of videos than our *AnnChor* dataset, and each video in Charades has a shorter temporal length. We follow the same approach that the authors in [13] used to train on the MultiTHUMOS dataset. We create a segmented version of the *AnnChor* dataset and limit the maximum temporal length of each video clip to 1280 frames. A two-stage process is followed for training, and both the Coarse and the Fine streams are initialized with the X3D backbone pre trained on Kinetics400 [15]. We use the same hyper-parameter configuration and training schedule as used in the implementation by [13].

Another method for benchmarking this dataset is the TriDet model [27], which relies on codebases that use MIT and Apache licenses. TriDet is based on a one-stage convolutional action localization approach that uses a novel Trident-head to model action boundaries to improve existing approaches for boundary prediction. The TriDet model has achieved state-of-the-art performance across benchmarks such as THUMOS14, HACS, and EPIC-KITCHEN 100. We follow a similar approach and configuration as the authors in [27] to benchmark on the THUMOS14 dataset. We kept our videos at their original length for the baseline experiment using the TriDet model and used extracted i3D features. More details on the implemented experiments for each of the models are provided in the following section.

4.2 AnnChor Benchmark

As mentioned above, in preparation for the implementation of the Coarse-Fine model, we create a segmented version of our *AnnChor* dataset which is similar to the approach taken by the authors in [13] to train on the *MultiTHUMOS dataset*. The sub-dataset's original 260 videos were segmented to produce 532 video clips and the complete dataset's original 1020 videos were segmented into 2004 video clips to align better with the video durations best suited to the Coarse-Fine model. We manually split the segmented dataset into 80% training and 20% testing sets. For the segmented sub-dataset 426 video clips were used for training and 106 were used for testing. For the segmented full dataset 1605 video clips were used for training and 399 for testing. For the Coarse-Fine implementation, all the videos were converted to a constant frame rate of 24 fps.

The TriDet model implementation required that we extract I3D features for all the videos in the dataset. All videos were converted to a constant frame-rate of 25 fps. Optical flow frames were obtained using the Recurrent All-Pairs Field Transforms (RAFT) architecture and subsequently, the I3D features were extracted. The number of videos used were 816 for training and 204 for testing. The TriDet paper [27] reports that the localization performance improves as the number of pyramid layers of the model increase. Their model achieves the best results on THUMOS14 with a pyramid level number of 6 and a bin number value of 16. We kept our parameters consistent with this approach and provided a maximum sequence length parameter of 4608 for our longer video samples.

5 Results and Analysis

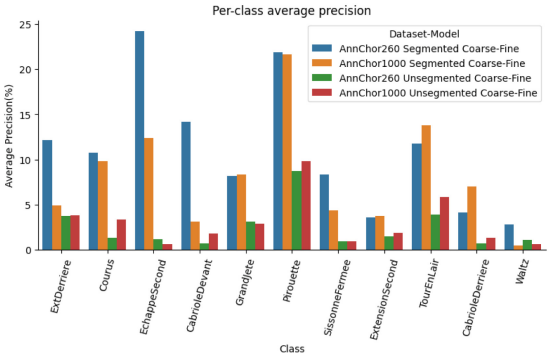


Fig. 9. Per-class average precision on the *AnnChor* dataset with Coarse-Fine models

Table 2. mAP Results

Dataset version	Model	mAP(%)
AnnChor260	Coarse-Fine Network	2.45
AnnChor1000	Coarse-Fine Network	2.99
AnnChor260 segmented	Coarse-Fine Network	11.09
AnnChor1000 segmented	Coarse-Fine Network	8.15
AnnChor 260	TriDet	2.81
AnnChor1000	TriDet	1.49

The baseline results that were obtained for the *AnnChor* dataset are shown in Fig. 9 along with Tables 2 and 3. The best-performing base-line result is for the segmented *AnnChor260* dataset with the Coarse-Fine Network model achieving

an mAP of 11.09%. This suggests that the *AnnChor* dataset and it's classical ballet context is a challenge for temporal action localization models.

During the *AnnChor* annotation process, we identified two action classes that could be problematic due to the varying approaches that dancers take in executing these movements, namely *Pirouette* and *Tour en l'air*. These classes involve turning actions where the start and end windows are nuanced. Figure 7 shows that these actions have varied segment lengths. However, as seen in Fig. 6 the *Pirouette* class had many more annotated instances which may account for the higher overall AP seen for this class in Fig. 9. The *Tour en l'air* class had fewer annotated instances and a lower AP score. Another challenging class was the *Waltz step* class which had the lowest number of annotated instances as well as a variance in the duration of segments as indicated by Figs. 6 and 7, resulting in an overall low AP score seen in Fig. 9.

Table 3. TriDet detailed evaluation results: mAP % at t-IoU values for the AnnChor260 (AC260) and AnnChor1000 (AC1000) datasets.

t-IoU	0.30	0.40	0.50	0.60	0.70	Avg
AC260 mAP	4.08	3.51	2.82	2.28	1.38	2.81
AC1000 mAP	2.54	1.81	1.36	1.04	0.72	1.49

Figure 9 presents generally higher average precision scores for the segmented *AnnChor260* sub-dataset for most classes except for the *Grand Jeté*, *Extension Second*, *Tour en l'air* and *Cabriolé Derriére* classes. This is an interesting result which shows that ballet is a complicated application domain, where more data may not guarantee better results for all action instances. Figure 9 points to the fact that there are many nuances in the data where different dancers and different versions of the same action introduce complexity for the temporal action localization task. The results point to an interesting case for analysis of similarity across action instances in this dataset.

The mAP results for the Coarse-Fine and TriDet models are presented in Tables 2 and 3. The results suggest that fine-grained temporal action localization is still difficult for video understanding models. The authors of FineGym [26] confirmed the difficulty of the temporal action localization task by reporting an mAP % of 9.6 for their more fine-grained sub-action classes. Similarly, the FineAction [20] authors reported mAP % values ranging between 5.31 and 9.25 for various state-of-the-art temporal action localization methods, with their own baseline method achieving an mAP % of 13.17. These results and current literature show that highly fine-grained actions are particularly challenging for action localization models. There is room for more comprehensive studies on understanding the boundaries of fine-grained actions.

6 Conclusion

We propose *AnnChor*, a dataset for the ballet domain, which provides fine-grained annotations for commonly occurring ballet actions. *AnnChor* is a novel and unique dataset with long-form videos filled with annotations for fine-grained actions. We provide a baseline study using Coarse-Fine Network and TriDet models. Our baseline results indicate that ballet is an exciting application field for the temporal action localization task which remains challenging, especially within fine-grained action contexts. We hope *AnnChor* provides the computer vision community with opportunities for further research and benchmarking in video action understanding.

Acknowledgements. This work is based on the research supported wholly/in part by the National Research Foundation of South Africa (Grant Number 138301).

References

1. Barioni, R.R., Costa, W., Aleluia, A., Teichrieb, V.: BalletVR: a virtual reality system for ballet arm positions training. In: 2019 21st Symposium on Virtual and Augmented Reality (SVR), pp. 10–16. IEEE (2019)
2. Bisig, D.: Generative dance - a taxonomy and survey. In: Proceedings of the 8th International Conference on Movement and Computing. MOCO 2022, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3537972.3537978
3. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–970 (2015)
4. Castro, D., et al.: Let's dance: learning from online dance videos. arXiv preprint arXiv:1801.07388 (2018)
5. Choi, E., Kim, N.Y.: Whole ballet education: exploring direct and indirect teaching methods. Res. Dance Educ. **16**(2), 142–160 (2015)
6. of Dance, R.A.: Benesh international: Benesh movement notation: Royal academy of dance (2022). https://www.royalacademyofdance.org/benesh-international-benesh-movement-notation/
7. Dutta, A., Gupta, A., Zissermann, A.: VGG image annotator (VIA) (2016). http://www.robots.ox.ac.uk/~vgg/software/via/
8. Dutta, A., Zisserman, A.: The via annotation software for images, audio and video. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2276–2279 (2019)
9. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
10. Gleason, J., Castillo, C.D., Chellappa, R.: Real-time detection of activities in untrimmed videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, pp. 117–125 (2020)

11. Gleason, J., Schwarcz, S., Ranjan, R., Castillo, C.D., Chen, J.C., Chellappa, R.: Activity detection in untrimmed videos using chunk-based classifiers. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, pp. 107–116 (2020)
12. Hutchinson, M.S., Gadepally, V.N.: Video action understanding. IEEE Access **9**, 134611–134637 (2021)
13. Kahatapitiya, K., Ryoo, M.S.: Coarse-fine networks for temporal activity detection in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8385–8394 (2021)
14. Kaspersen, E.T., Górny, D., Erkut, C., Palamas, G.: Generative choreographies: the performance dramaturgy of the machine. In: VISIGRAPP (1: GRAPP), pp. 319–326 (2020)
15. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
16. Kyan, M., et al.: An approach to ballet dance training through MS Kinect and visualization in a cave virtual reality environment. ACM Trans. Intell. Syst. Technol. (TIST) **6**(2), 1–37 (2015)
17. LaViers, A., Chen, Y., Belta, C., Egerstedt, M.: Automatic sequencing of ballet poses. IEEE Robot. Autom. Mag. **18**(3), 87–95 (2011). https://doi.org/10.1109/MRA.2011.942118
18. Li, Y., Li, Y., Vasconcelos, N.: Resound: towards action recognition without representation bias. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 513–528 (2018)
19. Li, Y., Chen, L., He, R., Wang, Z., Wu, G., Wang, L.: Multisports: a multi-person video dataset of spatio-temporally localized sports actions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13536–13545 (2021)
20. Liu, Y., Wang, L., Wang, Y., Ma, X., Qiao, Y.: FineAction: a fine-grained video dataset for temporal action localization. IEEE Trans. Image Process. **31**, 6937–6950 (2022)
21. Morris, G.: Artistry or mere technique? The value of the ballet competition. Res. Dance Educ. **9**(1), 39–54 (2008)
22. Neagle, R., Ng, K., Ruddle, R.: Developing a virtual ballet dancer to visualise choreography. In: Proceedings of the Symposium on Language, Speech and Gesture for Expressive Characters (2004)
23. Park, S.H., et al.: A feasibility study of ballet education using measurement and analysis on partial features of still scenes. Int. J. Distrib. Sens. Netw. **12**(12), 1550147716681794 (2016)
24. Paskevska, A.: Ballet: From the First Plié to Mastery, an Eight-Year Course (2002)
25. Ryman, R., of Dance Enterprises, R.A., of Dance Enterprises Staff, R.A., Chivers, F.: The Foundations of Classical Ballet Technique. Royal Academy of Dancing (2008). https://books.google.co.za/books?id=yGSKoAEACAAJ
26. Shao, D., Zhao, Y., Dai, B., Lin, D.: FineGym: a hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2616–2625 (2020)
27. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: TriDet: temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18857–18866 (2023)
28. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: crowdsourcing data collection for activity understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 510–526. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_31

29. Soomro, K., Zamir, A.R.: Action recognition in realistic sports videos. In: Moeslund, T.B., Thomas, G., Hilton, A. (eds.) Computer Vision in Sports. ACVPR, pp. 181–208. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09396-3_9
30. Sun, J.: A digital feature recognition technology used in ballet training action correction. Comput. Intell. Neurosci. **2022** (2022)
31. Trajkova, M., Cafaro, F.: E-ballet: designing for remote ballet learning. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, pp. 213–216 (2016)
32. Trajkova, M., Cafaro, F.: Takes tutu to ballet: designing visual and verbal feedback for augmented mirrors. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(1) (2018). https://doi.org/10.1145/3191770
33. Vishwakarma, D.K., Kapoor, R.: Hybrid classifier based human activity recognition using the silhouette and cells. Expert Syst. Appl. **42**(20), 6957–6965 (2015)
34. Vishwakarma, D.K., Singh, T.: A visual cognizance based multi-resolution descriptor for human action recognition using key pose. AEU-Int. J. Electron. Commun. **107**, 157–169 (2019)
35. Xia, H., Zhan, Y.: A survey on temporal action localization. IEEE Access **8**, 70477–70487 (2020)
36. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: FineDiving: a fine-grained dataset for procedure-aware action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2949–2958 (2022)
37. Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Fei-Fei, L.: Every moment counts: dense detailed labeling of actions in complex videos. Int. J. Comput. Visi. **126**(2), 375–389 (2018)
38. Yu, T., Johnson, P.: Tour jeté, pirouette: dance choreographing by computers. Lecture notes in computer science, pp. 156–157 (2003)
39. Zhao, H., Torralba, A., Torresani, L., Yan, Z.: HACS: human action clips and segments dataset for recognition and temporal localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8668–8678 (2019)

Privacy-Preserving Face Recognition with Adaptive Generative Perturbations

Delong Zhang[1], Yixing Peng[1], Ancong Wu[1(✉)], and Wei-shi Zheng[1,2,3]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
wuanc@mail.sysu.edu.cn
[2] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou, China
[3] Guangdong Province Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou, China

Abstract. For online face recognition services, the potential leakage of facial features and reconstruction techniques gives malicious attackers the opportunity to reconstruct face images, raising public concern about privacy. Previous privacy-preserving face recognition methods either require retraining the face recognition model or iterative perturbation that increases inference time. To overcome these limitations, we propose an efficient plug-and-play method that crafts **A**daptive **G**enerative perturbations for frozen **Face** recognition model (AGFace) to defend the reconstruction attacks. To generate perturbation with a single forward for shallow features extracted by frozen recognition model, we learn a Perturbation Generator to efficiently mine the adversarial perturbations by simultaneously minimizing the perturbation and maximizing the reconstruction error of a proxy reconstructor. To achieve privacy-utility trade-off, we propose the Adaptive Channel Selector to identify top-k reconstruction-sensitive channels for the features. By selecting these channels for perturbation, the dominant visual privacy information is protected with reconstruction-insensitive discriminative information preserved. Extensive experiments demonstrate that AGFace achieves state-of-the-art performance in terms of both privacy and utility among retraining-free methods and is comparable to retraining-dependent methods.

Keywords: Face Recognition · Reconstruction Attacks · Facial Feature Protection

1 Introduction

With the rapid development of deep learning, face recognition has been widely deployed in real-world applications with impressive recognition accuracy. Due to local computational resource constraints, dominant face recognition (FR) systems usually employ client-server mode: users of local clients send their face images to a cloud server for efficient online face recognition services. However, the potential data leakage [5] raises public concern about privacy.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 210–227, 2025.
https://doi.org/10.1007/978-3-031-78341-8_14

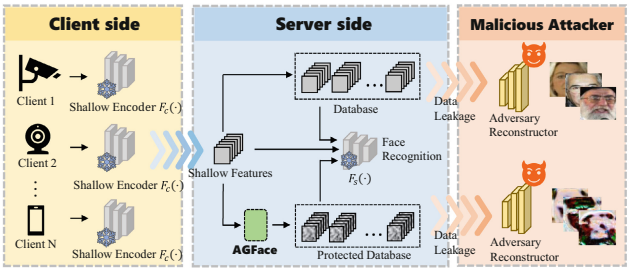


Fig. 1. Illustration of the client-server face recognition system and the potential malicious attacker with reconstruction attacks. Without retraining the face recognition (FR) model, our AGFace can be incorporated into the deployed face recognition system to protect shallow features to resist reconstruction attacks and maintain online face recognition accuracy at the same time.

To prevent privacy leakage, it is prevailing to upload face features instead of raw images to the cloud server. Specifically, the face recognition network can be divided into two parts. The first few layers, called Shallow Encoder, are installed on the client side to convert the users' face images into face features to compress the visual information. The remaining deep layers of the network are kept on the server. In this way, the visually indistinguishable shallow features extracted by the Shallow Encoder are transmitted to the server for subsequent online face recognition. However, it has been shown that attackers could use some recovery techniques to exploit eavesdropped features to reconstruct original images [6,15,30], exposing the privacy of users.

An expected privacy-preserving face recognition method should satisfy the three properties. 1) **Strong privacy protection**: The method should protect the features from reconstruction attacks. 2) **High task performance**: The method should not severely sacrifice the recognition accuracy. 3) **Low cost of training and inference**: The method should not bring significant overhead for training and inference. However, existing privacy-preserving methods are struggling to meet these properties. Encryption-based methods encrypt face images [1,11,14] and thus require high latency and expensive computation costs. Perturbation-based methods [13,16,34] transform face images or features into the privacy space through noisy perturbation. However, the privacy transformation compromises task performance, limiting their applications. Frequency-domain transformation [10,17,18,29] achieves a good trade-off between privacy and utility. Nevertheless, they require retraining the FR model and hence are maintaining considerable training overhead. Recently, the most related work [30] proposes a plug-and-play method that addresses the first three objectives. Unfortunately, it generates adversarial features in an iterative manner over several dozen steps, which leads to significantly slower inference.

To overcome the above challenges, we propose a novel method to craft ***A****daptive* ***G****enerative perturbations* for frozen **Face** recognition model (AGFace) to efficiently build adversarial features, which simultaneously resist reconstruc-

tion attacks and maintain face recognition performance. Inspired by [9], adversarial samples are features rather than bugs, we intend to mine adversarial perturbations from shallow features to resist reconstruction attacks. To this end, we first construct a proxy reconstructor to characterize a strong adversary model that maps the facial features to the original images. With the assistance of the proxy reconstructor, we then train a Perturbation Generator to produce adversarial perturbations by disturbing shallow features that cannot be effectively reconstructed to the original images. To maintain face recognition performance, we also regularize perturbations to keep it smooth and subtle. Since the Perturbation Generator produces perturbations by a single forward, our method is efficient and introduces negligible additional inference time.

To further refine generative perturbations for a better trade-off between privacy and utility, we propose the Adaptive Channel Selector to select reconstruction-sensitive channels for the adaptive addition of perturbations. In particular, with the prior knowledge regarding reconstruction mapping of the proxy reconstructor, the Adaptive Channel Selector learns a binary mask to find channels that are more sensitive to reconstruction. Afterward in the inference phase, the trained Adaptive Channel Selector is combined with the Perturbation Generator to craft adaptive generative perturbations. Specifically, the perturbations are added only to the reconstruction-sensitive channels of shallow features, while the other channels remain unprocessed. As a result, we obtain effective adversarial features by adaptively disturbing critical reconstruction-sensitive channels while maintaining discriminative information of shallow features, thereby achieving better performance on both accuracy and privacy.

To summarize, our contribution is three-fold:

1. We propose a novel method (namely AGFace) to efficiently craft adaptive generative perturbations, which protect feature privacy without causing severe deviation from the original features. Hence, AGFace can be incorporated into deployed FR systems without retraining the FR model for privacy-preserving face recognition.
2. For a better trade-off between privacy and utility, we develop an Adaptive Channel Selector to identify the sensitive channels that contain critical visual privacy.
3. Extensive experiments demonstrate that AGFace achieves state-of-the-art performance in terms of both privacy and utility among retraining-free methods and is comparable to retraining-dependent methods regarding recognition accuracy. Moreover, compared to FR baselines, AGFace brings a negligible drop in accuracy, along with small increases in inference time and parameter costs.

2 Related Work

2.1 Face Privacy Protection

We categorize existing face privacy protection methods into encryption-based methods, perturbation-based methods, frequency domain-based methods, and adversarial feature-based methods. **Encryption-based methods** [1,11,14] first

encrypt face images and perform face recognition model on the encrypted data, both in the training and testing phases. However, these methods face high latency and expensive computation costs. **Perturbation-based methods** [13,16,34,37] distort the images or features by adding noise to make them difficult to recognize or revert to the original image. Unfortunately, arbitrarily adding noise severely decreases the accuracy of face recognition. **Frequency domain-based methods** [10,17–19,29] transform RGB images into frequency-domain data and discard the DC component, which is particularly important for visual information. However, such methods need retraining the FR model to preserve satisfactory performance, thus incurring significant training overhead.

As the most related ones, **adversarial feature-based methods** [12,27,32] aim at generating adversarial features that thwart reconstruction attacks. These methods employ adversarial training to minimize the classification loss of the classifier while maximizing the reconstruction loss of the reconstruction network. However, since these methods are designed for attribute prediction, they cannot achieve satisfactory performance on face recognition. AdvFace [30] builds a shadow model to generate adversarial noise, which is added to the clean features to prevent features from restriction attacks. However, their adversarial noise from the shadow model requires iterative generation, significantly slowing down inference. Differently, our approach crafts adaptive generative perturbations with only single forward to generate, thus enabling significantly efficient inference and achieving better privacy-utility trade-off.

2.2 Reconstructing Face Images from Features

The optimization-based reconstruction attack methods [6,22] continue an optimization process until the output of the feature extractor is effectively reconstructed to resemble the original facial image. However, these interactive methods result in substantial computational expenses. Another effective and cost-saving way is learning-based approaches [4,15,21,26,30,36], which build a reconstruction network to learn the inverse mapping from the features to the original images. Dosovitskiy et al. [4] and Zhmoginov et al.. [36] aim to minimize the feature distance between the reconstructed images and the original images. Mai et al. [15] proposed a neighborly de-convolutional network to recover face images from corresponding features based on the assumption without knowledge of target networks. Similarly, Wang et al.. [30] established reconstruction networks based on three different architectures to carry out reconstruction attacks. Dosovitskiy et al. [21,26] proposed to learn a mapping from facial templates to the intermediate latent space of a pre-trained within a generative adversarial network (GAN)-based framework.

3 Adaptive Generative Perturbations

3.1 Threat Model Setup

In this work, The server provider is supposed to be **trustworthy** and aware of the architecture and weight of the face recognition model. As shown in Fig. 1,

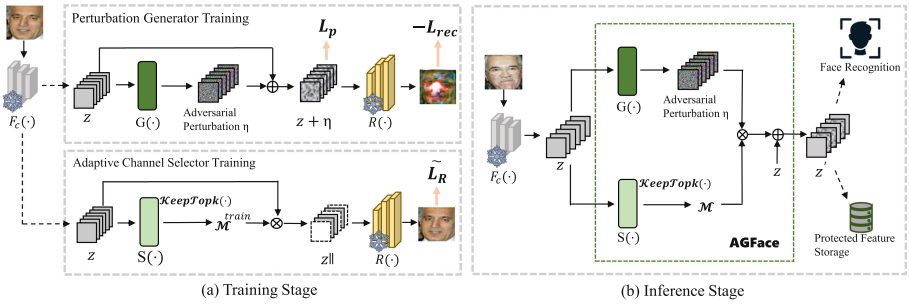


Fig. 2. Pipeline of AGFace. The Perturbations Generator $G(\cdot)$ and the Adaptive Channel Selector together constitute AGFace which is incorporated into the deployed face recognition system to protect facial features without retraining the FR model. Note that $R(\cdot)$ is a proxy reconstructor that is pre-trained and fixed to assist the training of AGFace.

the server provider divides the pre-trained face recognition network into two parts: Shallow Encoder $F_c(\cdot)$ on the resource-constraint client while remaining deep layers of the network $F_s(\cdot)$ on the server. Unfortunately, the database of face features still may be leaked to powerful attackers [5].

Following previous works [15,30], we consider that the powerful attacker has access to the black-box shallow Encoder $F_c(\cdot)$ and steals the features stored in the database. An attacker can launch reconstruction attacks by training a reconstructor $R(\cdot)$ that learns the mapping from features to face images. Suppose the attacker has face images from public datasets denoted as $X = \{x_1, \ldots, x_N\}$, and then the attacker can query $F_c(\cdot)$ to obtain facial shallow features $Z = \{z_1, \ldots, z_N\}$, where $z_i = F_c(x_i)$ and N is the number of facial images from the public dataset. After feeding z_i to $R(\cdot)$ to obtain reconstruction images $R(z_i)$, the attacker can train $R(\cdot)$ by optimizing: The attacker trains $R(\cdot)$ to reconstruct images from features z_i by optimizing:

$$\mathcal{L}_R = \sum_{i=1}^{N} \|x_i - R(z_i)\|_1 , \tag{1}$$

where $\|\cdot\|_1$ is the L_1-norm distance measuring the difference between reconstruction images and original images.

3.2 Overview of AGFace

Our goal is to mine adversarial perturbation to resist reconstruction attacks while maintaining the discriminative information of shallow features. Figure 2 shows the pipeline of AGFace, which consists of three stages: 1) In the Perturbations Generator Training stage, we aim to train the Perturbations Generator $G(\cdot)$ to generate perturbations to defend against reconstruction attack; 2) In the Adaptive Channel Selector Training stage, we train the Adaptive Channel

Selector $S(\cdot)$ to identify top-k channels that are sensitive to reconstruction; 3) In the Inference stage, we incorporate AGFace (consist of trained $G(\cdot)$ and $S(\cdot)$) into the deployed face recognition system to protect shallow features.

3.3 Perturbation Generation

The intuition of our method is to disrupt this mapping by adversarial perturbations. Inspired by [9] that adversarial samples are features rather than bugs, we intend to generate adversarial perturbations from shallow features. Then such perturbations are fused with shallow features to disrupt the mapping so as to resist reconstruction attacks.

Specifically, we initialize a shallow network as the Perturbations Generator $G(\cdot)$ to generate perturbation η from the shallow face feature z:

$$\eta = G(z), \tag{2}$$

where η has the same shape with z and each element in η is bounded in $[0, 1]$. Then z is fused with η to obtain adversarial features $z + \eta$.

To learn effective adversarial features, we feed $z + \eta$ to $R(\cdot)$ to obtain the reconstructed image $R(z + \eta)$, where $R(\cdot)$ is a fixed proxy reconstructor pre-trained by Eq. (1) to simulate a strong adversary. Since the proxy reconstructor maps shallow features to the original image, we aim to optimize η against the proxy reconstructor $R(\cdot)$ to maximize the distance between reconstructed images and original images:

$$\max_{\theta_G} \mathcal{L}_{\text{rec}} = \sum_{i=1}^{N} \|R(z_i + \eta_i) - x_i\|_1, \tag{3}$$

where θ_G is the parameter of $G(\cdot)$. During minimizing the $\mathcal{L}_{\text{rec}}$ and updating the parameters of $G(\cdot)$, the perturbation η is gradually optimized to become resistant to reconstruction attacks.

Moreover, in order to guarantee that perturbations have minimal impact on the original shadow features and distribute smoothly over all pixels, we constrain the L_2 norm of η by computing the perturbation loss:

$$\min_{\theta_G} \mathcal{L}_{\text{p}} = \sum_{i=1}^{N} \|\eta_i\|_2. \tag{4}$$

Overall, our training scheme simultaneously maximizes the reconstruction loss and constrains the L_2-norm of perturbations to maintain face recognition accuracy. Thus, the Perturbations Generator $G(\cdot)$ is optimized by:

$$\min_{\theta_G} \mathcal{L}_{\text{adv}} = -\mathcal{L}_{\text{rec}} + \lambda_p \mathcal{L}_{\text{p}}, \tag{5}$$

where λ_p is a trade-off parameter. After optimization, the privacy-enhanced adversarial feature $Z + \eta$ can perturb the reconstruction mapping from features to facial images and avoid severe deviation from the original features.

3.4 Adaptive Channel Selection

There is a dilemma that maintaining the recognition accuracy requires the perturbations to disturb the original features less while a large disturbance to the original features may contribute to protecting feature privacy. To better refine the perturbations for balancing the privacy protection ability and recognition accuracy, we further propose the Adaptive Channel Selector to identify top-k reconstruction-sensitive channels in the feature z. The design of the Adaptive Channel Selector starts from a simple insight that not all channels contribute equally to reconstruction. Thus, we only add perturbations to the sensitive channels to protect dominant privacy information while mitigating the server disturbance to the original features.

We propose the Adaptive Channel Selector $S(\cdot)$ to get a channel selective mask $\mathcal{M} \in \{0,1\}^{\mathcal{C}}$ ($\mathcal{C}$ are the number of channels of the shallow features) as a discrete vector that controls the selection of reconstruction-sensitive channels:

$$\mathcal{M} = \mathcal{K}eep\mathcal{T}opk(S(z)), \tag{6}$$

where $S(z) \in [0,1]^{\mathcal{C}}$ is the output of Channel Selector and $\mathcal{K}eep\mathcal{T}opk(\cdot)$ is a function where the top-k channels are retained in M while the other channels are dropped during the forward propagation. Formally, the c-th element in M is expressed as:

$$\mathcal{M}_c = \begin{cases} 1, & S(z)_c \in \mathcal{T}opk(S(z)) \\ 0, & S(z)_c \notin \mathcal{T}opk(S(z)) \end{cases}, \tag{7}$$

where $\mathcal{T}opk(.)$ are top-k reconstruction-sensitive channels with highest value and $S(z)_c$ is the c-th element in $S(z)$.

However, such a discretization operator will cause gradient vanishing if used in training (the derivative of all differentiable points is 0). To solve this, we adopt Gumbel Softmax to explore sensitive channels over all possible channels. Specially, from Gumbel distribution with $U \sim \text{Uniform}(0,1)$, we sample a random variable $\mathcal{N} = -\log(-\log(U))$ at each channel. We then pass $S(z) + \mathcal{N}$ through the softmax output to obtain the probabilities of a channel to be chosen:

$$p_c = \frac{\exp\left(\left(\log\left(S(z) + \mathcal{N}\right)\right)/\tau\right)}{\sum_{c=1}^{\mathcal{C}} \exp\left(\log\left(S(z) + \mathcal{N}\right)/\tau\right)}, \tag{8}$$

where τ is a temperature parameter. Then we can modify Eq. (7) as:

$$\mathcal{M}_c^{\text{train}} = \begin{cases} \mathcal{K}eep\mathcal{T}opk\left(p_c\right), & \text{in forward} \\ p_c, & \text{in backward} \end{cases} \tag{9}$$

where the difference between the forward and backward propagation ensures the differentiability. To mine the channels that are important for reconstruction, we modify the feature by masking out the channels that are unimportant for reconstruction according to the sensitive mask $\mathcal{M}^{\text{train}}$ and force the modified features being reconstructed to the original images by fixed $R(\cdot)$:

$$\tilde{z} = z \odot \mathcal{M}^{\text{train}}, \tag{10}$$

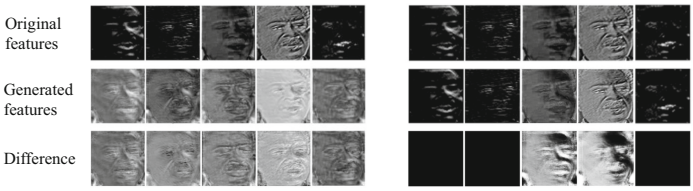


Fig. 3. The visualization of randomly chosen channels of facial shallow features from different methods. We adapt [27] as CPGAN-FR to generate adversarial features without retraining the FR model. (a) The generated features are predominant in CPGAN-FR; (b) The original features play a dominant role in our method.

$$\min_{\theta_S} \tilde{\mathcal{L}}_R = \sum_{i=1}^{N} \left\| R(\tilde{z}_i) - x_i \right\|_1 , \tag{11}$$

where θ_S is the parameter of $S(\cdot)$. If the original image can be recovered from the partially selected channel of features, it suggests that the Adaptive Channel selector has learned knowledge about seeking reconstruction-sensitive channels. Note that none of the above training processes change the FR model.

3.5 Inference

During inference, as a plug-and-play module, the AGFace (consists of $G(\cdot)$ and $S(\cdot)$) can be incorporated into the face recognition system to perform privacy enhancement to original shallow features with slightly modified. We freeze the Channel Selector $S(\cdot)$ to keep its' knowledge about seeking reconstruction-sensitive channels. Through Eq. (7), we can obtain the sensitive channel mask $\mathcal{M}$ to get final privacy-enhanced adversarial features with a controllable number of sensitive channels:

$$z' = z + \eta \odot \mathcal{M}. \tag{12}$$

The privacy-enhanced adversarial features z' rather than the original features z are stored in a database for subsequent online face recognition. In this way, even if the privacy-enhanced features are stolen by adversaries, they are still incapable of reconstructing them to original images.

Discussion. Recall that the intuition to resist reconstruction attacks is to build the adversarial features. A simple solution is to modify similar methods such as [12,27] to directly generate adversarial features by training a generator. However, directly generating adversarial features without retraining the FR model would corrupt the original feature distribution, which results in a drastic drop in recognition accuracy. For example, we adapt [27] as CPGAN-FR (See the Appendix for more details) to generate adversarial features, which leads to a significant accuracy drop (*e.g.*, 13.9% in CFP-FP dataset). Here we give a visualization and analysis to discuss the reason behind this phenomenon. As shown in the last row of Fig. 3(a), the difference between the original feature and the

generated feature is more similar to the generated feature, which suggests that the generated feature is dominant and nearly obliterates the expression of the original feature.

Differently, instead of directly generating adversarial features, we use $G(\cdot)$ and $S(\cdot)$ to craft adaptive generative perturbations and fuse them with the shallow features to obtain adversarial features. On the one hand, it maintains the discriminative information in shallow features so as to maintain accuracy. On the other hand, the critical top-k reconstruction-sensitive channels of shallow features are disturbed to resist reconstruction attacks.

4 Experiments

4.1 Experimental Setup

We adopt FaceNet [23], ArcFace [3], and CosFace [28] as baseline FR models for face recognition. Unless specifically stated, we used FaceNet for our main experiments and ablation studies. For a fair comparison, we follow previous work [30] to employ FaceNet with pre-trained Inception-ResNet-v1 [7] on CASIA-WebFace [33] dataset, and the first three convolutional layers are chosen as the Shallow Encoder $F_c(\cdot)$. Besides, for ArcFace and CosFace with pre-trained Resnet50 [7] pre-trained on VGGFace [2], the first convolutional layers and the first residual block are selected as $F_c(\cdot)$. The shallow feature z extracted by $F_c(\cdot)$ has 64 channels. We conduct experiments on three widely used face recognition datasets: LFW [8], AgeDB-30 [20], CFP-FP [24]. All images are cropped with MTCNN [35] to detect faces and facial landmarks in images and resize them to 160×160.

Metrics. For the utility of face recognition, we report the accuracy of identifying whether pairs of facial features belong to the same individual. For a quantitative evaluation of the defense efficacy against reconstruction attacks, we report the average Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index [31] (SSIM) of the reconstructed images. Furthermore, we conducted experiments with replay attacks, considering that the attacker feeds reconstructed images to the face recognition system for malicious authentication. We follow [25] to report the Success Attack Rate (SAR) at different False Match Rate. SAR(1%)/SAR(0.1%) indicates SAR at FMR = 1%/0.1%, respectively. Lower PSNR, SSIM, and SAR indicate stronger defense.

AGFace. We employ TransRec [30] as the backbone of the proxy reconstructor $R(\cdot)$. Both Perturbation Generator $G(\cdot)$ and Adaptive Channel Selector $S(\cdot)$ consist of two convolutional layers and one residual [7] block except that $S(\cdot)$ has an extra pooling layer. AGFace is trained on the CASIA-WebFace dataset.

Adversary Reconstruction Network. Following [15,30], the adversary is considered to adopt ResRec as the adversary reconstruction model, which takes a ResNet [7] structure. Additionally, we also consider the attacker constructs

Table 1. Quantitative results of the reconstructed images among retraining-free privacy-preserving methods. SAR(1%)/SAR(0.1%) indicates Success Attack Rate at False Match Rate = 1%/0.1%, respectively. The adversary reconstruction network is ResRec. The face recognition model is FaceNet [23].

Method	LFW				CFP-FP				AgeDB-30			
	PSNR	SSIM	SAR(1%)	SAR(0.1%)	PSNR	SSIM	SAR(1%)	SAR(0.1%)	PSNR	SSIM	SAR(1%)	SAR(0.1%)
Unprotected	27.87	0.93	75.23	73.41	22.89	0.83	69.88	67.51	23.96	0.87	65.23	60.40
random	22.81	0.90	72.14	70.29	20.73	0.79	65.66	62.13	21.68	0.86	55.31	50.14
DP	23.12	0.90	71.96	68.54	20.89	0.79	62.39	59.44	21.86	0.86	56.43	51.97
Deep-FR	14.34	0.25	51.40	49.23	12.21	0.22	49.13	46.12	12.32	0.25	52.13	45.26
CPGAN-FR	11.32	0.42	12.12	10.51	9.34	0.39	12.04	9.98	9.67	0.43	14.28	10.31
AdvFace	6.97	0.28	3.11	3.05	5.98	0.23	9.88	8.43	5.85	0.24	8.45	6.88
AGFace	**6.33**	**0.16**	**2.04**	**1.68**	**5.27**	**0.13**	**8.36**	**7.12**	**5.14**	**0.18**	**6.40**	**5.38**

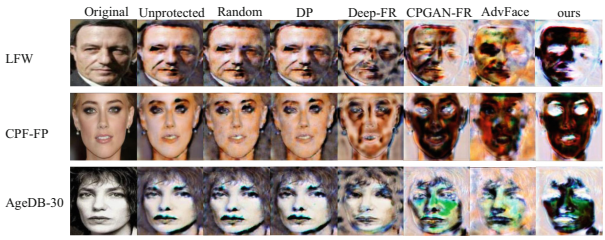


Fig. 4. Reconstructed images from different retraining-free privacy-preserving methods on three face datasets. Our AGFace hinders reconstruction attacks more effectively than the others.

three reconstruction networks (URec, NbRec-A, NbRec-B) to carry out a brute-force attack to simulate black-box attack scenarios. All reconstruction networks are trained on the CelebA dataset, which differs from the dataset used for our FR model.

Baseline Methods. According to the requirement of retraining the FR model, we categorize the baseline privacy-preserving FR methods into two groups.

For retraining-dependent methods, we choose three state-of-the-art methods for comparison: CPGAN [27], DCTDP [10], and DuetFace [17].

For retraining-free methods, we follow the same setting with [30] to implement **random** and **DP** to add noise to the features with a noise bound of 0.2. We modify two typical adversarial feature-based methods [12] [27] into **Deep-FR** and **CPGAN-FR** without re-train the FR model (refer to the appendix for more description). Moreover, **AdvFace** [30] iteratively generates adversarial noise by a shadow reconstruction network.

4.2 Resistance to Potential Attacks

Our main security goal is to defend two types of attacks on the leakage database of facial features: reconstruction attacks and replay attacks. Specifically, (1) for the **Reconstruction Attacks**, the adversary intends to recover images from leaked features, which would expose the privacy of face images; (2) for the

Replay Attacks, the adversary feeds the reconstructed facial image into the face recognition system for malicious face authentication purposes. To demonstrate the superior defense capabilities, we compare our AGFace with different retraining-free methods.

Effectiveness Against Reconstruction Attacks. Figure 4 illustrates the reconstructed images by the attacker. We can see that random (3th column) and DP (4th column) completely fail to resist the reconstruction attack, as the reconstructed images closely resemble the original ones. Meanwhile, the adversarial feature-based protection method including Deep-FR (5th column) and CPGAN (6th column) results in notable changes around the face in the reconstructed images. However, the key features of the person in the reconstructed images remain easily identifiable. In comparison, our AGFace (8th column) offers a considerably outperformed defense against reconstruction attacks and approximately 10 times faster inference than AdvFace (7th column). Table 1 provides quantitative metrics of reconstructed images on LFW, CFP-FP, AgeDB-30. We can see that the SSIM and PSNR of AGFace are lower than the other privacy-preserving methods, which means that our method achieves the best defense performance toward reconstruction attacks.

Effectiveness Against Replay Attacks. As Table 1 shown, our AGFace shows exceptional efficacy in thwarting replay attacks. Remarkably, the utilization of AGFace on unprotected features led to a substantial reduction in the SAR(1%) value, namely, 73.19%, 61.52%, and 58.83%, on LFW, CFP-FP, and AgeDB-30 datasets, respectively.

4.3 Recognition Accuracy Comparison

Table 2 shows the face recognition performance of different privacy-preserving methods. Our AGFace closely approximates the performance of the unprotected baseline FaceNet, with a negligible accuracy drop (about 0.2% $\sim$ 0.77% lower

Table 2. Comparison of the face recognition accuracy among privacy-preserving methods on three datasets. The face recognition is FaceNet and it is present as the accuracy upper bound without any privacy protection. "Re-training" indicates whether the FR model needs to be re-trained.

Method	Re-training	LFW	CFP-FP	AgeDB
CPGAN [27]	✓	97.31	89.20	86.32
DCTDP [10]	✓	97.32	91.24	87.03
DuetFace [17]	✓	98.02	84.37	87.10
random	✗	97.20	91.67	86.60
DP	✗	96.27	90.84	85.12
CPGAN-FR	✗	92.83	79.26	78.03
Deep-FR	✗	93.34	82.35	80.32
AdvFace [30]	✗	96.43	90.59	85.10
AGFace (ours)	✗	**97.93**	**92.39**	**87.12**
FaceNet (upper bound)	N/A	98.13	93.16	87.57

than FaceNet on the three datasets). Compared with **retraining-free privacy-preserving methods** (6th–11th row) which do not need retraining the FR model, our AGFace outperforms all SOTA methods, holding advantages on accuracy range from 0.6% to 15.09%. Compared to **retraining-dependent privacy-preserving methods** (2nd–5th row) which require additional training cost and deployment overhead, the performance of our AGFace is still comparable and even reaches SOTA on CFP-FP, AgeDB.

4.4 Generality of AGFace

Our method is effective and efficient to generalize to different face recognition models. In this subsection, we verify the generality of AGFace in three terms: Transferability, Compatibility, and inference overhead.

Transferability of Adversarial Features from AGFace. Since the attacker may build the reconstruction model with different networks to carry out a brute-force attack, it is necessary to study the defense effectiveness against different adversary reconstruction models. To this end, we employ adversary reconstruction models with three different backbones to simulate unseen adversaries. Figure 5 shows the reconstructed images from the adversarial features by different reconstruction networks that are remarkably indistinguishable from the original image. Table 3 provides a quantitative analysis of the average quality of reconstructed images as the SSIM and PSNR are low. According to the experimental results, it can be observed that AGFace is consistently effective against various adversary reconstruction models, suggesting the excellent transferability of its adversarial features.

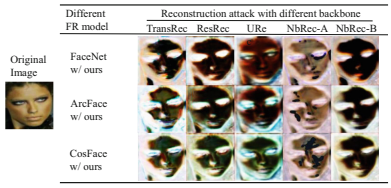


Fig. 5. Evaluation of AGFace with different FR models on LFW dataset. Our AGFace can combine with different SOTA FR models (ArcFace, CosFace) to generate adversarial features that defend against reconstruction attacks with different backbones (TransRec, ResRec, URec, NbRec-A, NbRec-B).

Table 3. Quantitative analysis of generality on LFW dataset.

Metric	Defense with	Reconstruction attack with different backbone				
	Different FR model	ResRec	TransRec	URec	NbRec-A	NbRec-B
PSNR	FaceNet w/ ours	6.33	6.21	7.63	6.93	6.72
	ArcFace w/ ours	6.16	6.04	7.24	6.57	6.43
	CosFace w/ ours	6.16	6.02	7.28	6.51	6.40
SSIM	FaceNet w/ ours	0.21	0.15	0.27	0.22	0.20
	ArcFace w/ ours	0.19	0.13	0.23	0.18	0.17
	CosFace w/ ours	0.18	0.13	0.22	0.18	0.16

Table 4. Compatability of AGFace on different FR methods. Integrating AGFace with ArcFace or CosFace incurs only a slight, acceptable decrease in face recognition performance.

Method	LFW	CFP-FP	AgeDB
ArcFace	99.60	98.32	95.88
ArcFace w/ ours	99.15	97.68	95.45
CosFace	99.63	98.52	95.83
CosFace w/ ours	99.35	97.33	95.35

Table 5. Training cost of different privacy-preserving face recognition methods on CASIA-WebFace. Our AGFace requires only lightweight training to save significant overhead.

Method	Iteration	Training time	Batch size	Memory	ACC
DCTDP	92k	173 min	256	44.52 MB	97.32
CPGAN	80K	164 min	256	42.13 MB	97.31
AGFace	1k	3 min	32	5.45 MB	97.93

Table 6. Inference cost of the plug-and-play methods compared to FR baseline. Experiments are conducted on LFW dataset.

Method	Inference time(s)	Param(Mb)
FaceNet	10.23	27.56
FaceNet w/ AdvFace [30]	113.04	27.67
FaceNet w/ AGFace(ours)	12.34	28.08
ArcFace	21.53	43.59
ArcFace w/ AdvFace [30]	173.39	43.70
ArcFace w/ AGFace(ours)	23.75	44.11

Compatibility of AGFace. A well-generalized privacy-preserving method should be compatible with different FR models. Therefore we incorporate AGFace with different SOTA FR methods to evaluate the accuracy and defense effectiveness. Table 4 shows the compatibility of AGFace which is incorporated with ArcFace and CosFace. Compared to unprotected features from ArcFace and CosFace, the adversarial features of AGFace cause a minor decrease in accuracy, specifically 0.45% and 0.43% on the LFW dataset. Such a minor decrease in accuracy is considered acceptable in light of the exceptional performance of AGFace in privacy protection as shown in Fig. 5 and Table 3.

Cost of Training and Inference. Unlike previous privacy-preserving FR methods that necessitate full training for performance assurance (see Table 6), our AGFace maintains the pre-trained FR model frozen and solely trains the lightweight networks, thereby significantly reducing training overhead.

Moreover, a widely used privacy enhancement module should incur low overhead in terms of inference speed and parameter cost. As shown in Table 5, compared to the FR baseline, AdvFace (another plug-and-play privacy-preserving method) requires approximately ten times longer for inference, whereas our AGFace introduces negligible time and parameter costs for inference. Therefore, our AGFace is resource-efficient.

4.5 Ablation Study

Hyperparameters for Trade-off. In our AGFace, both k and λ_p control the trade-off between privacy and utility. Table 7 presents an analysis of the remaining parameter choices. As k increases, more channels are disturbed with perturbations so that recognition accuracy decreases while privacy protection is enhanced (lower PSNR). λ_p controls the magnitude of the perturbation and it shows a similar trend when it decreases. We empirically search for the best combination of hyperparameters as our default setting.

Effectiveness of $S(\cdot)$. Table 8 demonstrate the ablation study of AGFace to verify the effectiveness of $S(\cdot)$. For the first row, We naively drop $S(\cdot)$, which suggests that the whole perturbations are added to shallow features, consequently leading to accuracy drops. For the second row, we replace $S(z)$ with a learnable

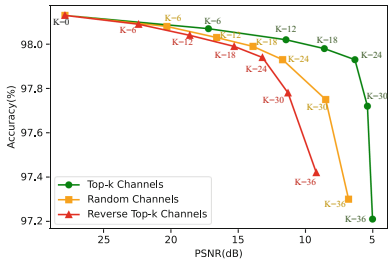


Fig. 6. Effectiveness of the Adaptive Channel Selector to identify reconstruction-sensitive channels. "Top-k Channels" indicates the original setting. "Random Channels" indicates randomly selecting k channels to add perturbations. "Reverse Top-k Channels" indicates the selection of top-k reconstruction-sensitive channels in reverse order.

Table 7. Ablation study of hyperparameters. When changing one of the parameters, we keep the other parameter unchanged.

Hyperarameter		ACC↑	PSNR↓	SSIM↓	SAR(1%)↓	SAR(0.1%)↓
k	12	98.02	11.45	0.42	14.13	12.98
	24	97.93	6.33	0.16	2.04	1.68
	36	97.21	5.23	0.13	1.89	1.57
	48	96.01	4.69	0.11	1.64	1.40
λ_p	10	97.99	13.24	0.47	16.44	15.90
	3	97.93	6.33	0.16	2.04	1.68
	1	97.34	6.21	0.14	1.86	1.54
	0.5	95.04	4.23	0.09	1.58	1.42

Table 8. Ablation study of AGFace. $G(\cdot)$: Perturbation Generator. $S(\cdot)$: Adaptive Channel Selector. "learnable vector" means that we replace $S(z)$ with a learnable vector of the same shape.

Setting	ACC↑	PSNR↓	SSIM↓	SAR(1%)↓	SAR(0.1%)↓
$G(\cdot)$	95.21	4.52	0.10	1.23	1.04
$G(\cdot)$ + learnable vector	96.83	7.34	0.34	8.34	7.20
$G(\cdot) + S(\cdot)$ (full)	97.93	6.33	0.16	2.04	1.68

vector which has a consistent shape with $S(z)$ in Eq. (8). We can see that it hurts both the accuracy and privacy compared with our full model (4th row), which suggests that the information in shallow features z is necessary for the adaptive channel selection.

To further evaluate whether our proposed Adaptive Channel Selector could effectively identify reconstruction-sensitive channels, Fig. 6 shows the experiment on disturbing different channels according to three different strategies. The upper right position of Fig. 6 showcases a superior balance, suggesting a strong performance both on utility and privacy. "Top-k channel" curve falls in the upper right of both the curve of "Random Channel" and "Reverse Top-k Channel", indicating Adaptive Channel Selector identifies meaningful channels to resolve the dilemma between utility and privacy.

4.6 Robustness Against Adaptive Attacks

Here we provide the robustness of our method against adaptive attacks. We assume that the adversary is aware of our method and tries to launch an adaptive attack. Since perturbations η are continuous variables, it is difficult to predict them. Therefore we assume that the adversary tries to find clean channels, which are not affected by perturbations, in order to reconstruct images.

Since the added noise is subtle, it is challenging for the adversary to confirm the number and location of the clean channels. Therefore we consider that the adversary adopts three strategies to predict clean channels and brutally launches attacks based on different numbers n of channels. The three strategies for predicting clean channels are as follows: **Random Channel**: Randomly selecting chan-

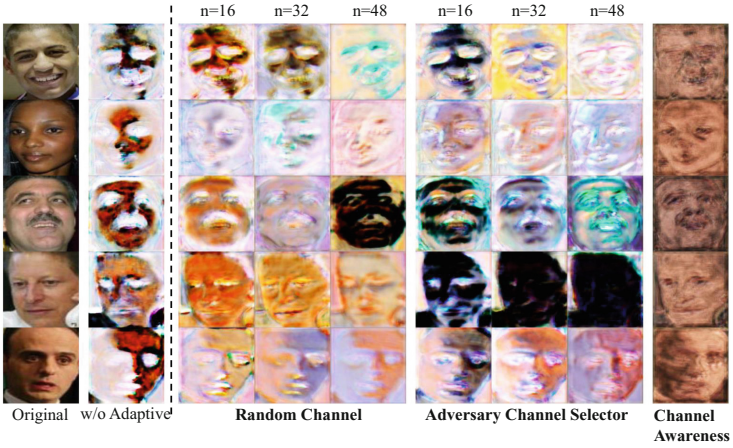


Fig. 7. Adaptive attacks on clean channels on LFW datasets. Here we use ArcFace to train AGface to protect 32 channels for each 64-channel shallow feature. The Adversarie uses ResRec to launch reconstruction attacks on n channels of each feature. For Advanced strategy, we assume the adversarie uses CosFace to train his own Channel Selector to simulate black-box scenarios.

nels for reconstructions. **Adversary Channel Selector**: the adversary trains his own Channel Selector to try to find clean channels on protected features. Note that in black-box scenarios, the adversary cannot know our face recognition (FR) model. Therefore, the adversary trains the Selector based on an arbitrary pretrained FR model (e.g., CosFace). **Channel Awareness**: The adversary is assumed to be fully aware of each shallow feature's clean channels.

After predicting clean channels, the adversary sets perturbed channels to zero value and trains the reconstruction networks to launch attacks. The results are shown in Fig. 7. For **Random Channel**, the adversary struggles to recover the images. For **Adversary Channel Selector**, the CosFace-based Selector (adversary) encounters difficulties in identifying the clean channels selected by the ArcFace-based Selector (ours). We speculate that it is due to the pattern gap between different FR models. For **Channel Awareness**, even if the adversary knows all clean channels (which is an exceedingly rare circumstance), complete recovery is still challenging. It remains **visually unclear** due to the insufficiency of available channels for reconstruction. The above analysis demonstrates the robustness of our AGFace to adaptive attacks.

5 Conclusions

In this work, we discuss the expected properties of the privacy-preserving face recognition (FR) methods and the shortages of previous works, especially the requirement of retraining the FR model and the long latency. To satisfy these properties simultaneously, we explore generating adversarial perturbations from

original features instead of directly generating adversarial features to avoid severe deviation from original features. Besides, we propose the Adaptive Channel Selector to identify the channels containing critical visual privacy for adaptive adversarial perturbations generation. Without retraining the FR model, extensive experiments demonstrate that our method *(AGFace)* achieves superior performance in terms of both privacy and utility with small increases in inference time and parameter costs.

Acknowledgements. This work was supported partially by the Guangdong NSF Project (No. 2023B1515040025), the Guangdong Basic and Applied Basic Research Foundation (2023A1515012974), the Guangzhou Basic and Applied Basic Research Scheme (2024A04J4066) and Fundamental Research Funds for the Central Universities, Sun Yat-sen University (23ptpy77).

References

1. Abdalla, M., Bourse, F., De Caro, A., Pointcheval, D.: Simple functional encryption schemes for inner products. In: Katz, J. (ed.) PKC 2015. LNCS, vol. 9020, pp. 733–751. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-46447-2_33
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE (2018)
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
4. Dosovitskiy, A., Brox, T.: Inverting visual representations with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4829–4837 (2016)
5. Drapkin, A.: Data breaches that have happened in 2022 and 2023 so far (2023). https://tech.co/news/data-breaches-updated-list
6. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322–1333 (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database forstudying face recognition in unconstrained environments. In: Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition (2008)
9. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
10. Ji, J., et al.: Privacy-preserving face recognition with learnable privacy budgets in frequency domain. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13672, pp. 475–491. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19775-8_28

11. Kou, X., Zhang, Z., Zhang, Y., Li, L.: Efficient and privacy-preserving distributed face recognition scheme via FaceNet. In: Proceedings of the ACM Turing Award Celebration Conference-China, pp. 110–115 (2021)
12. Li, A., Guo, J., Yang, H., Salim, F.D., Chen, Y.: Deepobfuscator: obfuscating intermediate representations with privacy-preserving adversarial learning on smartphones. In: Proceedings of the International Conference on Internet-of-Things Design and Implementation, pp. 28–39 (2021)
13. Li, Y., Wang, Y., Li, D.: Privacy-preserving lightweight face recognition. Neurocomputing **363**, 212–222 (2019)
14. Mai, G., Cao, K., Lan, X., Yuen, P.C.: Secureface: face template protection. IEEE Trans. Inf. Forensics Secur. **16**, 262–277 (2020)
15. Mai, G., Cao, K., Yuen, P.C., Jain, A.K.: On the reconstruction of face images from deep face templates. IEEE Trans. Pattern Anal. Mach. Intell. **41**(5), 1188–1202 (2018)
16. Mao, Y., Yi, S., Li, Q., Feng, J., Xu, F., Zhong, S.: A privacy-preserving deep learning approach for face recognition with edge computing. In: Proceedings of USENIX Workshop Hot Topics Edge Computing(HotEdge), pp. 1–6 (2018)
17. Mi, Y., et al.: DuetFace: collaborative privacy-preserving face recognition via channel splitting in the frequency domain. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 6755–6764 (2022)
18. Mi, Y., et al.: Privacy-preserving face recognition using random frequency components. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 19673–19684 (2023)
19. Mi, Y., et al.: Privacy-preserving face recognition using trainable feature subtraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 297–307 (2024)
20. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: AgeDB: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 51–59 (2017)
21. Otroshi Shahreza, H., Marcel, S.: Face reconstruction from facial templates by learning latent space of a generator network. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
22. Razzhigaev, A., Kireev, K., Kaziakhmedov, E., Tursynbek, N., Petiushko, A.: Black-box face recovery from identity features. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020. LNCS, vol. 12539, pp. 462–475. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-68238-5_34
23. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
24. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9. IEEE (2016)
25. Shahreza, H.O., Hahn, V.K., Marcel, S.: Vulnerability of state-of-the-art face recognition models to template inversion attack. IEEE Trans. Inf. Forensics Secur. (2024)
26. Shahreza, H.O., Marcel, S.: Comprehensive vulnerability evaluation of face recognition systems to template inversion attacks via 3D face reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
27. Tseng, B.W., Wu, P.Y.: Compressive privacy generative adversarial network. IEEE Trans. Inf. Forensics Secur. **15**, 2499–2513 (2020)

28. Wang, H., et al.: Cosface: large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274 (2018)
29. Wang, Y., Liu, J., Luo, M., Yang, L., Wang, L.: Privacy-preserving face recognition in the frequency domain. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2558–2566 (2022)
30. Wang, Z., et al.: Privacy-preserving adversarial facial features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8212–8221 (2023)
31. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
32. Xiao, T., Tsai, Y.H., Sohn, K., Chandraker, M., Yang, M.H.: Adversarial learning of privacy-preserving and task-oriented representations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12434–12441 (2020)
33. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
34. Zhang, C., Hu, X., Xie, Y., Gong, M., Yu, B.: A privacy-preserving multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. Front. Neurorobot. **13**, 112 (2020)
35. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
36. Zhmoginov, A., Sandler, M.: Inverting face embeddings with convolutional neural networks. arXiv preprint arXiv:1606.04189 (2016)
37. Zhong, Z., et al.: Slerpface: face template protection via spherical linear interpolation. arXiv preprint arXiv:2407.03043 (2024)

PhySU-Net: Long Temporal Context Transformer for rPPG with Self-supervised Pre-training

Marko Savic and Guoying Zhao(✉)

Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, 90014 Oulu, Finland
{marko.savic,guoying.zhao}@oulu.fi

Abstract. Remote photoplethysmography (rPPG) is a promising technology that consists of contactless measuring of cardiac activity from facial videos. However, current approaches are limited by data scarcity and environmental noise robustness. Most recent approaches utilize convolutional networks with limited temporal modeling capability or ignore long temporal context. Purely supervised rPPG methods are also severely limited by scarce data availability. In this work, we propose PhySU-Net, the first long temporal context rPPG transformer network and a novel self-supervised pre-training strategy that exploits unlabeled data to improve our model. Our strategy leverages traditional methods and image masking to provide pseudo-labels for physiologically relevant self-supervised pre-training. Our model is tested on three public benchmark datasets (OBF, VIPL-HR and MMSE-HR) and shows state-of-the-art performance in supervised training. Furthermore, we demonstrate that our self-supervised pre-training strategy further improves our model's performance by leveraging representations learned from unlabeled data. Our code is available at: https://github.com/marukosan93/PhySU-Net.

Keywords: Remote Photoplethysmography · Facial Videos · Transformer · Self-Supervised Learning

1 Introduction

Physiological signals like the blood volume pulse (BVP) are used to determine vital healthcare parameters such as heart rate (HR), heart rate variability (HRV), respiratory frequency (RF) and oxygen saturation (SpO2). Moreover, they are psychological indicators since they change accordingly with emotional states [20]. They are customarily measured with contact devices based on electrocardiography (ECG) [12] or photoplethysmography (PPG) [3]. A more convenient non-contact method that employs cheap and ubiquitous RGB cameras is remote photoplethysmography (rPPG) [39]. Similar to contact PPG, rPPG relies on capturing periodical variations in optical absorption of tissue caused by cardiac activity. The main differences between rPPG and contact PPG are in optical sensor (RGB camera vs. photodiode), distance from sensor (meters vs. millimeters) and lighting source

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 228–243, 2025.
https://doi.org/10.1007/978-3-031-78341-8_15

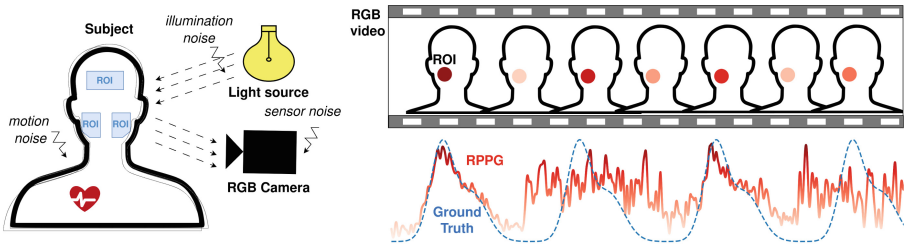


Fig. 1. (Left) Quasi-periodical variations in reflected light, which are caused by the heart's activity, are captured by the camera. Noise from the environment (from lighting, motion and sensor) is also acquired. (Right) A rough rPPG estimate can be obtained by averaging pixels from a ROI (e.g. cheeks or forehead) and the obtained temporal signal is highly susceptible to noise.

(LED vs complex environmental). Due to these differences in measuring environment, along with the weak signal derived from skin color variation, a camera also captures overwhelming environmental noise caused by lighting changes, subject movement and sensor variations. Accurately and robustly identifying the faint quasi-periodical rPPG signal is a challenging task. In Fig. 1 we provide a visualization on how rPPG data is acquired and of a simple extraction method.

Several early traditional methods had been proposed that relied on optical/physiological considerations expressed through mathematical models like CHROM [7], POS [42], PBV [8], LGI [33] or common blind source separation approaches such as ICA [34] and PCA [19]. However, since they lacked robustness in scenarios with variable light and movement, they were surpassed by deep learning approaches. Most deep learning methods employed Convolutional Neural Networks (CNN). Early 2D-CNN models extracted HR from adjacent frames, such as HR-CNN [35] and DeepPhys [4]. End-to-end 3D-CNN models such as PhysNet [45], rPPGNet [46] and AutoHR [44] exploited the temporal information. Non-end-to-end models used spatial-temporal maps, less affected by noise, like RhythmNet [30], CVD [31], Dual-GAN [25] and BVPNet [6], DRNet [10], ND-DeeprPPG [22] and NEST [26]. Particularly, CVD [31], Dual-GAN [25] and DRNet [10] and ND-DeeprPPG [22] focused on disentangling the useful physiological features from noise, and NEST [26] focused on maximizing the coverage of the feature space during training. Aside from CNNs, in GraphPhys [43] a graph convolutional neural network has been successfully utilised for rPPG.

Self-attention based Transformer networks such as ViT [11] and Swin [24] have shown to be superior to CNN in many computer vision problems due to their capabilities to capture global dependencies, but heavily rely on data quantity and pre-training on large-scale datasets. There have been several successful applications of transformer based architectures used in rPPG. EfficientPhys [23] utilizes Swin transformer for extracting rPPG signals and directly compares with a convolutional backbone. Physformer [47] used spatio-temporal input tokens, custom temporal difference guided global attention layers and an

elaborate dynamic supervision training scheme. In TransPPG [17] two signal map streams composed of foreground and background are fed to a transformer network. TransPhys [41] a hybrid CNN and Transformer architecture was proposed to pre-train contrastively and then fine-tune using labels. However, all of the aforementioned methods use small time segments as input (2 s-10 s) that do not fully exploit the strong long-range modeling capabilities of the transformer architecture. This is especially relevant with strong environmental noise, as the underlying rPPG signal has strong self-correlation over time, but is overshadowed by strong movement and illumination noise. However, modelling a longer time segment allows to better capture the correlation between the rPPG signal salient features over time, and can make the faint signal easier to distinguish from aleatory or impulsive noise. As an example, we can consider the signal shown in Fig. 1, where observing only the 2nd or 4th peak would lead to an inaccurate prediction, but modelling the whole time frame can allow us to capture the correlation between the cleaner parts of the signal.

Lack of labeled data is another major issue in rPPG, as data collection is costly, requires medical devices and presents privacy concerns. Supervised methods struggle to achieve robustness and high generalization capability when trained on small datasets with specific noise distributions. Data augmentation has been used to partially mitigate this issue, both with classic computer-vision augmentation strategies and with rPPG specific augmentations. For example, spatial-temporal augmentations have also been proposed [32,44] to extend the training set with extra samples containing borderline small or large HR values, by temporally up-sampling and down-sampling videos to achieve this. Augmenting datasets with synthetic videos generated from real data via video-to-video networks was also proposed in [37], RErPPGNet [16] and [1]. Natural images [30], synthetic signals [28] and synthetic avatar facial videos [27] have also been used to learn more generalized representations. Nonetheless, non-facial videos, simple augmentations and synthetic data cannot replicate signals and complex environmental noise present in real facial video data. Another direction towards data scarcity mitigation has been learning from facial videos without physiological annotations, which are plentiful in contrast to rPPG labeled data. Several self-supervised contrastive methods have also been proposed to learn from data without labels [13,36,40], but failed to reach close to supervised performance or to demonstrate their transferable capability. Moreover, they rely on weak constraints that do not hold for challenging scenarios.

The rPPG signal is quasi-periodical and drowned in environmental noise, but over a longer time frame it retains similar features (amplitude, frequency, dicrotic notch, systolic point, diastolic point) and has high self-correlation. Consequently, modeling a longer time frame can aid in distinguishing the rPPG signal from noise. Moreover, in a quantitative study [18] on HR measurement, the error increased hyperbolically with decreasing duration. We propose a method for rPPG that capitalizes on the long-range capabilities of the transformer architecture and makes use of compact noise-robust spatial-temporal maps. Our model can learn rich features from a long input of $\sim$20 s, making it more robust to

environmental noise. We design our framework for signal prediction as an image (multi-signal) reconstruction task, providing stronger supervision and enabling our model to train from scratch without any complex training strategies or data augmentation. Moreover, we propose a self-supervised pre-training framework that allows our model to seamlessly learn physiologically relevant representations by pre-training on unlabeled data. We leverage pseudo-labels that are generated via masking the input signals, enabling the model to learn temporal and frequency characteristics of rPPG signals from the input data alone, and use a traditional method to guide the model via an additional regression constraint. We clarify that our self-supervised pre-training is meant for subsequent fine-tuning and we produce useful representations for improving the downstream supervised learning task, instead of attempting to surpass purely self-supervised methods on unlabeled data. We summarise our contributions as follows:

- We propose PhySU-Net, the first long temporal context transformer network for rPPG, that is able to learn robust features from challenging data.
- We propose a novel self-supervised pre-training framework, that leverages the first image-based pretext task for rPPG and a novel regression constraint.
- We extensively evaluate our framework on three datasets (OBF [21], VIPL-HR [29], MMSE-HR [48]) and perform intra-dataset, cross-dataset, linear classification, transfer learning and ablation studies.

2 Methodology

To address the issue of non-robustness to environmental noise in rPPG, we propose PhySU-Net, a long temporal context transformer network with spatial-temporal map input. To mitigate data scarcity, a self-supervised pre-training scheme is proposed that can aid in learning useful representations from unlabeled data for improved generalization. Our complete framework is divided into preprocessing, PhySU-Net model and self-supervised pre-training. The complete overview is shown in Fig. 2.

2.1 Preprocessing

We preprocess the input videos into spatial-temporal maps because they partially exclude non-physiological noise and have a compact size. They are calculated by averaging pixels from regions of interest (ROI). This is a procedure that has been proven to successfully extract relevant features for rPPG as most traditional methods use average pooling of pixels. The additional computational burden is minimal compared to end-to-end methods that use cropped videos, since both end-to-end and non end-to-end methods need to detect landmarks, which is the heaviest step of the preprocessing. Additionally, as signal maps are more compact, they allow us to model longer temporal sequences. We calculate Multi-scale Spatial-Temporal Maps (MSTmaps) by following [31]. MSTMaps ate obtained by averaging pixels from combinations of ROIs that merge both local

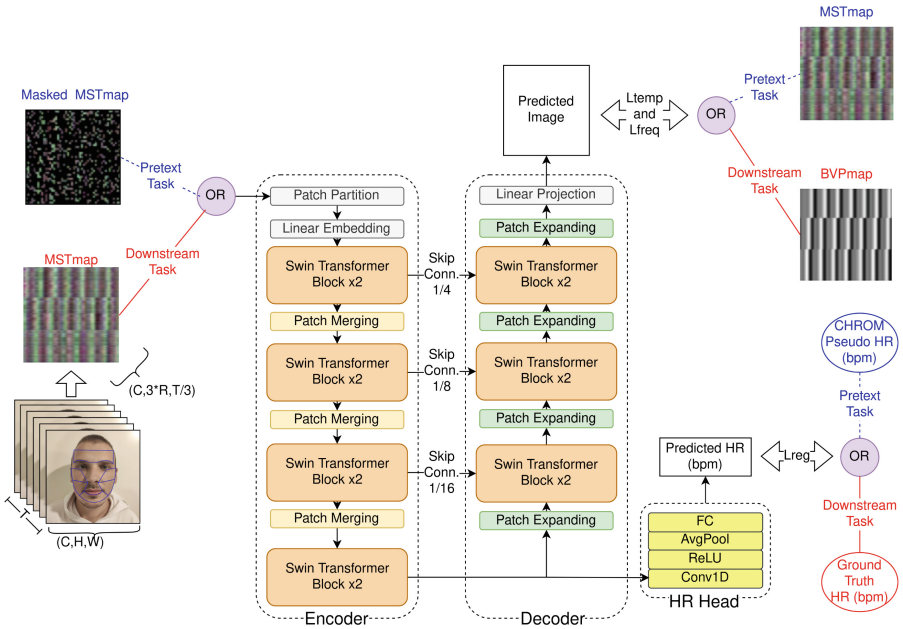


Fig. 2. Overview of our method: The input video is processed into a stacked MSTmap. For the supervised downstream task, the decoder reconstructs an image with similar temporal and frequency properties as the BVPmap label, and the HR head regresses the global HR value with the HR ground truth as label. For the self-supervised pretext task, only the input and the labels change. The input is a masked version of the MSTmap that the decoder attempts to reconstruct into a full MSTmap. For the HR regression, a CHROM [7] calculated pseudo-label is used.

and global physiological information present in the face. Firstly, to extract landmarks we utilize PyFeat [5] with RetinaFace [9] model for face detection and PFLD [14] model for landmark localization. The landmarks are then stabilized with a 5-point moving average filter. Six informative ROIs are defined within the face, which are joined in $R = 2^6 - 1$ ROI combinations, thus merging global and local information. For each ROI combination (R) and color channel (C) a temporal sequence is obtained by averaging the pixels for the whole video. Each of the $C * R$ sequences is then band pass filtered at $[0.7, 3]Hz$ to reduce interference of non-physiological signal components and is min-max normalized. In addition, the R dimension is resized from 63 to 64 for computational ease. The blood volume pulse (BVP) ground truth is also band pass filtered to cut off irrelevant frequencies, and HR regression targets are calculated via FFT. For the image reconstruction target, BVPmaps are generated by stacking $C * R$ copies of the BVP ground truth, so that each row of the MSTmap will correspond to the same BVP ground truth. Finally, both MSTmaps and BVPmaps are divided into three equal segments along the temporal axis and are stacked to form a square image.

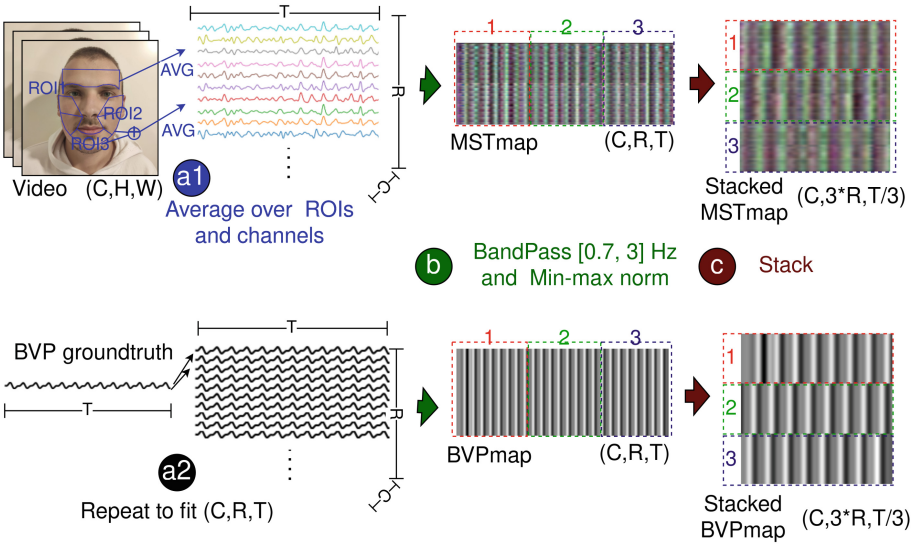


Fig. 3. Preprocessing: a1) $C \times R$ temporal sequences are extracted by averaging pixels for each channel and ROI combination. a2) The ground truth BVP is duplicated to fit the dimension of the MSTmap b) The sequences are filtered with a pass band of $[0.7, 3]Hz$ and min-max normalized. c) The MSTmap and BVPmap is temporally stacked to form square images.

This step will be discussed further in the next subsection. The preprocessing procedure is shown in Fig. 3.

2.2 PhySU-Net Model

We formulate a multitask learning problem comprised of HR regression and image-based rPPG signal prediction, proposing a rPPG extension of Swin-Unet [2] to solve it. Swin [24] transformer hierarchically builds feature maps by merging image patches and computes self-attention within local windows, allowing it to have linear complexity. It was adapted into a Swin-Unet [2] by constructing an encoder and decoder with skip connections and by adding a new patch expanding layer for up-sampling. This base architecture is particularly suited to the rPPG problem, as the self-attention mechanism can leverage long range dependencies of the quasi-periodical underlying rPPG signal. The rPPG signal prediction is framed as an MSTmap to BVPmap image reconstruction task [6], where we aim at a direct image-to-image correspondence between the input spatial-temporal map and ground truth signal map. Therefore, each coarse temporal input sequence is reconstructed into a cleaner rPPG signal, providing strong supervision and enabling our model to train easily from scratch. We further expand the base model to be better suited for rPPG with the following two additions. Firstly, the input MSTmap is stacked to a square size

as shown in Fig. 3(c). This strategy allows us to maintain the temporal consistency in the patch sampling, and to inject temporally distant patches in the local neighborhoods of windowed self-attention, encouraging the network to learn long range dependencies in earlier layers. Secondly, as the skip connections allow the network to reconstruct the final output from the earlier layers, we add an HR regression head that compels the encoder to learn global HR features as well. This strategy ensures that the network focuses both on global supervision via the HR regression constraint and fine-grained signal level supervision via the temporal and frequency constraints. The HR regression head is constructed of a 1d convolution, which aids faster convergence, followed by a ReLU non-linearity, adaptive average pooling and finally a fully connected layer regressing the HR value. The full architecture is shown in Fig. 2.

For the HR regression loss L_{reg}, we employ an L1 loss and compare the predicted value of the regression head with the ground truth HR. For the signal map reconstruction, we utilize both temporal and frequency based losses. The temporal loss promotes high correlation between prediction and label. Instead of the commonly used Pearson loss [45], which is sensitive to synchronization errors, we utilize the maximum cross-correlation (MCC) [13], which determines the correlation at an ideal offset and is invariant to phase differences between prediction and label.

$$MCC(x,y) = C_{pr} \times Max(\frac{\mathcal{F}^{-1}\{BPass(\mathcal{F}\{x\} \cdot \mathcal{F}^{*}\{y\})\}}{\sigma_x \times \sigma_y}) \tag{1}$$

The MCC, in Eq. 1, is calculated by firstly taking the FFT of the two signals and multiplying the one with the complex conjugate of the other. The result is band-pass filtered by zeroing the frequencies outside the HR relevant band $[0.7, 3]Hz$. Next, the IFFT is performed and divided by the standard deviation of the signals x and y, obtaining the cross-correlation calculated in the frequency domain. Then by taking the maximum of the cross-correlation, we obtain the correlation between the signals at the ideal offset, mitigating any synchronization errors between input videos and ground truth. Finally, the loss is scaled by C_{pr}, the ratio of power inside the HR relevant frequencies. Our temporal loss L_{temp}, in Eq. 2, is the mean of the negative MCC of each $C \times R$ temporal sequence from the predicted spatial-temporal map X and BVPmap label Y.

$$L_{temp}(X,Y) = 1 - \frac{\sum_{c=1}^{C}\sum_{r=1}^{R} MCC(X(c,r,t), Y(c,r,t))}{C \times R} \tag{2}$$

Our frequency loss L_{freq}, as shown in Eq. 3, is defined as the mean squared error between the Power Spectral Densities (PSD) of the prediction and label temporal sequences. The error is squared to accentuate the peaks in the spectrum, since they are of most relevance.

$$L_{freq}(X,Y) = \frac{\sum_{c=1}^{C}\sum_{r=1}^{R} (PSD(X(c,r,t)) - PSD(Y(c,r,t))^2}{C \times R} \tag{3}$$

The normalized PSD is defined in Eq. 4 with irrelevant values outside the $[0.7, 3]Hz$ band set to zero. It is calculated using the product of the FFT with its complex conjugate and normalized by dividing it by its summation, ensuring that all PSDs are on the same scale.

$$PSD(x) = \frac{\mathcal{F}\{x\} \cdot \mathcal{F}^*\{x\}}{\sum_{n=1}^{N} \mathcal{F}\{x\} \cdot \mathcal{F}^*\{x\}} \tag{4}$$

The final loss, in Eq. 5, is the weighted sum of the regression, temporal and frequency losses.

$$L = \alpha L_{reg} + \beta L_{temp} + \gamma L_{freq} \tag{5}$$

2.3 Self-supervised Pre-training

We propose a self-supervised pre-training framework that enables our network to learn useful representations on unlabeled data, which can be subsequently used to fine-tune with supervised learning on labeled data, offering better performance and generalization capability. Our self-supervised pre-training method consists of pseudo-HR regression and masked MSTmap reconstruction pretext tasks. Both pretext tasks utilize the same losses and network as the downstream task, the only difference being the input and labels. This allows to achieve quick adaptation from the pre-training to the downstream task. Inspired by Masked Auto Encoder [15], we input a masked version of the MSTmap and have the model reconstruct the missing parts, consequently making the model learn temporal and frequency characteristics (by minimizing L_{temp} and L_{freq}) of the coarse rPPG signals present in the MSTmap. We mask 75% of all 4×4 input patches for the image based task. For the HR regression task, we generate a pseudo-label with the traditional method CHROM [7], which serves as an additional constraint to guide the models' self-supervised learning towards learning features with global physiological meaning. Our framework can be easily adapted to other image-based pretext tasks and pseudo-labels obtained with different hand-crafted rPPG methods.

3 Experiments

We evaluate PhySU-Net on the OBF [21], VIPL-HR [29] and MMSE-HR [48] datasets, and show superior performance compared to state-of-the-art methods in intra and cross dataset testing. To prove the effectiveness of our self-supervised approach, a protocol similar to [40] is used, where linear classification (re-train only last fully-connected layer) and transfer learning (re-train whole network) are performed. Additionally, we provide ablation studies to analyze crucial network components and hyperparameters, demonstrate the effectiveness of longer temporal context and show generalizability of our self-supervised pre-training.

3.1 Experimental Setup

Datasets: OBF [21] contains 200 five-minute-long constant frame rate RGB videos with corresponding ground truth ECG and BVP. They were recorded from 100 diverse subjects with resting and elevated heart rates in an environment with stable lighting and minimal movement of the subjects. MMSE-HR [48] contains 102 videos of length 20–70 s recorded under stable lighting from 40 subjects in emotion elicitation experiments. It contains challenging motions as there are spontaneous facial expressions and head motions. VIPL-HR [29] contains 2,378 RGB videos of 20 s–30 s length with variable and unstable frame rate. It was recorded in a challenging environment with different devices, large movements and various lighting. It contains numerous sources of noise, making HR estimation very challenging. We assess the environmental noise robustness of our method by evaluating it on datasets with diverse environmental conditions. Based on the level of challenge, we can rank the datasets from most controlled conditions in OBF [21] (large dataset, good lighting, minimal movements, varied demographics, elevated HR), to less controlled in MMSE-HR [48] (small dataset, spontaneous movements) and with VIPL-HR [29] being the most challenging (large dataset, unstable fps, different recording devices, challenging lighting and movement).

Evaluation Metrics: We follow previous works by using absolute error (MAE), root-mean-square error (RMSE), standard deviation (SD) and Pearson's correlation coefficient (R).

Implementation: We choose T = 576 (19.2 s at 30fps) due to VIPL-HR videos being 20–30 s long and for computational ease. For fair comparison with other methods, a ten-fold subject exclusive cross validation is adopted for OBF, five-fold for VIPL-HR and three-fold for MMSE-HR in all experiments. In training, the AdamW optimizer is used with epsilon = 1e−8, betas = (0.9, 0.99), lr = 5e−5, wd = 0.05, batch = 8. Loss parameters are set at $\alpha = 5$, $\beta = 1$, $\gamma = 5$. For supervised experiments and self-supervised pre-training 50 epochs are used. For all fine-tuning (linear and transfer) on VIPL-HR, the epochs are lowered to 25. No data augmentation method is used, training samples are 576 frames long with a fixed sliding window of 30 frames, testing samples are 576 frames long with no overlap. All the implementation code will be shared upon acceptance.

3.2 Experimental Results

Supervised: We evaluate the supervised part of our method on the OBF and VIPL-HR datasets in intra-dataset testing, as shown in Table 1. We compare PhySU-Net to a wide array of previous supervised methods including traditional, convolutional and transformer based, showing that it reaches state-of-the-art performance with RMSE of 0.659 on OBF and 7.35 on VIPL-HR. Our method's long-range temporal modelling proves effective on extracting an accurate HR on the challenging VIPL-HR data, with more reliable predictions than other methods with notably lower SD. Following the protocol of [30,47] we also conduct cross-dataset evaluation by training the model on VIPL-HR and directly testing

Table 1. Supervised training results on OBF [21] and VIPL-HR [29]: divided in Traditional, Convolutional and Transformer based methods. Best results are marked in bold, second best are underlined.

Dataset: OBF [21]				
Type	Method	RMSE ↓	SD ↓	r ↑
Traditional	CHROM [7]	2.733	2.730	0.980
	POS [42]	1.906	1.899	0.991
Convolutional	rPPGNet [46]	1.800	1.756	0.992
	CVD [31]	1.260	1.257	0.996
Transformer	Physformer [47]	<u>0.804</u>	<u>0.804</u>	<u>0.998</u>
	PhySU-Net (Ours)	**0.659**	**0.618**	**0.999**

Dataset: VIPL-HR [29]					
Type	Method	MAE ↓	RMSE ↓	SD ↓	r ↑
Traditional	CHROM [7]	11.4	16.9	15.1	0.28
	POS [42]	11.5	17.2	15.3	0.30
Convolutional	DeepPhys [4]	11.0	13.8	13.6	0.11
	PhysNet [45]	10.8	14.8	14.9	0.20
	RhythmNet [30]	5.30	8.14	8.11	0.76
	CVD [31]	5.02	7.97	7.92	0.79
	Dual-GAN [25]	4.93	7.68	7.63	0.81
	BVPNet [6]	5.34	7.85	7.75	0.70
	NEST [26]	4.76	7.51	7.49	<u>0.84</u>
	GraphPhys [43]	6.69	9.70	9.30	0.48
	DRNet [10]	**4.18**	**6.78**	<u>6.75</u>	**0.85**
Transformer	Physformer [47]	4.97	7.79	7.74	0.78
	TransPPG [17]	4.94	7.42	7.44	0.79
	TransPhys [41]	5.09	10.6	10.12	0.69
	PhySU-Net (Ours)	<u>4.53</u>	<u>7.35</u>	**5.79**	0.80

on the MMSE-HR dataset. In the cross-dataset results shown in Table 2, our method obtains the best performance on all metrics and shows that it can adapt well to an unseen dataset.

Linear Classification and Transfer Learning: In Table 3, with transfer learning we obtain a notable performance increase with RMSE reduced from 7.35 to 7.07. This proves the transferable ability of our method, as it learns useful representations on unlabeled OBF data during pre-training that improve its performance when fine-tuning on VIPL-HR. Furthermore, in the linear classification test the representations learned with self-supervision on both unlabeled OBF (RMSE 9.16) and VIPL-HR (RMSE 9.28) are of good quality, as the performance is still satisfactory considering that only the last fully connected layer is re-trained on the challenging VIPL-HR data.

Input Sequence Length Ablation: In Table 4 rows 2 and 3 we show that reducing the temporal context length leads to lower performance. The best performance is obtained with the longest T = 576, showing that PhySU-Net is proficient at modelling long-temporal context and can learn more robust features with a longer context.

Table 2. Cross-dataset testing on MMSE-HR [48] after training on VIPL-HR [29]: divided in Traditional, Convolutional and Transformer based methods. Best results are marked in bold, second best are underlined.

Dataset: VIPL-HR [29] → MMSE-HR [48]					
Type	Method	MAE ↓	RMSE ↓	SD ↓	r ↑
Traditional	CHROM [7]	-	13.97	14.08	0.55
	Tulyakov2016 [38]	-	11.37	12.24	0.71
Convolutional	PhysNet [45]	-	13.25	12.76	0.44
	RhythmNet [30]	-	7.33	6.98	0.78
	CVD [31]	-	6.04	6.06	0.84
	AutoHR [44]	-	5.87	5.71	0.89
	ND-DeeprPPG [22]	**1.84**	<u>4.83</u>	-	<u>0.93</u>
Transformer	EfficientPhys-C [23]	2.91	5.43	-	0.92
	Physformer [47]	2.84	5.36	5.22	0.92
	PhySU-Net (Ours)	<u>2.28</u>	**4.43**	**3.75**	**0.94**

Table 3. Linear classification and Transfer learning on VIPL-HR [29]

Training Methods	Pre-train → Fine-tune	MAE ↓	RMSE ↓	SD ↓	r ↑
Purely Supervised	NONE → VIPL-HR	4.53	7.35	5.79	0.80
Linear classification (with Self-Supervision)	OBF → VIPL-HR	6.19	9.16	6.75	0.68
	VIPL-HR → VIPL-HR	6.30	9.28	6.81	0.67
Transfer learning (with Self-Supervision)	OBF → VIPL-HR	4.22	7.07	5.66	0.82

Network Components Ablation: In Table 4 rows 4 and 5, we show that both multitask learning components contribute to our method's effectiveness, as the HR head provides rough global supervision and the decoder performs signal level supervision on the signals. In row 6 of Table 4, performance declines without stacking, indicating that the network is encouraged to learn more informative features as attention windows also include signals that are further away in time, better exploiting the long-temporal context.

Transfer Learning Ablation and Generalization: As shown in rows 2, 3 and 4 of Table 5, both pretext tasks contribute to a better downstream prediction as there is a drop in performance when excluding either of them. Additionally, our method is generalizable as the pretext tasks can be easily changed. The HR regression obtained by using different traditional methods (CHROM [7], LGI [33], GREEN [39]) yields comparable results, as can be seen in Table 5 rows 4, 5 and 6. We also implement another image-based pretext task, in alternative to masking, called PBVPmap prediction. This task consists of predicting a PBVPmap constructed from pseudo-BVP signals obtained with CHROM [7] method. In row 7 of Table 5, we see that the PBVP task is also valid for pretraining, but masking yields stronger representations. Our self-supervised pretraining approach can be adapted to any kind of regression target and image-based task.

Table 4. Network design ablation on VIPL-HR [29]. Base method is marked in bold.

	MAE ↓	RMSE ↓	SD ↓	r ↑
Proposed (T = 576)	**4.53**	**7.35**	**5.79**	**0.80**
T = 384	4.80	7.80	6.15	0.78
T = 256	5.05	8.13	6.37	0.76
w/o HR head	4.72	7.85	6.28	0.78
w/o Decoder	5.20	7.93	5.98	0.77
w/o Stacking	5.62	8.35	6.17	0.74

Table 5. Transfer learning ablation study on VIPL-HR [29], with different self-supervised pre-training tasks on OBF [21]. Base method is marked in bold.

Tasks:		MAE ↓	RMSE ↓	SD ↓	r ↑
Regression	Image				
No task	No task	4.53	7.35	5.79	0.80
CHROM	No task	4.17	7.17	5.83	0.82
No task	Mask	4.46	7.32	5.80	0.81
CHROM	**Mask**	**4.22**	**7.07**	**5.66**	**0.82**
GREEN	Mask	4.15	7.06	5.71	0.82
LGI	Mask	4.17	7.08	5.72	0.82
CHROM	PBVP	4.35	7.18	5.71	0.82

Hyperparameter Sensitivity Analysis: As our loss function from Eq. 5 is composed of three weighted terms, we analyze our method's sensitivity to the parameters α, β, γ. In Fig. 4, we show that for α and β parameters, tied to the HR regression and temporal losses respectively, values within the same order of magnitude yield small variations in the range of $\approx 0.2RMSE$, while borderline values (0 or 20) lead to the final loss being severely unbalanced, thus resulting in suboptimal performance. Compared to β, γ is less impactful on the learning, meaning that, out of the two reconstruction losses, the temporal term is more significant than the frequency term. We conclude that our method's sensitivity to α, β, γ is low, as the performance significantly degrades only for largely unbalanced coefficients.

Visualization: From Table 3 and Table 5, quantitative results show that PhySU-Net benefits from both the self-supervised pre-training tasks. Moreover, in Fig. 5 we show a visualization of PhySU-Net features on the OBF [21] dataset after different self-supervised pre-training (separate tasks and then complete) and supervised training. When only pre-trained via the map masking task and, even though the features are useful for downstream learning, we can see that they still do not have a physiologically relevant structure yet. With only the pseudo-HR regression task, we notice that the physiological meaning is much higher

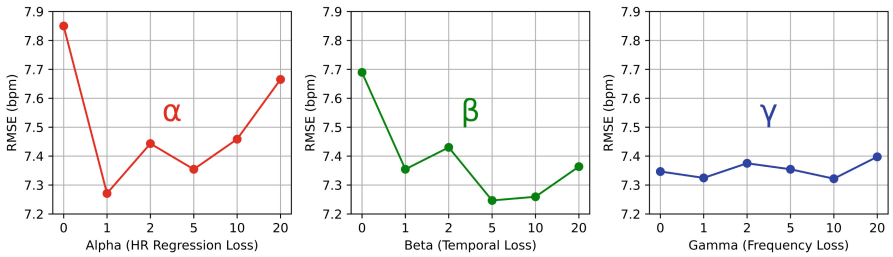


Fig. 4. Loss hyperparameter sensitivity analysis on VIPL-HR [29]

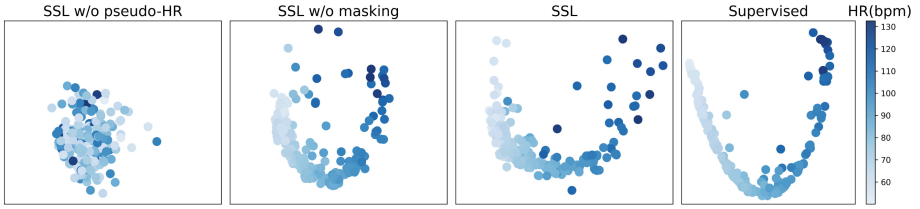


Fig. 5. PCA feature visualizations on OBF [21]. Self-supervised pre-training first with only masking (w/o pseudo-HR), second with only pseudo-HR (w/o masking), third with both tasks. The last visualization is from supervised training on the same data.

as the regression task gives more specific physiological context. Lastly, when pre-training with both tasks, we see that the feature structure is similar to supervised learning on the same data, with the self-supervised features having a looser distribution due to being more general than the strongly supervised features obtained when training with labels.

4 Conclusion

We propose PhySU-Net, a robust rPPG method that deals with challenging data by making full use of a long temporal context via our transformer model. With our proposed self-supervised pre-training, we further improve performance by leveraging unlabeled data. Experiments on three public domain datasets show that our supervised method is superior to most state-of-the-art methods. Moreover, with our generalized self-supervised pre-training the model can learn meaningful representations and transferable representations from unlabeled data. Future work can include the addition of new pretext tasks to our framework and usage of non-rPPG unlabeled data.

Acknowledgements. This work was supported by the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI (grants 336116, 345122, 359854), ICT 2023 project TrustFace (grant 345948), the University of Oulu & Research Council of Finland Profi 7 (grant 352788), and HORIZON-MSCA-SE-2022

project ACMod (grant 101130271). As well, the authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

1. Ba, Y., Wang, Z., Karinca, K.D., Bozkurt, O.D., Kadambi, A.: Style transfer with bio-realistic appearance manipulation for skin-tone inclusive RPPG. In: IEEE International Conference on Computational Photography (ICCP). IEEE (2022)
2. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13803, pp. 205–218. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-25066-8_9
3. Challoner, A., Ramsay, C.: A photoelectric plethysmograph for the measurement of cutaneous blood flow. Phys. Med. Biol. **19**(3), 317 (1974)
4. Chen, W., McDuff, D.: DeepPhys: video-based physiological measurement using convolutional attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 349–365 (2018)
5. Cheong, J.H., Xie, T., Byrne, S., Chang, L.J.: py-feat: Python facial expression analysis toolbox. CoRR **abs/2104.03509** (2021). https://arxiv.org/abs/2104.03509
6. Das, A., Lu, H., Han, H., Dantcheva, A., Shan, S., Chen, X.: BVPNet: video-to-BVP signal prediction for remote heart rate estimation. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 01–08. IEEE (2021)
7. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based RPPG. IEEE Trans. Biomed. Eng. **60**(10), 2878–2886 (2013)
8. De Haan, G., Van Leest, A.: Improved motion robustness of remote-PPG by using the blood volume pulse signature. Physiol. Meas. **35**(9), 1913 (2014)
9. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5203–5212 (2020)
10. Dong, Y., Yang, G., Yin, Y.: DRNet: decomposition and reconstruction network for remote physiological measurement. arXiv preprint arXiv:2206.05687 (2022)
11. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Fye, W.B.: A history of the origin, evolution, and impact of electrocardiography. Am. J. Cardiol. **73**(13), 937–949 (1994)
13. Gideon, J., Stent, S.: The way to my heart is through contrastive learning: remote photoplethysmography from unlabelled video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3995–4004 (2021)
14. Guo, X., et al.: PFLD: a practical facial landmark detector. arXiv preprint arXiv:1902.10859 (2019)
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
16. Hsieh, C.J., Chung, W.H., Hsu, C.T.: Augmentation of RPPG benchmark datasets: learning to remove and embed RPPG signals via double cycle consistent learning from unpaired facial videos. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13676, pp. 372–387. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19787-1_21

17. Kang, J., Yang, S., Zhang, W.: Transppg: two-stream transformer for remote heart rate estimate. CCF Trans. Pervasive Comput. Interact. (2024)
18. Kobayashi, H.: Effect of measurement duration on accuracy of pulse-counting. Ergonomics **56**(12), 1940–1944 (2013)
19. Lewandowska, M., Rumiński, J., Kocejko, T., Nowak, J.: Measuring pulse rate with a webcam-a non-contact method for evaluating cardiac activity. In: 2011 federated Conference on Computer Science and iNformation Systems (FedCSIS), pp. 405–410. IEEE (2011)
20. Li, C., Xie, L., Pan, H.: Branch-fusion-net for multi-modal continuous dimensional emotion recognition. IEEE Signal Process. Lett. **29**, 942–946 (2022). https://doi.org/10.1109/LSP.2022.3160373
21. Li, X., et al.: The OBF database: a large face video database for remote physiological signal measurement and atrial fibrillation detection. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 242–249. IEEE (2018)
22. Liu, S.Q., Yuen, P.C.: Robust remote photoplethysmography estimation with environmental noise disentanglement. IEEE Trans. Image Process. (2023)
23. Liu, X., Hill, B., Jiang, Z., Patel, S., McDuff, D.: Efficientphys: enabling simple, fast and accurate camera-based cardiac measurement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2023)
24. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
25. Lu, H., Han, H., Zhou, S.K.: Dual-GAN: Joint BVP and noise modeling for remote physiological measurement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12404–12413 (2021)
26. Lu, H., Yu, Z., Niu, X., Chen, Y.C.: Neuron structure modeling for generalizable remote physiological measurement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
27. McDuff, D., et al.: Scamps: Synthetics for camera measurement of physiological signals. Adv. Neural. Inf. Process. Syst. **35**, 3744–3757 (2022)
28. Niu, X., Han, H., Shan, S., Chen, X.: Synrhythm: learning a deep heart rate estimator from general to specific. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3580–3585. IEEE (2018)
29. Niu, X., Han, H., Shan, S., Chen, X.: VIPL-HR: a multi-modal database for pulse estimation from less-constrained face video. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018 Part V. LNCS, vol. 11365, pp. 562–576. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20873-8_36
30. Niu, X., Shan, S., Han, H., Chen, X.: RhythmNet: end-to-end heart rate estimation from face via spatial-temporal representation. IEEE Trans. Image Process. **29**, 2409–2423 (2019)
31. Niu, X., Yu, Z., Han, H., Li, X., Shan, S., Zhao, G.: Video-based remote physiological measurement via cross-verified feature disentangling. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020 Part II. LNCS, vol. 12347, pp. 295–310. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_18
32. Niu, X., et al.: Robust remote heart rate estimation from face utilizing spatial-temporal attention. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp. 1–8. IEEE (2019)
33. Pilz, C.S., Zaunseder, S., Krajewski, J., Blazek, V.: Local group invariance for heart rate estimation from face videos in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2018)

34. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans. Biomed. Eng. **58**(1), 7–11 (2010)
35. Špetlík, R., Franc, V., Matas, J.: Visual heart rate estimation with convolutional neural network. In: Proceedings of the british machine vision conference, Newcastle, UK, pp. 3–6 (2018)
36. Sun, Z., Li, X.: Contrast-phys: unsupervised video-based remote physiological measurement via spatiotemporal contrast. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022 Part XII. LNCS, vol. 13672, pp. 492–510. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19775-8_29
37. Tsou, Y.Y., Lee, Y.A., Hsu, C.T.: Multi-task learning for simultaneous video generation and remote photoplethysmography estimation. In: Proceedings of the Asian Conference on Computer Vision (2020)
38. Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J.F., Sebe, N.: Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2396–2404 (2016)
39. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Opt. Express **16**(26), 21434–21445 (2008)
40. Wang, H., Ahn, E., Kim, J.: Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2431–2439 (2022)
41. Wang, R.X., Sun, H.M., Hao, R.R., Pan, A., Jia, R.S.: TransPhys: transformer-based unsupervised contrastive learning for remote heart rate measurement. Biomed. Signal Process. Control **86**, 105058 (2023)
42. Wang, W., den Brinker, A.C., Stuijk, S., De Haan, G.: Algorithmic principles of remote PPG. IEEE Trans. Biomed. Eng. (2016)
43. Xiong, J., Ou, W., Liu, Z., Gou, J., Xiao, W., Liu, H.: GraphPhys: facial video-based physiological measurement with graph neural network. Comput. Electr. Eng. **113**, 109022 (2024)
44. Yu, Z., Li, X., Niu, X., Shi, J., Zhao, G.: AutoHR: a strong end-to-end baseline for remote heart rate measurement with neural searching. IEEE Signal Process. Lett. **27**, 1245–1249 (2020). https://doi.org/10.1109/LSP.2020.3007086
45. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In: British Machine Vision Conference (2019)
46. Yu, Z., Peng, W., Li, X., Hong, X., Zhao, G.: Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 151–160 (2019)
47. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P.H., Zhao, G.: Physformer: facial video-based physiological measurement with temporal difference transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4186–4196 (2022)
48. Zhang, Z., et al.: Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3438–3446 (2016)

SwiftFaceFormer: An Efficient and Lightweight Hybrid Architecture for Accurate Face Recognition Applications

Luis S. Luevano[1(✉)], Yoanna Martínez-Díaz[2], Heydi Méndez-Vázquez[2], Miguel González-Mendoza[3], and Davide Frey[1]

[1] Univ Rennes, Inria, CNRS, IRISA, Rennes, France
{luis-santiago.luevano-garcia,davide.frey}@inria.fr
[2] Advanced Technologies Application Center (CENATAV), 7A #21406 Siboney, 12200 Playa, Havana, Cuba
{ymartinez,hmendez}@cenatav.co.cu
[3] School of Engineering and Sciences, Tecnologico de Monterrey, 64849 Monterrey, Nuevo León, Mexico
mgonza@tec.mx

Abstract. With the growing breakthrough of deep learning-based face recognition, the development of lightweight models that achieve high accuracy with computational and memory efficiency has become paramount, especially for deployment on embedded domains. While Vision Transformers have shown significant promising results in various computer vision tasks, their adaptability to resource-constrained devices remains a significant challenge. This paper introduces SwiftFaceFormer, a new efficient, and lightweight family of face recognition models inspired by the hybrid SwiftFormer architecture. Our proposal not only retains the representational capacity of its predecessor but also introduces efficiency improvements, enabling enhanced face recognition performance at a fraction of the computational cost. We also propose to enhance the verification performance of our original most lightweight variant by using a training paradigm based on Knowledge Distillation. Through extensive experiments on several face benchmarks, the presented SwiftFaceFormer demonstrates high levels of accuracy compared to the original SwiftFormer model, and very competitive results with respect to state-of-the-art deep face recognition models, providing a suitable solution for real-time, on-device face recognition applications. Our code is available at https://github.com/Inria-CENATAV-Tec/SwiftFaceFormer.

Keywords: Lightweight Face Recognition · Efficient Vision Transformer · Knowledge Distillation · Efficient Face Transformer

1 Introduction

In the last decade, deep learning methods based on Convolutional Neural Networks (CNNs) have revolutionized the face recognition research landscape,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 244–258, 2025.
https://doi.org/10.1007/978-3-031-78341-8_16

achieving impressive levels of accuracy compared to "shallow" methods [29]. However, this increased performance often relies on a high model complexity, which makes it difficult to deploy on embedded devices or smartphones with memory and computational constraints, resulting in finding a suitable trade-off between speed and accuracy to be a significant challenge.

Designing efficient face recognition solutions, from lightweight deep learning architectures proposed for common computer vision tasks, has emerged as a great promising option. Models such as MobileFaceNet [5], ShuffleFaceNet [19], VarGFaceNet [31], MixFaceNets [2] and GhostFaceNets [1] have been built from MobileNetV2 [23], ShuffleNetV2 [17], VarGNet [32], MixNets [27] and GhostNets [11,28], respectively, reaching high levels of recognition accuracy with a low number of parameters and computational complexity.

On the other hand, it has been recently demonstrated that Transformer-based architectures can be incorporated into face recognition with promising results [35]. Although these methods are capable of capturing long-range relations among facial regions, the associated high computational costs due to the effective use of self-attention computation have restricted their usage in resource-limited domains. To address this issue, new hybrid models [9,12,15,26], that combine the strengths of both lightweight CNNs and Vision Transformer (ViT), have been introduced for face recognition, demonstrating that it is possible to meet real-time deployment in practical applications.

Recently, the SwiftFormer network [25] introduced a consistent hybrid design with an efficient additive attention mechanism to model the contextual information with linear complexity. Experiments on image classification, object detection, and segmentation tasks showed that this model achieves state-of-the-art (SOTA) performance with a good trade-off between accuracy and latency.

In this work, we present a new family of efficient and lightweight hybrid face models, namely SwiftFaceFormer, including five model variants with different levels of complexity. We adopt SwiftFormer [25] as a baseline network structure and adapt it for face recognition applications. Specifically, we leverage a Global Depthwise Convolution (GDC) layer followed by a convolution layer of size 1×1 and a batch normalization layer to produce a compact 512-dimensional feature vector in the embedding process. In addition to the four variants of the SwiftFormer model (XS, S, L1, L3), we introduce a new model variant (XXS), with lower computational complexity in terms of the number of floating-point operations (FLOPs), number of parameters, and model size. To enhance the recognition performance of this compact model, we apply hard knowledge distillation (KD) [4] to train our SwiftFaceFormer-XXS model to learn similar feature representations to the ones learned by a high-performance heavy network. Experiments on challenging benchmarks demonstrate the effectiveness and efficiency of SwiftFaceFormer in comparison to SOTA Lightweight CNNs, Vision Transformers, and hybrid models, showing its potential for deployment on resource-constrained face recognition applications.

The main contributions of our work are summarized as follows:

– We introduce a novel lightweight hybrid face architecture, called SwiftFaceFormer, which extends the efficient SwiftFormer network to the specific

domain of face recognition for real-time applications. The proposed hybrid network architecture leverages CNN and ViT capabilities through five model variants of different complexities.
- We extend the SwiftFaceFormer family of networks including an extremely-lightweight variant, named SwiftFaceFormer-XXS, that introduces an Efficient Convolutional Encoder with variable convolutional groups per stage. This approach heavily improves efficiency over the original SwiftFormer Convolutional Encoder and the rest of the SwiftFormer variants.
- To enhance the interpretation ability and the recognition performance of the most compact variant of our SwiftFaceFormer model (XXS), we apply the knowledge distillation paradigm. We provide two ablation studies about the effect of using different teacher models to learn feature representations and two different loss functions.
- We provide extensive experiments and comparisons with SOTA face models on different datasets including large-scale face recognition benchmarks such as IJB-B and IJB-C, showing the advantages of our proposed models in terms of both accuracy and efficiency.

The paper is organized as follows. Section 2 reviews the existing lightweight CNNs and ViT models for face recognition. Section 3 introduces the lightweight hybrid SwiftFaceFormer models tailored for face recognition. Experiments are presented in Sect. 4, followed by discussion and conclusion in Sect. 5.

2 Related Work

Here, we summarize existing approaches for developing face recognition models with low computational complexity that can be deployed on resource-restricted domains such as embedded devices or smartphones. We also give an overview of ViT models that have been proposed for face recognition, including those based on lightweight face recognition models.

2.1 Lightweight CNNs for Face Recognition

Designing small and efficient network architectures that reduce the computational effort in comparison to larger and more complex CNN models, has become a promising solution in recent years to achieve a better balance between speed and accuracy. In particular, for face recognition, the most common approach has been modifying lightweight networks originally designed for common computer vision tasks to the specific case of face recognition.

MobileFaceNet [5] and ShuffleFaceNet [19], which are based on MobileNetV2 [23] and ShuffleNetV2 [17], respectively, replace the Global Average Pooling (GAP) layer for a Global Depth-wise Convolution layer, and use the Parametric Rectified Linear Unit (PReLU) activation function instead of the Rectified Linear Unit (ReLU) function. Moreover, they adopt a fast downsampling strategy at the beginning of the networks, an early dimension-reduction strategy at the last

several convolutional layers, and a linear 1×1 convolution layer following a linear GDC layer as the feature output layer.

VarGFaceNet [31] improves the discriminative ability of VarGNet [32] by using an efficient variable group convolutional network for lightweight face recognition. In addition, to improve the interpretation ability of this lightweight network, a recursive knowledge distillation strategy is introduced. In a similar way, MixFaceNets [2] extend the MixConv [27] block with a channel shuffle operation aiming at increasing the discriminative ability. More recently, GhostFaceNets [1] extends two efficient neural architectures, GhostNetV1 [11] and GhostNetV2 [28], by replacing the GAP layer and the pointwise convolution layer with a modified GDC layer. They employ the PReLU activation function and replace the fully connected layers in the squeeze and excitation (SE) modules by convolution layers, to improve the discriminative power of their method.

Another strategy has been using Neural Architecture Search (NAS) [8] to automatically create efficient artificial neural networks specifically designed for face recognition. A family of extremely lightweight face models, namely PocketNets, was proposed in [4], aiming at automating the process of designing a neural network that achieved very competitive face recognition results. The authors also introduce a novel training paradigm based on knowledge distillation to ease the challenges caused by the significant gap between the teacher and student models, reducing the trade-off between model performance and compactness.

2.2 Vision Transformers for Face Recognition

In recent years, there has been a growing interest in the use of Vision Transformers (ViT) for different computer vision tasks, including face recognition. Face-Transformer [35] was the first attempt to investigate the performance of ViT models in face recognition, by introducing a Transformer model that uses sliding patches to capture inter-patch information from faces. Although this method achieves comparable performance to state-of-the-art CNNs, it is computationally heavy and unsuitable for low-resource environments.

Recently, CFormerFaceNet [12] combines a lightweight CNN face model with ViT. The authors designed a Group Depth-Wise Transpose Attention that used the CNN's ability to extract local facial features and the Transformer's capability to model global facial features, with lightweight modifications reducing computation requirements. In MobileFaceFormer [15], another hybrid method, both CNN and Transformer branches are parallelized in a dual branch design, and a bi-directional feature fusion bridge connecting dual branches is designed to concurrently retain local facial features and global facial interpretations. A convolutional token initialization method is proposed at the Transformer branch to perceive long-range facial information, enhancing feature interpretations. The CNN branch uses Depth-Wise Separable convolution and attention mechanisms are adopted to improve local facial feature extraction before an Attentive Global Depthwise Convolution (AGDC).

EdgeFace [9] presents a new hybrid model that adapts the EdgeNeXt architecture [18] for face recognition and introduces a Low Rank Linear (LoRaLin) module to further reduce the computation in linear layers while providing a minimal compromise to the performance of the network. In addition, a split depth-wise transpose attention (STDA) encoder is proposed to process input tensors and encode multi-scale facial features, while maintaining low computational costs and compact storage requirements.

HOTformer [26] is another novel face recognition model based on MobileFaceNets and ViTs that can effectively generate discriminative face representations by regional interaction of faces. The authors introduce two cooperation types of tokens named atomic tokens and holistic tokens to capture the region relationship of the face. Specifically, atomic tokens are generated by fixed-size patches to carry the fine-grained core representation, while holistic tokens are generated from adaptively spatial regions to aggregate information from several facial parts.

3 Approach

In this section, we detail the SwiftFaceFormer architecture specifically tailored for face recognition tasks. This approach is directly inspired by the SwiftFormer network [25], which achieves state-of-the-art performance in general-purpose computer vision tasks. Thus, we first describe the original SwiftFormer architecture, followed by the modifications introduced to make it an efficient and accurate face recognition model.

3.1 SwiftFormer Network

The SwiftFormer architecture [25] is a lightweight hybrid design that combines the strengths of CNNs and Transformers for real-time mobile vision applications. It builds on EfficientFormer [16] and improves the token mixing by using a simple yet effective Convolutional Encoder. This encoder replaces 3×3 average pooling layers used as a local token mixer by depth-wise convolutions, without increasing the parameters and latency. Moreover, SwiftFormer introduces an efficient additive attention module in the SwiftFormer Encoder module, to model the contextual information with linear complexity, that can be incorporated in all stages of the network. This leads to more consistent learning of local-global representations and significantly reduces the computational complexity.

The original SwiftFormer authors proposed four different configurations for this architecture, varying in complexity, named L3, L1, S, and XS. All of these versions use the same operators although with different depth and width levels. The depth configuration regulates the number of encoding operations, while the width level dictates the number of channels present in the feature map. In each stage, the network performs the Convolutional Encoder operations followed by the SwiftFormer Encoder step. After three stages of stacking multiple Encoder blocks, depending on the complexity, the feature map is averaged and its output

is sent to a Linear layer for classification tasks. This output before the linear layer is a Global Average Pooling (GAP) operation, common in other general-purpose architectures.

3.2 SwiftFaceFormer Architecture

To adapt the SwiftFormer architecture to the face recognition task, we introduce specific refinements to the original approach. As noted in previous works, the GAP operator is not suitable for the face recognition task, due to the averaging operation weighting all the inputs from the feature map equally, thus hindering the projection of non-linear features and the capacity to extract discriminative information present in open-set face recognition scenarios. This shortcoming is accentuated when using the GAP layer output before a Linear layer for final classification purposes. To mitigate this limitation, an alternative is to use an embedding head including a Batch Normalization and a Dropout step before the Linear layer [9].

Recent approaches [5,19,31] adopt the Global DepthWise Convolution to spatially reduce the feature map size and adjust the embedding dimension to the final embedding. To extend the SwiftFormer model to the face recognition scenario and improve its performance, we opt to adjust the output channels of the final SwiftFormer Encoding stage to the pre-defined face embedding dimension C using an efficient 1×1 point-wise convolution. We then employ the GDC layer with a 4×4 kernel size, reducing the spatial dimension from the SwiftFormer feature map to a vector, and employ another efficient 1×1 point-wise convolutional operator to output the final face embedding.

3.3 SwiftFaceFormer-XXS

When assessing SwiftFormer's efficiency performance, we noted a heavy load of convolutional operations in the Convolutional Encoders from the SwiftFormer architecture. As such, we analyzed the computation load in each one of the stages, noting that the earliest stages performed the most computations due to the larger spatial feature map sizes with two point-wise convolutional layers. Our approach consists of converting the last point-wise layer of the Convolutional Encoder into a grouped point-wise convolution. Our intuition for selecting the last point-wise convolutional layer for grouped convolutions instead of the first one, was to retain more input information with full convolutional operations and reduce the compromise on accuracy when using the following grouped convolution. We corroborated the effectiveness of our selection through experimentation.

We propose to employ a descending strategy for the number of groups at each stage. Using larger group sizes for the first stages heavily reduces the computation load and parameter count and leaving deeper stages less compromised achieves a reasonable balance between efficiency and accuracy. The last stage is left uncompromised with regular point-wise convolutions (groups $g = 1$) before the output to our face embedding head. Figure 1 shows our approach to this efficient Convolutional Encoder.

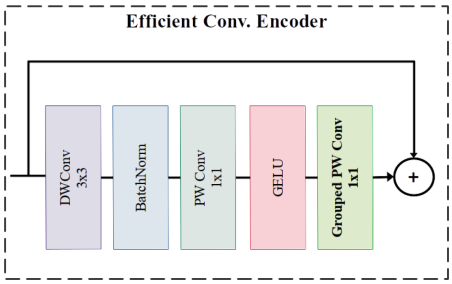


Fig. 1. Our Efficient Conv. Encoder in SwiftFaceFormer-XXS. We employ Grouped Point-Wise Convolutions only at the last layer for maximizing efficiency and mitigating accuracy penalties.

In addition, we reduced the depth regulating the number of encoding operations in stages 2 to 4, maintaining most of the operations in the third stage as in the original architecture. This further reduces the computation load with limited compromises to accuracy. Lastly, we adjusted the width (channels) of the feature maps starting from 16 in stage one to 128 in the last stage. We selected 128 as our final channel dimension as previous work [19] has suggested that 128 suffices for efficiently embedding facial features. Figure 2 illustrates the modifications for this efficient approach.

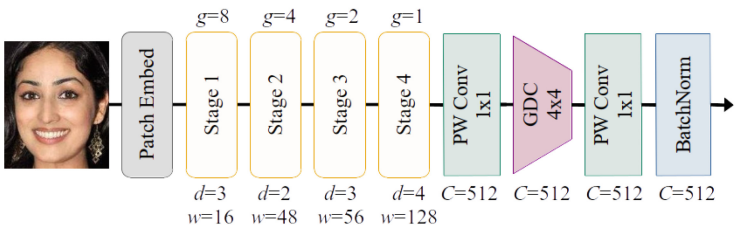


Fig. 2. SwiftFaceFormer-XXS overall architecture. Consistent with the original SwiftFormer notation for the stages, the complexity is expressed as depth d for the number of encoding operations and width w for the number of feature map channels. C denotes the embedding channel dimension for our face recognition head.

4 Experiments

In this section, we introduce the experimental setup of our proposed SwiftFaceFormer models and evaluate their recognition performance over several face benchmark datasets. In addition, we validate the accuracy improvements of the designed SwiftFaceFormer-XXS architecture through two ablation studies.

4.1 Datasets

We used the MS1M-RetinaFace dataset (MS1MV3) [6,7] for fine-tuning our SwiftFaceFormer models (XXS, XS, S, L1, L3) to the face recognition task. We choose to use this particular dataset to allow a fair evaluation of our method with the rest of the state-of-the-art face recognition benchmarks [7,14]. This dataset is a clean version of the MS-Celeb-1M dataset [10], which contains 5.1 million of face images collected from 93,431 identities. To evaluate the effectiveness and robustness of trained SwiftFaceFormer models, we employed several benchmarks including Labeled Faces in the Wild (LFW) [13], Celebrities in Frontal-Profile in the Wild (CFP-FP) [24], AgeDB-30 [22], Cross-age LFW (CALFW) [34], Cross-Pose LFW (CPLFW) [33], IARPA Janus Benchmark-B (IJB-B) [30] and IARPA Janus Benchmark-C (IJB-C) [21].

4.2 Implementation Details

For training our approach, we adopt a Stochastic Gradient Descent (SGD) optimizer with a batch size of 3×128 to improve training stability. We perform training on three Nvidia GeForce GTX A6000 GPUs. The learning rate is initialized to 0.05 and decreased by a factor of 10 periodically at epochs 8, 20, 25, and 30, training for a total of 34 epochs as in [4]. The momentum parameter is set to 0.9 and weight decay at 5e−4. The parameter-initialization method for convolutions is Xavier with random sampling from a Gaussian normal distribution. We use the ArcFace [6] loss function with an angular margin $m = 0.5$, which turned out to be the best for face recognition. All experiments are implemented on the Pytorch framework. We adopted the pre-trained weights on ImageNet from the original SwiftFormer models [25] to initialize our networks to achieve the best performance on face recognition tasks. During inference, the classification head of the SwiftFaceFormer models is removed and the resulting 512-D embedding is used for the comparisons.

On our Knowledge-Distillation approach for SwiftFormer-XXS, we employed hard-sample distillation with two separate headers, as in [25]. For optimizing the embeddings, we used the Mean Squared Error Loss scaled to 10^4 and the Cosine Distance scaled to 64, when applicable. A second separate header is added to compute the ArcFace loss. Each loss value is scaled to 0.5 for our optimization process. For verification, the embeddings from both headers are averaged and used for similarity scoring.

All face images used for both training and testing are detected and aligned as in [6], generating generate the face crops of 112×112, where each pixel (in [0, 255]) is normalized by subtracting 127.5 and then dividing by 128.

4.3 Results

We now present and discuss our experimental results focusing on comparisons with the original general-purpose SwiftFormer architecture and comparing our

approach with state-of-the-art face recognition models. We also conduct an ablation study of our Knowledge Distillation approach for bridging the accuracy gap between SwiftFormer-XXS and more computationally expensive models.

Comparison with SwiftFormer Architecture. To show the advantages of our proposed SwiftFaceFormer architecture for the specific case of face recognition, we compare it with the original SwiftFormer network. For a fair comparison, we trained SwiftFormer models (XS, S, L1, L3) under the same training setting as our SwiftFaceFormer models. In Table 1, we show the verification accuracy of the models on LFW, CFP-FP, AgeDB-30, CALFW and CPLFW datasets. In addition, the number of parameters (Params.) and the MFLOPs are given. It can be seen from the table that, for all variants (XS, S, L1, L3), the proposed SwiftFaceFormer outperforms the original SwiftFormer models, maintaining a very similar computational complexity. To validate our intuition for selecting the second point-wise convolutional layer as a grouped convolution instead of the first layer, in our Efficient Convolutional Encoder, we trained the SwiftFormer-XXS approach on MS1MV3 and tested face verification on the same datasets as above. Testing both possibilities, we found our approach yields an average increase of 0.5%, with a more notable verification accuracy difference of 1.55% and 1.27% on CPLFW and CFP-FP, respectively.

Table 1. Comparison of the recognition rates proposed SwiftFaceFormer with the original SwiftFormer models on popular face recognition benchmarks.

Method	LFW (%)	CFP-FP (%)	AgeDB-30 (%)	CALFW (%)	CPLFW (%)
SwiftFormer-L3	99.67	96.91	97.02	95.68	90.37
SwiftFaceFormer-L3 (ours)	**99.75**	**97.80**	**97.55**	**96.03**	**90.70**
SwiftFormer-L1	99.63	96.37	95.58	95.53	89.77
SwiftFaceFormer-L1 (ours)	**99.68**	**96.61**	**96.95**	**95.80**	**90.10**
SwiftFormer-S	99.42	95.40	95.90	95.25	88.72
SwiftFaceFormer-S (ours)	**99.60**	**96.49**	**96.83**	**95.78**	**90.00**
SwiftFormer-XS	99.45	95.10	94.92	94.98	88.55
SwiftFaceFormer-XS (ours)	**99.60**	**95.47**	**96.35**	**95.35**	**88.65**

Comparison with the State-of-the-Art. Table 2 presents a comparison between our proposed SwiftFaceFormer models (XXS, XS, S, L1, and L3) and previous state-of-the-art CNNs, Transformer-based, and hybrid models on seven face recognition benchmarks. The existing models are ordered according to the number of parameters (compactness), showing those with less than 4M parameters. Our most efficient SwiftFaceFormer models are presented at the end of the table. SwiftFaceFormer-XS and SwiftFaceFormer-XXS-KD are in with less than 4M. In the case of HOTformer-Net models [26], the number of parameters is unknown, however, the authors used HOTformer-Net (base) and HOTformer-Net (small) for comparisons with state-of-the-art lightweight models.

Verification results from this table reveal that our SwiftFaceFormer models obtain comparable performance to SOTA face recognition models from the literature. Among our models belonging to the first category, SwiftFaceFormer-L3 achieves the best performance. Although it is the most complex of our models, SwiftFaceFormer-L3 (28M parameters) achieves comparable results to other deeper CNN and ViT models with more than twice number of parameters and 10 times more FLOPs. For example, SwiftFaceFormer-L3 outperformed T2T-ViT, ViT-P10S8, and ViT-P8S8 models on the challenging CFP-FP and CALFW datasets. Our other two models, SwiftFaceFormer-L1 (11.8M parameters) and SwiftFaceFormer-S (6M parameters) perform very similarly to ResNet18-Q8-bit (24M parameters), obtaining even higher verification scores under pose variations from CFP-FP and CPLFW. Moreover, the SwiftFaceFormer-S model reaches the accuracy levels of GhostFaceNetV2-2.

Table 2. Comparison with the state-of-the-artViT, Hybrid, and CNN models on popular face recognition benchmarks. The models are ordered based on the number of parameters per approach type, and divided into $>$ 4M and $<$ 4M parameters. The FLOPs column shows the number of Floating Operations Per Second in millions. The IJB-B and IJB-C columns correspond to the verification TAR at FAR = 1e−4 on the IJB-B and IJB-C datasets, while the rest show verification accuracy (%).

Method	Type	Params. (M)	FLOPs (M)	LFW (%)	CFP-FP (%)	AgeDB-30 (%)	CALFW (%)	CPLFW (%)	IJB-B (%)	IJB-C (%)
T2T-ViT [35]	ViT	63.5	25,400	99.82	96.59	98.07	95.85	93.00	–	95.67
ViT-P10S8 [35]	ViT	63.3	24,800	99.77	96.43	97.83	95.95	92.93	–	96.06
ViT-P12S8 [35]	ViT	63.3	24,800	99.80	96.77	98.05	**96.18**	93.08	–	96.31
ViT-P8S8 [35]	ViT	63.2	24,800	99.83	96.19	97.82	95.92	92.55	–	95.96
ResNet100-ElasticFace [3,14]	CNN	65.2	24,211.8	99.80	98.73	**98.28**	**96.18**	93.23	95.43	96.65
ResNet100-ArcFace [14]	CNN	65.2	24,211.8	99.82	98.27	98.15	95.45	92.08	94.20	95.60
ResNet50-Q8-bit [14]	CNN	43.6	–	99.78	97.70	98.00	96.00	92.17	94.15	95.66
ResNet18-Q8-bit [14]	CNN	24.0	1,810	99.63	94.46	97.03	95.72	89.48	91.57	93.56
GhostFaceNetV2-1 [1]	CNN	6.9	272.1	**99.86**	**99.33**	98.62	96.11	**94.65**	**96.48**	**97.75**
GhostFaceNetV2-2 [1]	CNN	6.8	76.5	99.68	94.28	96.83	95.73	90.16	91.88	93.15
VarGFaceNet [14,31]	CNN	5.0	1,022	99.85	98.50	98.15	95.15	88.55	92.94	94.70
GhostFaceNetV1-1 [1,14]	CNN	**4.1**	215.7	99.73	96.83	98.00	95.93	91.93	93.12	94.94
GhostFaceNetV1-2 [1,14]	CNN	**4.1**	**60.3**	99.68	93.31	96.92	95.60	90.07	91.25	93.45
HOTformer-Net (large) [26]	Hybrid	–	2,840	99.80	98.80	98.20	95.90	92.90	95.30	96.60
MixFaceNet-M [2]	CNN	3.9	626.1	99.68	–	97.05	–	–	91.55	93.42
MixFaceNet-S [2]	CNN	3.1	451.7	99.60	–	96.63	–	–	90.17	92.30
ShuffleFaceNet [14,19,20]	CNN	2.6	577.5	99.67	97.26	97.32	95.05	88.50	92.25	94.30
MobileFaceNet [5,14,20]	CNN	2.0	933.3	99.70	96.90	97.60	95.20	89.22	92.83	94.70
PocketNetM-128-KD [4]	CNN	1.7	1,099	99.65	95.07	96.78	95.67	90.00	90.63	92.63
MixFaceNet-XS [2]	CNN	1.0	161.9	99.60	–	95.85	–	–	88.48	90.73
PocketNetS-128 [4]	CNN	**0.9**	587.1	99.50	93.78	95.88	95.01	88.93	88.29	90.79
PocketNetS-128-KD [4]	CNN	**0.9**	587.1	99.55	93.82	96.50	95.15	89.13	89.23	91.47
HOTformer-Net (base) [26]	Hybrid	–	1,301	99.70	**97.80**	97.60	**96.00**	91.90	**93.80**	95.50
HOTformer-Net (small) [26]	Hybrid	–	765	99.70	96.50	96.90	95.60	91.10	92.50	94.50
EdgeFace-S [9]	Hybrid	3.7	306.1	**99.78**	95.81	96.93	95.71	92.56	93.58	**95.63**
EdgeFace-XS [9]	Hybrid	1.8	154	99.73	94.37	96.00	95.28	91.82	92.67	94.85
CFormerFaceNet [12]	Hybrid	1.7	**40.0**	99.73	95.06	97.12	95.80	90.20	–	–
MobileFaceFormer [15]	Hybrid	1.4	–	99.60	96.79	**97.69**	95.98	**98.43**	–	–
SwiftFaceFormer-L3 (ours)	Hybrid	28.0	2,015.6	99.75	97.80	97.55	96.03	90.70	92.92	94.70
SwiftFaceFormer-L1 (ours)	Hybrid	11.8	804.6	99.68	96.61	96.95	95.80	90.10	91.81	93.82
SwiftFaceFormer-S (ours)	Hybrid	6.0	485.2	99.60	96.49	96.83	95.78	90.00	91.56	93.54
SwiftFaceFormer-XS (ours)	Hybrid	3.4	293.7	99.60	95.47	96.35	95.35	88.65	90.20	92.32
SwiftFaceFormer-XXS-KD (ours)	Hybrid	1.5	64.1	99.43	92.50	94.82	94.78	86.97	87.81	90.28

For the second category, the performance of our SwiftFaceFormer-XS and SwiftFaceFormer-XXS-KD models, demonstrate promising results on the evaluated benchmarks. For instance, SwiftFaceFormer-XS obtains as good verification

results as the hybrid EdgeFace-S model and the lightweight MixFaceNet-S CNN model. Also, it is able to achieve competitive results with respect to ResNet18-Q8-bit, which belongs to the first category. The use of the KD paradigm allows us to enhance the performance of our compact SwiftFaceFormer-XXS model, offering a good trade-off between efficiency and accuracy for deploying it in limited-resource devices.

As it can be appreciated, in general, we have developed novel hybrid face recognition models that perform well compared to the state-of-the-art, which demonstrates that combining the strengths of both lightweight CNNs and Transformers makes it possible to reduce the computational requirements for practical applications.

4.4 Ablation Study

The knowledge distillation (KD) paradigm enables a student model to learn from a teacher model, making it a popular technique for training lightweight models from more complex ones. Intending to enhance the performance of our compact model, SwiftFaceFormer-XXS, this section presents two ablation studies based on the hard simple distillation method. First, we show the effect of using different teacher networks for transferring their interpretation capabilities. Then, we evaluate the impact of using different loss functions during KD training.

Using Different Teacher Networks. We conduct experiments to investigate the effect of different teacher models on SwiftFaceFormer-XXS. We employ two pretrained and fully converged teacher networks, ResNet100-ArcFace [6] and SwiftFaceFormer-L3, respectively, to measure the relationship between teachers and student structures. Table 3 presents the recognition results on several face datasets. We observe that, introducing KD into the SwiftFaceFormer-XXS training phase improves its achieved verification performance on all evaluation benchmarks, especially on the large-scale IJB-B and IJB-C databases. Although the performance of teacher models is very similar, when SwiftFaceFormer-XXS is trained with KD using SwiftFaceFormer-L3 as the teacher network, the verification results are higher. This shows that, using a simplified version of the teacher network as student, reduces the model capacity gap between a large deep neural network and a small student neural network.

Using Different Loss Functions. To investigate the effect of loss functions, we train SwiftFaceFormer-XXS models with KD using Mean Square Error (MSE) and Cosine (COS) loss functions, respectively. Comparing the obtained results in Table 4, we can appreciate that for both loss functions, SwiftFaceFormer-XXS achieves very similar results. We choose the SwiftFaceFormer-XXS model trained with the MSE loss function since it offers more stable results during training.

Efficiency Assessment. To support our claim on real-time inference performance on edge device hardware, we performed latency experiments on our proposal. Table 5 presents the latency and single image throughput (FPS) for our SwiftFaceFormer models on the Nvidia Jetson Nano edge device, TensorRT, warmup time of 200 ms, and at least 10 inferences. We note that, our XXS-KD variant exhibits the lowest latency and the highest FPS. We also included the Average FR accuracy of the benchmarks of our method from Table 2 and divided it for the inference latency of our methods, calculating an "Accuracy per latency" score, to better assess the performance gains of our proposed method. We note a huge improvement of Accuracy per latency points with the XXS-KD variant, demonstrating its feasibility for usage on real-time hardware-constrained deployments, even though this could potentially increase training time if a teacher network is not already trained. Measuring the efficiency impact of our XXS design decisions, we observed that the width (w) adjustments resulted in a 71.36% reduction in FLOPs, grouped convs. amounted to a 11.23% reduction, and depth changes with a 12.81% reduction. Notably, the width adjustment in Stage 1 alone led to a 36.99% FLOPs reduction.

Table 3. Recognition rates and TAR@FAR = 1e−4 on IJB-B and IJB-C obtained by using KD for training SwiftFaceFormer-XXS model with different teacher models (R100-ArcFace and SwiftFaceFormer-L3) on popular face recognition benchmarks.

Method	LFW (%)	CFP-FP (%)	AgeDB-30 (%)	CALFW (%)	CPLFW (%)	IJB-B (%)	IJB-C (%)
SwiftFaceFormer-XXS (no KD)	99.22	90.94	92.75	94.02	85.70	81.07	82.80
ResNet100-ArcFace (Teacher)	99.82	98.27	98.15	95.45	92.08	94.20	95.60
SwiftFaceFormer-XXS (student)	**99.42**	92.00	**94.92**	**94.80**	86.42	87.30	89.80
SwiftFaceFormer-L3 (Teacher)	99.67	96.91	97.02	95.68	90.37	92.92	94.70
SwiftFaceFormer-XXS (student)	**99.43**	**92.50**	94.82	**94.80**	**86.97**	**87.81**	**90.28**

Table 4. Verification results obtained by using different loss functions (MSE and COS) for KD training of SwiftFaceFormer-XXS model on popular face recognition benchmarks. IJB-B and IJB-C report TAR@FAR = 1e−4.

Method	LFW (%)	CFP-FP (%)	AgeDB-30 (%)	CALFW (%)	CPLFW (%)	IJB-B (%)	IJB-C (%)
SwiftFaceFormer-L3 (Teacher)	99.67	96.91	97.02	95.68	90.37	92.92	94.70
SwiftFaceFormer-XXS (student)-MSE	99.43	**92.50**	94.82	**94.80**	**86.97**	**87.81**	90.28
SwiftFaceFormer-XXS (student)-COS	**99.48**	92.30	**95.20**	**94.80**	86.62	87.80	**90.30**

Table 5. Efficiency metrics in terms of latency, FPS throughput, number of parameters, Floating Operations Per Second (FLOPs), Average FR accuracy, and Accuracy per latency tested on the Nvidia Jetson Nano platform. Our XXS-KD variant shows remarkable efficiency performance across all metrics.

Method	Latency (ms)	FPS throughput	Params (M)	FLOPs (M)	Avg. FR Acc. (%)	Acc. per latency (%/ms)
SwiftFaceFormer-L3	36.9	27.1	28.0	2,015.6	**95.6**	2.6
SwiftFaceFormer-L1	18.0	55.3	11.8	804.6	95.0	5.3
SwiftFaceFormer-S	12.8	77.7	6.0	485.2	94.8	7.4
SwiftFaceFormer-XS	9.1	109.6	3.4	293.7	94.0	10.3
SwiftFaceFormer-XXS-KD	**4.6**	**215.5**	**1.5**	**64.1**	92.4	**20.1**

5 Conclusion

We introduce SwiftFaceFormer, a novel family of hybrid models using Lightweight Face CNN and Transformer architectures tailored for face recognition by adapting the SwiftFormer model and incorporating a Global Depth-Wise Convolution layer, followed by a 1×1 convolution layer and batch normalization to produce a compact 512-dimensional feature vector. Our most notable contribution is the design of our lightest version, SwiftFaceFormer-XXS, using grouped point-wise convolutions in specific sections of SwiftFormer's Convolution Encoders and progressively decrease the groups and width per stage for maximizing efficiency. Finally, using Knowledge Distillation on the XXS variant, we achieve a remarkable accuracy and efficiency trade-off for real-time resource-constrained devices.

Acknowledgment. This work was partially funded by the SOTERIA H2020 project. SOTERIA received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No101018342. This content reflects only the author's view. The European Agency is not responsible for any use that may be made of the information it contains.

The authors would like to thank the financial support from Tecnologico de Monterrey through the "Challenge-Based Research Funding Program 2022". Project ID # E120 - EIC-GI06 - B-T3 - D.

References

1. Alansari, M., Hay, O.A., Javed, S., Shoufan, A., Zweiri, Y., Werghi, N.: GhostFaceNets: lightweight face recognition model from cheap operations. IEEE Access **11**, 35429–35446 (2023)
2. Boutros, F., Damer, N., Fang, M., Kirchbuchner, F., Kuijper, A.: MixFaceNets: extremely efficient face recognition networks. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE (2021)
3. Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A.: ElasticFace: elastic margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1578–1587 (2022)
4. Boutros, F., Siebke, P., Klemt, M., Damer, N., Kirchbuchner, F., Kuijper, A.: PocketNet: extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. IEEE Access **10**, 46823–46833 (2022)
5. Chen, S., Liu, Y., Gao, X., Han, Z.: MobileFaceNets: efficient CNNs for accurate real-time face verification on mobile devices. In: Zhou, J., et al. (eds.) CCBR 2018. LNCS, vol. 10996, pp. 428–438. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97909-0_46
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and pattern Recognition, pp. 4690–4699 (2019)
7. Deng, J., Guo, J., Zhang, D., Deng, Y., Lu, X., Shi, S.: Lightweight face recognition challenge. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)

8. Elsken, T., Metzen, J.H., Hutter, F.: Neural architecture search: a survey. J. Mach. Learn. Res. **20**(1), 1997–2017 (2019)
9. George, A., Ecabert, C., Shahreza, H.O., Kotwal, K., Marcel, S.: EdgeFace: efficient face recognition model for edge devices. IEEE Trans. Biometr. Behav. Identity Sci. **6**(2), 158–168 (2024)
10. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 87–102. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_6
11. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: GhostNet: more features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1580–1589 (2020)
12. He, L., He, L., Peng, L.: CFormerFaceNet: efficient lightweight network merging a CNN and transformer for face recognition. Appl. Sci. **13**(11), 6506 (2023)
13. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition (2008)
14. Kolf, J.N., et al.: EFaR 2023: efficient face recognition competition. In: 2023 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–12. IEEE (2023)
15. Li, J., Zhou, L., Chen, J.: MobileFaceFormer: a lightweight face recognition model against face variations. Multimedia Tools Appl. **83**(5), 12669–12685 (2024)
16. Li, Y., et al.: EfficientFormer: vision transformers at MobileNet speed. In: Advances in Neural Information Processing Systems 35, pp. 12934–12949 (2022)
17. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 116–131 (2018)
18. Maaz, M., et al.: EdgeNeXt: efficiently amalgamated CNN-transformer architecture for mobile vision applications. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13807, pp. 3–20. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-25082-8_1
19. Martinez-Diaz, Y., Luevano, L.S., Mendez-Vazquez, H., Nicolas-Diaz, M., Chang, L., Gonzalez-Mendoza, M.: ShuffleFaceNet: a lightweight face architecture for efficient and highly-accurate face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
20. Martinez-Diaz, Y., et al.: Benchmarking lightweight face architectures on specific face recognition scenarios. Artif. Intell. Rev. **54**(8), 6201–6244 (2021)
21. Maze, B., et al.: IARPA Janus Benchmark-C: face dataset and protocol. In: 2018 International Conference on Biometrics (ICB), pp. 158–165. IEEE (2018)
22. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: AgeDB: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 51–59 (2017)
23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
24. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9. IEEE (2016)
25. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: SwiftFormer: efficient additive attention for transformer-based real-time mobile vision applica-

tions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17425–17436 (2023)
26. Su, W., Wang, Y., Li, K., Gao, P., Qiao, Y.: Hybrid token transformer for deep face recognition. Pattern Recogn. **139**, 109443 (2023)
27. Tan, M., Le, Q.V.: MixConv: mixed depthwise convolutional kernels. arXiv preprint arXiv:1907.09595 (2019)
28. Tang, Y., Han, K., Guo, J., Xu, C., Xu, C., Wang, Y.: GhostNetV2: enhance cheap operation with long-range attention. In: Advances in Neural Information Processing Systems 35, pp. 9969–9982 (2022)
29. Wang, M., Deng, W.: Deep face recognition: a survey. Neurocomputing **429**, 215–244 (2021)
30. Whitelam, C., et al.: IARPA Janus Benchmark-B face dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 90–98 (2017)
31. Yan, M., Zhao, M., Xu, Z., Zhang, Q., Wang, G., Su, Z.: VarGFaceNet: an efficient variable group convolutional neural network for lightweight face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
32. Zhang, Q., et al.: VarGNet: variable group convolutional neural network for efficient embedded computing. arXiv preprint arXiv:1907.05653 (2019)
33. Zheng, T., Deng, W.: Cross-Pose LFW: a database for studying cross-pose face recognition in unconstrained environments. Technical report **5**(7), Beijing University of Posts and Telecommunications (2018)
34. Zheng, T., Deng, W., Hu, J.: Cross-Age LFW: a database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)
35. Zhong, Y., Deng, W.: Face transformer for recognition. arXiv preprint arXiv:2103.14803 (2021)

MIHNet: Multi-scale Intra-layer Fusion with Hybrid Structure for Low-Quality 3D Face Recognition

Yuting Hu, Yue Ming(✉), Panzi Zhao, and Jiangwan Zhou

Beijing Key Laboratory of Work Safety Intelligent Monitoring, School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
{huyuting,yming,zpz,zhoujiangwan}@bupt.edu.cn

Abstract. Low-quality 3D face recognition (FR) is a crucial application in human-computer interaction. However, it is sensitive to changes in 3D shapes and textures caused by time, resulting in poorer recognition performance on the time subset. To solve the problem, we first propose a novel time robust feature extractor (TFE), which introduces an improved Transformer block to mitigate interference from texture and shape changes. Additionally, TFE includes a novel feature mixer (GL-Mixer), which effectively integrates local and global features while reducing redundancy. Finally, we utilize TFE to construct MIHNet, a multi-scale intra-layer fusion network with a hybrid structure for low-quality 3D FR. Experiments on two publicly available low-quality datasets and one cross-quality dataset demonstrate that MIHNet achieves competitive recognition accuracy, particularly attaining state-of-the-art (SOTA) performance on the time subset.

Keywords: Low-quality 3D FR · Time robust feature extractor · Hybrid structure · Intra-layer fusion

1 Introduction

Low-quality 3D FR, profiting from the 3D faces' rich geometric information, has become popular in biometrics. However, most existing 3D FR [1–5] focus on improving recognition performance under various poses and occlusions while overlooking the impact of time variations on recognition accuracy. Specifically, facial changes over time can result in a decline in recognition performance. Therefore, investigating how to tackle the challenges posed by temporal variations is essential for improving the practicality and reliability of 3D FR technology.

The work presented in this paper was partly supported by Natural Science Foundation of China (Grant No. 62076030), Beijing Natural Science Foundation (Grant No. L241011) and basic research fees of Beijing University of Posts and Telecommunications (Grant No. 2023ZCJH08).

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 259–274, 2025.
https://doi.org/10.1007/978-3-031-78341-8_17

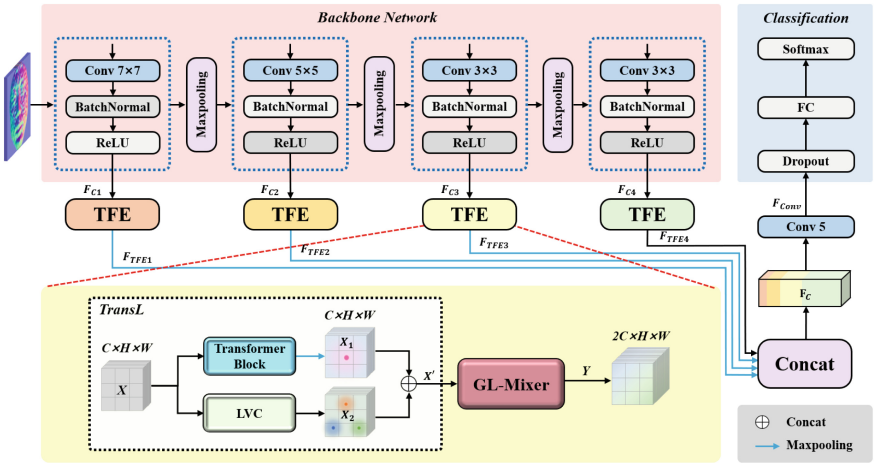


Fig. 1. MIHNet consists of three parts: backbone network, TFE module, and classification. The TFE module includes a dual-stream module, TransL, for extracting global and local features, and a feature fusion module, GL-Mixer.

To enhance the performance of the time subset, multi-modal low-quality 3D FR utilizes RGB face images to compensate for the deficiency of texture information in 3D face data. Uppal et al. [2] proposed a deep-guided attention mechanism, enabling the network to focus on critical facial regions in RGB images with the assistance of depth features. Nonetheless, there is interference in the matching and fusion of information between modalities, which can lead to insufficient acquisition of discriminative features. Some methods employed single-modal 3D face data as input to avoid interference. MQFNet [3] combined high-quality 3D face image output from the pix2pix network with low-quality 3D face images, feeding them into a multi-quality fusion network to improve the performance of 3D FR. LMFNet [4] integrated high-level and low-level features separately, effectively reducing the interference of information between layers. While these single-modal low-quality 3D FR models mitigate inter-modal interference, they lose a lot of discriminant facial details information. Consequently, the total and subsets recognition performance fails to meet expectations, particularly when dealing with time subsets.

Inspired by recent studies [9–11] applying CNN combined with Transformer [12] to the field of image classification, we propose a multi-scale intra-layer fusion network (MIHNet) based on a CNN-Transformer hybrid structure for low-quality 3D FR, as illustrated in Fig. 1. Firstly, we constructed a time robust feature extractor (TFE) which employs joint guidance of global and local information, consisting of a feature extraction module (TransL) and a feature mixer (GL-Mixer). TansL is a dual-stream feature extraction module, with one branch incorporating the LVC module proposed by [13] and the other branch featuring an improved Transformer Block, which leverages the designed differentiated spatial

reduction attention module (DSRA) to amplify the distinctions between layers. Secondly, a novel feature mixer (GL-Mixer) is designed to minimize redundancy in global and local feature fusion while enhancing the expressive ability of the intra-layer fused features. Finally, building upon the TFE, time robust features at different scales are collected and fused to capture abundant and discriminative texture and shape information, improving performance specifically on the time subset. The experiments conducted on the low-quality datasets Lock3DFace, KinectFaceDB and the cross-quality dataset formed by Bosphorus show that our proposed model achieves competitive total recognition accuracy and the SOTA recognition accuracy on the time subset.

In summary, our contributions can be summarized as follows:

- We propose a multi-scale intra-layer feature fusion network, which is based on the structure of a CNN-Transformer. Through the joint guidance of global and local features, the network extracts robust features at different scales that remain invariant over time and improves the performance of low-quality 3D FR.
- A novel time robust feature extractor (TFE) is developed that contains TransL and GL-Mixer module. TransL leverages a dual-stream network to extract rich global and local features while enhancing inter-layer feature diversity. GL-Mixer, with its residual structure and spatial-channel reconfiguration units, fortifies the intra-layer fusion features' representation ability and diminishes redundancy in the fusion of global and local features.
- Extensive experiments on two low-quality datasets and one cross-quality dataset validate the competitiveness of our approach in total recognition accuracy and achieve the highest recognition accuracy on the time subset.

2 Related Works

In this section, we review some methods for low-quality 3D FR. Then, we introduce recent works that apply Transformer to FR.

2.1 Low Quality 3D FR

Research on deep learning-based low-quality 3D FR is relatively scarce. Mu et al. [14] constructed a lightweight network to achieve real-time low-quality 3D FR. To diminish the RGB-based FR methods' sensitivity to geometric information changes induced by facial variations, Uppal et al. [2] utilized depth features extracted by CNNs to guide the network's attention towards discriminative facial regions in RGB images. Zhao et al. [4] developed a lightweight multi-scale fusion network with a hierarchical structure (LMFNet), achieving a balance between model performance and lightweight. Zhu et al. [6] proposed a progressive multi-modal fusion framework (PMMF) that combines features from both RGB and depth images after refining low-quality depth. Niu et al. [7] introduced an improved residual network based on image and curvature, which

enhances recognition accuracy under low-light or no-light conditions. Xu et al. [8] proposed an innovative depth map denoising network (DMDNet) and utilized Led3D [14] to extract complementary features from two modalities. While existing methods have enhanced total and subset recognition accuracy, the recognition accuracy for the time subset remains comparatively low. The introduction of Transformer [12] holds the potential to address this problem.

2.2 Transformer in FR

Transformer was originally developed to address natural language processing tasks. Due to its outstanding performance, Dosvitskiy et al. [15] applied Transformer to image classification, sparking subsequent research in the field of face-related studies. Sun et al. [16] proposed an efficient self-supervised framework for deep facial expression recognition, which involves explicitly modeling time facial movements during the fine-tuning process of ViTs [15], aiming to reduce computational costs while maintaining high performance. Nguyen et al. [17] introduced a diagonal micro attention (DMA) mechanism that accurately identifies subtle facial changes between consecutive video frames. Zhong et al. [18] refined the token generation method in ViTs, showcasing that Transformer model attains comparable performance to CNNs when possessing similar numbers of parameters and computational complexity. He et al. [44] incorporated Transformer into CNN networks to compensate for the CNNs' shortcomings in attending to critical facial feature regions. Dan et al. [19] introduced a patch-level data augmentation strategy to increase sample diversity, encouraging ViTs to utilize patches that are easily overlooked by deep networks, effectively improving facial recognition performance. These models are all proposed for 2D FR. To achieve 3D FR, Zheng et al. [1] subsequently proposed a complementary multi-modal fusion Transformer (CMMF-Trans) network, enhancing the robustness of FR to variations in lighting, pose, and time. Although these methods achieve FR with high accuracy, compared with the single-mode low-quality 3D FR model, the multi-modal CNN-Transform method has very high parameters and computational complexity, which makes it unable to be applied in practice. However, there is limited research on applying transformers to single-modal low-quality 3D FR. Therefore, we propose a single-mode intra-layer feature fusion network based on CNN-Transformer structure for low-quality 3D FR.

3 Method

In this section, we describe the entire network architecture and provide a detailed explanation of its constituent module: time robust feature extractor (TFE), as well as the key module Transformer Block and GL-Mixer in TFE.

3.1 Total Network Architecture

As depicted in Fig. 1, MIHNet comprises three components: the backbone network, the time robust feature extractor (TFE) guided by both global and

local information, and the classification. Firstly, to present a good compromise between performance and model complexity, we employ four feature extraction blocks as the backbone network, each block containing a convolutional layer, a batch normalization layer, and a ReLU activation function. We opt for kernel sizes of 7 and 5 for the convolutions in the first and second feature blocks, respectively, aiming to obtain more facial detail information from the input images, while the kernel sizes in the third and fourth feature blocks are set to 3. Secondly, to extract more effective time-robust features from feature maps of various scales, we utilize the four most critical time robust feature extractors (TFE) in MIHNet. These extractors process the features $\mathbf{F}_{C1}$, $\mathbf{F}_{C2}$, $\mathbf{F}_{C3}$, and $\mathbf{F}_{C4}$ of each layer output by the backbone network individually, resulting in the processed features $\mathbf{F}_{TFE1}$, $\mathbf{F}_{TFE2}$, $\mathbf{F}_{TFE3}$, and $\mathbf{F}_{TFE4}$. Subsequently, the four processed features are concatenated according to the channel dimension to obtain the fused feature $\mathbf{F}_C$. Finally, $\mathbf{F}_C$ is fed into Conv5 for final fusion to obtain $\mathbf{F}_{Conv}$, which is then passed into the classifier for model training.

3.2 Time Robust Feature Extractor (TFE)

To strengthen model's capacity to learn robust features from images captured at intervals in time. In this work, we propose a module for extracting time robust features, termed the time robust feature extractor (TFE). It captures both global and local information from the input feature map, guiding the model to learn invariance from geometric changes in face images. TFE consists of a dual-stream feature extraction module (TransL) and a feature mixer (GL-Mixer), where the TransL module comprises two branches: a LVC module for extracting local features and a Transformer Block for extracting global features. The entire workflow of the proposed TFE is depicted in Fig. 1. Specifically, the feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ obtained from the feature extraction block is fed into both the LVC module and our improved Transformer Block, resulting in corresponding feature maps $\{\mathbf{X}_1, \mathbf{X}_2\} \in \mathbb{R}^{(C\times H\times W)}$. Subsequently, these feature maps are concatenated along the channel dimension to yield the feature map $\mathbf{X'} \in \mathbb{R}^{2C\times H\times W}$. Finally, the fused feature $\mathbf{Y}$ is accomplished using GL-Mixer. The entire process of TFE can be formulated as:

$$\begin{aligned} \mathbf{X}_1 &= \text{Trans}(\mathbf{X}) \\ \mathbf{X}_2 &= \text{LVC}(\mathbf{X}) \\ \mathbf{X'} &= \text{Concat}(\mathbf{X}_1, \mathbf{X}_2) \\ \mathbf{Y} &= \text{Mixer}(\mathbf{X'}) \end{aligned} \tag{1}$$

where Trans$(\cdot)$ and LVC$(\cdot)$ refer to transformer block and LVC module, Concat$(\cdot)$ denotes connection along channel, and Mixer$(\cdot)$ stands for the GL-Mixer.

3.3 The Transformer Block

The texture and shape variations due to the time interval are one of the factors leading to the performance degradation of low-quality 3D FR time subsets. To

address this problem, we followed the construction paradigm of the Transformer block to created a novel Transformer Block consisting of patch embedding and stacked three residual blocks, as shown in Fig. 2, which includes dynamic positional encoding (DPE) [21], differential spatial reduction attention (DSRA), and inverted residual FFN (IRFFN) [22]. Initially, a patch embedding layer, consisting of a convolutional layer and a batch bormalization layer, is employed to downsize the intermediate features, facilitating better processing of large-sized image inputs. DPE subsequently integrates the 3D positional information from the feature maps into all tokens. Then, we utilize our designed DSRA to capture relationships between different positions. It allows the model to enhance the distinction between features across different layers, thereby reducing the interference of redundant inter-layer features and extracting more discriminative semantic information. Finally, IRFFN enhances gradient propagation across layers without significantly increasing computational cost and outputs the global features $\mathbf{F}_G$.

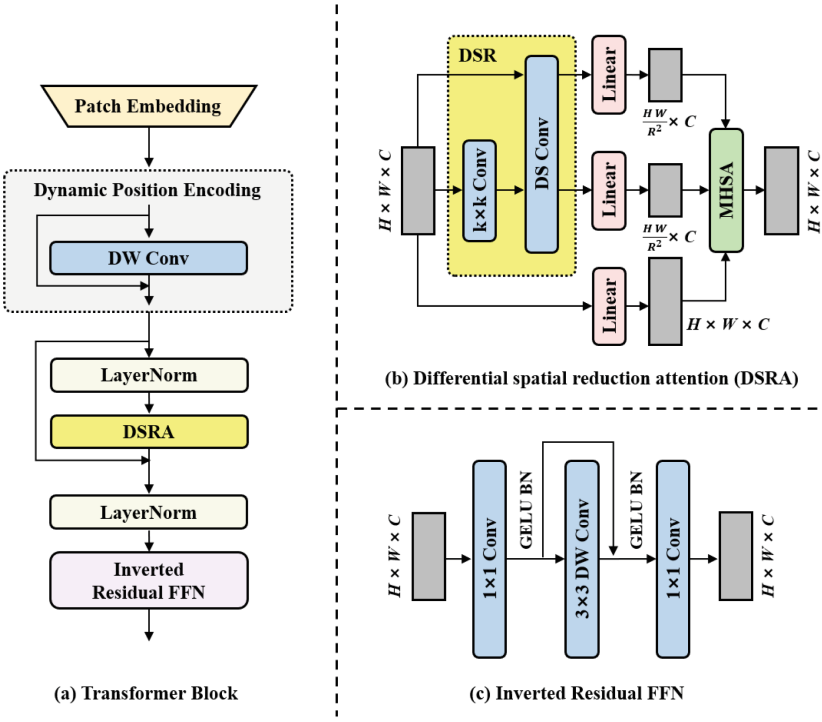


Fig. 2. Workflow of the proposed TFE module.

3.4 Differential Spatial Reduction Attention (DSRA)

For specific 3D FR tasks, it's crucial to extract multi-scale features at different layers while enhancing the diversity among these features. Considering the advantages of spatial reduction attention (SRA) [23] and its improved methods [22–25], we propose a differential spatial reduction attention (DSRA) to meet this demand, as depicted in Fig. 2(b), which similarly involves linearly transforming the input $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ into query $\mathbf{Q} \in \mathbb{R}^{\frac{HW}{R^2} \times C_q}$, key $\mathbf{K} \in \mathbb{R}^{\frac{HW}{R^2} \times C_k}$, value $\mathbf{V} \in \mathbb{R}^{\frac{HW}{R^2} \times C_v}$. The symbols C_q, C_k, and C_v represent the dimensions of the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, correspondingly. R denotes the reduction ratio of the DSA.

In contrast, to enhance feature disparities across various layers and reduce inter-layer interference, we apply a multi-scale operation to X before projecting it into V. This operation can be implemented as a convolution, with each layer using convolutional kernels of varying sizes $\{7, 5, 3, 1\}$. To mitigate computational complexity, we utilize depth-wise separable convolutions to reduce the spatial scale of $\mathbf{K}$ and $\mathbf{V}$. The DSRA is calculated by:

$$\begin{aligned} (\mathbf{K}', \mathbf{V}') &= \mathrm{DSR}(\mathbf{X}) \\ \mathbf{Q} &= \mathrm{Linear}(\mathbf{X}) \\ \mathbf{K} &= \mathrm{Linear}(\mathbf{K}') \\ \mathbf{V} &= \mathrm{Linear}(\mathbf{V}') \\ \mathrm{DSRA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{B}\right)\mathbf{V} \end{aligned} \tag{2}$$

where $\mathrm{Linear}(\cdot)$ represents a linear transformation. $\mathbf{B}$ denotes a learnable attention bias matrix used for positional encoding. $\mathbf{K}' \in \mathbb{R}^{\frac{HW}{R^2} \times C_k}$ and $\mathbf{V}' \in \mathbb{R}^{\frac{HW}{R^2} \times C_v}$ are the outputs obtained after spatial $\mathrm{DSR}(\cdot)$. $\mathrm{DSR}(\cdot)$ denotes the operation of enhancing feature differences and reducing spatial dimensions of the input, which can be written as:

$$\mathrm{DSR}(\mathbf{X}) = (\mathrm{DSConv}(\mathrm{kConv}(\mathbf{X})), \mathrm{DSConv}(\mathbf{X})) \tag{3}$$

where $\mathrm{kConv}(\cdot)$ indicates convolution with $k \times k$ kernel, while $\mathrm{DSConv}(\cdot)$ refers to depthwise separable convolution.

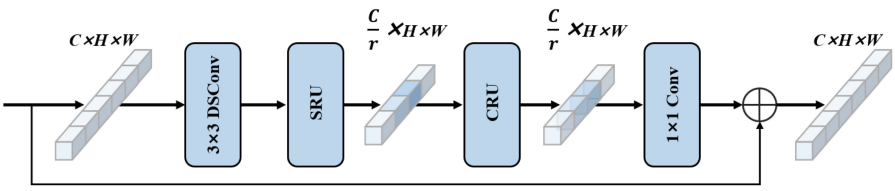


Fig. 3. The architecture of the GL-Mixer which is responsible for fusing global and local features.

3.5 The Global and Local Feature Mixer (GL-Mixer)

We construct a novel feature fusion mixer (GL-Mixer) with a residual structure in Fig. 3, aiming to eliminate redundant features while leveraging both global features for rich contextual information and local features for fine-grained details to learn time robust features. To reduce computational complexity and channel dimensions, we process the input feature maps utilizing a 3×3 depth-wise separable convolution. The spatial reconstruction unit (SRU) and channel reconstruction unit (CRU) proposed in [26] are employed here to enhance feature expression and reduce redundancy. We employ a 1×1 convolution to restore the number of channels. Additionally, the added residual structure effectively preserves the representational capacity of the original features. The above processes are expressed as:

$$\text{Mixer}(\mathbf{X}') = \text{Conv}_{1\times1}^{(\frac{C}{r}\to C)}\left(\text{CRU}(\text{SRU}(\text{DSConv}_{3\times3}^{(C\to\frac{C}{r})}(\mathbf{X}')))\right) + \mathbf{X}' \quad (4)$$

where r is the channel reduction ratio in GL-Mixer, $\text{SRU}(\cdot)$ and $\text{CRU}(\cdot)$ represents the spatial reconstruction unit and the channel reconstruction unit.

4 Experiments

In this section, we introduce the datasets utilized, training specifics, and experimental results across various datasets to validate the effectiveness of time robust feature extraction.

4.1 Datasets

We evaluate our approach on four public datasets: Lock3Dface [27], KinectFaceDB [28], FRGC v2.0 [29], Bosphorus [30,39]. The first two are low-quality 3D face datasets, while the last two are a high-quality 3D face datasets. Lock3DFace dataset, captured by Kinect V2 camera, stands as the largest low-quality 3D facial dataset currently available, comprising 5,671 videos sequence from 509 subjects, with each video sequence containing 59 frames. This dataset covers diverse facial variations such as neutral, expression, occlusion, pose and time changes. KinectFaceDB comprises 936 images of 52 subjects from six different ethnicities, and exhibits diverse facial types, including neutral, smile, mouth opening, illumination, eye occlusion, mouth occlusion, right face occlusion, left contour, right contour and time changes. FRGC v2.0 is exclusively used for training and Bosphorus is a high-quality dataset that focuses on facial variations related to expression, pose, and occlusion.

4.2 Preprocessing and Implementation Details

Preprocessing: For Lock3DFace, we follow the preprocessing steps outlined in [14], where we generate a 3D point cloud of the face from the original depth map,

Table 1. The recognition accuracy on Protocol 1 of the Lock3DFace dataset, where NU, FE, PS, OC, TM represents neutral, expression, pose, occlusion, time, respectively. AVG denotes the average rate and bold indicates the best result, double underline indicates the second best result, and tilde indicates the third best result.

Year	Methods	Accuracy				
		FE	PS	OC	TM	AVG
2018	MobileNet-V2 [40]	85.38%	32.77%	28.30%	10.60%	44.92%
2019	Led3D [14]	86.94%	48.01%	37.67%	26.12%	54.28%
2021	Zhang et al. [20]	92.38%	49.30%	43.34%	31.80%	58.68%
2021	Xiao et al. [43]	89.88%	45.32%	47.04%	38.76%	59.03%
2021	MQFNet [3]	90.55%	**52.81%**	44.64%	22.65%	61.04%
2022	LMFNet [4]	90.56%	49.02%	53.14%	31.97%	62.11%
2023	DSNet [38]	92.49%	49.14%	**53.29%**	33.18%	62.26%
2024	**Ours**	**92.57%**	50.17%	52.96%	**39.46%**	**62.98%**

centered around the tip of the nose. Subsequently, we perform outlier removal, hole filling, and normalization to produce 128×128 normal maps for training and testing. To mitigate overfitting due to limited 3D data, we additionally employ data augmentation techniques in [14], including pose generation, shape jittering, and scaling. Processing KinectFace involves generating depth, azimuth, and elevation maps using the method described in [41], followed by synthesizing three-channel images to augment the data. The method in [14] is utilized to synthesize cross-quality datasets from Bosphorus.

Implementation Details: We train and evaluate the MIHNet model using the PyTorch platform equipped with NVIDIA GeForce RTX 4090. The SGD optimizer with an initial learning rate of 10^{-2} is chosed and the learning rate decreases by 10% every 10 epochs during training. We set the batchsize to 32 and epoch to 100.

4.3 Experiments on Three Datasets

Results on Lock3DFace: To validate the effectiveness of MIHNet, we employ the Lock3DFace dataset following the protocols proposed in [14]. Protocol 1 involved training solely on the first neutral video from 509 subjects along with its augmented data. Subsequently, six frames were selected from each remaining video for testing the model. Protocol 2 divided the dataset into training and testing sets based on the number of subjects. Specifically, the model was trained on the original data and corresponding augmented data of 340 subjects, while the remaining 169 subjects had six frames from each original video used for testing. Table 1 and Table 2 respectively display the results of Protocol 1 and Protocol 2.

As summarized in Table 1, the proposed model outperforms the current existing SOTA methods. The average (AVG), expression (FE) and time (TM) recognition rates are 0.72%, 0.08% and 0.7% higher than the second-best results. The performance of the pose (PS) subset is slightly lower than [3], while the accuracy of the occlusion (OC) subset falls slightly behind [38]. This may be due to MQFNet [3] incorporating both the high-quality 3D face recovered from low-quality 3D faces and the original low-quality 3D face into the model, enhanc-

Table 2. The recognition accuracy on Protocol 2 of the Lock3DFace dataset. Total denotes the total rate, bold indicates the best result, double underline indicates the second best result, and tilde indicates the third best result.

Year	Methods	Input	Rank-1 Accuracy					
			NU	FE	PS	OC	TM	Total
2018	MobileNet-V2 [40]	Depth	98.91%	95.74%	69.92%	61.44%	43.00%	79.49%
2018	Cui et al. [32]	Depth	99.55%	98.03%	65.26%	81.62%	55.79%	79.85%
2019	Hu et al. [33]	Depth	90.90%	-	-	-	-	76.70%
2019	Led3D [14]	Depth	99.62%	97.62%	64.81%	68.93%	64.97%	81.02%
2021	3D-FRM [34]	Depth	99.93%	96.77%	69.06%	65.16%	74.19%	82.27%
2021	MQFNet [3]	Normal	99.95%	97.31%	73.61%	80.97%	61.67%	86.55%
2022	LMFNet [4]	Normal	99.95%	99.01%	78.61%	83.08%	75.60%	87.49%
2022	PointFace [5]	XYZ	99.35%	97.93%	72.03%	77.06%	66.33%	84.78%
2023	Niu et al. [7]	Depth	-	-	-	-	-	78.24%
2023	LDNFNet [8]	Depth	**100%**	99.37%	**82.94%**	**85.39%**	71.95%	**87.71%**
2024	**Ours**	Normal	99.95%	**99.43%**	80.02%	83.10%	**76.37%**	87.52%

Table 3. The results on the KinectFaceDB dataset. P1 consists of seven variations obtained from session 2, including neutral, smile, mouth opening, illumination, and three types of occlusions: paper occlusion, mouth occlusion, and eye occlusion. P2 comprises three variations without occlusions in session 2. P3 exclusively contains neutral faces in session 2. Total denotes the total rate, bold indicates the best result, double underline indicates the second best result, and tilde indicates the third best result.

Year	Ref.	Accuracy		
		P1	P2	P3
2019	3DLBP+CNN [31]	80.9%	90.2%	96.1%
2019	Led3D [14]	76.2%	87.5%	89.7%
2020	CABNet-FE [35]	67.3%	79.4%	82.7%
2020	CNN+BSIF [36]	93.6%	99.4%	90.2%
2022	LMFNet [4]	94.9%	95.5%	96.3%
2023	Neto et al. [37]	73.4%	82.2%	86.5%
2023	DSNet [38]	95.0%	**99.5%**	96.3%
2024	**Ours**	**95.3%**	**99.5%**	**96.9%**

ing the expression of 3D facial shape information. Unfortunately, the fusion of these two types of 3D facial features may introduce significant interference in facial texture and detail information, resulting in a performance drop of 2.02% and 16.81% on the FE and TM subsets, respectively, compared to our model. DSNet [38] adopted a dense connectivity approach to obtain richer multi-scale information, enabling the learning of more features details and contour information from facial images with local regions missing. Likewise, DSNet [38] overlooked interference between inter-layer features and was sensitive to shape and texture features. Our approach outperforms DSNet by 6.28% on the TM subset, significantly achieves an improvement of 0.7% compared to the top-performing result on the time (TM) subset, significantly demonstrating the superiority of MIHNet in handling face images with time change.

Table 4. The results on the cross-quality dataset. HL: the gallery is high-quality and the probe is low-quality. LL: gallery and probe are both low-quality. Total denotes the total rate, bold indicates the best result, double underline indicates the second best result, and tilde indicates the third best result.

Year	Ref.	Accuracy	
		HL	LL
2018	MobileNet-V2 [40]	79.12%	78.09%
2019	Led3D [14]	91.27%	90.70%
2021	PointFace [5]	92.96%	91.86%
2022	LMFNet [4]	93.64%	92.42%
2023	DSNet [38]	93.71%	92.46%
2024	Ours	**94.13%**	**92.85%**

Table 2 shows that our model attains recognition accuracy that is on par with the most advanced model currently. Specifically, MIHNet performs exceptionally well in FE and TM, surpassing the second-best recognition accuracy by 0.06% and 0.77%, respectively. For scenes with missing local regions, our model exhibits the second-best performance, trailing behind LDNFNet [8] on the PS subset and on the OC subset. LDNFNet [8] added a denoising network before performing the FR task, which can effectively reduced the interference of noise on extracting discriminative features. However, this approach lose part of the valuable facial information, such as texture and details, resulting in a recognition rate of 4.42% lower than our MIHNet on the TM subset.

Results on KinectFaceDB: We conduct experiments on KinectFace, another low-quality 3D face dataset with lower face image quality than in Lock3DFace, to further validate the effectiveness of our proposed model. The setting of this experiment follows [31], and the results are listed in Table 3. Our model achieves the highest recognition accuracy on subsets P1 and P3, with an improvement of 0.3% and 0.6% over the second-best DSNet [38], and the performance is comparable to DSNet [38] on the P2 subset.

Results on Cross-Quality Dataset: We additionally verify the generalization of our model on the cross-quality dataset. By introducing four different levels of noise intensity to FRGC v2.0, we generate cross-quality data, supplemented with data augmentation during model training. Table 4 illustrates the model's performance on the synthesized cross-quality data derived from Bosphorus. The proposed model achieves a recognition accuracy of 94.13% on HL and 92.85% on LL, outperforming the second-best DSNet [38] by 0.42% and 0.39%, respectively.

These aforementioned results of Table 1, Table 2, Table 3, and Table 4 demonstrate that MIHNet is not only effective in recognizing faces with time variations, but also exhibits strong generalization capability.

Table 5. Ablation study of TransL module and GL-Mixer on the Lock3DFace dataset, with indication of module presence denoted by ✓ and bold indicates the best result.

Model	TansL	GL-Mixer	Rank-1 Accuracy					
			NU	FE	PS	OC	TM	Total
Model 1			98.64%	97.54%	68.62%	71.27%	65.41%	81.57%
Model 2	✓		98.64%	97.66%	77.33%	77.84%	69.57%	85.32%
Model 3	✓	✓	**99.95%**	**99.43%**	**80.02%**	**83.10%**	**76.37%**	**87.52%**

4.4 Ablation Study

To study the contributions of the proposed TransL and GL-Mixer, we conduct a series of ablation experiments on the Lock3DFace dataset, with the results presented in Table 5.

We conduct experiments on three models to prove the efficacy of the proposed TransL and GL-Mixer modules. Model 1 excludes the TransL and GL-Mixer modules, using the feature fusion approach from the baseline [14]. Model 2 integrates the TransL module into Model 1, while Model 3 added the GL-Mixer module to Model 2. Table 5 presents the Rank-1 accuracy of these models on the Lock3Dface dataset. From the table, it is evident that Model 2 exhibits a 3.75% improvement over Model 1, and Model 3 further improves by 2.2% compared to Model 2. This indicates that the TransL module can extract both global and local feature information from the feature maps, and the GL-Mixer effectively combines the features output by the TransL module to achieve more discriminative features.

To evaluate the efficacy of our proposed differentiated spatial reduction attention (DSRA) within the Transformer Block, we substituted the DSRA with SRA from Wang et al. [23], and OSRA from Lou et al. [42] to construct Model-SRA and Model-OSRA. The experimental results, depicted in Fig. 4, reveal that incorporating DSRA into the model led to superior total performance, and performance across individual subsets compared to Model-SRA and Model-OSRA. Notably, there was a notable enhancement in recognition accuracy on the OC and TM subsets, surpassing Model-OSRA by 0.81% and 1.06%, respectively.

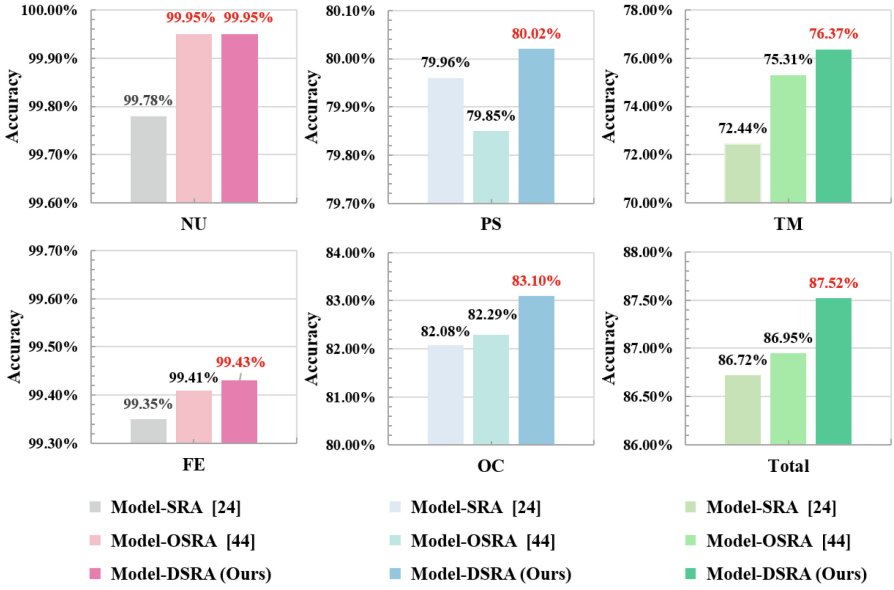


Fig. 4. The performance comparison of different subsets of the Lock3DFace dataset, where the attention mechanism of the model is replaced with SRA, OSRA, and our proposed DSRA, respectively. The red font indicates the best result. (Color figure online)

This effectively illustrates the ability of DSRA to magnify feature distinctions between layers, enabling the model to extract more discriminative information.

5 Conclusion

In this paper, we propose a multi-scale intra-layer fusion network with a hybrid structure, MIHNet, for low-quality 3D FR to address the problem of model's sensitivity to changes in 3D shapes and textures caused by time. A dual-stream feature extraction module is designed to extract both global and local information, promoting the network's expressive ability. Additionally, a feature fusion module named GL-Mixer has been introduced, enabling the network to learn discriminative features that are robust to time variations while reducing redundancy. Extensive experiments on three challenging low-quality 3D face datasets demonstrate that our model successfully extracts time robust features and improving the recognition accuracy of low-quality 3D FR.

References

1. Zheng, H., Wang, W., Wen, F., Liu, P.: A complementary fusion strategy for RGB-D face recognition. In: Pór Jónsson, B., et al. (eds.) MMM 2022. LNCS,

vol. 13141, pp. 339–351. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98358-1_27
2. Uppal, H., Sepas-Moghaddam, A., Greenspan, M., Etemad, A.: Depth as attention for face representation learning. IEEE Trans. Inf. Forensics Secur. **16**, 2461–2476 (2021)
3. Lin, S., Jiang, C., Liu, F., Shen, L.: High quality facial data synthesis and fusion for 3D low-quality face recognition. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE (2021)
4. Zhao, P., Ming, Y., Meng, X., Yu, H.: LMFNet: a lightweight multiscale fusion network with hierarchical structure for low-quality 3-D face recognition. IEEE Trans. Hum.-Mach. Syst. **53**(1), 239–252 (2022)
5. Jiang, C., Lin, S., Chen, W., Liu, F., Shen, L.: PointFace: point cloud encoder-based feature embedding for 3-D face recognition. IEEE Trans. Biom. Behav. Identity Sci. **4**(4), 486–497 (2022)
6. Zhu, Y., Gao, J., Wu, T., Liu, Q., Zhou, X.: Exploiting enhanced and robust RGB-D face representation via progressive multi-modal learning. Pattern Recogn. Lett. **166**, 38–45 (2023)
7. Niu, W., Zhao, Y., Yu, Z., Liu, Y., Gong, Y.: Research on a face recognition algorithm based on 3D face data and 2D face image matching. J. Vis. Commun. Image Represent. **91**, 103757 (2023)
8. Xu, R., et al.: Depth map denoising network and lightweight fusion network for enhanced 3D face recognition. Pattern Recogn. **145**, 109936 (2024)
9. Fan, Q., Huang, H., Chen, M., Liu, H., He, R.: RMT: retentive networks meet vision transformers. arXiv preprint arXiv:2309.11523 (2023)
10. Vasu, P.K.A., Gabriel, J., Zhu, J., Tuzel, O., Ranjan, A.: FastViT: a fast hybrid vision transformer using structural reparameterization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5785–5795 (2023)
11. Lin, W., Wu, Z., Chen, J., Huang, J., Jin, L.: Scale-aware modulation meet transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6015–6026 (2023)
12. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
13. Quan, Y., et al.: Centralized feature pyramid for object detection. IEEE Trans. Image Process. **32**, 4341–4354 (2023)
14. Mu, G., Huang, D., Hu, G., Sun, J., Wang, Y.: Led3D: a lightweight and efficient deep approach to recognizing low-quality 3D faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5773–5782 (2019)
15. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Sun, L., Lian, Z., Liu, B., Tao, J.: MAE-DFER: efficient masked autoencoder for self-supervised dynamic facial expression recognition. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 6110–6121 (2023)
17. Nguyen, X.-B., Duong, C.N., Li, X., Gauch, S., Seo, H.-S., Luu, K.: Micron-BERT: BERT-based facial micro-expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1482–1492 (2023)
18. Zhong, Y., Deng, W.: Face transformer for recognition. arXiv preprint arXiv:2103.14803 (2021)
19. Dan, J., et al.: TransFace: calibrating transformer training for face recognition from a data-centric perspective. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20642–20653 (2023)

20. Zhang, Z., et al.: Learning flexibly distributional representation for low-quality 3D face recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 4, pp. 3465–3473 (2021)
21. Li, K., et al.: UniFormer: unified transformer for efficient spatiotemporal representation learning. arXiv preprint arXiv:2201.04676 (2022)
22. Guo, J., et al.: CMT: convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12175–12185 (2022)
23. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
24. Ren, S., et al.: Shunted self-attention via multi-scale token aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10853–10862 (2022)
25. Chu, X., et al.: Twins: revisiting the design of spatial attention in vision transformers. In: Advances in Neural Information Processing Systems 34, pp. 9355–9366 (2021)
26. Li, J., Wen, Y., He, L.: SCConv: spatial and channel reconstruction convolution for feature redundancy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6153–6162 (2023)
27. Zhang, J., Huang, D., Wang, Y., Sun, J.: Lock3DFace: a large-scale database of low-cost kinect 3D faces. In: 2016 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2016)
28. Min, R., Kose, N., Dugelay, J.-L.: KinectFaceDB: a kinect database for face recognition. IEEE Trans. Syst. Man Cybern. Syst. **44**(11), 1534–1548 (2014)
29. Phillips, P.J., et al.: Overview of the face recognition grand challenge. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 947–954 (2005)
30. Alyüz, N., et al.: 3D face recognition benchmarks on the bosphorus database with focus on facial expressions. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M. (eds.) BioID 2008. LNCS, vol. 5372, pp. 57–66. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89991-4_7
31. Neto, J.B.C., Marana, A.N., Ferrari, C., Berretti, S., Del Bimbo, A.: Depth-based face recognition by learning from 3D-LBP images. In: 3DOR@ Eurographics, pp. 55–62 (2019)
32. Cui, J., Zhang, H., Han, H., Shan, S., Chen, X.: Improving 2D face recognition via discriminative face depth estimation. In: 2018 International Conference on Biometrics (ICB), pp. 140–147. IEEE (2018)
33. Hu, Z., Zhao, Q., Liu, F.: Revisiting depth-based face recognition from a quality perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
34. Mu, G., Huang, D., Li, W., Hu, G., Wang, Y.: Refining single low-quality facial depth map by lightweight and efficient deep model. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8 (2021)
35. Cardia Neto, J.B.: 3D face recognition with descriptor images and shallow convolutional neural networks. Universidade Federal de São Carlos (2020)
36. Grati, N., Ben-Hamadou, A., Hammami, M.: Learning local representations for scalable RGB-D face recognition. Expert Syst. Appl. **150**, 113319 (2020)
37. Neto, J.B.C., et al.: Learning streamed attention network from descriptor images for cross-resolution 3D face recognition. ACM Trans. Multimedia Comput. Commun. Appl. **19**(1s), 1–20 (2023)

38. Zhao, P., et al.: DSNet: dual-stream multi-scale fusion network for low-quality 3D face recognition. AIP Adv. **13**(8), 085310 (2023)
39. Savran, A., Sankur, B., Bilge, M.T.: Comparative evaluation of 3D vs. 2D modality for automatic detection of facial action units. Pattern Recogn. **45**(2), 767–782 (2012)
40. Sandler, M., et al.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
41. Gilani, S.Z., Mian, A.: Learning from millions of 3D scans for large-scale 3D face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1896–1905 (2018)
42. Lou, M., Zhou, H.-Y., Yang, S., Yu, Y.: TransXNet: learning both global and local dynamics with a dual dynamic token mixer for visual recognition. arXiv preprint arXiv:2310.19380 (2023)
43. Xiao, S., Li, S., Zhao, Q.: Low-quality 3D face recognition with soft thresholding. In: Feng, J., Zhang, J., Liu, M., Fang, Y. (eds.) CCBR 2021. LNCS, vol. 12878, pp. 419–427. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86608-2_46
44. He, L., He, L., Peng, L.: CFormerFaceNet: efficient lightweight network merging a CNN and transformer for face recognition. Appl. Sci. **13**(11), 6506 (2023)

Infant Cry Classification Using Modified Group Delay Cepstral Coefficients

Arth J. Shah(✉), Hiya Chaudhari, and Hemant A. Patil

Speech Research Lab, DA-IICT, Gandhinagar, India
{202101154,202101047,hemant_patil}@daiict.ac.in

Abstract. Classification of pathological *vs.* normal infant cries is used to infer the infant's health conditions. Such an approach can be beneficial in many situations and even to save infants' lives. In this paper, we propose a novel classification system based on the Modified Group Delay Cepstral Coefficients (MGDCC), for classifying infant cries. We investigate generalizability of proposed MGDCC features. The Convolutional Neural Network (CNN) was used as a pattern classifier in this study. Proposed MGDCC features are found to perform better than widely used spectral features, such as Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), and Group Delay Cepstral Coefficients (GDCC). Experiments are performed on two datasets namely, Baby Chillanto (D1) dataset, and DA-IICT Infant Cry (D2) corpus and for various experimental evaluation factors, such as noise robustness under signal degradation conditions, cross-database scenario, and analysis of latency period. We obtained 2.25% increase accuracy as compared to existing optimal accuracy for proposed task. Better performance of MGDCC is may be due to its capability to implicitly capture time dependencies in the sequence of audio samples via fourier transform phase information.

Keywords: Infant Cry Classification · Modified Group Delay Cepstral Coefficients · Convolutional Neural Network · Modified Group Delay Function

1 Introduction

Infancy is the period between birth and the acquisition of language. Infants communicate with us through crying, expressions and babbling making it essential to analyse and understand their cries. Previous studies have shown, the ability of infants to distinguish between two different languages just after four days of birth [1]. Recently, biometrics based on infants' fingerprints have been adapted to improve infant security, but they come with multiple challenges as aquisition of an infant's fingerprint is difficult as they commonly suck their fingers leading to wet fingers, and also keep their fists closed [2]. The infants who suffered from Autism Spectrum Disorder (ASD), have higher fundamental frequency (F_0) (the frequency of oscillation of vocal folds) during cry production mechanism in

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 275–289, 2025.
https://doi.org/10.1007/978-3-031-78341-8_18

infants, which is responsible for controlling pathological cries [3]. Due to a lack of knowledge about the ailment, many newborns die from the disease that strikes them in the first quarter of their lives. Sudden Infant Death Syndrome (SIDS) is the primary cause of newborn mortality when it comes to diseases, birth asphyxia, and associated anomalies [4]. For the infants who died of SIDS, landmark studies funded by the National Institutes of Health (NIH), USA, revealed evidence of abnormalities in the brain stem (specifically, the medulla oblongata), which is known to control breathing functions.

Collecting infant cry samples is a difficult task, as only a few parents/guardians cooperate and permit due to apprehension of recordings. Although large scale data on infant cry classification is not available (due to data imbalance, ethical issues, data privacy, etc.), the market continues to demand better approaches due to solid relevance of this problem. In order for doctors to have control over concerns pertaining to infants' assistance, a system that can more accurately distinguish between pathological and normal cries must be secured. This enhances the quality of life for newborns.

2 Related Works

Many attempts have been made to develop an optimal system that classifies pathological *vs.* healthy cries. Due to recent advancement in machine learning and deep learning, many previous studies have employed such processes for infant cry classification. In [5], the authors used magnitude spectrum-based features with Convolutional Neural Network (CNN) as classifiers for infant cry classification task. Many such studies have been reported in recent time, such as in [6] and [7], authors used Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC) for infant cry classification. In [8], the authors used MFCC as features and machine learning approaches, such as Multi-Layer Perception (MLP) and CNN as classifier for infant cry classification task.

However, most of the studies fail to explore the characteristics of infant cry signals. According to authors' best knowledge and belief, this is the second study to employ phase-based features for infant cry classification task. In particular, Modified Group Delay Cepstral Coefficients (MGDCC) have been employed in many speech tasks previously, such as speech recognition, phoneme recognition, and isolated word recognition [9–12]. The success of MGDCC, in existing studies motivated us to explore its properties for classification of infant cries with standard data and advanced DL models, such as CNN [13]. In [13], the authors did not fine-tune model systematically, nor they did experiment on any standard dataset. Newly proposed approach (this system) is able to obtain 9.18% increase in accuracy than the one proposed previously [13]. For fair comparison of our system, we also compared our work with existing approaches, such as Whisper features, which are currently most renounced features, for various speech application, such as, speech recognition [14], emotion classification [15], Deepfake Audio Detection (ADD) [16], and infant cry classification [17]. For optimizing time and space complexity, we also conducted latency analysis (to be discussed soon in

Sub-Sect. 5.4). We have also conducted experiments based on noise robustness for additive babble noise, which helps us to check the robustness of the proposed system, as compared to other feature vectors.

Figure 1 (a) represents the time-domain waveform, (b) represents short-time waveforms, and (c) presents modified group delay function (MGDF) for the normal *vs.* pathological cries signal. It can be observed from Fig. 1 (c) that the in normal cry signal, we get higher number of poles of MGDF. On the other hand, we obtain well separated poles in the MGDF in the pathological cry signal. These poles in the MGDF correspond to formants of the infants. As a result, the formants and their harmonics of the infants may differ in ways that are indicative of an underlying condition.

The remaining part of the paper is organized as follows: Sect. 3 presents computational aspects for proposed phase-based features. Section 4 contains information about the dataset, and other details of experimental setup. Section 5 contains detailed explanation about the experimental results obtained. Section 6 concludes the paper by summarizing the findings and suggesting possible future research directions.

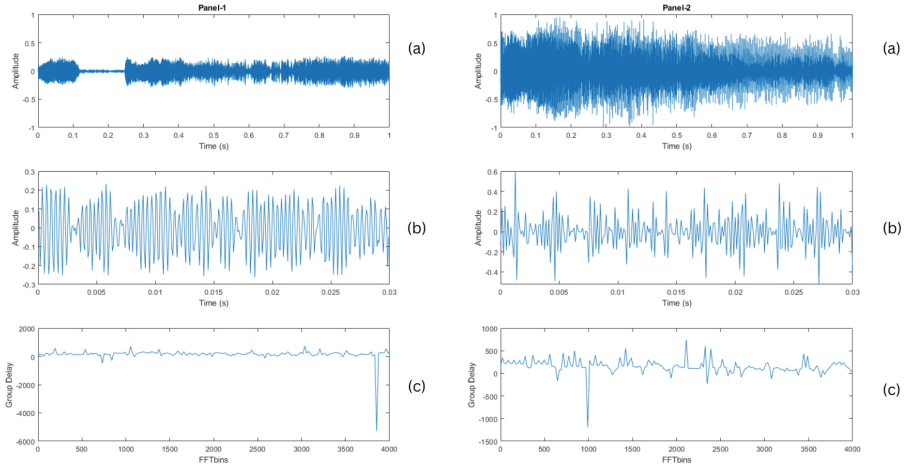


Fig. 1. Panel I (healthy signal), and Panel II (pathological signal): (a) time-domain waveforms, (b) short-time segment, and (c) MGDF.

3 Phase Based Approach

Information obtained by any spectral-based feature vector may not be sufficient enough to categorize an infant cry, which motivated us to explore the temporal characteristics in the form of signals' Fourier transform phase. Group Delay Function (GDF), which is one of the most well known phase-based feature, is known for its ability to capture *temporal* characteristics (in frequency domain) of a speech signal. The unit circle (in Z-plane), zeros (valleys), and poles (spikes) of

a given speech segment are the three main focal points of the GDF concept. However, the vocal tract system and the excitation source also greatly contribute to envelop and fine structure of speech segment. The magnitude spectrum aims to find spectral envelope, which is not sufficient knowledge from speech signal. This proposed study is intended to abstract more information about speech signal by integrating its spectral and temporal characteristics. This additional information helps us for better classification of infant cries. While the information derived from a magnitude-based spectrum can be easily extracted, phase-based information for the classification of infant cries is still not well-explored. The phase-based information is obtained after unwrapping of phase spectrum to invert signal processing artifacts of the arc-tangent function.

Speech signals are a combination of magnitude and phase spectrum, which prompted us to investigate phase-based characteristics of the speech spectrum. After investigating the GDF, the authors were prepared to explore other phase-based features, which led us to investigate the modified group delay features. The MGDF is a modified version of GDF. The MGDF feature is achieved by reducing the number of spurious *spikes* in GDF [10]. The meaningful feature emerges by reducing the number of poles for a specific speech segment. MGDF has previously been investigated for many tasks, as mentioned in Sect. 1. In this context, the GDF refers to the delay experienced by a *group* of frequencies, when provided as input to the system [18]. For discrete-time signal $p(n)$, its Discrete-Time Fourier Transform (DTFT) can be given by $P(e^{j\omega})$. Further, its magnitude-phase representation can be given by [19]:

$$P(e^{j\omega}) =\mid P(e^{j\omega}) \mid e^{j\phi(e^{j\omega})}, \tag{1}$$

where $|P(e^{j\omega})|$ is the magnitude spectrum, and $\phi(e^{j\omega})$ is the phase spectrum of the signal at frequency, ω. Neglecting the magnitude spectrum, GDF can be represented in terms of phase spectrum as:

$$\tau(e^{j\omega}) = -\frac{d}{d\omega}\phi(e^{j\omega}) = -j\,\mathrm{mag}\left[\frac{d}{d\omega}\log(P(e^{j\omega}))\right]. \tag{2}$$

Alternatively, Eq. (2) can be also represented in form of frequency-domain as:

$$\tau(e^{j\omega}) = \frac{P_R(e^{j\omega})Q_R(e^{j\omega}) + P_I(e^{j\omega})Q_I(e^{j\omega})}{|P(e^{j\omega})|^2}, \tag{3}$$

where $P(e^{j\omega}) \rightarrow$ DTFT of $p(n)$, and $Q(e^{j\omega}) \rightarrow$ DTFT of $n * p(n)$, R and I are the real and imaginary parts of the Fourier transform, respectively. MGDF is created after adding γ parameters for restoring dynamic range of speech and α for α to reduce the number of *spikes* in Eq. (3), and modifying it, we get equation of MGDF as [20]:

$$\tau_{mgd}\left(e^{j\omega}\right) = \frac{\tau_m\left(e^{j\omega}\right)}{|\tau_m\left(e^{j\omega}\right)|}\left|\tau_m\left(e^{j\omega}\right)\right|^{\alpha}, \tag{4}$$

where,

$$\tau_m(e^{j\omega}) = \frac{P_R(e^{j\omega})Q_R(e^{j\omega}) + P_I(e^{j\omega})Q_I(e^{j\omega})}{|P_c(e^{j\omega})|^{2\gamma}}, \tag{5}$$

where $|P_c(e^{j\omega})|$ represents cepstrally smooth version of $|P(\omega)|$. Figure 2 represents the steps for extraction of proposed MGDCC features, along with the functional block diagram of feature extraction and pattern classification process. The classification of this feature vector shows promising results due to their ability to capture temporal characteristics of audio signal in frequency domain. For the extraction of features, Hamming window of duration 25 ms was chosen, the shift was taken to be 10 ms, and the number of coefficients per frame (i.e., dimention of feature vector) was taken as 20.

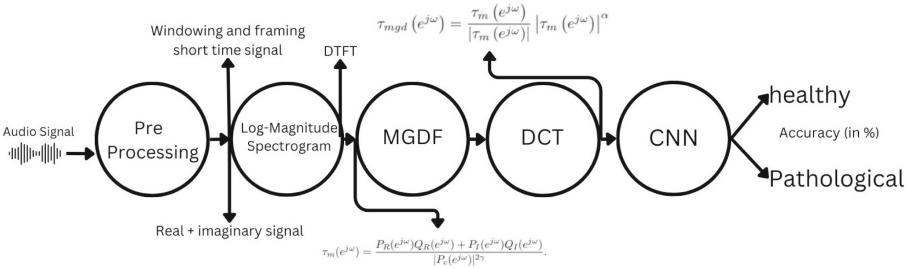


Fig. 2. Functional block diagram of proposed MGDCC based classification.

4 Experimental Setup

4.1 Dataset Used

The dataset being used is Baby Chillanto dataset (D1) was created by recordings, which were made by doctors of NIAOE-CONACYT, located in Mexico [21,22]. Chillanto is a Spanish word meaning screaming or crying. It consists a total of 1049 healthy infant cries, and 1219 pathological cries. This dataset has classification of 5 classes, which can be classified into 2 categories as mentioned in Table 1. Another database used for this study was DA-IICT Infant Cry Database (D2) [23], which contains a total of 793 healthy and 416 pathological cries [23]. This dataset has a total of 3 class audios, which can be further merged and classified in 2 categories as mentioned in Table 1. For fair comparison, both datasets were resampled at 16 kHz sampling rate. Table 1 indicates more details about the number of samples in both datasets.

4.2 Pattern Classifier Used

For this study, we employed CNN as pattern classifier because it captures spatial and temporal dependencies in the audio signals. The MGDCC features of the cry signal, which carry significant discriminative information, can be effectively

Table 1. Audio sample numbers for datasets used [22,23]

Class	Sub-class	Baby Chillanto	DA-IICT
Pathological	Asthma	–	182
	Asphyxia	340	215
	Deaf	879	–
Healthy	Pain	192	–
	Hunger	350	–
	Normal	507	793

learned by CNN due to their hierarchical feature learning capability. The CNN architecture was built with a sigmoidal activation layer, and 3 ReLU activation layers. CNN consists of five convolution blocks and three fully-connected layers. Each layer is made up of *2* -D convolution layers, a ReLU activation layer, and a batch normalization layer. At the end of each layer, max-pooling is used to downsample feature maps. The final dense layer has a single unit with a sigmoid activation function, producing a binary classification output (0 or 1) that indicates whether the input belongs to class 0 or 1. Learning rate was taken as 0.003 and optimizer was chosen to be Adam. Input shape was taken to $20 \times 893 \times 1$. For the dimensions less than 893, the rows were padded with extra zeros until column 893. Learning rate was selected as 0.003, with batch-size of 64. Adams optimizer were used for this paper. The architecture and structure-code implemented are openly available[1].

4.3 Other Spectral Features Used

MFCC and LFCC. Two cepstral features, namely, MFCC and LFCC were used for comparison with phase-based features. *20* -D MFCC and *20* -D LFCC features were extracted using *25* ms window length and *10* ms of window overlap. LFCC captures information w.r.t. linear frequency scale, unlike MFCC, which employs a logarithmic Mel frequency scale. These two features were chosen in order to compare phase-based features with the other magnitude spectrum-based features [24,25].

4.4 Performance Metrics

F1-Score. It is a useful metric to balance the trade-off between precision and recall for this problem. F1-score can be described as harmonic mean of precision and recall, providing a balance between these two metrics. It ranges from 0 to 1. If F1-score reaches closer to 1, then it's believed to have best precision and recall for any problem. F1-score is given by:

$$F1 - Score = \frac{2TP}{2TP + FN + FP}. \tag{6}$$

where FP, and FN denote false positives and false negatives, respectively.

[1] github repo: "https://github.com/ARTHARKING55/CNN_ICPR_MGDCC".

5 Experimental Results

5.1 Fine-Tuning of Parameters for MGDCC

In this sub-Section, we fine-tune two parameters of MGDCC, namely, α and γ, which are responsible for reducing the spurious spikes of MGDCC feature vector. A mixed-phase system can be made into a minimum phase system by minimizing the number of spikes in the MGDF spectrum. This implies that all of the poles and zeros of the z-domain system function are located inside the unit circle in the z-plane [11]. For creating an optimal feature vector for any task, we need to find what exact number of parameters and data are required in order to get the best possible results. The GDF becomes a mixed phase system if the number of spikes is increased; conversely, if the number of spikes is drastically decreased, the necessary quantity of data can be lost. In order to create balance of number of spikes and to obtain maximum results, we need to find the optimal values of α and γ parameters of MGDCC, for the infant cry classification. Figure 3 represents testing accuracy obtained when we vary the parameters of MGDCC. It can be noted that as we move towards 1 for parameter γ, the accuracy of model degrades lately. This phenomenon occurs due to heavy number of unwanted spikes present while extraction of GDCC. On the other hand, as we move value of towards α towards 1, the number accuracy degrades less, however, at low value of α, we achieve absolute maximum accuracy of 99.56% at $\alpha = 0.2$ and $\gamma = 0.2$, which suggests that, for the specific infant cry classification problem, the optimum number of spikes is needed. Further examination of the other point values of α and γ helps us to obtain the exact number of spikes needed. Greedy search algorithm was used for the task of fine tuning. For the experiments, the range of α and γ, which lies between 0 and 1, was varied by step size 0.1.

5.2 Comparison with Cepstral Features

In this study, we compare phase-based features to magnitude-based features, i.e., MFCC, and LFCC. We used two datasets for a fair comparison of features. The idea to test our findings in different acoustical and environmental settings inspired this work. The results shows in Table 2 indicate that the results of MFCC are better than LFCC under both datasets conditions, which specifies the significance of low frequency region resolution for infant cry classification. Additionally, it states that less information is needed to classify high frequency region data as healthy *vs.* pathological cries. Similar results have been observed for MFCC and LFCC features w.r.t. CNN as classifier in previous studies [5], where authors employed different structure of CNN without fine-tuning each layer from scrap.

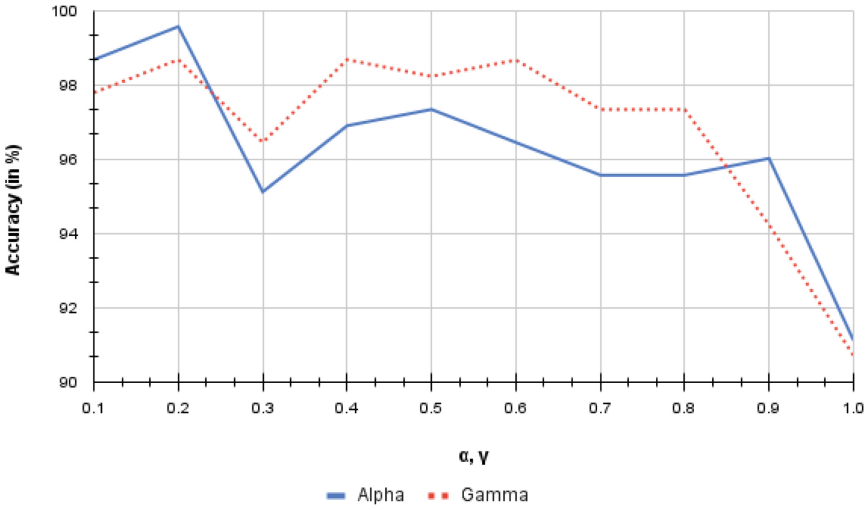


Fig. 3. Fine-tuning of α and γ parameters of MGDCC on Baby Chillanto (D1) dataset.

On the other hand, GDCC performs poorly than MFCC, indicating that there are more spurious spikes included in the extraction of phase-based features. MGDCC, a modified version of GDCC, outperforms all other features for the same pattern classifier and parameters, indicating the features ability (i.e., their discrimination power) to capture temporal characteristics, and convert a maximum phase system to a minimum phase system. The high F1-Score for the MGDCC feature vector highlights its high precision and recall. It further indicates well balanced performance of the model w.r.t. proposed feature vector. Baby Chillanto (D1) being a balanced dataset gives much high F1-Score compared to DA-IICT Infant Cry Database (D2) Corpus, which is an unbalanced dataset. The poor EER and F1-Score ratios are caused by the same unbalanced dataset. For the D1 dataset, MGDCC had 1.78% higher accuracy than MFCC, and 2.67% higher accuracy than LFCC features. For the D2 corpus, MGDCC had a 4.06% increase in accuracy over MFCC and an 18.04% increase over LFCC features. Poor performance on DA-IICT Infant Cry dataset (D2), is due to its recording setup, i.e., the recordings were done in open, making them wild, and thereby evaluating proposed methodology on most realistic scenerio possible.

5.3 Evaluation on Cross-Dataset Scenarios

In order to test the performance of our model in realistic testing scenarios, we tested models of all four feature sets with alternate dataset than on which they were trained, i.e., realistic cross-database scenario. This section of the study looks

Table 2. Comparison of Features w.r.t. Different Datasets

Dataset	Features	Accuracy (in %)	EER (in %)	F1-Score
DA-IICT	MFCC	79.55	20.49	66.67
	LFCC	65.57	34.42	52.27
	GDCC	79.51	20.49	61.54
	MGDCC	**83.61**	16.39	69.71
Bably Chillanto	MFCC	97.78	2.22	97.91
	LFCC	96.89	3.12	97.12
	GDCC	91.56	8.44	92.12
	MGDCC	**99.56**	0.97	99.11

into the ability of the MGDCC feature set to perform in the random datasets testing portion, where the data has not been trained. Results shown in Table 3 indicate that even when evaluated across different datasets, the MGDCC feature performs better than other feature vectors. It can be observed that the % accuracy reduces significantly for each feature by performing cross-evaluation, which indicates we need more large and balanced database in order to identify infant's cry as healthy or pathological cry. We obtained total of 1.58% increase in accuracy than MFCC feature, and 9.78% accuracy when we test the model on D1 dataset. We obtained 10.45% increase in accuracy than MFCC, and 9.11% increase than LFCC when model is trained on D2 corpus and testing on Baby Chillanto (D1) corpus. The better performance of MGDCC for cross-database may be attenuated to the fact that it may capture class-specific characteristics more dominantly due to its mathematical structure of representing Fourier phase information, which further helps in synthesizing given pattern-specific characteristics (i.e., class information).

Table 3. Results on Cross-Database Scenarios

test ==>	DA-IICT			Baby Chillanto		
train ==>	Baby Chillanto			DA-IICT		
Performance Metrics	Accuracy (in %)	EER	F1-Score	Accuracy (in %)	EER	F1-Score
MFCC	68.03	30.02	68.23	45.77	43.32	46.05
LFCC	59.83	37.34	61.11	47.11	44.12	47.38
GDCC	34.42	47.86	33.98	53.77	37.76	54.37
MGDCC	**69.61**	**29.87**	**71.77**	**56.22**	**35.13**	**55.92**

5.4 Analysis of Latency Period

Latency period is the least amount of frames necessary for achieving the highest classification accuracy [26]. We have evaluated the accuracy against the latency

introduced by the frame count. The frames are varied from a range 50–300. The results show that MGDCC consistently performs better than other features because of the ability to capture phase-based and temporal attributes. Analysis of latency period is carried out to calculate minimum time period of data required for given classification task. This further helps to optimize space complexity, along with time complexity, in order to make system reliable and efficient. Latency analysis was performed on D1 dataset (Fig. 4).

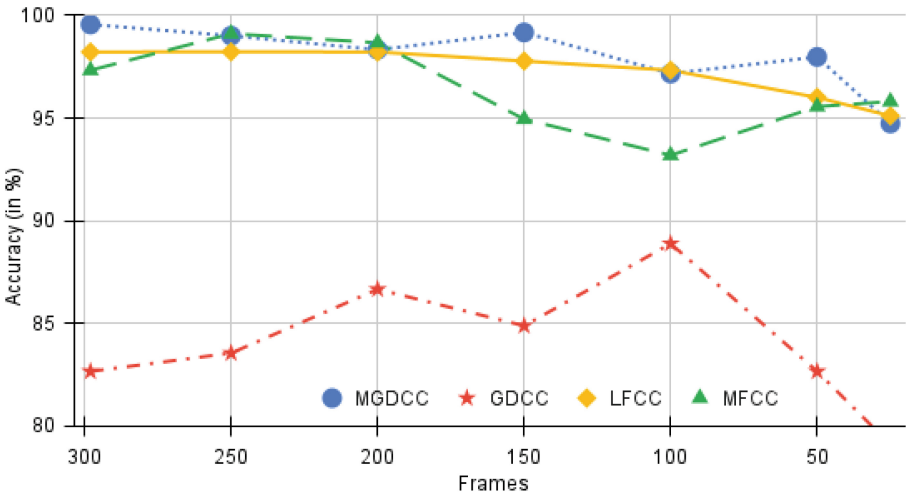


Fig. 4. Analysis of Latency Period on Baby Chillanto (D1) Dataset.

5.5 Robustness of MGDCC for Signal Degradation Conclusions

In this study, we examine how adding babble noise to the MGDCC proposed features affects their noise robustness. All experiments of this sub-Section were performed on standard D1 dataset. We also performance of these features in comparison to MFCC, LFCC, and GDCC. Additionally, the effects of noise power on the proposed method are examined by taking several SNR levels for babble noise into account, such as −5 dB, 0 dB, 5 dB, and 10 dB. These experiments assist us for analyzing the model's performance in real-world scenarios. We obtained better results using MGDCC than MFCC and LFCC features, which shows the superiority of MGDCC features even under noisy conditions. In addition, this experiment also helps us to get estimation of models accuracy in day-to-day life scenario. We were unable to perform experiments on other types of noise due to limited time and storage resources (Fig. 5).

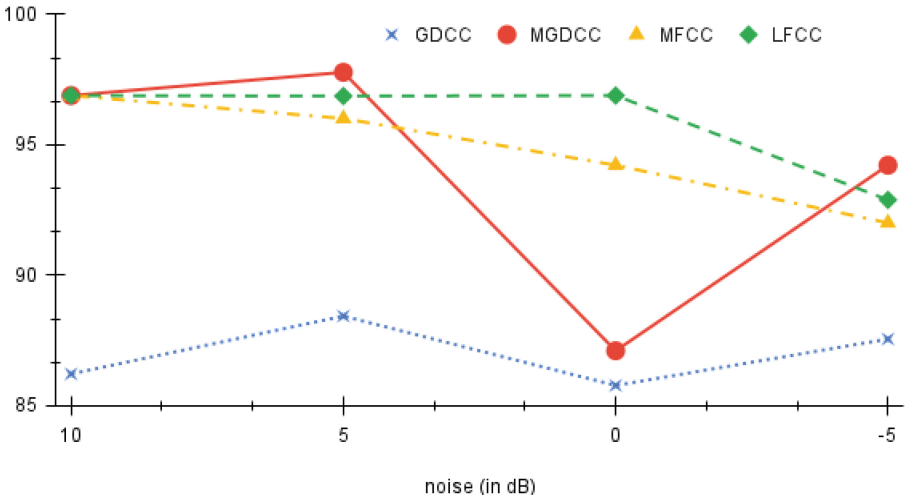


Fig. 5. Results under noise degradation using Additive Babble Noise on Baby Chillanto (D1) dataset.

5.6 Comparison With Existing Works

This sub-Section compares the proposed approach w.r.t. existing approaches. We compare our results with state-of-the-art (SOTA) OpenAI's Whisper, which is one of the most popular feature in recent days. Furthermore, we compare our findings with alternative approaches based on MFCC and spectrogram characteristics. It can be observed that the results for MGDCC with CNN classifier, is better than the existing approaches. This indicates the importance of phasebased properties and poles and zeros in audio signal. We obtained 12.53% higher accuracy than existing baseline work [27] proving the superiority of MGDCC over spectrograms. In order to do fair comparison with baseline, kept the classifier same (layers may differ in minor) and just altered feature set. Whisper being a large scale pre-trained model, is also not able to capture particular information which MGDCC is able to capture. Appendix gives mathematical proof justifying noise robustness of MGDCC features. Only accuracy was employed as comparison factor in this sub-Section, as most of the studies in the literature have reported only accuracy as their evaluation factor (Table 4).

Table 4. Comparison With Different Existing Works on D1 Dataset.

Source	Features	Pattern Classifiers	Accuracy (in %)
[27]	Spectrogram	CNN	87.03
[27]	Spectrogram	TLCNN & SVM	90.8
[27]	Spectrogram	SVM	90.10
[27]	Spectrogram	TL_ResNet50	90.80
[27]	Spectrogram	SVM + TL_ResNet50	91.10
[28]	Spectrogram → TL CNN	GCN (supervised/unsupervised)	92.01/94.39
[17]	MFCC	Bi-LSTM	97.17
[17]	Whisper	CNN	97.31
[17]	Whisper	Bi-LSTM	97.31
[**Proposed**]	**MGDCC**	**CNN**	**99.56**

6 Summary and Conclusions

In this work, we presented significance of Fourier transform phase-based features to classify infant cries. We enquired MGDF-based features for identifying healthy *vs.* pathological cries. This study investigated two group delay-based features, namely, GDCC and MGDCC, using two different datasets. The primary goal of this study is to emphasize the significance of phase-based features for the chosen task. The features extracted by signal processing concepts were then fed into the CNN classifier for classification. For this study, we made a variety of observations based on testing accuracy, EER, and F1-score. In comparison to existing widely used features, we achieved significantly better results for the task selected. We also discussed the difference and improvement between GDCC and MGDCC feature vectors. The proposed system have been explored for only one type of noise, which we aim to extend the work to various different types of noise, to analyze effect of different types of noise on model, as a future task. Future works also involve more detailed exploration on phase-based features, and fusion of features analysis of proposed approach.

Acknowledgements. The authors specially thank Mr. Aditya PSS. (JRF at Speech Lab DA-IICT), and DA-IICT authorities for his helpful support in this study. The authors sincerely thank the MeitY, for funding this study under project 'BHASHINI', (Grant ID: 11(1)2022-HCC(TDIL)).

Appendix: Noise Robustness of MGDCC

Let clean $_{\text{signal}}(x)$ be a clean signal, degraded by adding uncorrelated, additive noise (x) with 0 mean and σ^2 variance. Then, the noisy $_{\text{signal}}(x)$ can be represented as,

$$\text{noisy}_{\text{signal}}(x) = \text{clean}_{\text{signal}}(x) + \text{noise}(x). \tag{7}$$

Obtaining the power spectrum, and taking the Fourier transform, we get,

$$P_{\text{noisy}}\left(e^{j\omega_0}\right) = P_{\text{clean}}\left(e^{j\omega_0}\right) + P_{\text{noise}}\left(e^{j\omega_0}\right). \tag{8}$$

Two frequency regions, which are mutually exclusive (higher and lower SNR), can be obtained from Eq. (8). For the scenario of lower signal-to-noise ratio (SNR), we examine frequencies ω_0 satisfying $P_{\text{clean}}\left(e^{j\omega_0}\right) \ll \sigma^2\left(\omega_0\right)$, while for higher SNR, we focus on frequencies ω_0, where $P_{\text{clean}}\left(e^{j\omega_0}\right) \gg \sigma^2\left(\omega_0\right)$ [29]. For low SNR, we have:

$$P_{\text{noisy}}\left(e^{j\omega_0}\right) = \sigma^2\left(\omega_0\right)\left(1 + \frac{P_{\text{clean}}\left(e^{j\omega_0}\right)}{\sigma^2\left(\omega_0\right)}\right). \tag{9}$$

Solving Eq. (9), and neglecting higher order terms, we get:

$$\ln\left(P_{\text{noisy}}\left(e^{j\omega_0}\right)\right) \approx \ln\left(\sigma^2\left(\omega_0\right)\right) + \frac{1}{\sigma^2\left(\omega_0\right)}\left[d_0 + \sum_{x=1}^{+\infty} d_x \cos\left(\frac{2\pi}{\omega_0}\omega_0 x\right)\right]. \tag{10}$$

Equation (10) can be further solved and GDF can be obtained as mentioned in [29]:

$$\tau\left(e^{j\omega_0}\right) \approx \frac{1}{\sigma^2\left(\omega_0\right)} \sum_{x=1}^{+\infty} x d_x \cos\left(\omega_0 x\right). \tag{11}$$

Similarly for higher SNR, we have:

$$P_{\text{noisy}}\left(e^{j\omega_0}\right) = P_{\text{clean}}\left(e^{j\omega_0}\right)\left(1 + \frac{\sigma^2\left(\omega_0\right)}{P_{\text{clean}}\left(\omega_0\right)}\right). \tag{12}$$

Taking the logarithm on both sides of Eq. (12) and using the Taylor series expansion results in expanded term as:

$$\ln\left(P_{\text{noisy}}\left(e^{j\omega_0}\right)\right) \approx \frac{d_0}{2} + \frac{\sigma^2\left(\omega_0\right) e_0}{2} + \sum_{x=1}^{+\infty}\left(d_x + \sigma^2\left(\omega_0\right) e_x\right)\cos\left(\omega_0 x\right). \tag{13}$$

Equation (13) can be solved to GDF and the term obtained can be represented as [29]:

$$\tau\left(e^{j\omega_0}\right) \approx \sum_{x=1}^{+\infty} x\left(d_x + \sigma^2\left(\omega_0\right) e_x\right)\cos\left(\omega_0 x\right) \tag{14}$$

The respective GDF for these cases (Eq. (11), and Eq. (14)) summarized and represented as follows [29]:

$$\tau_{GDF}\left(e^{j\omega_0}\right) \approx \begin{cases} \frac{1}{\sigma^2\left(\omega_0\right)} \sum_{x=1}^{+\infty} x d_x \cos\left(\omega_0 x\right), & \text{for lower SNR,} \\ \sum_{x=1}^{+\infty} x\left(d_x + \sigma^2\left(\omega_0\right) e_x\right)\cos\left(\omega_0 x\right), & \text{for higher SNR,} \end{cases} \tag{15}$$

The Fourier series coefficients of $\ln\left(P_{\text{noisy}}\left(e^{j\omega_0}\right)\right)$ and $\frac{1}{P_{cle}\ \mathrm{a}_n(e^{j\omega_0})}$ are denoted by d_x's and e_x's, respectively. Equation (15) reveals that in the lower SNR scenario, the GDF is inversely proportional to the noise power, suggesting that the GDF effectively preserves peaks and valleys amidst additive noise. Conversely, for higher SNR values, the GDF is proportional to noise power, although the noise power is lower than the signal power. These findings imply that the GDF tracks the signal spectrum rather than the noise spectrum.

References

1. Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., Amiel-Tison, C.: A precursor of language acquisition in young infants. Cognition **29**(2), 143–178 (1988)
2. Engelsma, J.J., Deb, D., Cao, K., Bhatnagar, A., Sudhish, P.S., Jain, A.K.: Infant-ID: fingerprints for global good. IEEE Trans. Pattern Anal. Mach. Intell. **44**(7), 3543–3559 (2021)
3. Bonneh, Y.S., Levanon, Y., Dean-Pardo, O., Lossos, L., Adini, Y.: Abnormal speech spectrum and increased pitch variability in young autistic children. Front. Hum. Neurosci. **4**, 237 (2011)
4. Makhoul, J.: Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
5. Uthiraa, S., Kachhi, A., Patil, H.A.: Linear frequency residual features for infant cry classification. In: Karpov, A., Samudravijaya, K., Deepak, K.T., Hegde, R.M., Agrawal, S.S., Prasanna, S.R.M. (eds.) SPECOM 2023. LNCS, vol. 14338, pp. 550–561. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-48309-7_44
6. Dewi, S.P., Prasasti, A.L., Irawan, B.: The study of baby crying analysis using MFCC and LFCC in different classification methods. In: 2019 IEEE International Conference on Signals and Systems (ICSigSys), Bandung, Indonesia, pp. 18–23 (2019)
7. Pusuluri, A., Kachhi, A., Patil, H.A.: Analysis of time-averaged feature extraction techniques on infant cry classification. In: Prasanna, S.R.M., Karpov, A., Samudravijaya, K., Agrawal, S.S. (eds.) SPECOM 2022. LNCS, vol. 13721, pp. 590–603. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20980-2_50
8. Abbaskhah, A., Sedighi, H., Marvi, H.: Infant cry classification by MFCC feature extraction with MLP and CNN structures. Biomed. Signal Process. Control **86**, 105–261 (2023)
9. Zhu, D., Paliwal, K.K.: Product of power spectrum and group delay function for speech recognition. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, p. I-125 (2004)
10. Murthy, H.A., Gadde, V.: The modified group delay function and its application to phoneme recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, vol. 1, p. I-68 (2003)
11. Hegde, R.M., Murthy, H.A., Gadde, V.R.R.: Significance of the modified group delay feature in speech recognition. IEEE Trans. Audio Speech Lang. Process. **15**(1), 190–202 (2006)
12. Alsteris, L.D., Paliwal, K.K.: Evaluation of the modified group delay feature for isolated word recognition. In: Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, Sydney, Australia, vol. 2, pp. 715–718 (2005)

13. Chittora, A., Patil, H.A.: Modified group delay based features for asthma and HIE infant cries classification. In: Král, P., Matoušek, V. (eds.) TSD 2015. LNCS (LNAI), vol. 9302, pp. 595–602. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24033-6_67
14. O'Sullivan, J., et al.: Automatic speech recognition for ASD using the open-source whisper model from OpenAI (2023)
15. Feng, T., Narayanan, S.: Foundation model assisted automatic speech emotion recognition: transcribing, annotating, and augmenting. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, pp. 12116–12120 (2024)
16. Yang, Y., et al.: A robust audio deepfake detection system via multi-view feature. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, pp. 13131–13135 (2024)
17. Charola, M., Kachhi, A., Patil, H.A.: Whisper encoder features for infant cry classification. In: Proceedings of INTERSPEECH, Dublin, Ireland, vol. 2023, pp. 1773–1777 (2023)
18. Hannan, E., Thomson, P.: Estimating group delay. Biometrika **60**(2), 241–253 (1973)
19. Murthy, H.A., Yegnanarayana, B.: Group delay functions and its applications in speech technology. Sadhana **36**, 745–782 (2011)
20. Zhu, X., Li, Y., Yong, S., Zhuang, Z.: A novel definition and measurement method of group delay and its application. IEEE Trans. Instrum. Meas. **58**(1), 229–233 (2008)
21. Reyes-Galaviz, O.F., Cano-Ortiz, S.D., Reyes-García, C.A.: Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies. In: 2008 Seventh Mexican International Conference on Artificial Intelligence, pp. 330–335. Cambridge (2008)
22. Reyes-Galaviz, O.F., Cano-Ortiz, S.D., Reyes-García, C.A.: Validation of the cry unit as primary element for cry analysis using an evolutionary-neural approach. In: 2008 Mexican International Conference on Computer Science, Baja California, Mexico, pp. 261–267 (2008)
23. Chittora, A., Patil, H.A.: Data collection and corpus design for analysis of nonnal and pathological infant cry. In: 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Gurugram, India, pp. 1–6 (2013)
24. Zheng, F., Zhang, G., Song, Z.: Comparison of different implementations of MFCC. J. Comput. Sci. Technol. **16**, 582–589 (2001). Accessed 16 Apr 2024
25. Dewi, S.P., Prasasti, A.L., Irawan, B.: Analysis of LFCC feature extraction in baby crying classification using KNN. In: IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), Hong Kong, pp. 86–91 (2019)
26. Joukov, N., Traeger, A., Iyer, R., Wright, C.P., Zadok, E.: Operating system profiling via latency analysis. In: OSDI, Seattle, WA, vol. 6, pp. 89–102 (2006)
27. Le, L., Kabir, A.N.M., Ji, C., Basodi, S., Pan, Y.: Using transfer learning, SVM, and ensemble classification to classify baby cries based on their spectrogram images. In: 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW), Monterey, CA, USA, pp. 106–110 (2019)
28. Chunyan, J., Chen, M., Bin, L., Pan, Y.: Infant cry classification with graph convolutional networks. In: 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS), Las Vegas, USA, pp. 322–327 (2021)
29. Parthasarathi, S.H.K., Padmanabhan, R., Murthy, H.A.: Robustness of group delay representations for noisy speech signals. Int. J. Speech Technol. **14**, 361–368 (2011)

Vascular Biometrics Experiments on Candy – A New Contactless Finger-Vein Dataset

Sushil Bhattacharjee(✉), David Geissbuehler, Guillaume Clivaz, Ketan Kotwal, and Sébastien Marcel

Idiap Research Institute, Martigny, Switzerland
sushil.bhattacharjee@idiap.ch

Abstract. Finger-vein (FV) biometrics is an active and growing topic of research. Most FV systems available today rely on contact sensors that capture vein patterns of a single finger at a time. We have recently completed a project aimed at designing a contactless vein sensing platform, named *sweet*. In this paper we present a new FV dataset collected using *sweet*. The dataset includes multiple FV samples from 120 subjects and 280 presentation attack instruments (PAI), captured in a contactless manner. Further, we present baseline FV authentication (FVA) results achieved for proposed dataset. The *sweet* platform is equipped to capture a sequence of images suitable for photometric-stereo (PS) reconstruction of 3D surfaces. We present a FV presentation attack detection (PAD) method based on PS reconstruction, and the corresponding baseline FV PAD results on the proposed dataset. (See Footnote 4.)

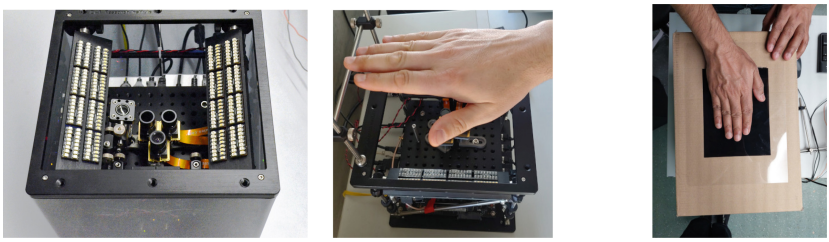
1 Introduction

Vascular biometrics offers several advantages over other biometrics modalities, such as high accuracy, robustness to variations in ambient illumination and robustness to presentation attacks (PA). Finger-vein (FV) sensors are the most common type of vascular-biometrics devices. Most commercially available FV sensors today use transmissive near-infrared (NIR) illumination – the camera captures the light passing through the finger. These FV devices rely on contact-imaging, where the user is required to place a finger on a transparent support, to provide a biometric sample.

The main benefit of such a contact-imaging apparatus is the high quality of vascular images. In contact-imaging based FV devices, the distance between the presented finger and the camera is fixed *a priori*, and therefore the camera parameters can be adjusted accordingly, to obtain sharp images. The NIR illumination intensity may also be controlled adequately, so as to provide sufficient contrast between the blood-vessels and the surrounding tissue. Contact-imaging

This work was supported by the Innosuisse project Candy, and the Swiss Center for Biometrics Research and Testing.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 290–308, 2025.
https://doi.org/10.1007/978-3-031-78341-8_19



(a) *sweet* platform (b) *Bona-fide* presentation (c) PA using a PAI of species B4

Fig. 1. (a) Prototype of the *sweet* platform. (b) Contactless *bona-fide* presentation. (c) Apparatus for recording certain species of presentation attacks (PA). Note, in (b), that the horizontal bar has been placed only to indicate to data-subjects the approximate height at which to present the hand. The finger-tips are not actually resting on the bar.

for biometric sensors, however, may be undesirable in certain deployment scenarios where hygiene is critical, such as hospitals, or even commercial points of sale.

We have developed a platform for contactless hand-vascular biometrics. The open-source design of this hardware platform, named *sweet*, is presented in a separate paper [3]. The prototype *sweet* platform is shown in Fig. 1(a). Here, the hand is presented at a distance of between 10 and 15 cm from the sensor, as illustrated in Fig. 1(b). *sweet* can capture biometrics samples under a variety of illumination-modalities. At present it is equipped to capture images under white LED light using a color (RGB) camera, and two NIR wavelengths, 850 nm and 950 nm, using two NIR cameras. The NIR cameras record vascular data. Each NIR camera also records data for Photometric-Stereo (PS) reconstruction of the target-surface, by capturing a set of images under a controlled illumination-sequence. This PS capability has been included for detecting presentation attacks (PA). Because of its contactless design, the *sweet* platform also supports vascular biometrics for palm and wrist modalities. The present work is concerned with FV biometrics only.

In this work we present a new dataset, named CandyFV, collected using the *sweet* platform. This dataset, which includes several FV samples per subject from a gender-balanced test crew of 120 subjects could support research on a variety of questions related to FV biometrics. In this work we have described our baseline FV authentication (FVA) method, and have provided results of this method on the new dataset.

The proposed dataset also includes PAs made using PA instruments (PAI) of 14 PAI species. We also present a FV PAD method that utilizes the surface-normal map reconstructed using PS. The main contributions of this work are (*i*) a new FV dataset containing 1200 genuine FV samples and 1400 PA samples from 14 PAI species; (*ii*) baseline FVA performance using a well-understood FVA pipeline; and (*iii*) a new baseline FV PAD method based on PS reconstruction, accompanied by its performance on the new dataset.

We begin, in Sect. 2, with a brief review of related research. The new dataset, including the data collection process, is described in Sect. 3. The baseline FVA method, and the PS based FV PAD method are described in Sect. 4. Then, the experimental methodology and results for both FVA and FV PAD are discussed in Sect. 5. A summary of this work and some perspectives on future work are presented in Sect. 6.

2 Related Research

For a comprehensive overview of FV biometrics, we point the interested reader to a recent review of the subject compiled by Hou *et al.* [4]. In this section we discuss selected works related to FVA and FV PAD that are relevant to the present study.

Finger-Vein Authentication Datasets: Newly proposed FVA algorithms can be compared to the state of the art using publicly available datasets such as SDUMLA-HMT [22], MMCBNU_6000 [9], VERA-finger [19], UTFVP [18], and SCUT-SFVD [13]. One common characteristic of these datasets is that the biometric samples represent only single fingers. In contrast, FV samples in the presented dataset show four fingers together, which enables finger-fusion for more robust FVA and FV PAD.

Finger-Vein Authentication Systems: Mainstream vascular biometrics systems still rely on hand-crafted features such as Repeated Line-Tracking (RLT) [10], maximum curvature (MC) [11], wide-line detection (WLD) [5]. These algorithms extract binary pixel-maps representing the vein-network in the biometric sample allowing to compare them. In this work we have used MC features.

Frequency-domain methods for FVA has also been proposed. Yang *et al.* [21] have used a bank of Gabor filters to enhance veins at different scales and then construct a set of FVCodes that are compared using a Cosine-similarity function. Yang *et al.* claim that their method performs better than MC features [11]. These results, however, have been estimated over a proprietary, unpublished dataset. More recently, Kovač and Marák [8] have used Gabor filters to detect feature-points in vein-images.

Publicly available FV datasets are not large enough to train a convolutional neural network (CNN) from scratch. Up to now, deep-learning based FVA approaches have adapted pre-trained CNNs through transfer-learning on FV datasets to construct feature-extractors. Besides FVA, deep-learning based methods have also been developed for other purposes such as vein enhancement, vein segmentation, and even encryption (see [23]). Kotwal and Marcel proposed a lightweight module – Residual Feature Pyramid Network (ResFPN) – for enhancement of vascular structures [7]. Bros *et al.* [2] propose a Residual Convolutional Autoencoder (RCAE) for vein-enhancement that reduces the classification error on the UTVFP dataset from 2.1% to 1%. In the present work we have used this RCAE in our FVA pipeline as well.

Research on FV PAD: Since the first comprehensive review of FV PAD methods [14], several new FV PAD methods have been proposed. Qui *et al.* [13] use total variational (TV) decomposition to decompose a FV image into structure and noise components. The two components are then modeled separately using local binary patterns (LBP), and the LBP histograms are subsequently classified using a cascade-classifier. This method (TV-LBP) achieves nearly perfect PAD on two public datasets – IDIAP VERA-finger [19] and SCUT-FVD [13].

Singh *et al.* [17] have used the SfSNet [16] to reconstruct a surface-normal map as well as a diffusion-map from a single (transmissive) image. The key difference between their work and our proposed FV-PAD is that in our case the surface-normal map is computed from photometry-data collected explicitly for this purpose (PS), whereas Singh *et al.* decompose a single image into a surface-normal map and a diffusion map using a pre-trained SfSNet. Surface-normal maps reconstructed using SfSNet do not always reflect the reality.

Raghavendra *et al.* [15] have used transfer learning to adapt a pre-trained Alex-Net for FV PAD. They demonstrate that their method consistently achieves a BPCER of 0 on several public datasets. The APCER for their method varies from 0.5% to 3.5%, depending on the species of presentation attack instrument (PAI) used.

From the analysis presented in this section we see that the idea of using Photometric-Stereo for FV PAD has not been explored previously (only for biometric recognition).

3 Data Collection

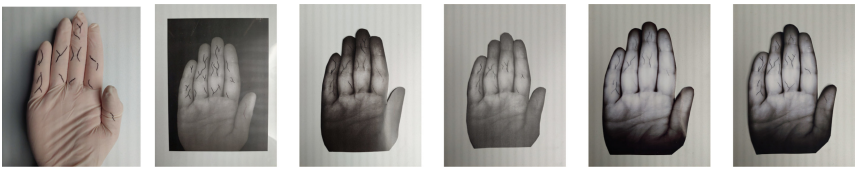
Using the *sweet* platform we have collected a new dataset, named CandyFV for FVA and PAD experiments. The proposed dataset is described in this section.

(a) 850 nm (b) 950 nm (c) PS

Fig. 2. Finger-vein data captured in various channels using the *sweet* platform. (a) Fingers illuminated under 850 nm NIR; (b) fingers illuminated under 950 nm NIR; (c) pseudo-color map showing the 3D surface-normals reconstructed from the photometric stereo (PS) data.

***Bona Fide* FV Samples:** FV samples from 120 subjects (62 male and 58 female) comprise the CandyFV dataset. The two gender groups are approximately evenly distributed over three age-groups: 18–30 years, 31–50 years, and 51 and above. Five samples for each hand are recorded for each subject. As illustrated in Fig. 2, the subject presents the hand with four fingers close together (named the *fingers-closed* modality), over the three cameras, at a distance of roughly 10–15 cm from the cameras. (Note that, the presented hand is not in contact with the horizontal steel bar; it just appears so in the picture due to the perspective.) Thumbs are not used in our experiments.

Each sample includes 20 usable images per camera (2 NIR cameras, and one color camera), captured under a variety of illuminations. For FVA, we use the FV images recorded by the two NIR cameras under 850 nm ('NIR-850') and 950 nm ('NIR-950') illumination. Each sample yields three images captured under NIR-850 illumination for each of the two (left, right) NIR cameras, and similarly, three images with NIR-950 illumination. The sample also includes four images captured specifically for PS reconstruction. These have been used in FV PAD experiments.



(a) B3 (b) B4 (c) B5 (d) B6 (e) B9 (f) B10

Fig. 3. Illustrations of some PAIs used in this work. The PAI species (described in Table 1) shown in each image is indicated in the caption.

Presentation Attacks: As is well known, vascular biometrics are very secure in the context of PAs. Unlike for superficial biometrics traits, such as face or fingerprint, it is almost impossible to construct a FV PAI without the cooperation of the textittarget subject. Here we have simulated the situation where a malevolent actor may gain unauthorized access to a database of FV biometrics samples. In other words, our PAIs have been constructed from the *bona-fide* samples recorded previously using the *sweet* platform. To construct our PAIs we have arbitrarily selected 20 target subjects. For each target subject a single NIR-850 image of the right hand captured by the left NIR camera has been used to create PAIs of several species, grouped into two levels, A and B:

Level A: PAIs can be created with minimal effort, within a day.

Level B: PAIs may take between one and three days to create, and require some expertise or specialized equipment.

The various PAI species are described in Table 1. As indicated in this table, we have considered four Level A and 10 Level B PAI species. The PAIs of Level

A are made by printing the source-images on two kinds of photo-paper (glossy and matte) using two different kinds of printers – a laser printer and an inkjet printer. The toner used in laser printers typically has NIR-absorbent properties. The PAIs in the various Level B species have been made either by enhancing the source-images before printing them, or by creating artefacts in other ways that may confound the proposed FV-PAD method.

For each PAI, again, five samples have been recorded using the *sweet* platform. In most cases, the PAI is presented to the *sweet* platform in the same way as a *bona-fide* presentation. Care is taken to hold the PAI over the cameras, as flat and horizontal as possible.

A different method has been used to capture PA samples of PAIs of the species B4. PAIs of this species are created by printing the source-image on a transparent sheet of plastic. To perform the PA, a *bona-fide* hand is placed over the image printed on transparent plastic. The idea is to see if the 3D shape of the *bona-fide* hand may fool the PS-based FV PAD method. Here we have placed a cardboard box over the *sweet* platform. The box has a rectangular hole cut out on the top surface. A transparent plastic PAI is placed on the box such that the printed FV image is visible to the cameras through the hole in the box. The attacker then places his or her (*bona-fide*) hand over the PAI, so that it is visible to the cameras through the PAI. This setup is illustrated in Fig. 1(c). Examples of selected PAI artefacts are shown in Fig. 3.

4 Baseline Algorithms

In this section we describe the FVA and FV PAD algorithms used in this study to generate baseline results for the CandyFV dataset.

Table 1. Descriptions of the PAI species of levels A and B. The letters A and B in the first column indicate the level of the PAI species.

Species	Description
A1	Source image printed on glossy paper on a laser printer
A2	Source image printed on matte paper on a laser printer
A3	Source image printed on glossy paper on an inkjet printer
A4	Source image printed on matte paper on an inkjet printer
B1	Image printed on glossy paper on a laser printer, with veins enhanced manually using a black marker
B2	Image printed on matte paper on a laser printer, with veins enhanced manually using a black marker
B3	Vein pattern drawn with black marker on a latex glove
B4	Image printed on plastic transparency, with a *bona fide* hand presented over the transparency
B5	Hand-region cut out from image printed on glossy paper on a laser printer
B6	Hand-region cut out from image printed on matte paper on a laser printer
B7	Image printed on glossy paper on an inkjet printer, with veins enhanced manually using a black marker
B8	Image printed on matte paper on an inkjet printer, with veins enhanced manually using a black marker
B9	Hand-region cut out from image printed on glossy paper on an inkjet printer
B10	Hand-region cut out from image printed on matte paper on an inkjet printer

4.1 FV Authentication Baseline

Finger-vein samples are compared based on *templates*. To enroll a new subject in the biometrics verification system, the subject first provides a biometric sample. A template constructed from this sample is stored in the biometrics system, associated with the subject's identity. During the probe phase, the subject claims a certain identity, and provides a new biometric probe sample. The system then compares the probe-template (derived from the probe sample) with the template previously enrolled for the claimed identity. If the two templates are sufficiently similar (*i.e.*, the match-score is above a predetermined threshold), we consider that the probe sample indeed corresponds to the claimed identity. In this section we describe the template creation process used in our baseline FVA method, as well as the method used here for comparing FV templates.

4.1.1 Finger-Vein Template Creation

The flowchart of the FV template creation process is shown in Fig. 4. Each input FV-sample is an image corresponding to a presentation, showing all fingers of the presented hand. First, the four fingers – index-, middle-, ring-, and little-finger – are segmented out from the input image. One template is constructed for each separate finger. The various steps indicated in Fig. 4 are described below.

Fig. 4. Flowchart for constructing a FV-template from a FV-sample.

Finger Segmentation: First we generate a *foreground mask* using adaptive thresholding (Otsu's method [12]) to detect the hand-region (foreground object) in the image. Small regions in the resulting binary image are deleted using morphological opening. We then scan the foreground mask along the horizontal axis for the first foreground pixel (assumed to belong to the hand in the image). The location of detected pixel is assumed to correspond to the tip of one finger (the tallest finger). The left and right boundaries of the finger are obtained by scanning the input image horizontally for finger-edges on both sides. This scanning process is repeated for each row in the image, as long as the left and right finger-boundaries extracted correspond to a reasonable finger-width determined empirically from few samples. The scanning process terminates when the finger-width estimated for a given row exceeds the nominal finger-width value by a factor of 1.25. At this point we assume that we have identified all pixels representing a single finger – currently the tallest finger in the foreground region. Then we remove this finger from the binary mask (set all finger-pixels to background) and repeat the scanning process again, this time to find another finger. This procedure is repeated four times, to detect four fingers in the image.

In this approach, fingers are detected in order of their height in the input image (the finger closest to the top-edge of the image is detected first, followed by the second-tallest, and so on). We use the relative coordinates of the center-of-gravity of each finger-mask to renumber the fingers in a natural order from index- to little-finger. This procedure of renumbering the fingers works correctly only when all four fingers have been detected. (If, for example, only three fingers have been detected, then we cannot tell whether these are index-, middle- and ring-finger, or middle-, ring-, and little-finger.) For this reason, images where all four fingers are not detected, are excluded from further processing. The process of finger-segmentation is illustrated in Fig. 5, and an example of an extracted finger-region is shown in Fig. 6(a).

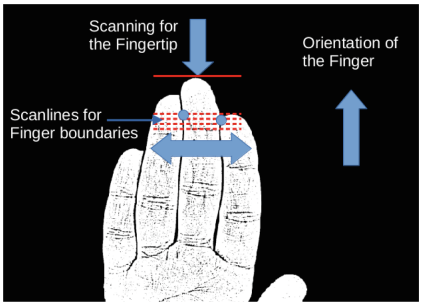


Fig. 5. Illustration of the finger-segmentation process.

Finger Normalization: Next, a normalization step proposed by Huang *et al.* [5] is applied to each individual finger-image. This step simply rotates the finger-image to align the longitudinal axis of the finger to the vertical axis as best as possible. The purpose of this normalization is to correct for minor hand-rotations ('yaw') during presentation. Compare the finger-orientations in Fig. 6(a) and (b) to understand the effect of this normalization step.

Finger-Vein Enhancement: We use a pre-trained autoencoder [2] to enhance the vascular structures in the input (normalized finger) image. Preliminary experiments showed that FV-enhancement improves the FV recognition accuracy significantly. Hence, we have included the FV-enhancement module in our processing pipeline. A sample result of the vein enhancement process is shown in Fig. 6(c).

Vascular Feature Extraction: FV patterns are compared based on a set of image-features extracted from the two vein-images being compared. In this work we have used the Maximum-Curvature (MC) features [11]. A sample result of the finger-vein MC feature-extraction process is shown in Fig. 6(d). The MC feature-map extracted for a single finger is considered as the biometric template for the finger.

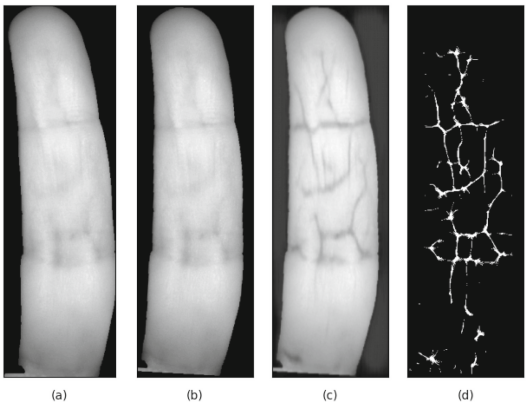


Fig. 6. Example result of the vein-enhancement. (a) Extracted finger-image; (b) normalized finger-image; (c) Vein-enhanced finger-image; (d) MC-feature-map extracted from (c). Note the slight rotation towards the vertical axis in (b) w.r.t. (a). The normalized finger-image, (b), forms the input to the vein-enhancement autoencoder [2].

4.1.2 Finger-Vein Matching

We have used the method proposed by Miura *et al.* [11] to compare two MC-feature based templates. This method uses cross-correlation (computed in the frequency domain) to find the position of best match of the two input feature-maps. The cross-correlation coefficient at the best-match position is taken as the match-score between the two templates.

4.2 FV PAD Using Photometric Stereo

Our baseline method for FV-PAD relies on a dense surface-normal map computed from the PS data recorded by each NIR camera. Before presenting our FV-PAD method, we briefly describe the processing of recovering the surface-normal map using PS.

4.2.1 Photometric Stereo Reconstruction

The LEDs in the *sweet* platform are organized in four banks. The four LED banks, when fired in a pre-defined sequence, illuminate the presented hand from different angles. The resulting set of four images can be used to recover the surface-normal map of the hand [20]. This PS reconstruction produces a pixel-wise surface-normal map. The method also produces an albedo image of the presented hand. Figure 7 shows an example of PS reconstruction using data from the *sweet* platform.

(a) (b) (c) (d) (e)

Fig. 7. Sample recovered 3D surface-normal map shown as a false-color composite (a), and its individual directional components: (b) X; (c) Y; and (d) Z. (e) Recovered albedo image.

Fig. 8. Flowchart of the proposed FV-PAD algorithm.

4.2.2 Proposed FV-PAD Method

The processing pipeline of the proposed FV PAD method is shown in Fig. 8. The various steps in the proposed FV PAD method are described here.

Surface-Normal Map: Recall that, in every sample, the *sweet* platform captures a sequence of four images where the presented hand is illuminated from a different angle in every image. The first step is to process these four input images to generate the surface-normal map for the hand (using the method described in Sect. 4.2.1). One surface-normal map is reconstructed for each NIR camera. Each map consists of surface-normal vectors computed for each pixel-position.

Hand-Region Detection: K-means clustering [1] is used to group pixels into two clusters. The input to the clustering method consists of the three components of the surface-normal as well as the albedo value of each pixel. The cluster with the higher mean-albedo is considered as the foreground cluster, which may include multiple isolated foregound regions. Based on the assumption that the largest foreground region corresponds to the presented hand, we use connected-component analysis to retain only the largest isolated region in the foreground cluster as the hand-region.

Feature Extraction: The feature-vector that we use for FV PAD is derived from a two-dimensional (2D) histogram of surface-normal angles. Each surface-normal vector, $(x,\ y,\ z)$, of the retained hand-region is transformed into an equivalent representation $(\theta_1,\ \theta_2,\ M)$ where θ_1 represents the angle of the vector with respect to the x-y plane and θ_2 represents the angle of the vector with respect to the y-z plane. The angles θ_1 and θ_2 are naturally bounded in the range $[-\pi/2,\ \pi/2]$. The magnitude, M, of the surface-normal vector is ignored. For each surface-normal map we construct a 2D histogram by quantizing θ_1 and

θ_2 each into N bins over the detected hand-region. ($N = 41$ in this study.) The N bins cover the range $[-\pi/2, \pi/2]$ in each dimension. The 2D histograms are used to construct feature-vectors to distinguish between *bona fide* and PA classes. For both classes, a large majority of the histogram-bins are empty (bin count = 0). Such empty bins act as confounders during the classification process. To mitigate this problem, we reduce the dimensionality of the feature-vectors using Principal Components Analysis (PCA) [1].

Feature Classification: We have used Gaussian Mixture Models (GMM) [1] to classify the feature-vectors generated in the previous step. One GMM is trained for each class (*bona fide*, PA). Given an input probe feature-vector, the likelihood that it comes from each of the two classes is evaluated using the GMMs corresponding to each class. Specifically, for a given feature-vector, let ll_{BF} and ll_{PA} be the log-likelihood values returned by the GMMs for the *bona fide* and PA classes, respectively. The difference $(ll_{BF} - ll_{PA})$ is taken as the classification-score for the probe feature-vector.

5 Baseline Experiments Using CandyFV Dataset

Baseline FVA and FV PAD results for the CandyFV dataset are discussed here. First we describe the methodology used in our baseline FVA experiments (Sect. 5.1), followed by the baseline FVA results on the CandyFV dataset (Sect. 5.2). In Sect. 5.3 we present the methodology adopted for our baseline FV PAD experiments. Baseline FV PAD results are presented in Sect. 5.4.

5.1 Methodology for FVA Experiments

Two sets of FVA experiments are presented here – single-finger FVA, and multi-finger FVA based on score-fusion. For the FVA experiments, first we group the subjects into two disjoint subsets, named the *development* ('Dev') set and the *evaluation* ('Eval') set. The Dev set is used for tuning hyper-parameters of the FVA system for the desired performance. The performance of the tuned FVA system is then evaluated using the Eval set, to quantify the performance of the system. The Dev and Eval sets have been constructed arbitrarily – data for the first 60 subjects has been assigned to the Dev set and data for the remaining subjects has been assigned to the Eval set.

Recall that for the FVA experiments, only the NIR-850 and NIR-950 images from CandyFV dataset are used. From each image captured by the *sweet* platform, we extract three individual finger-vein images, corresponding to the index-, middle- and ring-finger recorded in the image. Within each set (Dev or Eval), we have five FV samples for each subject. That is, for each camera we have 15 NIR-850 images and similarly 15 NIR-950 images, per hand, per subject. Considering three fingers (index, middle, and ring) per hand, we have 90 single-finger vein images for each camera and each NIR-illumination for each subject.

In single-finger FVA experiments, FV templates have been compared under eight different protocols, listed in Table 2.[1] The table shows the number of enrollment and probe samples in each subset, for the eight protocols. In each protocol, we have arbitrarily selected one sample of each subject as the enrollment sample. The remaining samples have been designated as probe-samples. For single-finger FVA, each enrolled sample is considered a unique identity. Each probe-sample has been used for four comparisons – one genuine comparison (with the correctly matched identity), and three *zero-effort-impostor* (ZEI) comparisons (with non-matched identities). In each ZEI comparison, the claimed-identity for a given probe-template is selected randomly.

Table 2. List of protocols for finger-comparison experiments. For each protocol, the number of enrollment images and probe images in the development (Dev) set, as well as in the evaluation (Eval) set are also listed. Each protocol name consists of three strings indicating the hand ('LH' or 'RH'), the NIR camera used ('left' or 'right'), and the NIR illumination (850 nm or 950 nm).

Id.	Protocol Name	Dev Set		Eval Set	
		Num. Images		Num. Images	
		Enrol.	Probe	Enrol.	Probe
P1	LH_left_850	159	8064	159	7836
P2	LH_left_950	141	4908	138	4665
P3	LH_right_850	159	7968	156	7791
P4	LH_right_950	147	6372	144	6300
P5	RH_left_850	156	7200	156	6998
P6	RH_left_950	153	6138	136	4644
P7	RH_right_850	156	7116	156	7032
P8	RH_right_950	144	4842	123	3969

Next, we consider each hand of a subject as a unique identity. We have implemented a finger-score fusion based method to identify each hand based on three fingers of the hand: index-, middle-, and ring-finger. Thus, each hand-probe is represented by a 3-D feature-vector consisting of single-finger FVA scores. While constructing these feature-vectors, finger-FVA scores are selected either only from genuine-probes of a given hand identity, or only from ZEI-probes of the hand in question. In this way we obtain, for each hand-identity, a set of genuine-probe ('match') feature-vectors, and another set of ZEI ('non-match') feature-vectors.

[1] Each protocol name is composed of three elements: <Hand>_<Camera>_<NIR>. The Hand may be 'LH' (Left hand) or 'RH' (Right hand). The Camera component ('left' or 'right') indicates the NIR camera from which the template has been derived. The NIR component may be '850' or '950', indicating the illumination used to capture the image-sample.

A two-class classifier is then constructed using the feature-vectors in the Dev set. This classifier is used to label the hand feature-vectors of the Eval set. In this study, we have used Support Vector Machines (SVM) with RBF (radial basis function) kernel [1], for the hand-identification experiments. In each protocol in Table 2, we fuse the FVA-scores of the three fingers of the hand. Thus, all probe feature-vectors used in a given experiment represent information from the same hand, captured by the same NIR camera, under the same NIR illumination.

5.2 FVA Performance at Fixed FMR

In this section we first present the results of single finger recognition, followed by hand-recognition results based on finger-score fusion. In all our experiments, we have estimated the recognition performance at the operating point corresponding to a specific FMR of 0.1%.[2] That is, the score-threshold is selected such that the FMR over the Dev set does not exceed the desired FMR limit. This score-threshold is then applied the Dev set and the Eval set, to determine the actual FMR and FNMR rates over each dataset.

5.2.1 Single Finger-Vein Recognition

In Table 3 we summarize the FMR and FNMR achieved for various evaluation protocols, for single-finger FVA, for the FMR ceiling of 0.1%. The results show that the FVA performance is significantly better for the right-hand fingers (protocols P5–P8) than for the left-hand fingers. We do not have any logical explanation for this phenomenon. We assume that right hand presentations by most subjects may have been more consistent (less variability) than left hand presentations, for one of two reasons: (1) increased familiarity with the data-capture

Table 3. Finger-vein recognition performance (expressed as percentages) at False-Match rate (FMR) of 0.1% (for the Dev set). The table shows results for single-finger recognition as well as hand-recognition based on finger-score fusion. The lowest HTER values are highlighted in bold characters.

Protocol	Single-Finger Recognition						Multi-Finger Hand Recognition					
	Dev Set			Eval Set			Dev Set			Eval Set		
	FMR	FNMR	HTER	FMR	FNMR	HTER	FMR	FNMR	HTER	FMR	FNMR	HTER
P1	0.1	3.82	1.96	0.17	7.3	3.74	0.1	2.68	1.39	0.20	4.29	2.25
P2	0.08	8.7	4.39	0.09	10.05	5.07	0.08	4.2	2.14	0.54	3.57	2.06
P3	0.08	5.77	2.93	0.0	6.46	3.23	0.05	2.71	1.38	0.05	3.93	1.99
P4	0.09	7.74	3.91	0.02	9.45	4.74	0.06	3.04	1.55	0.6	2.67	1.66
P5	0.09	0.66	0.38	0.79	0.61	0.7	0.06	0.0	0.03	0.46	0.0	0.23
P6	0.09	0.57	0.33	0.66	0.60	0.63	0.07	0.0	0.03	0.45	0.0	0.23
P7	0.09	0.33	0.21	0.68	0.45	**0.57**	0.06	0.0	0.03	0.11	0.0	**0.06**
P8	0.09	1.39	0.74	0.04	2.27	1.15	0.08	0.46	0.27	0.0	1.51	0.76

[2] In some experiments, the actual FMR for the Dev set may be slightly lower than the desired FMR ceiling of 0.1%.

procedure – subjects were consistently asked to present the left-hand first, or (2) simply due to right-handedness of most subjects.

The single-finger FVA results in Table 3 that the recognition-rates achieved for protocols involving 850 nm NIR illumination are usually somewhat better than the corresponding (*i.e.*, same hand, same camera) protocols involving 950 nm illumination. This result is counterintuitive. In theory, we expect 950 nm illumination to provide better results than 850 nm, because 950 nm NIR penetrates the soft-tissue of the fingers to a deeper extent than 850 nm NIR. Also, 850 nm NIR tends to produce more speckle noise on the skin-surface. On the other hand, much more power is needed for the 950 nm illumination. Our conjecture is that in the *sweet* platform the 950 nm illumination may not be sufficiently powerful.

5.2.2 Hand-Recognition Based on Finger-Score Fusion

The results of FVA-score fusion within each FVA protocol are also shown in Table 3. These numbers quantify the performance of the score-fusion system corresponding to the classification-score-threshold that limits the FMR over the Dev set to 0.1%. We note that FVA-score fusion improves the hand-recognition performance compared to single finger FVA. For the left hand, the single FVA error-rates (HTER in Table 3) range from 3.5% to 5% for each of the individual fingers. Finger-fusion reduces the left-hand recognition error-rates to about 2% or lower in all four left-hand protocols. For the right-hand, single-finger FVA performance is already very high (FVA in protocols P5–P8 in Table 3). Multi-finger FVA performance for the right hand still reduces the classification error. The best performance for multi-finger FVA, with a HTER of 0.06% for the 'RH_right_850', is almost a 10-fold improvement over single-finger FVA in the same protocol.

5.3 Methodology for FV PAD Experiments

In this section we explain the methodology for the FV PAD experiments using the method described in Sect. 4.2, on the CandyFV dataset. For the classification experiments reported here, the PCA model for dimensionality reduction has been trained to retain 99% of the total information in the input space (for the Training set). The trained PCA model is applied to transform the input histogram-features into a new 10-D feature-space. Thus, in the classification step, each presentation is represented by a feature-vector of length 10.

Data Partitions: We have randomly grouped the subjects into two non-overlapping sets: *Training* and *Test*[3]. Each set consists of data from 60 subjects – 10 target subjects, and 50 non-target subjects. In each set (training, test), the *bona fide* feature-vectors correspond to all 60 subjects comprising the set, whereas the PA feature-vectors come from the 10 target subjects in the set.

[3] Due to the small number of target-subjects in the CandyFV dataset, no separate development set has been used for tuning classifier hyper-parameters.

For each subject we consider only the feature-vectors representing the surface-normal map of the right hand, computed for one camera – the 'left' camera. The Training set consists of 590 *bona fide* and 1390 PA vectors. The Test set consists of 608 *bona fide* and 1382 PA vectors.

PAD Classification: Two GMMs have been trained, one using the *bona fide* feature-vectors and the other using the PA feature-vectors (of the Training set). For each GMM we select the number of Gaussians such that the Akaike Information Criterion (AIC) is minimum [1], based on trials with different numbers of Gaussians. Thus, we obtain a GMM with two Gaussians for the *bona fide* class and a GMM with 26 Gaussians for the PA class.

PAD Performance Metrics: We report the PAD results using the following figures of merit defined in the ISO/IEC 30107-3 standard [6]:

APCER: Attack Presentation Classification Error Rate – the proportion of PAs that are misclassified, and

BPCER: *Bona fide* Presentation Classification Error Rate – proportion of *bona fide* presentations that are misclassified.

The Average Classification Error Rate (ACER) – the mean of the APCER and BPCER – summarizes the two ISO figures of merit with a single number. APCER, BPCER, and ACER are expressed as percentages.

5.4 Baseline FV PAD Results

In this section we present the classification results produced by our baseline FV PAD algorithm at two classification-score-thresholds: (1) the threshold corresponding to the EER on the Training set, and (2) the threshold that limits the APCER to a maximum of 1% on the Training set. FV PAD results for both operating points are shown in Table 4. The table shows overall PAD results for each operating point, PAD results for each PA level (Level A and Level B), as well as for each individual PAI species.

5.4.1 PAD Classification at EER

Here we have selected the classification-score-threshold, $\mathcal{T}_{EER} = -1.389$, corresponding to the EER (ACER = 2.4%) over the Training set. The APCER and BPCER values achieved for the Test set using $\mathcal{T}_{EER}$ are shown in the left half of Table 4. Following ISO convention [6], when several PAI species of the same Level are used, the performance metrics of the worst performing species are to be taken as the aggregate PAD performance for the Level. Therefore, PAD performance for Level A corresponds to the PAD performance obtained for PAI species A1, and the performance metrics stated for Level B correspond to the species B3. We note that, apart from PAI species B3, B5 and B6, the ACER for the other species are low.

5.4.2 PAD Classification at Fixed APCER (1%)

Next we examine the performance of the proposed FV PAD method when the APCER is limited to 1% (this operating point is also referred to as the $BPCER_{100}$ point). Note that APCER $\leq 1\%$ is a very stringent constraint. The corresponding score-threshold, $\mathcal{T}_{1\%} = 0.464$, is determined from the Training set.

The PAD performance metrics for the Test set at this operating point are shown in the right half of Table 4. We note, again, that the APCER is high for three PAI species, namely B3, B5 and B6, but quite low for the remaining PAI species. We note that in this experiment $BPCER_{100}$ remains at an acceptable level for most applications. We also note that, except for the PAI species B3, B5 and B6, the APCER for the various PAI species does not exceed 3%.

5.4.3 Analysis of FV-PAD Results

The results in Table 4 validate the PS based surface-shape reconstruction approach for FV PAD, in general. The proposed method is clearly not adequate for PAs of species B3 (PAs where the hand is covered with a latex glove on

Table 4. Performance of the proposed FV-PAD method (on the Test set) at two operating points: the EER for the Training set (*i.e.*, score-threshold $\mathcal{T}_{EER} = -1.839$), and $BPCER_{100}$ (*i.e.*, $\mathcal{T}_{1\%} = 0.464$, which limits the APCER on the Training set to 1%). Note that the APCER observed for the Test set may be higher from the desired limit.

Level/Species	EER ($\mathcal{T}_{EER}$)			APCER $\leq 1\%$ ($\mathcal{T}_{1\%}$)		
	APCER	BPCER	ACER	APCER	BPCER	ACER
Overall	5.2	1.2	3.2	3.1	9.4	6.2
Level **A**	1.0	1.2	1.1	0.0	9.4	4.7
Level **B**	31.7	1.2	16.4	15.0	9.4	12.2
A1	1.0	1.2	1.1	0.0	9.4	4.7
A2	0.0	1.2	0.6	0.0	9.4	4.7
A3	0.0	1.2	0.6	0.0	9.4	4.7
A4	0.0	1.2	0.6	0.0	9.4	4.7
B1	1.0	1.2	1.1	0.0	9.4	4.7
B2	1.0	1.2	1.1	1.0	9.4	5.2
B3	31.7	1.2	16.4	12.2	9.4	10.8
B4	0.0	1.2	0.6	0.0	9.4	4.7
B5	18.0	1.2	9.6	15.0	9.4	12.2
B6	16.0	1.2	8.6	12.0	9.4	10.7
B7	0.0	1.2	0.6	0.0	9.4	4.7
B8	0.0	1.2	0.6	0.0	9.4	4.7
B9	5.0	1.2	3.1	2.0	9.4	5.7
B10	4.0	1.2	2.6	3.0	9.4	6.2

which vein-patterns have been drawn with dark, NIR-visible, ink). Surface-normal statistics alone are not sufficient to distinguish this class of PAs from *bona fide* presentations. Additional cues are required to detect PAIs of species B3.

PAI species B5, B6, B9, and B10 include PAIs made from printed vein-pattern images, where the hand region has been cut out following the outline of the fingers (see Fig. 3). Our experiments shows that the APCER is high for the two PAI species printed on laser printer (B5, B6) but significantly lower for B9 and B10, which have been printed on an inkjet printer. In general, we believe that for these four PAI species, additional data, in the form of more PAIs in the training set, may help improve the PAD results to some degree.

6 Conclusions

We present a new finger-vein (FV) dataset, named CandyFV, collected using a newly designed contactless finger-vein sensor named *sweet*. Our open-source sensor design is described in a separate paper. Unlike with most FV sensors currently available, to use *sweet* the user presents a hand at a distance of 10–15 cm. The platform can record FV samples of multiple fingers simultaneously, using multiple illumination regimes. At present the platform illuminates the presented hand with two different NIR wavelengths (850 nm and 950 nm) as well as white light. Among other imaging modes, it also collects a sequence of four images where the presented hand is illuminated from a different angle for every image. Photometric Stereo (PS) reconstruction may be applied to this image-sequence to compute a surface-normal map of the hand.

CandyFV includes data for both FV authentication (FVA) as well as FV PAD experiments. For FVA experiments, the dataset includes FV samples from 120 subjects. For FV PAD studies, samples from 20 subjects have been used to construct PAIs of 14 PAI species of Levels A and B. In this work we have also provided some baseline FVA and FV PAD results for the proposed dataset[4].

The baseline FVA pipeline used here relies on FV features extracted using the maximum curvature (MC) method. Our experiments show that, by combining three fingers of a hand, FVA accuracy of 0.06% FNMR can be achieved at the operating point corresponding to a FMR of 0.1%.

The baseline FV PAD method presented in this work analyzes the surface-normal map reconstructed using PS, to determine whether the corresponding presentation is *bona fide* or a PA. Overall FV PAD performance, summarized by the ACER for the test set, was 3.2% at EER and 6.2% when APCER was constrained to 1%. The CandyFV dataset presented here is rich in FV information, and can be used to further the state of the art in FV biometrics. In future

[4] The CandyFV dataset is available for research purposes under license via the following URL: https://www.idiap.ch/dataset/candyfv. Python code for our baseline methods for FVA and FV PAD, and our experimental protocols are available, under license, via the following URL: https://gitlab.idiap.ch/bob/bob.paper.icpr2024.candy_fv_pad.

work, we plan to develop a vision-transformer based method for FVA using this dataset, as well as more effective FV PAD approaches.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning, 1st edn. Springer (2007)
2. Bros, V., et al.: Vein enhancement with deep auto-encoders to improve finger vein recognition. In: Proceedings of IEEE International Conference BIOSIG (2021)
3. Geissbühler, D., et al.: *sweet* – an open source modular platform for contactless hand vascular biometric experiments (2024). https://arxiv.org/abs/2404.09376
4. Hou, B., Zhang, H., Yan, R.: Finger-vein biometric recognition: a review. IEEE Trans. Instrum. Measur. **71**, 1–26 (2022)
5. Huang, B., et al.: Finger-vein authentication based on wide line detector and pattern normalization. In: 2010 20th International Conference on Pattern Recognition (2010)
6. ISO/IEC 30107-3. Information Technology—Biometric Presentation Attack Detection – Part 3: Testing and Reporting. Standard, International Organization for Standardization, Geneva (2023). https://www.iso.org/standard/79520.html
7. Kotwal, K., Marcel, S.: Residual feature pyramid network for enhancement of vascular patterns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1588–1595 (2022)
8. Kovač, I., Marák, P.: Finger vein recognition: utilization of adaptive gabor filters in the enhancement stage combined with sift/surf-based feature extraction. SIViP **17**, 635–641 (2023)
9. Lu, Y., et al.: An available database for the research of finger vein recognition, vol. 1, pp. 410–415 (2013)
10. Miura, M., et al.: Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification. In: Proceedings on IAPR Conference on Machine Vision and Applications, vol. 15 (2004)
11. Miura, M., et al.: Extraction of finger-vein pattern using maximum curvature points in image profiles. In: Proceedings on IAPR Conference on Machine Vision Applications (2005)
12. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1) (1979)
13. Qiu, X., Kang, W., Tian, S., Jia, W., Huang, Z.: Finger vein presentation attack detection using total variation decomposition. IEEE Trans. Inf. Forensics Secur. **13**(2), 465–477 (2018)
14. Raghavendra, R., Busch, C.: Presentation attack detection algorithms for finger vein biometrics: a comprehensive study. In: Proceedings of the 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (2015)
15. Raghavendra, R., et al.: Transferable deep convolutional neural network features for fingervein presentation attack detection. In: 5th International Workshop on Biometrics and Forensics (IWBF) (2017)
16. Sengupta, S., et al.: SfSNet: learning shape, reflectance and illuminance of faces in the wild. In: Proceedings of the IEEE Conference on CVPR (2018)
17. Singh, J.M., et al.: Detecting finger-vein presentation attacks using 3D shape & diffuse reflectance decomposition. In: 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (2019)

18. Ton, B.T., Veldhuis, R.N.J.: A high quality finger vascular pattern dataset collected using a custom designed capturing device. In: Proceedings of 2013 International Conference on Biometrics (ICB), pp. 1–5 (2013)
19. Vanoni, M., et al.: Cross-database evaluation using an open finger vein sensor. In: 2014 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS) Proceedings, pp. 30–35. IEEE (2014)
20. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Opt. Eng. **19**(1) (1980)
21. Yang, J., Shi, Y., Wu, R.: Finger-vein recognition based on gabor features. In: Riaz, Z. (ed.) Biometric Systems, Chap. 2. IntechOpen, Rijeka (2011)
22. Yin, Y., et al.: SDUMLA-HMT: a multimodal biometric database. In: Biometric Recognition, pp. 260–268. Springer, Heidelberg (2011)
23. Zhang, R., et al.: Deep learning for finger vein recognition: a brief survey of recent trend (2022). https://arxiv.org/pdf/2207.02148.pdf

FedFVIQA: Personalized Federated Learning for Two-Stage Finger Vein Image Quality Assessment

Xingli Liu, Jian Guo(✉), Hengyu Mu, Chong Han, and Lijuan Sun

Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China
{1222045705,guoj,1022041005,hc}@njupt.edu.cn, sunlijuan_nupt@163.com

Abstract. Finger vein recognition systems are widely used in several fields, and existing methods usually require many high-quality images for training to ensure accuracy. As a result, research on the finger vein image quality assessment (FVIQA) has received considerable attention. However, in reality, the available finger vein images are often distributed across multiple organizations or companies. Due to insufficient data and a shortage of quality labels, it is difficult for such organizations and companies to independently train accurate FVIQA models. At the same time, due to user privacy and ownership constraints, it is typically not practical to pool data from multiple organizations or companies for model training. To address this problem, this paper introduces federated learning into FVIQA for the first time and proposes a personalized federated learning method for two-stage FVIQA (FedFVIQA). In the first stage, each client labels the quality of unlabeled finger vein images based on their similarity distribution for personalized scoring. In the second stage, the clients collaborate with a server for the training of quality classification model, thereby producing optimal personalized models. Finally, this paper reports extensive experiments conducted on the SDUMLA-HMT, NJUPT-FVP, HKPU-FV datasets to verify the superiority of the proposed method.

Keywords: Finger vein image quality assessment · Federated learning · Biometrics

1 Introduction

With the ongoing development of deep learning technology, the performance of finger vein recognition techniques based on deep learning has continuously improved, and these techniques have been widely applied in many fields, such as finance, healthcare, and logistics. Such finger vein recognition methods usually require many high-quality images for training. Moreover, attempting to use low-quality images for this purpose can cause the network to fail to extract stable and reliable features, ultimately leading to higher rejection and misidentification rates and a decrease in recognition accuracy. To enhance the performance of finger vein recognition, assessing the quality of finger vein images to identify

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 309–325, 2025.
https://doi.org/10.1007/978-3-031-78341-8_20

and reject such low-quality images has gradually become a popular topic of research.

Finger vein image quality assessment is one of the most effective ways to improve the accuracy and efficiency of recognition systems. The existing quality assessment methods for finger vein images can be divided into two categories: methods based on hand-designed features [3,19] and methods based on deep learning [15,24]. Quality assessment methods based on hand-designed features rely on the extraction of inherent features of finger vein images based on a priori knowledge as the basis for assessing image quality. These methods are effective, but the selection and design of the corresponding features rely on the empirical knowledge of domain experts, and the extracted features may not sufficiently capture the deep and complex features of an image. In contrast, quality assessment methods based on deep learning usually rely on the application of CNNs for finger vein image quality assessment (FVIQA) to achieve deep mining of the image features. Deep learning models can learn features of an image at different levels of abstraction. This multilevel learning allows deep learning-based methods to outperform methods belonging to the first category, but they often require training on large amounts of finger vein data and associated quality labels in order to correctly judge image quality. Currently, due to the insufficient amount of local data held by each organization or company and the small number of associated quality labels, it is difficult for any individual organization or company to train an effective and stable FVIQA model. Moreover, due to increasing privacy protection concerns worldwide, the data of each organization or company tends to be isolated in its own independent databases or applications. This problem of "data isolation" further exacerbates the difficulty of related research.

Federated learning (FL), a distributed learning framework designed with the aim of protecting user privacy achieves collaborative learning among individual clients. FL has been proposed to solve the distributed learning problem in scenarios requiring the protection of user privacy and has found numerous applications in the field of biometric identification [2,5,9]. However, in the context of FVIQA, further study is still needed for the following reasons: (1) There is a lack of sufficient labeled data. (2) The data heterogeneity across clients makes the traditional FL framework inapplicable. To address the above issues, this paper proposes a personalized FL method for two-stage FVIQA (FedFVIQA). In the first stage, each client implements quality annotation of its own local images. In the second stage, the clients and the server collaborate for the training of quality classification networks based on the obtained quality labels. Experimental results show that FedFVIQA has high stability and accuracy. The specific contributions of this paper are as follows:

– This paper proposes a new FL framework for image quality assessment (FedFVIQA). To our knowledge, this paper presents the first exploration of FVIQA in the FL setting. In addition, there is very little existing work on image quality assessment (IQA) based on FL in the field of biometrics in general, including face recognition, iris recognition, and fingerprint recognition.

Therefore, the method proposed in this paper is also informative for research in related fields.
- This paper proposes a personalized federated quality classification mechanism, which enables the creation of an optimal personalized model for each client to avoid the model offset problem.
- This paper proposes an adaptive finger vein image quality labeling mechanism that enables accurate scoring of image quality for automatic image quality labeling, thereby solving the problem of the shortage of quality labels for finger vein images.
- The results of extensive experiments conducted on the SDUMLA-HMT [22], NJUPT-FVP [16], HKPU-FV [6] datasets demonstrate the superiority of the proposed method.

2 Related Work

2.1 Finger Vein Image Quality Assessment

The existing FVIQA methods can be broadly classified into two categories: quality assessment methods based on hand-designed features and quality assessment methods based on deep learning. Quality assessment methods based on hand-designed features rely mainly on expert experience and aim to construct FVIQA models that conform to the characteristics of the human visual system. Representative methods of this type include Literature [3,12,14,18]. This type of method is useful, but its applicability is usually limited to a single dataset. In addition, this type of method relies on the empirical knowledge of domain experts, and researchers usually cannot access all the attributes related to image quality in finger vein images. Deep learning-based quality assessment methods mainly involve performing IQA by means of neural networks, aiming to establish recognition result-oriented FVIQA models. Representative methods in this category include methods Literature [15,17,23,25]. This class of methods performs better than the first class overall, but such methods usually require a large amount of high-quality data for training. In practice, however, the available data are typically limited and of varying quality. Sharing and access among different data sources are also limited due to incompatibility, privacy and ownership protection, and other factors.

2.2 The Application of Federated Learning in the Field of Biometrics

FL allows multiple devices or nodes to train a model together without sharing their original data, thus guaranteeing the privacy and security of the data. Since it was proposed in 2016, many experiments have proven the effectiveness and security of FL [21]. In the field of biometric identification in particular, FL technology also has numerous successful applications. This is because it addresses two pain points faced by traditional centralized training methods: privacy protection

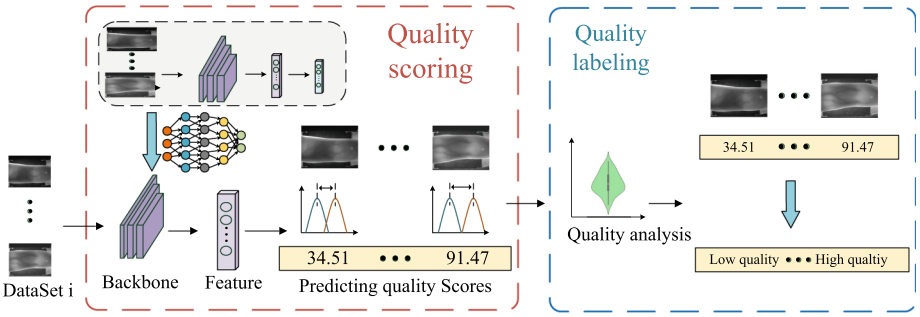


Fig. 1. First stage: adaptive image quality annotation. This stage is deployed on the client side. It consists of two parts: quality scoring (depicted in the red box) and image quality labeling (depicted in the blue box). (Color figure online)

and a lack of data. It allows clients to collaboratively optimize models without sharing their data, thereby indirectly expanding the training dataset while ensuring data privacy. Consequently, since its proposal, FL has attracted extensive attention from researchers in the field of biometrics. Related works have been proposed in various areas, such as face recognition [9], iris recognition [11], fingerprint recognition [1] and finger vein recognition [8]. Nevertheless, compared with other applications in biometrics, work on FL in the area of biometric IQA is relatively scarce, and related research is still in its infancy. This is mainly due to two reasons: the lack of available labeled data and the serious heterogeneity of the data. This prevents the classical FL methods cannot simply be directly applied for biometric IQA. To overcome the above challenges, this paper designs a personalized FL method for two-stage FVIQA (FedFVIQA).

3 Proposed Method

In this work, a new two-stage framework, FedFVIQA, is constructed for the FL task of FVIQA. FedFVIQA consists of two stages. The first stage, which is deployed locally on each client, consists of the automatic labeling of images with quality labels to solve the problem of insufficient quality-labeled finger vein images. The second stage is deployed both locally and on a server to enable multiple clients to collaborate for IQA. Each client locally trains a quality classification network, and personalized model aggregation is implemented on the server side. The details of the first and second stages are described in the following subsections.

3.1 First Stage: Adaptive Image Quality Annotation

This section describes the first stage of the FedFVIQA framework, in which each client implements the quality annotation of unlabeled finger vein images. Its specific flow is shown in Fig. 1. This stage is divided into two main parts: quality scoring and image quality annotation.

Quality Scoring. This section describes the implementation of the image quality score calculation. First, image features are extracted locally from training data. Then, the intraclass and interclass similarities of each image are calculated based on the obtained image features. Finally, the image similarity is statistically analyzed and scored.

First, the image features are extracted. To ensure the consistency and accuracy of the assessment, the feature extraction model used here is consistent with the finger vein recognition model used in the subsequent experimental tests, although with some modifications. Specifically, in this work, a 512-dimensional linear layer is deployed after the backbone network. This linear layer is designed as a feature embedding layer for extracting the feature vectors of images. Ultimately, the output of the recognition model is jointly determined by the results of the feature embedding layer and a classification layer. Accordingly, the model is trained to acquire image features.

Next, the similarities between images are calculated from the obtained image features, and the calculated similarity values are used to obtain the intraclass and interclass similarity distributions of the images. In this paper, the cosine similarity is used to measure the similarity between images. In Eq. 1 and Eq. 2, $T_{m_k}^{in}$ denotes the intraclass similarity distribution of a set of images m_k, $T_{m_k}^{out}$ denotes the interclass similarity distribution of the set of images m_k, and n_k and n_g denote image identity labels, $\cos(f(m_k), f(m_g))$ denotes the cosine similarity between the two feature vectors of images m_k and m_g.

$$T_{m_k}^{in} = \{cos\left(f\left(m_k\right), f\left(m_g\right)\right) | n_k = n_g\} \tag{1}$$

$$T_{m_k}^{out} = \{cos\left(f\left(m_k\right), f\left(m_g\right)\right) | n_k \neq n_g\} \tag{2}$$

$$T_{m_k}^{in'} = \left\{T_{m_k(v-1)\%b}^{in} | v = 1, 2, \ldots, a\right\} \tag{3}$$

However, since the lengths of $T_{m_k}^{in}$ and $T_{m_k}^{out}$ are not consistent, with the length of $T_{m_k}^{out}$ generally being greater than that of $T_{m_k}^{in}$, it is impossible to derive the distribution of the images by directly comparing $T_{m_k}^{in}$ with $T_{m_k}^{out}$; instead, it is necessary to expand the length of $T_{m_k}^{in}$ to be consistent with that of $T_{m_k}^{out}$. The expansion calculation is expressed in Eq. 3. In Eq. 3, $T_{m_k}^{in'}$ denotes the expanded $T_{m_k}^{in}$, v denotes the index of an element in $T_{m_k}^{in'}$, b denotes the length of $T_{m_k}^{in}$, and a denotes the length of $T_{m_k}^{out}$.

$$Distance_{m_k}\left(T_{m_k}^{in'} \middle|\middle| T_{m_k}^{out}\right) = inf_{\gamma \in \prod(T_{m_k}^{in'}, T_{m_k}^{out})} E_{(T_{m_k}^{in'}, T_{m_k}^{out}) \sim \gamma}\left[\left\|T_{m_k}^{in'} - T_{m_k}^{out}\right\|\right] \tag{4}$$

$$Q_{Score_{m_k}} = 100 \times \frac{Distance_{m_k}\left(T_{m_k}^{in'} \middle|\middle| T_{m_k}^{out}\right) - min(Distance)}{\max\left(Distance\right) - min(Distance)} \tag{5}$$

Finally, the image quality is calculated. The method of quality score calculation used here is borrowed from the method of quality label generation presented in [13]. In this paper, the Wasserstein distance is used to calculate the distance

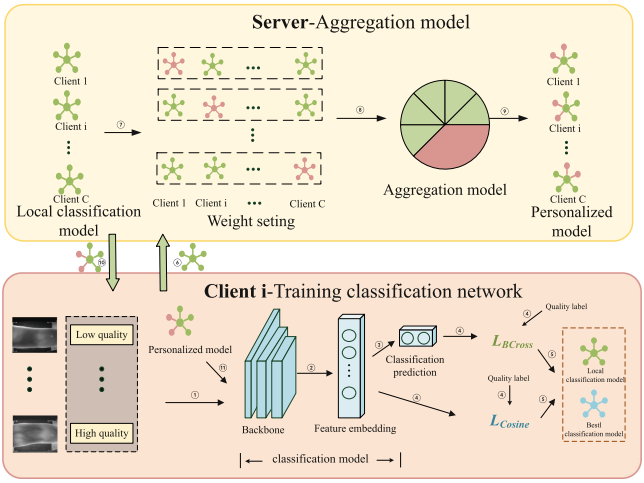


Fig. 2. Second stage: personalized federated quality classification. This stage is deployed on both the clients and the server. The client side performs quality classification model training, and the server side performs personalized model aggregation.

between intraclass samples and interclass samples, denoted by $Distance_{m_k}$, which is calculated as shown in Eq. 4. Then, this score is normalized to the range of [0,100], as shown in Eq. 5. In Eq. 4, *inf* stands for the infimum, $\prod\left(T_{m_k}^{in'}, T_{m_k}^{out}\right)$ denotes the set of probabilities of the joint distribution of $T_{m_k}^{in'}$ and $T_{m_k}^{out}$, E stands for the expectation, and $\left\|T_{m_k}^{in'} - T_{m_k}^{out}\right\|$ is the L2 norm difference between $T_{m_k}^{in'}$ and $T_{m_k}^{out}$, which is defined as the square root of the sum of the squared differences between the two distributions at corresponding positions. In Eq. 5, min() is the minimum value function, max() is the maximum value function, and distance represents the set of all possible values of $Distance_{m_k}$.

Quality Annotation. This section describes how each image is annotated as being of either high-quality or low-quality. These labels should not be determined based only on the raw scores. This is because the feature and quality distributions of each client's data differ, and the score distributions obtained based on the Wasserstein distance will also differ. Consequently, the same score may indicate a low-quality image for a client with a high overall image quality but a high-quality image for a client with a low overall image quality. As a result, using the same scoring standard for all clients may lead to misclassification and inaccurate recognition.

Therefore, this paper proposes an adaptive threshold calculation method to adjust the quality annotation for each client. This calculation method is shown in Eq. 6. In Eq. 6, *th* denotes the calculated threshold, *quality* denotes the set of quality scores for this client's images, d is a weighting coefficient, mean() denotes the mean function, and std() denotes the standard deviation function.

$$th = mean(quality) - std(quality)d \tag{6}$$

Finally, the image quality scores for each client data source are statistically analyzed to calculate a corresponding score threshold, and images with scores higher than *th* are tagged as high-quality images, whereas those with scores lower than *th* are tagged as low-quality images.

3.2 Second Stage: Personalized Federated Quality Classification

This section describes the second stage of the FedFVIQA framework, in which personalized federated quality classification is implemented on the client and server sides. The specific process is shown in Fig. 2. This stage is divided into two main parts: client-side quality classification model training and server-side personalized model aggregation. The specific process is described as follows.

Client-Side Quality Classification Model Training. This section describes the implementation of the client-side quality classification model training. Each client first constructs a quality classification model and then feeds its images labeled with quality tags into this classification model for training, thereby generating a local classification model. At this time, the optimal classification model is also initialized. Then, the clients upload their local classification models to the server for model aggregation. Once its aggregated model is returned, each client continues training on the basis of this aggregated model. This process is repeated until finally, the optimal trained model is obtained for each client and is saved to be used for image quality prediction.

The first step is to construct a quality classification model. To ensure the consistency and accuracy of the evaluation, the quality classification model used here is consistent with the third-party finger vein recognition model used later for experimental testing, although with some modifications. Specifically, in this paper, a 512-dimensional linear layer is deployed as a feature embedding layer after the backbone network of the recognition model, and the classification layer of the recognition model is removed and replaced with a 2-dimensional linear layer to serve as a classification prediction layer. The feature embedding layer is used to extract the feature vectors of the input images, and the classification prediction layer is used to predict whether each image is of high or low quality. Ultimately, the output of the recognition model is jointly determined by the results of the feature embedding layer and the classification prediction layer.

Next, the constructed quality classification model is locally trained to achieve image quality prediction. Images annotated with quality labels are fed into the quality classification model to obtain their feature vectors and corresponding classification predictions. In this way, a local classification model is generated, and at the same time, an optimal model is locally initialized. Specifically, the local model is the classification model from the last epoch of each round of local training, and the optimal model is the model obtained during the local training process that achieves the minimum loss value. To improve the performance of

the quality classification model in accurately predicting image quality, a loss function L_{loss} is designed for network training in this paper.

The loss function L_{loss} consists of the binary cross-entropy loss $L_{BinaryCrossLoss}$ and the cosine loss $L_{CosineLoss}$, as shown in Eq. 7. When the model is trained, the outputs of the classification prediction layer and the quality labels of the images are input together into $L_{BinaryCrossLoss}$, which is used to reduce the distance between the real image labels and the prediction probabilities. Similarly, the outputs of the feature embedding layer and the quality labels of the images are used together as the input to $L_{CosineLoss}$, which works to reduce the distance between similar samples and increase the distance between dissimilar samples.

$$L_{Loss} = L_{BinaryCrossLoss} + L_{CosineLoss} \tag{7}$$

There are generally more high-quality data than low-quality data in a dataset, so there will be an imbalance in the numbers of samples belonging to the high-quality and low-quality image categories. To solve this problem, in the method proposed in this paper, different weights are assigned to different sample categories when using the cross-entropy loss function to balance the difference in the amount of data between the different categories. The binary cross-entropy loss represents the difference between the true image labels and the predicted probabilities.

The cosine loss represents the feature distance between images in the same class and different classes and is calculated as shown in Eq. 8. The cosine loss value is the sum of the loss values obtained for the current image and each other image. In Eq. 8, I is the total number of samples, and $comp_{A,i}$ is the loss value for image A and other image i. The formula for $comp_{A,i}$ is shown in Eq. 9. In Eq. 9, $c_{A,i}$ denotes the cosine distance between the two feature vectors of images. And m denotes a boundary value; *similar* denotes images. A and i are similar samples; *dissimilar* denotes images. A and i are dissimilar samples.

$$L_{CosineLoss} = \sum_{i}^{I} comp_{A,i} \tag{8}$$

$$comp_{A,i} = \begin{cases} c_{A,i}^2, \ if \quad c_{A,i} \geq m \ and \ similar \\ (m - c_{A,i})^2, \ if \quad m \geq c_{A,i} \ and \ dissimilar \\ 0, \quad otherwise \end{cases} \tag{9}$$

At the completion of each round of local training, each client uploads its classification model to a central server. The server then sends the corresponding aggregated model back to each client. In turn, the individual clients continue training on their local data based on these aggregated models to further optimize their local optimal models. Finally, each client saves its optimal model obtained at the end of the training process. This optimal model is the one that performs best on the client's local data, i.e., the one that obtains the smallest loss value on these local data. These optimal models will then be used for image quality prediction.

Server-Side Personalized Aggregation. This section describes the implementation of personalized FL to generate an optimal personalized model for each client. The quality classification models are aggregated on a third-party server, which generates a more accurate personalized model for each client by integrating the quality classification models of all clients. The aggregation method for the personalized model $\Phi'_{g(n)}$ is shown in Eq. 10.

$$\Phi'_{g(n)} = \frac{(1-\theta)}{(num-1)} \sum_{u}^{U} \left(\Phi_{g(u)}\right) + \theta \Phi_{g(n)} \tag{10}$$

In Eq. 10, θ denotes a threshold used when aggregating models. num denotes the number of clients. U denotes the set of clients other than n, i.e., $U = N - n$, and N denotes the set of all clients. $\Phi_{g(n)}$ and $\Phi_{g(u)}$ denote the models sent by the nth client and the u th client, respectively. A different global model, $\Phi'_{g(n)}$, can be aggregated for each client through this personalized aggregation mechanism.

Finally, the server sends the corresponding aggregated model back to each client. Once a client receives its personalized aggregated model, the client continues local training on the basis of this model. In this way, each client can take advantage of the knowledge contained in the global model while optimizing a personalized model based on its own local data.

4 Experiments

4.1 Experimental Setup

In these experiments, validation was carried out on three public finger vein datasets: SDUMLA-HMT from the Machine Learning and Mining Laboratory of Shandong University (short for SDUMLA), NJUPT-FVP from Nanjing University of Posts and Telecommunication (short for NJUPT), HKPU-FV from the Hong Kong Polytechnic University (short for HKPU). The SDUMLA-HMT dataset consists of 636 classes, with 6 images per class, corresponding to a total of 3816 finger vein images. The NJUPT-FVP dataset consists of 840 classes of 10 images each, for a total of 8400 images. The HKPU-FV dataset consists of 312 classes, with 6 images per class, corresponding to a total of 1872 finger vein images. The samples in SDUMLA-HMT, NJUPT-FVP, HKPU-FV were divided into a training set and a test set at a 1:1 ratio, and each dataset was further divided into different groups of the same size to serve as independent datasets for different clients.

All the experiments reported in this paper were run on an octa-core Windows 11 machine with an Intel i7-9700K processor. The code was built and implemented using the PyTorch deep learning framework, and the models were trained using NVIDIA GeForce RTX 2080 Ti GPUs. For the finger vein recognition model in the experiments, ResNet18 [4] was used. All local clients performed 1 round of FL for every 3 epochs. The base learning rate was 0.01, and the batch size was 8.

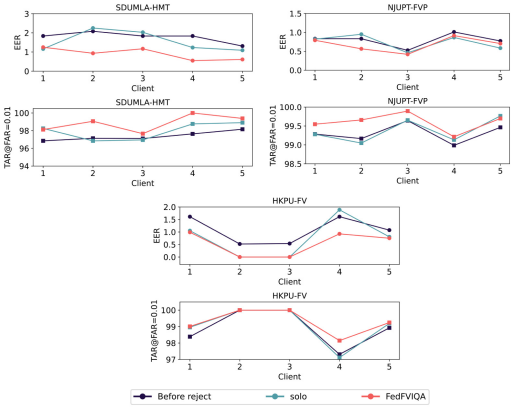


Fig. 3. Performance comparison of FedFVIQA with solo training. The black lines represent the performance of both FedFVIQA and solo training before the rejection of low-quality images, whereas the red line represents the performance of FedFVIQA after the rejection of low-quality images, and the blue line represents the performance of solo training after the rejection of low-quality images. (Color figure online)

To demonstrate the effectiveness and superiority of the proposed method, we analyze the identification of low-quality images by comparing the EER and TAR@FAR metrics before and after their rejection. By adjusting the threshold, the False Rejection Rate (FRR) is made equal to the False Acceptance Rate (FAR). The point at which FRR equals FAR is called the Equal Error Rate (EER), denoted as EER = FRR = FAR. EER is commonly used to evaluate the overall performance of the system, with a lower EER indicating better performance. If two samples are from the same class but are incorrectly identified by the system as being from different classes, it constitutes a false rejection. The FRR is the proportion of false rejections out of all genuine match attempts. Conversely, if two samples are from different classes but are incorrectly identified by the system as being from the same class, it constitutes a false acceptance. The FAR is the proportion of false acceptances out of all impostor match attempts. The True Accept Rate (TAR) represents the proportion of correctly accepted genuine matches. TAR@FAR = 0.01 indicates the TAR value when FAR is set to 0.01. A higher TAR value indicates greater system effectiveness. In our experiments, we set FAR to 0.01, meaning that one false acceptance is allowed for every 100 impostor attempts. Both EER and TAR results are presented as percentages.

4.2 Experimental Results and Analyses

To verify the performance of the method proposed in this paper, four groups of experiments were designed, and the results are discussed and analyzed here. The four groups of experiments are described as follows: (1) validate the effectiveness of the FedFVIQA framework; (2) validation of the need for a personalized aggregation approach; (3) ablation experiments; (4) comparative Experiments.

Table 1. Comparisons of solo centralized training and FedFVIQA.

Dataset	Methods	Rejection	EER	TAR
SDUMLA	Centralize	No	1.572	98.166
		Yes	1.400	98.488
	Solo	No	1.781	97.380
		Yes	1.555	97.946
	FedFVIQA	No	1.781	97.380
		Yes	0.904	98.847
NJUPT	Centralize	No	0.512	99.738
		Yes	0.451	99.865
	Solo	No	0.798	99.310
		Yes	0.737	99.377
	FedFVIQA	No	0.798	99.310
		Yes	0.680	99.603
HKPU	Centralize	No	0.748	99.252
		Yes	0.641	100.000
	Solo	No	1.072	98.925
		Yes	0.749	99.053
	FedFVIQA	No	1.072	98.925
		Yes	0.534	99.281

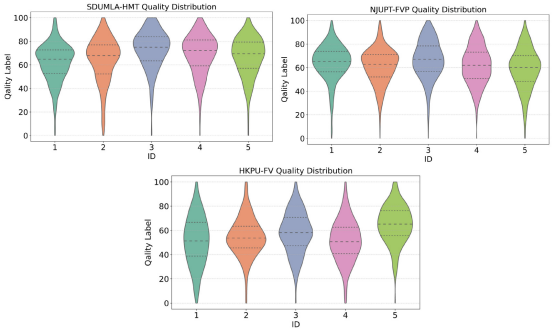


Fig. 4. Violin plots of the quality distributions.

Validation of the Effectiveness of the FedFVIQA Framework. FedFVIQA was compared with the solo and centralized training methods. Table 1 presents the results of the FedFVIQA, centralized, and solo methods for comparison, where the average of the client outcomes is reported for FedFVIQA and solo training. Moreover, Fig. 3 displays the specific FedFVIQA and solo results for each client.

As seen from Table 1, FedFVIQA outperforms solo training in terms of both metrics on average. Compared to centralized training, FedFVIQA achieves higher performance improvements in both metrics. On the SDUMLA dataset, the EER performance is improved by 49.24% with FedFVIQA, by 12.69% with solo training, and by 10.94% with centralized training; the TAR performance is improved

Table 2. Performance changes before and after using the personalized aggregation mechanism.

Dataset	Methods	EER	TAR
SDUMLA	Before rejection	1.781	97.380
	Without the mechanism	1.289	98.711
	With the mechanism	0.904	98.847
NJUPT	Before rejection	0.798	99.310
	Without the mechanism	0.684	99.532
	With the mechanism	0.680	99.603
HKPU	Before rejection	1.072	98.925
	Without the mechanism	0.746	99.184
	With the mechanism	0.534	99.281

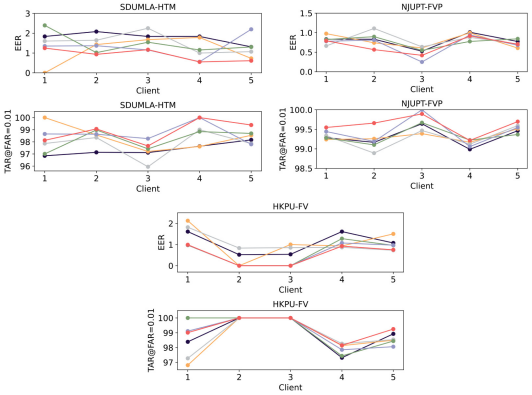


Fig. 5. Comparison of the method proposed with other methods for each individual client. The black lines represent the performance for all methods before the rejection of low-quality images.

by 1.51% with FedFVIQA, by 0.58% with solo training, and by 0.33% with centralized training. On the NJUPT dataset, the EER performance is improved by 14.79% with FedFVIQA, by 7.64% with solo training, and by 11.91% with centralized training; the TAR performance is improved by 0.30% with FedFVIQA, by 0.07% with solo training, and by 0.13% with centralized training. Figure 3 shows that the FedFVIQA method has stable performance on each client, and the performance on all of them is improved after the rejection of low-quality images. In contrast, solo training shows unstable performance among the clients, and on some clients, the performance even decreases after the rejection of low-quality images. For example, after the rejection of low-quality images, the EERs achieved with solo training for clients with IDs 2 and 3 on the SDUMLA dataset are higher than their EERs before the rejection of these images. This is because of the small amount of data available for each client, which makes effective model training impossible when using the solo approach. From the experimental results, it can be concluded that the FedFVIQA method has high robustness and accuracy.

Table 3. Results of the ablation experiments.

Dataset	Adaptive threshold	Personalized aggregation	L_{loss}	EER	TAR
SDUMLA	×	×	×	1.904	97.130
	√	×	×	1.840	97.322
	√	√	×	1.345	98.310
	√	√	√	0.904	98.847
NJUPT	×	×	×	0.795	99.351
	√	×	×	0.743	99.383
	√	√	×	0.722	99.552
	√	√	√	0.680	99.603
HKPU	×	×	×	.128	98.823
	√	×	×	0.771	99.213
	√	√	×	0.599	99.191
	√	√	√	0.534	99.281

Validation of the Need for a Personalized Aggregation Approach. To validate the need for the personalized aggregation method proposed in this paper, the performance of the FedFVIQA method was compared before and after the personalized aggregation mechanism was applied. Additionally, the quality distribution was visualized for each client, demonstrating that the data quality varied among the clients. The mean values of the results before and after the use of the personalized aggregation mechanism are recorded in Table 2, and Fig. 4 illustrates the quality distributions for each client.

As shown in Table 2, the use of the personalized aggregation mechanism results in an increase in the mean values of both metrics. On the SDUMLA dataset, the performance improvements in the EER and TAR metrics are 27.62% and 1.37%, respectively, before using the mechanism and 49.24% and 1.51%, respectively, after using the mechanism. On the NJUPT dataset, the respective performance improvements in the EER and TAR metrics are 14.29% and 0.22% before and 14.79% and 0.30% after. On the HKPU dataset, the respective performance improvements in the EER and TAR metrics are 30.41% and 0.26% before and 50.19% and 0.36% after. Figure 4 illustrates the quality distributions of the SDUMLA dataset, the NJUPT dataset and the HKPU dataset on the five clients. Figure 4 shows that the quality distributions of the different datasets are different, with the SDUMLA dataset having a higher overall quality, the NJUPT dataset having a more concentrated overall quality and the HKPU dataset having a lower overall quality. Moreover, the distributions of the same dataset on different clients also differ; for example, the quality scores of the data from the SDUMLA dataset held by the client with ID 1 are basically concentrated in the range of [50, 73], while the quality scores of the data held by the client with ID 3 are basically concentrated in the range of [63, 85]. For the NJUPT dataset, the quality scores of the data held by the client with ID 1 are basically concentrated in the range of [57, 72], while those of the client with ID 2 are basically concentrated in the range of [50, 70]. Similarly, for the HKPU dataset, the quality scores of the data held by the client with ID 4 are basically concentrated in the range of [41, 64], while those of the client with ID 5 are basically concentrated in the range of [52, 77]. These experimental results show that the data

quality distribution of each client may be different, meaning that a model that is well trained for one client may not achieve the same performance on other clients; therefore, it is necessary to propose a personalized aggregation method.

Ablation Experiments. The efficacy of the various components of the framework proposed in this paper was verified through ablation experiments, in which the base FedAvg [21] architecture was gradually supplemented with the following three components: the adaptive threshold calculation, the personalized model aggregation mechanism, and L_{loss}.

Table 3 shows that after the addition of the adaptive threshold calculation, the EER and TAR values are improved on both finger vein datasets, especially the HKPU dataset, where the EER is reduced by 31.65% and the TAR is improved by 0.39%. To test the effectiveness of the proposed personalized aggregation mechanism, this mechanism was implemented on the server in place of the traditional model averaging method. With the addition of this mechanism, the performance is further improved for almost all clients. In this case, the greater improvement occurs on the SDUMLA dataset, where the EER is reduced by 26.90% and the TAR is improved by 1.02%. Finally, L_{loss} denotes the use of the combined loss function designed in this paper instead of the traditional cross-entropy loss function used for performing network training. The experimental results show that the model performance reaches the optimal level for almost all clients after the introduction of the L_{loss} function. Especially on the SDUMLA dataset, the EER is reduced by 32.79%, and the TAR is improved by 0.55%. These results indicate that L_{loss} function designed in this paper is more effective for FL based on finger vein images. In summary, all three main components of the method proposed in this paper effectively improve the accuracy and robustness of FL for FVIQA.

Comparative Experiments. Finally, FedFVIQA was compared with other state-of-the-art methods. Because no FL method has previously been proposed for IQA, the classical FedAvg [21] and FedProx [7] methods were applied for FVIQA to serve as references in this evaluation. Moreover, the method proposed in this paper was compared with two traditional centralized IQA methods, EQ [10] and SER [20]. Because the FL approach is fundamentally different from the centralized learning approach, for the comparisons with the traditional centralized IQA methods, each centralized method was implemented locally at each client, and the average value across the clients was considered for comparison as well as the value for each client. Notably, since the SER and EQ methods also yield a quality score for each image, the same number of images as in the FedFVIQA method were rejected for each client in these comparisons. Table 4 shows the mean values of the results of each method across the five clients on three datasets, and the data for each client is shown in Fig. 5. As shown in Table 4, the FedFVIQA method outperforms the existing methods on three datasets. As shown in Fig. 5, the FedFVIQA method exhibits more stable performance among the individual clients than the other methods do, and in all cases, the perfor-

Table 4. Comparison with other methods.

Dataset	Methods	EER	TAR
SDUMLA	Before reject	1.781	97.380
	EQ	1.514	97.804
	SER	1.124	98.387
	FedAvg	1.485	98.189
	FedProx	1.327	98.669
	FedFVIQA(Our)	0.904	98.847
NJUPT	Before reject	0.798	99.310
	EQ	0.800	99.282
	SER	0.786	99.315
	FedAvg	0.778	99.329
	FedProx	0.693	99.447
	FedFVIQA(Our)	0.680	99.603
HKPU	Before reject	1.072	98.925
	EQ	1.021	99.282
	SER	1.113	98.687
	FedAvg	0.643	99.175
	FedProx	0.605	98.985
	FedFVIQA(Our)	0.534	99.281

mance is improved and more robust after the rejection of low-quality images. In summary, the method proposed in this paper outperforms both existing FL methods and traditional IQA methods with high performance.

5 Conclusion

In this work, we have proposed a personalized federated learning framework for two-stage finger vein image quality assessment (FedFVIQA). This framework improves the applicability of federated learning to finger vein image quality assessment tasks through the design of a personalized aggregation mechanism to accommodate client finger vein image datasets with different quality distributions. Moreover, we propose an adaptive scoring threshold calculation method for the automatic quality labeling of unlabeled finger vein images. In addition, we report extensive experiments conducted to evaluate the proposed method. The experimental results show that the FedFVIQA framework outperforms both existing federated learning methods and traditional image quality assessment methods with high robustness and accuracy.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62272242, and the Postgraduate Research

& Practice Innovation Program of Jiangsu Province under Grant KYCX23_1074 and Grant SJCX23_0268.

References

1. Che, T., et al.: Federated fingerprint learning with heterogeneous architectures. In: 2022 IEEE International Conference on Data Mining (ICDM), pp. 31–40. IEEE, Orlando, FL, USA (2022)
2. Guliani, D., Beaufays, F., Motta, G.: Training speech recognition models with federated learning: a quality/cost framework. In: ICASSP 2021–2021 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), pp. 3080–3084. IEEE, Toronto, Canada (2021)
3. Hartung, D., Martin, S., Busch, C.: Quality estimation for vascular pattern recognition. In: 2011 International Conference on Hand-Based Biometrics, pp. 1–6. IEEE, December 2011
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Kim, J., Park, T., Kim, H., Kim, S.: Federated learning for face recognition. In: 2021 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–2. IEEE, Las Vegas, NV, USA (2021)
6. Kumar, A., Zhou, Y.: Human identification using finger images. IEEE Trans. Image Process. **21**, 2228–2244 (2012)
7. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proc. Mach. Learn. Syst. **2**, 429–450 (2020)
8. Lian, F., Huang, J., Liu, J., Chen, G., Zhao, J., Kang, W.: FedFV: a personalized federated learning framework for finger vein authentication. Mach. Intell. Res. **20**(5), 683–696 (2023)
9. Liu, C., Wang, C., Chien, S., Lai, S.: Fedfr: joint optimization federated framework for generic and personalized face recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1656–1664. AAAI Press (2022)
10. Liu, R., Tan, W.: Eqface: a simple explicit quality network for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1482–1490, June 2021
11. Luo, Z., Wang, Y., Wang, Z., Sun, Z., Tan, T.: Fediris: towards more accurate and privacy-preserving iris recognition via federated template communication. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3357–3366 (2022)
12. Nguyen, D.T., Park, Y.H., Shin, K.Y., Park, K.R.: New finger-vein recognition method based on image quality assessment. KSII Trans. Internet Inf. Syst. **7**(2), 347–365 (2013)
13. Ou, F., et al.: SDD-FIQA: unsupervised face image quality assessment with similarity distribution distance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7670–7679, June 2021
14. Peng, J., Li, Q., Niu, X.: A novel finger vein image quality evaluation method based on triangular norm. In: 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 239–242. IEEE, Kitakyushu, Japan, August 2014

15. Qin, H., El-Yacoubi, M.A.: Deep representation for finger-vein image-quality assessment. IEEE Trans. Circuits Syst. Video Technol. **28**(8), 1677–1693 (2017)
16. Ren, H., Sun, L., Guo, J., Han, C.: A dataset and benchmark for multimodal biometric recognition based on fingerprint and finger vein. IEEE Trans. Inf. Forensics Secur. **17**, 2030–2043 (2022)
17. Ren, H., Sun, L., Guo, J., Han, C., Cao, Y.: A high compatibility finger vein image quality assessment system based on deep learning. Expert Syst. Appl. **196** (2022)
18. Shaheed, K., Qureshi, I.: A hybrid proposed image quality assessment and enhancement framework for finger vein recognition. Multimed. Tools Appl. 1–26 (2022)
19. Waluś, M., Bernacki, K., Popowicz, A.: Quality assessment of NIR finger vascular images for exposure parameter optimization. Biomed. Res. **27**(2), 383–391 (2016)
20. Yang, C., Guo, J., Zhang, Y., Huang, Y., Guo, Z.: Finger vein image quality assessment based on stochastic embedding robustness. In: Proceedings of the 42nd Chinese Control Conference, pp. 7889–7894. IEEE, Tianjin, China, 24–26 July 2023
21. Ye, D., Yu, R., Pan, M., Han, Z.: Federated learning in vehicular edge computing: a selective model aggregation approach. IEEE Access **8**, 23920–23935 (2020)
22. Yin, Y., Liu, L., Sun, X.: SDUMLA-HMT: a multimodal biometric database. In: Sun, Z., Lai, J., Chen, X., Tan, T. (eds.) CCBR 2011. LNCS, vol. 7098, pp. 260–268. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25449-9_33
23. Zeng, J., Chen, Y., Zhai, Y., Gan, J., Feng, W., Wang, F.: A novel finger-vein recognition based on quality assessment and multi-scale histogram of oriented gradients feature. Int. J. Enterp. Inf. Syst. (IJEIS) **15**(1), 100–115 (2019)
24. Zeng, J., Chen, Y., Qin, C.: Finger-vein image quality assessment based on light-cnn. In: 2018 14th IEEE International Conference on Signal Processing (ICSP), pp. 768–773. IEEE, February 2018
25. Zeng, J., Chen, Y., Qin, C., Gan, J., Zhai, Y., Feng, W.: Finger vein image quality assessment based on cascaded fine-tuning convolutional neural network. J. Image Graph. **24**(6), 271–315 (2019)

Infrared Database for Gait Recognition in Dynamic Outdoor Environment

Sonam Nahar[1(✉)] and Sasan Mahmoodi[2]

[1] School of Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India
sonam.nahar@sot.pdpu.ac.in
[2] School of Electronics and Computer Science, University of Southampton, Southampton, UK
sm3y07@soton.ac.uk

Abstract. Gait serves as an effective biometric for long-distance identification, particularly in scenarios where other biometric techniques present limited results. Most of the current gait recognition research relies on gait videos captured in controlled settings, predominantly using RGB cameras, while only a minority employ infrared cameras. There is a notable demand for real-time gait recognition in uncontrolled environments, especially utilizing infrared cameras for security and surveillance purposes. This study introduces a multi-frequency gait database constructed from long, medium, and short wavelength infrared (LWIR, MWIR, and SWIR) as well as visible (RGB) cameras in uncontrolled outdoor settings. The database encompasses recordings of individuals engaged in four distinct activities: normal walking, walking with a coat, carrying a backpack, and holding a briefcase. Additionally, it uses a knowledge-based system for silhouette extraction in dynamic environments. This research evaluates the robustness of state-of-the-art gait recognition methods to changes in environmental conditions, clothing, and carrying covariates by utilizing our dataset to establish a benchmark for databases captured across various frequency bands. Furthermore, it assesses gait recognition performance at lower scales (up to 0.05).

Keywords: Gait Recognition · Gait Database · Infrared and RGB · Gait Representation

1 Introduction

'Gait' is defined as the way people walk and is used as a behavioral biometric for human identification [6]. The suitability of gait as biometrics emerges from the fact that gait patterns can be captured and perceived from a distance in an unconstrained background. It does not require subject cooperation and can operate without interfering with the subject's activity, unlike other kinds of biometric features such as face, ear, iris, and fingerprint. With this advantage, gait-based recognition systems have been widely used in video surveillance and criminal

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 326–341, 2025.
https://doi.org/10.1007/978-3-031-78341-8_21

investigations [2]. However, gait can be altered by variations in appearance (such as clothing, carrying, view, etc.) and environmental conditions which makes the recognition task difficult. In recent years, there has been notable progress in the development of gait recognition methods that demonstrate robustness against these covariates [10,13].

The construction of a common gait database is essential for the development and fair evaluation of gait recognition approaches by considering various walking conditions (such as view, clothing, carrying objects, environment conditions, speed), the number and diversity in subjects, and the spectrum of the camera. Although several gait datasets have been constructed with diverse and large sets of subjects with varying views, clothing, and carrying conditions [4,9,12,14,15,17,18,20]. However, the variations in environmental conditions and spectrum of camera systems are still insufficiently addressed. Most of these databases capture the data in controlled environments, such as laboratories where each individual walks on a treadmill or a marked path. In laboratory settings, the illumination and background are controlled and only one individual walks into the scene. Few existing datasets were constructed in outdoor settings but they considered a static background. However, real-life problems require recognition systems that can accurately recognize individuals in more natural environments with multiple persons and objects in the scene. Another common aspect of the existing dataset is that they are mostly captured in the visible (RGB) spectrum. There is a demand for infrared gait datasets as infrared gait videos offer certain advantages over RGB videos in specific scenarios. Infrared cameras record the heat emitted by the human body rather than the colors and hence can capture images in low light conditions and are less susceptible to changes in the environment and appearance of the subject. These advantages make infrared gait videos more attractive in real-time security applications. There is still a lack of work on infrared-based gait recognition. CASIA-C [16] is the first thermal imagery dataset collected for gait recognition, whereas only the silhouettes of the dataset are publicly available. Recently, thermal gait datasets, IOTG [19] and CASIA-E [14] have been created for deep learning based gait recognition. Nonetheless, these thermal gait datasets were collected in a controlled environment using a 'near-infrared camera'. Therefore, it becomes essential to establish an infrared gait database for infrared-based gait recognition in uncontrolled outdoor environments, particularly for surveillance and security purposes.

In this paper, we address these limitations and construct a new real-time gait database for human identification. The key contributions of the paper are summarized as follows:

1. We introduce a multi-frequency gait database, **MultiFreqGAIT**[1], comprises of gait videos captured using a short-wave infrared (SWIR), a medium-wave infrared (MWIR), and a long-wave infrared (LWIR) thermal camera, along with a visible band color camera, in an uncontrolled and dynamic outdoor

[1] To access this database, please send emails to `sm3y07@soton.ac.uk`.

environment. In this context, "dynamic and uncontrolled" implies the presence of multiple individuals and objects in the background, with no control over illumination conditions. It contains gait recordings in four scenarios: normal walking, walking while wearing a coat, carrying a backpack, and holding a briefcase.
2. We demonstrate the validity of our dataset through experiments with state-of-the-art gait representations for human identification. Our results show the advantage of employing infrared gait videos for recognition in dynamic scenes. In addition, we assess the gait recognition performance at lower scales. This demonstrates the capability of gait recognition even from a distance.

Therefore, our dataset holds significant potential for advancing gait recognition in real-time scenarios. To the best of our knowledge, this is the first gait database comprising both infrared and RGB videos captured in uncontrolled and dynamic outdoor environments. No prior research has utilized three distinct wavelengths of infrared cameras for gait recordings. In gait recognition, the optimal camera angle is 90° as it provides the maximum information. Therefore, we capture gait sequences exclusively from a 90° perspective with respect to the cameras. In the future, we plan to expand the database to include various viewing angles, laying the groundwork for cross-view gait recognition in uncontrolled outdoor environments.

The remaining sections of the paper are structured as follows: In Sect. 2, an overview of existing gait databases is presented. Section 3 explains the developed dataset and uses a knowledge-based method to segment the silhouettes of subjects. Section 4 delves into the state-of-the-art gait representations. The experimental setup and results are discussed in Sect. 5. Finally, in Sect. 6, the paper concludes and mentions potential directions for future work.

2 Overview of Existing Gait Datasets

In Table 1, we present an overview of the popular and publicly available gait databases, detailing their characteristics such as spectrum (RGB/infrared), environment (indoor/outdoor), background (static/dynamic), as well as covariates such as clothing, footwear, carrying bags, walking speeds, viewing angles, etc. Our MultiFreqGAIT dataset specifications are also included for comparison.

3 MultiFreqGAIT Dataset

In this section, we explore the details of our data collection process and the extraction of silhouettes.

3.1 Data Collection

Gait videos were collected using long-wave infrared (LWIR), medium-wave infrared (MWIR), and short-wave infrared (SWIR) thermal cameras, as well as

Table 1. Overview of popular and publicly available gait datasets used in the literature

Dataset	Spectrum	No. of Subjects	Environment	Covariates
SOTON [12]	RGB	115	Static Indoor, Static Outdoor	Normal Walking; Shoes; 2 views; Clothing and Carrying.
USF HumanID [9]	RGB	122	Static Outdoor	Normal Walking; 2 Views; Carrying a Briefcase
CASIA-B [20]	RGB	124	Static Indoor	Normal Walking; 11 Views; Carrying a Bag; Wearing a Coat
CASIA-C [16]	Infrared	153	Static Outdoor	Three Walking Speeds; Carrying a Bag
CASIA-E [14]	RGB, Infrared	1014	Static Indoor, Static Outdoor	Carrying a Bag; Wearing a Coat
OU-ISIR Speed [17]	RGB	34	Static Indoor	Nine Walking Speeds
OU-ISIR Clothing [4]	RGB	68	Static Indoor	Up to 32 combinations of clothing
OU-ISIR LP Bag [18]	RGB	62,528	Static Indoor	Seven different carried objects
OU-ISIR MVLP [15]	RGB	10,307	Static Indoor	Normal Walking; 14 views
MultiFreqGAIT	**Infrared (LWIR, MWIR, SWIR) and RGB**	**31**	**Urban and Rural Outdoor Scenes with Dynamic Background**	**Normal Walking; Carrying a Bag; Carrying a Briefcase; Wearing a coat**

a color visible band camera, separately in "urban" and "rural" outdoor settings. LWIR, MWIR, and SWIR are distinct spectral bands in the infrared (IR) region. LWIR cameras function in the long-wave infrared spectrum, typically from 8 to 15 μm in wavelength, offering thermal sensitivity. MWIR cameras operate in the mid-wave infrared, spanning 3 to 8 μm, providing improved resolution and longer-range capabilities. In contrast, SWIR cameras, with wavelengths from 0.9 to 1.7 μm, detect reflected radiation. LWIR cameras are ideal for non-invasive gait analysis in low light, while MWIR cameras offer clear subject tracking.

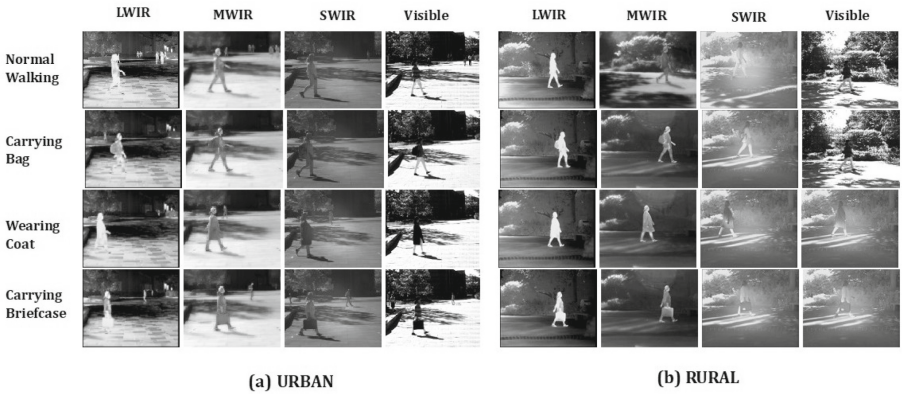


Fig. 1. Examples of gait images captured in (a) urban and (b) rural outdoor scenes, showcasing various walking conditions (normal walking, walking with a bag, walking with a coat, and walking with a briefcase) using LWIR, MWIR, SWIR, and visible cameras, as part of the MultiFreqGAIT dataset

SWIR cameras excel in material differentiation and dynamic imaging, suitable for detailed gait analysis in diverse lighting conditions. On the other hand, RGB camera captures the gait in the visible spectrum of light and are susceptible to appearance and illumination changes.

Urban recordings were conducted at the Highfield campus near building 32 at the University of Southampton, UK, while rural scenes were captured at Valley Garden near building 46 in Highfield campus. Both environments featured dynamic backgrounds with no control over moving objects or changes in illumination in the background. Data were collected from university students and staff of diverse ethnicities, religions, and body types. We considered side-view walking, where the subject walks perpendicular to the camera's viewpoint. The subject moved in a straight line from left to right and then back from right to left, and the camera was positioned at a distance of ten meters from the walking subject. All subjects walked wearing their clothing and footwear. Subjects were recorded as they walked normally along a straight line, six times w.r.t. each camera. Subsequently, subjects wore a long coat and walked twice along the same path, and similarly, two videos were recorded for each subject while carrying a bag and briefcase, respectively. The recording setup remained consistent across both urban and rural environments. Figure 1 displays examples of gait images captured in urban and rural outdoor scenes, showcasing various walking conditions (normal walking, walking with a bag, walking with a coat, and walking with a briefcase) using LWIR, MWIR, SWIR, and visible cameras. The video files had different frame sizes: 694×576 for LWIR, 640×512 for MWIR and SWIR, and 1440×1080 for the visible camera. They all maintained a frame rate of 30 frames per second (fps). Finally, gait data from 31 subjects were recorded in urban and rural settings, with each subject completing 12 walks in the scene

(6 normal walks, 2 wearing a coat, 2 carrying a bag, and 2 carrying a briefcase). This results in a total of $12 * 31 = 372$ gait videos for each camera in each scene. Thus, our database comprises eight subsets, each containing gait videos of 31 subjects in different environments using various camera types. The first four subsets (LW-U, MW-U, SW-U, and Vi-U) record gait videos of subjects in urban areas, employing LWIR, MWIR, SWIR, and Visible cameras respectively. While the remaining four subsets (LW-R, MW-R, SW-R, and Vi-R) contain videos of subjects in rural settings, captured using the same camera types. The collection of video data was approved by the university's ethical committee, and each subject provided informed consent for the use of the collected data for research purposes.

3.2 Silhouette Extraction

Silhouette extraction serves as the initial step in gait recognition, typically involving the application of background subtraction methods to isolate silhouettes in gait sequences. While this approach is effective in ideal scenarios where the foreground is in motion against a static background, but real-time situations as seen in our dataset, often feature dynamic backgrounds with multiple moving objects. In such cases, direct application of background subtraction results in the extraction of both foreground objects and moving elements in the background, leading to poor recognition accuracy. We introduce a silhouette extraction method that isolates the silhouette of the foreground person while filtering out other moving individuals in the background, which are regarded as artifacts and clutter.

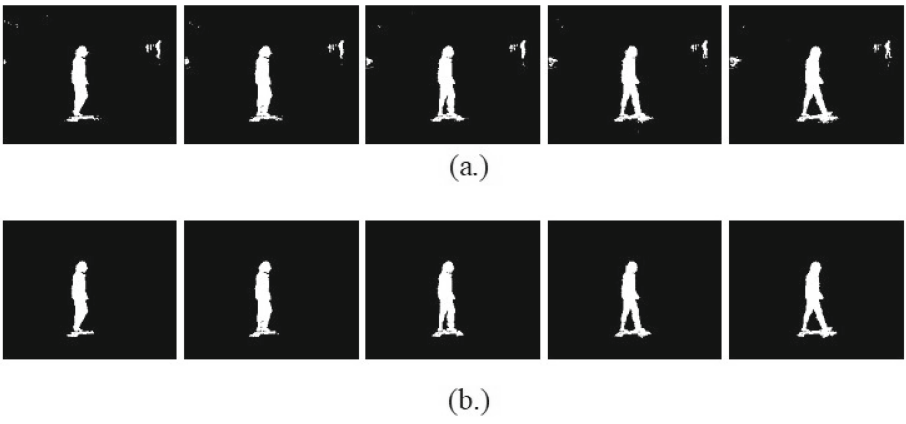


Fig. 2. Silhouette extraction results for a sample LWIR gait sequence in an urban scene, comparing (a) the direct background subtraction (GMM) method with (b) the knowledge-based approach

We outline our silhouette extraction method as follows: Initially, objects are detected in each frame of a gait video using YOLO-based object detection [7]. Each detected object is enclosed within a rectangular box and labeled by YOLO. Our objective is to extract the silhouette of the foreground person (subject of interest) in each frame. We record the bounding box locations labeled as "person" computed by YOLO to detect the walking person and by disregarding others. Employing a knowledge-based system, incorporating factors like the number of individuals of interest and their proximity to cameras, we differentiate between foreground and background persons, discarding boxes belonging to the latter. Consequently, only one box corresponding to the foreground person is recorded in each frame. The limitation of YOLO-based person detection arises when it fails to detect the subject of interest in certain frames. In such instances, the background person with the largest bounding box is incorrectly identified as the foreground. To address this, we calculate the distance of the bounding box in the X and Y directions relative to the box in the previous frame, ensuring its accuracy. If the distance exceeds a predefined threshold, indicating a false detection, the frame is marked as "no detection". Additionally, frames where YOLO fails to detect any person are also labeled as "no detection". To estimate the box for the subject in these missed frames, we employ the linear interpolation technique.

To extract the silhouettes, we now apply the background subtraction on the original gait video using the Gaussian Mixture Model (GMM) [21], a widely-used technique in computer vision. We then locate bounding boxes in each resulting frame using the above-described YOLO-based method. The binary region within these bounding boxes in the background-subtracted frames represents the subject's silhouette. We set all pixels outside the bounding box to zero. Figure 2 illustrates the effectiveness of our approach compared to direct background subtraction for silhouette extraction, demonstrating notably superior results.

4 Gait Recognition

To assess the gait recognition performance using our dataset, we consider state-of-the-art model-free gait recognition methods such as gait energy image (GEI [3], gait entropy image (GEnI) [1], frequency domain gait entropy (EnDFT) [8], effective parts frequency domain gait entropy (EP-EnDFT) [8] and GEI-Net, a deep learning based method [11]. In these methods, gait features are represented using the silhouettes of a walking subject. Initially, the silhouettes are pre-processed using size normalization and center alignment. In our experimental setup, we standardize the silhouettes to dimensions of 390×260 pixels. Gait cycles are then estimated using the method of autocorrelation presented in [5]. A gait cycle is defined as the time interval between the same repetitive events of walking that generally start when one foot is in contact with the ground.

The GEI is obtained by averaging silhouettes over a gait cycle as [3]:

$$GEI(i,j) = \frac{1}{F}\sum_{f=1}^{F} B(i,j,f). \tag{1}$$

Here, F stands for the total number of frames in a single gait cycle, B symbolizes a silhouette where pixel coordinates are indicated by i and j, and f denotes the frame index within the gait cycle. The silhouettes are binary images, meaning that $B(i,j,f)$ is either 0 or 1. GEI captures both the static components (head and torso) and dynamic regions (leg and arm movements) of a person's gait. Nonetheless, these components are vulnerable to changes in appearance caused by clothing, carrying objects, and environmental factors. Given that dynamic regions hold the most valuable information about human gait and are less influenced by these external factors, entropy-based gait features have been suggested to extract motion details from human silhouettes.

To generate a GEnI, the Shannon entropy is computed for every pixel across the silhouette images B throughout an entire gait cycle as follows [1]:

$$GEnI(i,j) = -(p(i,j)\log_2 p(i,j) + (1-p(i,j))\log_2(1-p(i,j)). \tag{2}$$

Here, $p(i,j)$ denotes the probability at pixel (i,j), representing the frequency of occurrences where $B(i,j)$ equals 1 throughout a gait cycle. Another variant of entropy-based gait features is the frequency domain gait entropy (EnDFT) [8]. To derive the EnDFT features, the discrete Fourier transform (DFT) of silhouette images is computed pixel by pixel across time (frame), after which the entropy of the DFT is calculated at each pixel over a gait cycle, focusing on the first three frequency components. EnDFT gives more weight to dynamic areas and less weight to static areas.

The gait signature encompasses various body parts, and the impact of different covariates doesn't uniformly affect all parts. While it may modify certain parts of the entire gait, others crucial for gait recognition remain unaffected. Based on this, an efficient part-based approach was introduced [8]. It involves dividing the entire body into five parts and utilizing only three parts that significantly enhance recognition rates. These segments, contributing positively, predominantly consist of dynamic elements engaged in motion throughout a gait cycle. Consequently, these effective (dynamic) parts are automatically selected by the algorithm proposed in [8], and the EnDFT features are employed to represent these effective (dynamic) parts. Among the five body parts, only the EnDFT features of the three most effective body parts are retained for recognition, while the remaining two parts are discarded. Deep learning has recently gained popularity in gait recognition, achieving state-of-the-art results. Therefore, we also evaluate our dataset using GEI-Net [11], a prominent deep learning method. A convolutional neural network is trained on gait energy images from our dataset to learn gait features. GEI-Net consists of two triplets of convolution, pooling, and normalization layers, followed by two fully connected layers that output similarities to individual training subjects. Recognition is performed using a simple

Euclidean distance measure between gallery and probe features. Figure 3 shows examples of GEI, GEnI, EnDFT, and effective part-based EnDFT gait features, respectively from our dataset. One can see that dynamic areas such as legs and arms are represented by higher intensity values while the static areas such as the head and torso have low values in the GEnI and EnDFT images. This is because silhouette pixel values in the dynamic areas are more uncertain and thus more informative, leading to higher entropy values. It can also be observed that the effect of appearance changes caused by carrying a bag and briefcase, and wearing a coat is significant in GEIs whereas it is reduced in entropy-based features. The identified effective parts primarily encompass dynamic regions, which are less influenced by changes in appearance (refer to Fig. 3(d)). We've illustrated these effective parts in the context of normal walking. These same body parts are employed for other variations in walking as well.

By employing principle component analysis (PCA) on these gait representations, we reduce the dimensionality of the data. The coefficients of the modes calculated in PCA are then used as feature vectors to represent gaits. With such feature vectors, gait recognition is achieved by using a K-NN classifier. The K-NN classifier is trained using the features of gait sequences in the gallery set and the subject/class recognition performance is evaluated for every gait sequence in the probe set.

5 Experimental Results and Discussion

This section shows experimental results using our dataset. Our gait database comprises 8 datasets: LW-U, MW-U, SW-U, Vi-U, LW-R, MW-R, SW-R, and Vi-R. The experiments are carried out separately for each dataset. Firstly, we extract the silhouettes from each video of a subject within every dataset, employing the method outlined in Sect. 3.2. Subsequently, we manually collect the silhouette sequences depicting the subject walking in either direction. For our experiments, we consider the left-to-right walking direction. In this context, a silhouette sequence refers to a collection of silhouette images depicting a subject's movement in one direction, also referred to as a "gait sequence". We capture 6

Table 2. Recognition accuracies in both urban and rural scenes using LWIR, MWIR, SWIR, and visible data when both gallery and probe contain normal walking sequences

	Urban				Rural			
	LWIR	MWIR	SWIR	Visible	LWIR	MWIR	SWIR	Visible
GEI	74.6	88.1	89.6	94.7	92.4	94.3	85.1	96.9
GEnI	72.0	89.2	89.8	93.8	91.2	94.8	85.1	95.8
EnDFT	95.2	90.0	89.9	98.4	98.0	95.0	91.2	100
EP	98.4	91.2	91.5	98.4	100	96.2	93.5	100
GEI-Net	98.5	92.4	92.5	98.6	98.2	96.2	94.8	100

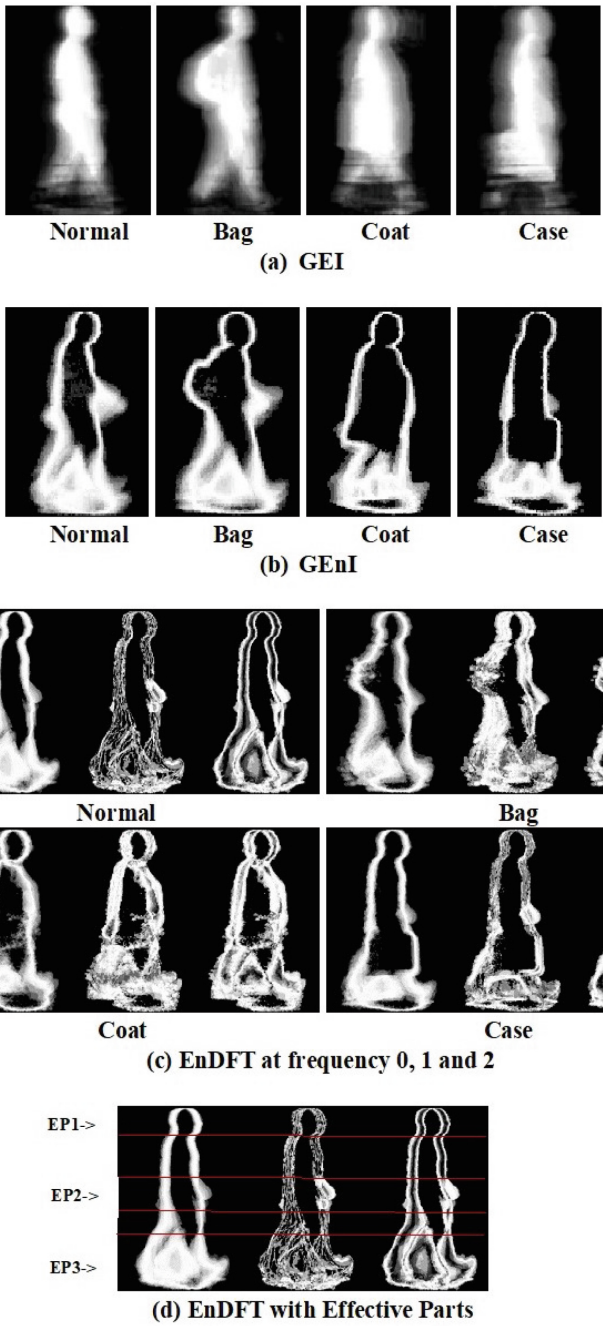


Fig. 3. Examples of (a) GEI, (b) GEnI, and (c) EnDFT gait features at frequencies 0, 1, and 2, showcasing various walking conditions including normal, carrying a bag, wearing a coat, and holding a briefcase. Additionally, (d) displays EnDFT features with three effective parts during a regular walking pattern

normal, 2 bag, 2 briefcase, and 2 coat gait sequences i.e., total 12 sequences for every subject in each dataset. Each gait sequence comprises 2 to 4 gait cycles. Every dataset is divided into two subsets: the gallery set and the probe set. The gallery set is constructed by taking the first four normal walking sequences (nm-01, nm-02, nm-3, nm-04) of each subject. The probe set is the rest of two normal (nm-05, nm-06), two bags (bg-01, bg-02), two briefcases (cs-01, cs-02), and two coat (ct-01, ct-02) sequences.

We evaluate the performance of state-of-the-art gait recognition methods (discussed in Sect. 4) using our dataset in terms of rank-1 recognition accuracy. Table 2 shows the results using LWIR, MWIR, SWIR, and visible gait data in both urban and rural scenarios where both the gallery and probe gait sequences belong to normal walking. Table 3, Table 4 and 5 demonstrates the recognition accuracies when the probe contains bag, coat, and briefcase sequences, respectively. We conduct these experiments to assess the robustness of state-of-the-art gait methods to changes in environmental conditions, clothing, and carrying covariates by utilizing our dataset to establish a benchmark for databases captured in various frequency bands. Note that all experiments were conducted on a machine with 11th Gen Intel(R) Core(TM) i7-1165G7, 2.80 GHz frequency, SSD - 512 MB, RAM - 8 GB, and System Type - 64-bit operating system.

Table 3. Recognition accuracies when gallery contains normal and probe with bag sequences

	Urban				Rural			
	LWIR	MWIR	SWIR	Visible	LWIR	MWIR	SWIR	Visible
GEI	22.9	41.1	57.2	45.4	39.3	45.6	55.3	83.4
GEnI	24	48.5	64.1	45.3	54.8	50.1	54.0	77.0
EnDFT	43.6	45.3	64.3	54.4	66.0	54.2	57.5	80.5
EP	45.2	65.7	65.1	61.7	75.0	65.3	63.1	84.8
GEI-Net	50.64	59.68	72.8	62.5	65.7	62.7	63.5	80.8

From the results, the performance of GEI and GEnI are nearly similar, while the EnDFT performs significantly better. EnDFT with effective parts (EP) and GEI-Net perform better in both thermal and visible bands, across both urban and rural settings. Despite the challenging background conditions, we achieve over 95% accuracy for normal sequences in urban scenes, with the highest recognition rate reaching 100% in rural scenes, particularly noticeable in LWIR and visible bands (see Table 2). As indicated in Table 3, the EP and GEI-Net methods attain higher accuracies, reaching approximately 65% in MWIR and SWIR in urban settings, but surpasses 80% in the visible band within rural areas, when the probe includes bag sequences. Likewise, we achieve over 65% accuracy in MWIR and SWIR bands within urban settings, with approximately 70% accuracy in the visible band in rural areas for both coat (Table 4) and briefcase

(Table 5) sequences in the probe. Overall, gait recognition methods show better performance when utilizing thermal data, particularly in MWIR and SWIR for urban scenarios. Conversely, performance is enhanced with the utilization of visible data in rural settings. Generally, urban environments present more complex dynamic backgrounds compared to rural ones. Our infrared gait data demonstrates superior performance over visible data in urban settings, particularly when dealing with covariates. This shows the advantage of employing infrared imaging for real-time gait recognition applications, such as security and surveillance. The state-of-the-art gait methods exhibit excellent performance when applied to our outdoor infrared and RGB dataset to showcase robustness against covariates such as clothing variations and carrying objects.

Table 4. Recognition accuracies when gallery contains normal and probe with coat sequences

	Urban				Rural			
	LWIR	MWIR	SWIR	Visible	LWIR	MWIR	SWIR	Visible
GEI	28.9	38.4	27.3	17.0	49.4	26.9	16.7	38.9
GEnI	21.7	43.2	24.3	25.0	45.4	32.7	25.9	45.9
EnDFT	27.5	32.3	35.3	39.6	56.2	36.3	42.4	63.7
EP	41.9	68.9	62.5	58.7	66.7	39.2	47.9	70.5
GEI-Net	39.9	58.3	60.5	55.7	54.7	35.2	45.2	65.5

Table 5. Recognition accuracies when gallery contains normal and probe with briefcase sequences

	Urban				Rural			
	LWIR	MWIR	SWIR	Visible	LWIR	MWIR	SWIR	Visible
GEI	24.8	36.5	35.4	27.4	42.7	35.8	34.9	62.7
GEnI	26.9	35.8	55.4	42.2	47.9	32.7	31.9	50.7
EnDFT	46.8	41.1	54.2	55.0	50.0	35.2	33.5	65.9
EP	48.4	51.8	69.7	56.7	53.0	39.5	39.9	70.5
GEI-Net	40.4	44.4	60.7	51.7	49.1	34.1	55.1	65.5

The characteristic of gait is that it can be efficiently recognized at a distance. Hence, we perform experiments to assess the gait recognition performance at lower image sizes (scales). We directly subsample the silhouette images of the gait images because we observed in our experiments that the visual quality of the scaled silhouette images was similar to the silhouettes obtained from the scaled gait images. Therefore, to reduce the time complexity, we directly subsample or scale down the silhouette images of the original gait images using

the nearest neighbour interpolation method. With the use of nearest neighbour interpolation, scaled-down images remain binary. We choose EnDFT features with effective parts to evaluate the gait recognition performance at smaller scales due to their superior performance in both thermal and visible bands, across both urban and rural settings, as evidenced in our results.

We consider various scale factors for our experiments such as 0.7, 0.5, 0.3, 0.1, and 0.05. For a fair comparison, the gait images without scaling i.e., scale factor = 1 are also considered. Tables 6, 7, 8 and 9 depict the recognition accuracies across various scale factors in urban and rural settings for LWIR, MWIR, SWIR, and visible gait images, respectively. The results in these tables include various walking conditions in the probe, such as normal, bag, coat, and briefcase.

Table 6. Experimental results for EnDFT with effective part-based gait recognition across various scales using LWIR gait data in both urban and rural scenes

Scale	LWIR-Urban				LWIR-Rural			
	Normal	Bag	Case	Coat	Normal	Bag	Case	Coat
1	98.4	**45.2**	**48.4**	**41.9**	**100**	**75.0**	53.0	**66.7**
0.7	95.2	32.6	46.8	29.1	97.9	45.8	**54.3**	58.3
0.5	**100.0**	35.8	45.2	30.7	97.9	54.1	**54.3**	64.5
0.3	98.4	35.5	**48.4**	37.0	95.8	45.8	**54.3**	60.4
0.1	98.4	40.4	50.0	38.7	83.3	33.3	36.9	52.0
0.05	70.9	30.7	27.5	27.4	70.8	39.5	32.6	35.4

From Tables 6, 7, 8 and 9, we can see that in many of the cases, the recognition accuracies significantly increase concerning the lower scale factors. This is because when we down-sample the gait images, the noise and artefacts are lowered and hence we obtain better recognition rates at lower scales. We achieve a 100% accuracy rate for normal walking sequences in urban environments using

Table 7. Experimental results for EnDFT with effective part-based gait recognition across various scales using MWIR gait data in both urban and rural scenes

Scale	MWIR-Urban				MWIR-Rural			
	Normal	Bag	Case	Coat	Normal	Bag	Case	Coat
1	**91.2**	**65.7**	51.8	**68.9**	**96.2**	**65.3**	39.5	39.2
0.7	84.4	44.6	51.7	**68.9**	84.7	60.8	36.9	47.8
0.5	84.4	55.3	55.1	62.0	86.9	58.6	41.3	45.6
0.3	84.4	44.6	55.1	56.8	86.9	54.3	50.0	41.3
0.1	87.9	50.0	**58.6**	65.5	91.3	56.5	**58.6**	**60.8**
0.05	86.2	42.8	44.8	50.0	82.6	50.0	43.4	41.3

Table 8. Experimental results for EnDFT with effective part-based gait recognition across various scales using SWIR gait data in both urban and rural scenes

Scale	SWIR-Urban				SWIR-Rural			
	Normal	Bag	Case	Coat	Normal	Bag	Case	Coat
1	91.5	65.1	69.7	62.5	**93.5**	63.1	**39.9**	47.9
0.7	91.0	57.1	67.8	64.2	91.3	**69.5**	31.8	**52.1**
0.5	91.0	58.9	67.8	**67.8**	93.4	69.5	34.0	**52.1**
0.3	**92.8**	64.2	**71.4**	66.0	91.3	67.3	36.3	47.8
0.1	82.1	**75.0**	60.7	66.0	93.4	58.6	36.3	50.0
0.05	57.1	23.2	28.5	28.5	65.2	44.3	28.7	37.8

Table 9. Experimental results for EnDFT with effective part-based gait recognition across various scales using Visible gait data in both urban and rural scenes.

Scale	Visible-Urban				Visible-Rural			
	Normal	Bag	Case	Coat	Normal	Bag	Case	Coat
1	**98.4**	**61.7**	56.7	**58.7**	**100.0**	84.8	**70.5**	70.5
0.7	93.3	53.3	**68.3**	48.2	**100.0**	78.2	63.6	70.4
0.5	98.3	48.3	**68.3**	44.8	**100.0**	78.2	65.9	**72.7**
0.3	93.3	58.3	**68.3**	50.0	**100.0**	**86.9**	68.1	68.1
0.1	90.0	55.0	63.3	53.4	**100.0**	84.7	70.4	56.8
0.05	88.3	41.6	60.0	48.2	87.5	71.7	65.9	56.8

LWIR data at a scale of 0.5, and in rural settings using visible data up to a scale factor of 0.1. Typically, the SWIR camera produces high-resolution images, thereby maintaining recognition performance up to a scale factor of 0.1 without degradation, as indicated in Table 8. We can see in our results that 0.05 is the lowest scaling factor where the recognition performance starts to degrade across all scenarios. The results show the advantage of gait in real-time video surveillance systems, as it can be recognized even at low image resolutions or scales while the other biometrics like face, iris, fingerprint, etc. require relatively high image resolutions for person identification.

6 Conclusions and Future Work

We introduced a multi-frequency gait database constructed from long, medium, and short wavelength infrared (LWIR, MWIR, and SWIR) as well as visible (RGB) cameras in uncontrolled outdoor scenes. It contains recordings of 31 persons, walking in four variations: walking normally, carrying a backpack, wearing a coat, and holding a briefcase in urban and rural scenes. A knowledge-based system is used here for silhouette extraction in dynamic environments.

Our experimental results indicate that state-of-the-art gait recognition methods perform effectively when applied to our infrared and RGB datasets, demonstrating robustness against appearance variations caused by changes in clothing and carrying covariates. The results show the importance of infrared gait recognition, especially in real-time urban scenarios. In addition, our results demonstrate the capability of gait recognition even from a distance, specifically at reduced scales (up to 0.05).

In the future, we plan to extend our dataset to include at least 300 subjects with various viewing angles. This extensive dataset will subsequently serve as the foundation for cross-view and deep learning-based gait recognition in uncontrolled outdoor settings.

Acknowledgement. This contribution is a part of the "Gait Biometric System" project, supported by the Defence Science and Technology Laboratory (DSTL) of the United Kingdom government. We extend our gratitude to the DSTL team for their efforts in establishing the camera system and recording gait videos at the University of Southampton. We would like to express our appreciation to Prof Mark Nixon, Co-Principal Investigator of the project, for his invaluable guidance and support. Additionally, we wish to thank all volunteers who participated in the data collection process.

References

1. Bashir, K., Xiang, T., Gong, S.: Gait recognition using gait entropy image. In: 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009). pp. 1–6 (2009)
2. Bouchrika, I., Goffredo, M., Carter, J., Nixon, M.: On using gait in forensic biometrics. J. Forensic Sci. **56**(4), 882–889 (2011)
3. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Trans. Pattern Anal. Mach. Intell. **28**(2), 316–322 (2006)
4. Hossain, M.A., Makihara, Y., Wang, J., Yagi, Y.: Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. Pattern Recogn. **43**(6), 2281–2291 (2010)
5. Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., Yagi, Y.: Gait recognition using a view transformation model in the frequency domain. In: Proceedings of the 9th European Conference on Computer Vision - Volume Part III. p. 151-163. ECCV'06, Springer-Verlag, Berlin, Heidelberg (2006)
6. Murray., M.: Gait as a total pattern of movement. American Journal of Physical Medicine **1**(46), 290–333 (1967)
7. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection (2016)
8. Rokanujjaman, M., Islam, M.S., Hossain, M.A., Islam, M.R., Makihara, Y., Yagi, Y.: Effective part-based gait identification using frequency-domain gait entropy features. Multimedia Tools Appl. **74**(9), 3099-3120 (may 2015)
9. Sarkar, S., Phillips, P., Liu, Z., Vega, I., Grother, P., Bowyer, K.: The humanid gait challenge problem: data sets, performance, and analysis. IEEE Trans. Pattern Anal. Mach. Intell. **27**(2), 162–177 (2005)
10. Sepas-Moghaddam, A., Etemad, A.: Deep gait recognition: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 264–284 (2023)

11. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: View-invariant gait recognition using a convolutional neural network. In: 2016 International Conference on Biometrics (ICB). pp. 1–8 (2016)
12. Shutler, J.D., Grant, M.G., Nixon, M.S., Carter, J.N.: On a large sequence-based human gait database. In: Applications and Science in Soft Computing, pp. 339–346. Springer (2004)
13. Singh, J.P., Jain, S., Arora, S., Singh, U.P.: Vision-based gait recognition: A survey. IEEE Access **6**, 70497–70527 (2018)
14. Song, C., Huang, Y., Wang, W., Wang, L.: Casia-e: A large comprehensive dataset for gait recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 2801–2815 (2023)
15. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Transactions on Computer Vision and Applications **10**(1), 4 (2018)
16. Tan, D., Huang, K., Yu, S., Tan, T.: Efficient night gait recognition based on template matching. In: International Conference on Pattern Recognition. Hong Kong, China (August 2006)
17. Tsuji, A., Makihara, Y., Yagi, Y.: Silhouette transformation based on walking speed for gait identification. In: Computer Vision and Pattern Recognition. San Francisco, CA, USA (June 2010)
18. Uddin, M.Z., Ngo, T.T., Makihara, Y., Takemura, N., Li, X., Muramatsu, D., Yagi, Y.: The ou-isir large population gait database with real-life carried object and its performance evaluation. IPSJ Transactions on Computer Vision and Applications **10**(1), 1–11 (2018)
19. Youssef, F., El-Mahdy, A., Ogawa, T., Gomaa, W.: Thermal gait dataset for deep learning-oriented gait recognition. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2023)
20. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th International Conference on Pattern Recognition (ICPR'06). vol. 4, pp. 441–444 (2006)
21. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. **27**(7), 773–780 (2006)

Recurrent Graph Convolutional Network for Sequential Pose Prediction from 3D Human Skeleton Sequence

Tomohiro Fujita[1,2(✉)] and Yasutomo Kawanishi[2]

[1] Doshisha University, 610-0394 Kyoto, Japan
[2] Guardian Robot Project, RIKEN, IH, 619-0288 Kyoto, Japan
tofujita@mail.doshisha.ac.jp, yasutomo.kawanishi@riken.jp

Abstract. Temporal human pose prediction from a 3D human skeleton sequence is vital for robot applications such as autonomous control and human-robot interaction. Recent pose prediction methods generally make predictions using GCN. However, because all frames of human poses are processed at once using a GCN, it is necessary to wait for preprocessing until all input frames are available, and intermediate predictions cannot be obtained until post-processing is complete. In addition, when predicting not in the time domain but in the frequency domain, if the input/output time is less than a few seconds, the number of sampling points is extremely small and the frequency resolution is low. In this study, we propose Recurrent Graph Convolutional Network (RGCN) and its application to a pose prediction. The advantages of RNN and GCN for sequential predictions in the time domain without frequency transformation are combined into RGCN to address the problem of existing methods. Through evaluation on public datasets, we confirmed that the accuracy of the proposed model using RGCN, which predicts sequentially in the time domain with simple processing and fewer parameters, is comparable to that of latest prediction method.

Keywords: Pose prediction · sequential prediction · Recurrent Graph Convolutional Network (RGCN)

1 Introduction

Predicting human pose and location in future seconds is essential task for robot applications such as robot-human interaction [10,15], human tracking [9], and autonomous control [5]. Even if we can predict the future poses of a target person only one second in advance, the prediction is useful for robot applications. Therefore, we aim to predict the future human poses as accurately as possible for about one second.

When a robot automatically controls or interacts with humans, it is essential to be robust to the surrounding environment to increase the stability of the system. Therefore, human poses are generally represented by a time series of

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 342–358, 2025.
https://doi.org/10.1007/978-3-031-78341-8_22

a set of the body-joint locations on 3D coordinates because 3D human skeleton sequences are robust to environmental factors such as person's clothing, surrounding situation, and background.

Traditionally, machine learning models, such as hidden Markov model [2] and restricted Boltzmann machine [24], have been used for pose prediction. However, in recent years, deep learning models with higher prediction accuracy, such as recurrent neural networks (RNNs) and graph convolutional networks (GCNs), have been used [17].

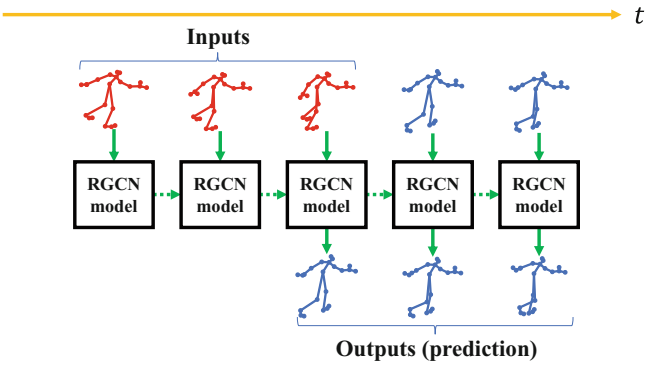


Fig. 1. Concept of sequential pose prediction using the proposed Recurrent Graph Convolutional Network (RGCN). The proposed model can explicitly learn the relationships between body joints and predict future poses by sequentially processing pose sequences.

Since 3D human skeletons can be regarded as graphs, pose prediction methods that use GCN, which is suitable for processing graph structures, are increasing. Some prediction methods that consider human body joints as a graph as it is [3,22] and other methods that convert the movement of each body joint into frequency coefficients using Discrete Cosine Transform (DCT) [4,18] have been proposed. However, although the methods using GCN can explicitly learn the relationship between body joints and have high accuracy, they consider the body joints of all frames as one graph; they must wait for pre-processing until all input frames are available, and prediction results cannot be obtained until all post-processing is completed. Therefore, there is a problem that the waiting time required for input/output processing is long. In addition, when predicting in the frequency domain using DCT, the number of sampling points is extremely small for input/output of several seconds or less, and the frequency resolution is low.

On the other hand, methods using RNN predict by treating human poses as time-series data. The advantage of using RNN is that the next joint coordinates can be simply obtained by the relative distance from the previous coordinate through sequential processing, and the input/output latency is small. However,

methods with only RNN is not possible to explicitly learn the correlation between body joints as with GCN, and the problem is how to represent a single feature vector of the body joints, which results in low accuracy or high calculation cost.

In this study, we propose a Recurrent Graph Convolutional Network (RGCN) that has the advantages of both RNN, which can process time-series data sequentially, and GCN, which can convolve graph structures, and a pose prediction method that uses the proposed RGCN and predicts human poses in the time domain. Figure 1 shows conceptual figure of the pose prediction.

Our contributions are summarized as follows:

- We propose Recurrent Graph Convolutional Network (RGCN) that can sequentially process time-series data with a graph structure.
- We propose a pose prediction model using RGCN and its learning method, and the model predicts human poses with short pre- and post-processing latency through sequential prediction in the time domain.

The rest of this paper is organized as follows: In Sect. 2, recent work on human pose prediction is summarized. In Sect. 3, the details of the proposed sequencial human pose prediction method are described. In Sect. 4, experimental results are presented. Finally, we conclude the paper in Sect. 5.

2 Related Work

Machine learning methods such as hidden Markov model [2], restricted Boltzmann machine [24], and Gaussian process latent variable model [26] had been used in pose prediction, however, it is difficult to capture complex human motions using these methods. Recently, pose prediction methods that use deep learning such as RNN [6,14,19,25] and GCN [3,4,7,14,18,22] are widely used.

Since RNNs have recursive structures, they can efficiently process time-series data, and when predicting human poses, it is possible to make predictions by sequentially processing the pose sequence. In terms of the advantage of RNN, sequential processing prediction allows the next pose to be easily obtained from the previous pose and has small input/output delays. However, because RNN alone cannot explicitly learn the correlation between body joints, the problem is how to represent a single feature vector representing the pose. Many existing methods using RNN represented joint features as one-dimensional feature vector [6,19,25], however, this leads to increase in the cost of feature extraction or a decrease in accuracy.

On the other hand, since the human body can be regarded as a graph structure, GCNs, which can extract features using convolution operations from arbitrary graph structures, have been used for the pose prediction in recent years. An advantage of using GCN is that the relationships between body joints, which are the vertices of the skeleton graph, can be learned explicitly. Prediction methods using GCN include methods that consider temporal changes in the 3D coordinates of body joints directly as a graph [3,14,22] and other methods that convert the movement of each body joint into frequency coefficients using Discrete Cosine

Transform (DCT) [4,8,18]. As examples, Cui et al. considered a 3D human skeleton sequence as a graph with the coordinate values of each body joints as features, and used GCN with learnable adjacency matrices to capture the spatial dependence between each joints [3]. Li et al. proposed an Encoder-Decoder-based method with GCN, CNN and RNN that predicts multi-scale poses by grouping body joints that have a close positional relationship [14]. Whereas, Mao et al. proposed a prediction method using DCT and GCN with learnable adjacency matrices [18] which converts the temporal information of a 3D skeleton sequence in the time domain into information in the frequency domain. This method represents the temporal changes of each body joint in terms of frequency coefficients and predicts human poses as a set of vectors of frequency coefficients corresponding to each body joint in the frequency domain. Dang et al. proposed a GCN-based method that groups closely positioned body joints and gradually predicts multi-scale poses in the frequency domain [4]. Fujita et al. proposed a pose prediction method using DCT and GCN that progressively predicts the frequency coefficients by multi-scaling the frequency domain into high-, medium-, and low-frequency domains [8].

The accuracy of the methods using GCN described above is high because they can explicitly learn the correlation between the body joints. However, it is necessary to process all frames at once; pre-processing must wait until all input frames are available, and intermediate prediction results cannot be obtained until post-processing is completed. Therefore, there is a problem that the waiting time for input/output latency is long. In addition, when predicting in the frequency domain using DCT, the number of sampling points is very small for input/outputs of less than a few seconds, and the frequency resolution is low.

Other prediction methods have been proposed, such as a pose prediction method using a simple multi-layer perceptron [11] and a method based on Message Passing Neural Network (MPNN), which is designed to have equivariance and invariance [28]. Even with these methods, prediction is performed in the frequency domain, and there are the problems of the frequency resolution and latency to batch processing for all frames. Also, other prediction methods using Transformer have been proposed [1,27]; however, the calculation costs are relatively high, and the methods using Transformer require batch processing of all input/output frames.

In this study, to solve the problems of existing methods, we focus on the advantages of both RNN, which can process time series sequentially, and GCN, which can perform graph convolution, and propose a simple and novel Recurrent Graph Convolutional Network (RGCN) that can sequentially process time-series data with graph structure, and a pose prediction method using the RGCN. There are also studies on combining RNN and GCN in the field of pose estimation rather than pose prediction [20,21]. However, in our study, we generalize the formula for combining RNN and GCN and construct an RGCN with a more concise expression to be useful for pose prediction.

3 Recurrent Graph Convolutional Network and Pose Prediction

3.1 Overview of the Proposed Method

Methods using GCN have delays for input/output processing due to batch processing of all frames. In this study, we propose a Recurrent Graph Convolutional Network (RGCN) that can sequentially process time-series data with a graph structure, which has two characteristics: RNN that can process time series sequentially, and GCN that can convolve graph structures. Then, we attempt to reduce the delays for input/output processing by sequentially predicting the pose sequence using a pose prediction model with RGCN.

3.2 Recurrent Graph Convolutional Network

First, the input format of the network is a matrix H_{in} in which the feature vector corresponding to each graph node is stacked. A graph convolution layer with a learnable adjacency matrix is defined by the following equation:

$$H_{\text{out}} = \sigma(GH_{\text{in}}W), \tag{1}$$

where $H_{\text{in}} \in \mathbb{R}^{n \times F_{\text{in}}}$ is the input, $G \in \mathbb{R}^{n \times n}$ is the learnable adjacency matrix, $W \in \mathbb{R}^{F_{\text{in}} \times F_{\text{out}}}$ is the weight matrix, $H_{\text{out}} \in \mathbb{R}^{n \times F_{\text{out}}}$ is the output graph feature, and σ is an activation function, respectively. F_{in} and F_{out} are feature dimensions, n is the number of nodes. Note that the bias, which is added after calculating the weight matrix, is omitted in Equation (1).

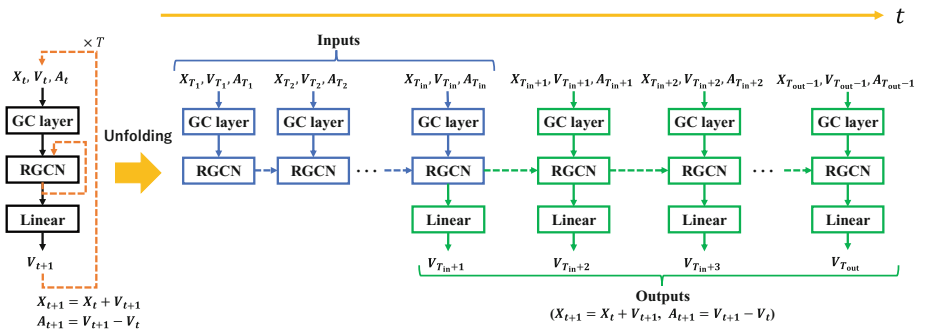


Fig. 2. Pose prediction model using RGCN and sequential process. The proposed model receives location, velocity, and acceleration information as inputs and outputs velocity information (relative distance) to the next location. The next position and acceleration information are then calculated using the predicted velocity information (relative distance).

We define a function f as matrix product with W and the process of activation function in Equation (1) and get

$$H_{\text{out}} = f(GH_{\text{in}}). \tag{2}$$

Here, by replacing f with RNN process,

$$H_{\text{out}} = \text{RNN}(GH_{\text{in}}). \tag{3}$$

Therefore, the Equation (1) can be replaced by an RNN with recursive structure. Although it is possible to use existing RNNs such as vanilla RNN and LSTM, in this study, we use a simplified RNN to reduce the number of parameters. The proposed RGCN in this study is defined by,

$$S_t = \sigma((GH_t + aS_{t-1})W), \tag{4}$$

where $H_t \in \mathbb{R}^{n \times F_{\text{out}}}$ is the input, $G \in \mathbb{R}^{n \times n}$ is the learnable adjacency matrix, $W \in \mathbb{R}^{F_{\text{out}} \times F_{\text{out}}}$ is the weight matrix, $S_t, S_{t-1} \in \mathbb{R}^{n \times F_{\text{out}}}$ are the output graph features, and σ is an activation function, respectively. In addition, a is a scalar value and learnable parameter that controls the degree of information transmission in time steps. Note that the bias, which is added after calculating the weight matrix, is omitted in Eq. (4).

3.3 Pose Prediction Using RGCN

The pose at a time step t is represented by a graph, and it is represented as a matrix X_t in which the feature vectors corresponding to each body joint are stacked. These feature vectors are sets of 3D coordinates of human body joints obtained using sensors, such as infrared or inertial motion capture sensors. We assume that $\mathcal{X}^{\text{in}} = (X_{T_1}, \ldots, X_{T_{\text{in}}})$ is the input 3D human pose sequence, and $\mathcal{X}^{\text{out}} = (X_{T_{\text{in}}+1}, \ldots, X_{T_{\text{out}}})$ is the sequence of prediction outputs. The purpose is to predict $\mathcal{X}^{\text{out}}$ from $\mathcal{X}^{\text{in}}$. Therefore, the proposed model with the RGCN receives one human pose in order from the beginning of the input sequence, updates the hidden state, and outputs the next human poses from the last input sequentially.

In order to add velocity information as the input to the pose prediction model, $\mathcal{V}^{\text{in}} = (V_{T_1}, \ldots, V_{T_{\text{in}}})$ is calculated by taking the difference $V_t = X_t - X_{t-1}$ in the 3D pose sequence. Here, we set $V_{T_1} = O$ because X_{T_0} is not available.

Similarly, in order to add acceleration information, we also take the difference in the obtained velocity information and calculate $\mathcal{A}^{\text{in}} = (A_{T_1}, \ldots, A_{T_{\text{in}}})$, $A_{T_1} = O$. Note that the obtained values are not differential values; strictly speaking, they are not velocity and acceleration. However, since they contain information on velocity and acceleration, we represent velocity and acceleration hereafter.

Thus, the three obtained feature matrices of $\mathcal{X}^{\text{in}}$, $\mathcal{V}^{\text{in}}$, $\mathcal{A}^{\text{in}}$ are concatenated and arranged into a feature matrix as the input. The pose prediction model f_p

outputs the next location, velocity, and acceleration using the following equation:

$$\widehat{V}_t = f_p(X_{t-1}, V_{t-1}, A_{t-1}), \tag{5}$$

$$\widehat{X}_t = X_{t-1} + \widehat{V}_t, \tag{6}$$

$$\widehat{A}_t = \widehat{V}_t - V_{t-1}. \tag{7}$$

This calculation is performed sequentially for each time step t. As shown in Fig. 2, the prediction model consists of an input layer (graph convolution layer), stacked RGCN layers, and an output layer (linear layer). After time $T_{\text{in}} + 1$, the output of the model is used as the next input because observations have not yet been obtained. For this, it is set as $V_t \leftarrow \widehat{V}_t$, $X_t \leftarrow \widehat{X}_t$, $A_t \leftarrow \widehat{A}_t$.

3.4 Training Method for Prediction Model

We used two losses to train the proposed pose prediction model. First, similar to the training method of the existing methods [4,18,28], the loss is calculated by Mean Per Joint Position Error (MPJPE), which is the mean of all Euclidean distances for each joint in the predicted and ground truth skeleton pairs, for the predicted pose sequence. Normal MPJPE (L_{S}) is defined by the following equation:

$$L_{\text{S}} = \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| \widehat{\mathbf{p}}_{j,t} - \mathbf{p}_{j,t} \right\|_2, \tag{8}$$

where J denotes the number of human body joints in the 3D coordinates and T denotes the number of time step. In addition, $\widehat{\mathbf{p}}_{j,t} \in \mathbb{R}^3$ and $\mathbf{p}_{j,t} \in \mathbb{R}^3$ represent the predicted location for j-th joint at the time step t and the ground truth, respectively. However, since the our prediction model processes the pose sequence sequentially by differential calculation, the overall accuracy may decrease if mispredictions are generated in the early stage. Considering this fact, we used weighted loss function (L_{SW}) that weights the MPJPE in sequential order.

$$L_{\text{SW}} = \frac{1}{TJ} \sum_{t=1}^{T} T e^{-\frac{2}{T}t} \cdot \left(\sum_{j=1}^{J} \left\| \widehat{\mathbf{p}}_{j,t} - \mathbf{p}_{j,t} \right\|_2 \right). \tag{9}$$

Since low-frequency components (smooth motions) are considered important for natural human motions [18], the proposed model using RGCN, unlike other methods, does not predict in the frequency domain, making it difficult to emphasize the low-frequency components during the training. Therefore, as the second loss function, we calculated MPJPE of moving average after calculating the simple moving average of the predicted and ground truth pose sequences. The moving average acts as a low-pass filter and is expected to remove high-frequency components from body joint movements. The moving average of the prediction is calculated as follows:

$$\mathrm{M}^{\widehat{\mathbf{p}}}_{j,t} = \frac{\widehat{\mathbf{p}}_{j,t} + \widehat{\mathbf{p}}_{j,t+1} + \widehat{\mathbf{p}}_{j,t+2}}{3}. \tag{10}$$

Note that this was not calculated if there were missing values. After performing the same calculation for the ground truth, we computed the MPJPE for the sequences of moving average. The moving average MPJPE (L_{mv}) is

$$L_{\mathrm{mv}} = \frac{1}{T_{\mathrm{mv}} J} \sum_{t=1}^{T_{\mathrm{mv}}} \left(\sum_{j=1}^{J} \left\| \mathrm{M}^{\widehat{\mathbf{P}}}_{j,t} - \mathrm{M}^{\mathbf{P}}_{j,t} \right\|_2 \right), \tag{11}$$

where T_{mv} is the number of time step of the moving average. Also, $\mathrm{M}^{\mathbf{P}}$ is moving average for ground truth.

Therefore, the final loss L used for the training is as follows:

$$L = L_{\mathrm{sw}} + L_{\mathrm{mv}} \tag{12}$$

4 Evaluation

4.1 Outline of Experiments

We compared the proposed method with existing methods on two public datasets. In the experiments, we used Python 3.11, PyTorch 2.1.0, CUDA11, NVIDIA Tesla V100. As the existing methods, we used the prediction method using GCN, Traj-GCN [18] and MSR-GCN [4]. We also used EqMotion [28], which currently has the highest prediction accuracy, for the comparison.

In addition, to confirm the effectiveness of the proposed RGCN, we also experimented with a simple sequential prediction model using Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) and confirmed these accuracy. The features for each body joint were concatenated and flattened to a one-dimensional feature vectors. For fair comparison, we also applied the learning method described in Sect. 3.4 to the models using GRU and LSTM.

Table 1. Comparison of the number of model parameters in Human 3.6 M

Model (units, layers)	Total number of learnable parameters ↓
GRU model (256, 3)	1,252,162
GRU model (512, 3)	4,863,554
LSTM model (256, 3)	1,646,914
LSTM model (512, 3)	6,439,490
Traj-GCN	2,555,275
MSR-GCN	6,298,185
EqMotion	641,892
Ours (256, 5)	**335,200**
Ours (512, 5)	1,322,848

Human3.6M: The Human3.6M is a dataset for human motion analysis [12]. This dataset contains 7 participants and 15 action categories. Like existing method [4], we used the data of {S1, S6, S7, S8, S9} as the training set and the data for S5 and S11 as the test set and validation set respectively. Also, following the existing method MSR-GCN [4], we chose 22 body joints from the original 32 joints. For MSR-GCN, we also prepared multi-scale skeletons with 12, 7, and 4 joints. Each sequence was downsampled from 50 Hz to 25 Hz, and we used 10 frames for 0.4 s as input and 25 frames for 1.0 s as output by using a sliding window of 35 frames.

CMU Mocap: The CMU Mocap dataset[1] is often used for pose prediction. We experimented using a subset of the dataset with the 8 action classes used in [4,13]. Following the existing method [4,13], we split the training set and test set. Then, we split one-fifth of the training set as the validation set. Also, we chose 25 body joints from the original 38 joints and prepared multi-scale skeletons with 12, 7, and 4 joints. Each sequence was downsampled from 120 Hz to 30 Hz, and we used 12 frames for 0.4 s as input and 30 frames for 1.0 s as output by using a sliding window of 42 frames.

4.2 Configuration and Parameters

Regarding the sequence length, the input and output are 10 and 25 in Human3.6M, and 12 and 30 in CMU Mocap, respectively. In addition, in the proposed method, the number of hidden layer units was set to 256 or 512 units, and the number of RGCN layers was set to five layers. The size of the adjacency matrices is the number of body joints. In the models using GRU and LSTM,

Table 2. Average MPJPE of 5 trials and its standard deviation on the Human3.6M.

Model (units, layers)	Average MPJPE (mm) ↓	
	Validation	Test
GRU model (256, 3)	65.38 ± 0.21	75.42 ± 0.19
GRU model (512, 3)	65.55 ± 0.53	75.62 ± 0.69
LSTM model (256, 3)	67.32 ± 0.30	76.31 ± 0.37
LSTM model (512, 3)	68.76 ± 0.20	77.96 ± 0.49
Traj-GCN	71.57 ± 0.21	85.02 ± 0.77
MSR-GCN	69.48 ± 1.15	82.39 ± 2.50
EqMotion	$\underline{62.47 \pm 0.50}$	72.59 ± 0.66
Ours (256, 5)	62.50 ± 0.48	$\underline{72.26 \pm 0.73}$
Ours (512, 5)	$\mathbf{61.40 \pm 0.25}$	$\mathbf{70.89 \pm 0.65}$

[1] The data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

the number of hidden layer units was also set to 256 or 512 units, and the number of layers was set to three layers. The models using GRU and LSTM have a input layer (linear layer with activation) and a output layer (linear layer). The parameters of MSR-GCN, Traj-GCN, and EqMotion were set by following the existing studies. The number of parameters to be learned for each model is listed in Table 1.

All learnable parameters were updated by using RAdam optimizer [16] with the default hyperparameters. To prevent gradient explosion, we set the gradient clipping values to 1.0. We also set the dropout ratio to 0.2, batch size to 256, and the maximum number of epoch to 200. As the activation function of RGCN, we used Leaky ReLU with a slope of 0.2. When the validation loss did not decrease for five epochs, the learning rate was multiplied by 0.1. When the loss did not decrease for 11 epochs, the training was terminated.

The center positions of the data were normalized so that the base of the spine in the first frame of the input became the origin. In order to learn rotation in 3D coordinates in the proposed method, data augmentation was performed by rotating each sequence using rotation matrices with random angles in [0, 360) during the training.

We used MPJPE [12] explained in the Sect. 3.4 as the evaluation metric, and evaluated the predictions based on how close the predicted 3D future pose sequence is to the ground truth. We performed the training with five trials with different initial weights for each method and evaluated the performance by the average.

4.3 Results and Discussion

Table 2 and 3 show the average MPJPE of five trials and its standard deviation for each method on the Human3.6M and CMU Mocap, respectively. These values in the tables are rounded off to the third decimal place. Figure 3 shows examples

Table 3. Average MPJPE of 5 trials and its standard deviation on the CMU Mocap.

Model (units, layers)	Average MPJPE (mm) ↓	
	Validation	Test
GRU model (256, 3)	69.60 ± 0.97	79.21 ± 1.28
GRU model (512, 3)	64.07 ± 2.78	75.87 ± 1.30
LSTM model (256, 3)	69.23 ± 0.92	83.94 ± 1.27
LSTM model (512, 3)	70.02 ± 1.50	83.85 ± 0.94
Traj-GCN	65.82 ± 0.90	76.61 ± 1.07
MSR-GCN	70.99 ± 16.87	81.79 ± 11.24
EqMotion	$\mathbf{58.81 \pm 0.87}$	$\mathbf{67.00 \pm 0.71}$
Ours (256, 5)	$\underline{58.93 \pm 1.05}$	$\underline{67.79 \pm 1.00}$
Ours (512, 5)	64.65 ± 4.81	72.99 ± 3.96

Table 4. MACs for each model on the Human3.6M.

Model	MACs (G) ↓
MSR-GCN	0.101
EqMotion	0.196
Ours (256, 5)	0.278
Ours (512, 5)	1.060

Table 5. Processing time and frames per second (FPS) required to predict 25 frames from 10 frames of input.

Model	Device name (time (s) ↓ / FPS ↑)		
(units, layers)	RTX4090	i9-13900K (24 core)	N100 (4 core)
MSR-GCN	2.38×10^{-2} / 42	7.96×10^{-3} / 125	1.70×10^{-1} / 5
EqMotion	1.44×10^{-2} / 69	3.12×10^{-2} / 32	2.74×10^{-1} / 3
Ours (256, 5)	3.51×10^{-2} / 28	7.56×10^{-3} / 132	2.47×10^{-2} / 40
Ours (512, 5)	3.45×10^{-2} / 28	1.19×10^{-2} / 83	4.39×10^{-2} / 22

Table 6. Average MPJPE of 5 trials and its standard deviation in the ablation study on the Human3.6M. Comparison of 256 units and 5 layers.

Method	Average MPJPE (mm) ↓	
	Validation	Test
baseline	63.05 ± 0.29	72.57 ± 0.33
+ weighted loss	62.41 ± 0.52	72.26 ± 0.34
+ moving average loss	62.50 ± 0.48	72.26 ± 0.73

Table 7. Average MPJPE of 5 trials and its standard deviation in the ablation study on the CMU Mocap. Comparison of 256 units and 5 layers.

Method	Average MPJPE (mm) ↓	
	Validation	Test
baseline	61.59 ± 1.55	72.37 ± 1.65
+ weighted loss	59.45 ± 1.22	68.61 ± 0.89
+ moving average loss	58.93 ± 1.05	67.79 ± 1.00

of a comparison between the proposed method using RGCN (512 units) and EqMotion for pose prediction with relatively large movements in Human3.6M. Similarly, Fig. 4 shows examples of prediction results involving relatively small movements in Human3.6M.

From Table 2, we confirmed that the average MPJPE of the proposed method is higher than that of the existing method in the case of Human3.6M. Meanwhile,

from Table 3, the average MPJPE of the proposed method was slightly lower than that of EqMotion in CMU Mocap. We consider that the reason for this is that the size of the CMU Mocap subset used in the experiment was smaller than that of Human3.6M, which caused overfitting. In fact, comparing the MPJPEs for the test data that had the highest score among the five trials in CMU Mocap, EqMotion had a score of 66.30, while the proposed method had a similar score of 66.23. In addition, in the case of 512 units, the accuracy worsens because of overfitting.

Comparing the normal RNN and RGCN, from Tables 1–3, the method using RGCN, which can learn joint relationships explicitly, has higher scores than the normal RNN models, despite fewer parameters. Even if the number of layers of the normal RNN models was increased to four or more, the accuracy did not improve. Therefore, RGCN can process time-series data with graph structure more efficiently than normal RNN.

From Tables 1 and 2, in the case of the proposed method with 256 units, the average MPJPE is comparable to the average MPJPE of EqMotion, indicating that the parameters were reduced without decreasing the accuracy, despite the number of parameters being approximately 0.53 times larger. Additionally, there is no need to wait until all input frames are available to start pre-processing, and intermediate prediction results can be returned through sequential processing. Therefore, the waiting time is short, and faster speeds can be expected in other applications that use prediction results.

The results of the comparison of the computational complexity and prediction speed between the models are shown in Tables 4 and 5. The MACs are rounded off to the fourth decimal place, and the processing time is rounded off to the third decimal place, and the MACs were calculated using a tool called ptflops [23]. MACs is roughly GMACs = 0.5×GFLOPs [23]. The comparison of the prediction speed in Table 5 shows the results when using an NVIDIA RTX4090 GPU and intel i9-13900K and when all processing was performed on an intel i9-13900K or N100 CPU. In addition, each processing time includes the time required for pre- and post-processing of the skeleton sequence and transfer overhead from CPU to GPU, and is shown as an average of 1000 times. From Table 4, it can be seen that the proposed method, which performs calculations sequentially each time, requires a larger calculations than the existing methods. On the other hand, from Table 5, when using the RTX4090 GPU, the existing method EqMotion was the fastest, but when using CPUs, the proposed method had the fastest prediction speed. In particular, the proposed model (256 units, 5 layers) can be executed at 40 FPS, even on the N100 (4-core CPU), making it possible to predict more than seven times faster. In addition, in the existing methods, when the skeletons are acquired from the sensor at 25 frames per second, a delay of about 0.36 s occurs for the input of 10 frames; therefore, the speed difference becomes even larger. From Figs. 3, 4, and Table 2, we also confirmed that by increasing the number of units, the proposed method was able to predict better than existing methods, whether the motion is small or large.

We performed ablation studies to verify the effectiveness of the learning method described in Sect. 3.4. Tables 6 and 7 show the results of the ablation

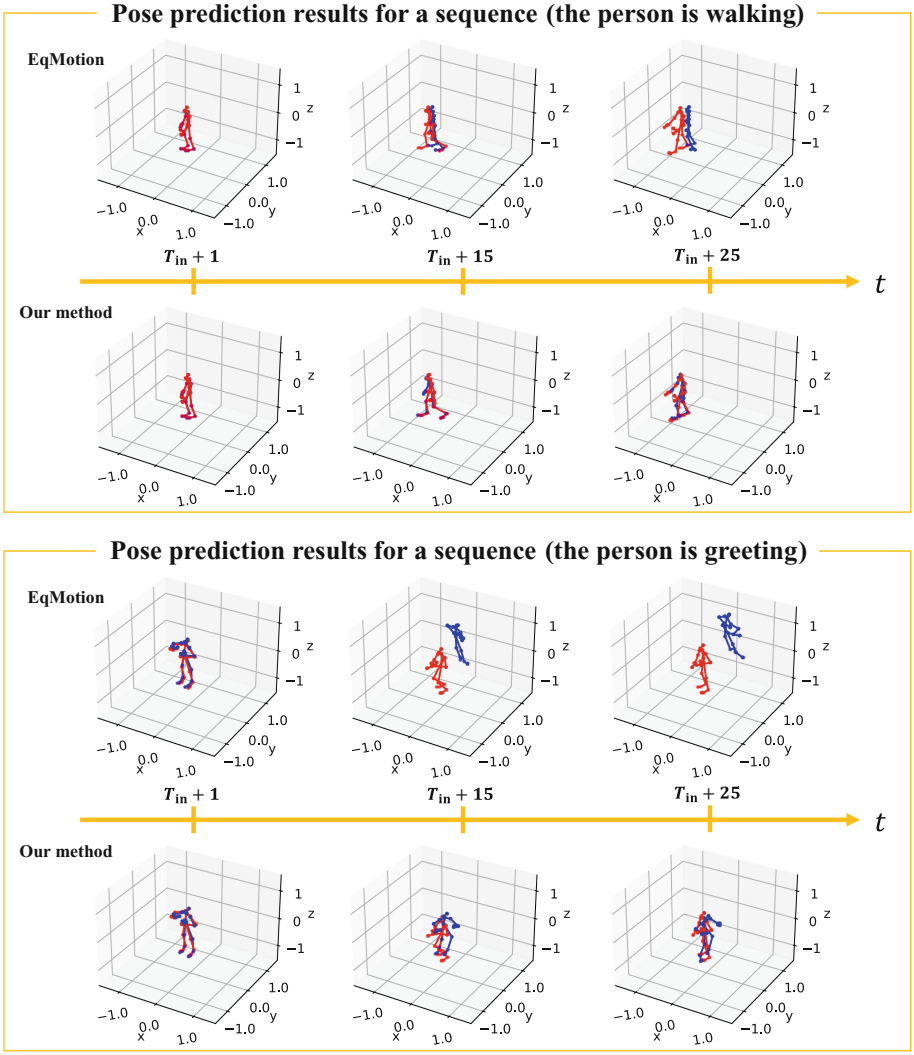


Fig. 3. Examples of prediction (blue) and ground truth (red) on Human3.6M. In Human3.6M, the proposed method could predict more accurately than the existing method, even if the movements were relatively large. In these two examples, the person is walking and greeting, respectively. (Color figure online)

studies. These values in the tables are rounded off to the third decimal place. For Human3.6M, the prediction accuracy improved only when weighted loss was added; however, adding moving average loss did not improve accuracy. On the other hand, for CMU Mocap, the accuracy improved with both weighted loss and moving average loss. RGCN has a disadvantage in that it is sensitive to noise and

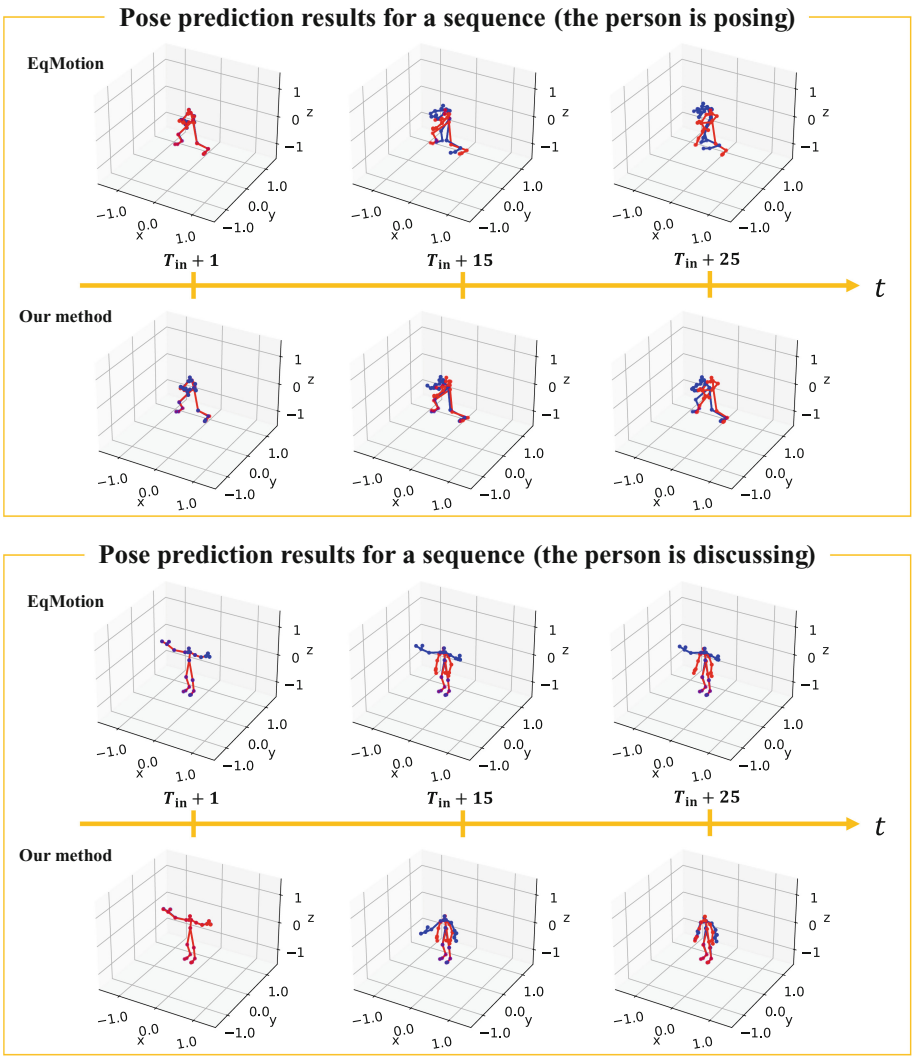


Fig. 4. Examples of prediction (blue) and ground truth (red) on Human3.6M. In Human3.6M, the proposed method could predict more accurately than the existing methods, even when the movements were relatively small. In these two examples, the person is posing and discussing, respectively. (Color figure online)

accumulation prediction errors due to its simple prediction mechanism. Therefore, weighted loss is very effective, given the learning characteristics of RGCN. Also, moving average loss is considered to be effective when the dataset contains many high-frequency components caused by noise, and it may not be effective depending on the dataset with little noise.

5 Conclusion and Future Work

In this study, we proposed a Recurrent Graph Convolutional Network (RGCN) that can sequentially process time-series data with a graph structure and a pose prediction method using RGCN, and attempted to improve the processing latency of the input and output in pose prediction. The proposed method can achieve the same level of prediction accuracy as existing methods, even with approximately half the number of parameters, and pre-processing is simple and the waiting time for pre- and post-processing is short, making it possible to make predictions with reduced overhead.

Because the increase in parameters is smaller when the number of layers is increased than when the number of units is increased, it may be possible to further improve the accuracy and reduce the number of parameters. However, multi-layering may increase the overhead required for each calculation. Analysis of larger dataset is also a future task. In the future, we would like to work on the remaining issues to improve the accuracy and further reduce the calculation costs.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Number JP23K16914. The computation was carried out using the General Projects on supercomputer "Flow" at Information Technology Center, Nagoya University.

References

1. Aksan, E., Kaufmann, M., Cao, P., Hilliges, O.: A spatio-temporal transformer for 3D human motion prediction. In: Proceedings of the 2021 International Conference on 3D Vision, pp. 565–574, December 2021
2. Brand, M., Hertzmann, A.: Style machines. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 183–192, July 2000
3. Cui, Q., Sun, H., Yang, F.: Learning dynamic relationships for 3D human motion prediction. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6519–6527, June 2020
4. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In: Proceedings of the 18th IEEE/CVF International Conference on Computer Vision, pp. 11467–11476, October 2021
5. Foka, A.F., Trahanias, P.E.: Probabilistic autonomous robot navigation in dynamic environments with human motion prediction. Int. J. Soc. Robot. **2**, 79–94 (2010)
6. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the 15th IEEE International Conference on Computer Vision, pp. 4346–4354, December 2015
7. Fujita, T., Kawanishi, Y.: Future pose prediction from 3D human skeleton sequence with surrounding situation. Sensors **23**(2), 876 (2023)
8. Fujita, T., Kawanishi, Y.: Human pose prediction by progressive generation in multi-scale frequency domain. In: Proceedings of the 2023 18th International Conference on Machine Vision and Applications, pp. 1–5, July 2023

9. Gong, H., Sim, J., Likhachev, M., Shi, J.: Multi-hypothesis motion planning for visual object tracking. In: Proceedings of the 2011 International Conference on Computer Vision, pp. 619–626, November 2011
10. Gui, L.Y., Zhang, K., Wang, Y.X., Liang, X., Moura, J.M., Veloso, M.: Teaching robots to predict human motion. In: Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 562–567, October 2018
11. Guo, W., Du, Y., Shen, X., Lepetit, V., Alameda-Pineda, X., Moreno-Noguer, F.: Back to mlp: a simple baseline for human motion prediction. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4809–4819, January 2023
12. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
13. Li, C., Zhang, Z., Lee, W.S., Lee, G.H.: Convolutional sequence to sequence model for human dynamics. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, pp. 5226–5234, June 2018
14. Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., Tian, Q.: Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 214–223, June 2020
15. Liu, H., Wang, L.: Human motion prediction for human-robot collaboration. J. Manuf. Syst. **44**, 287–294 (2017)
16. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. In: International Conference on Learning Representations, pp. 1–13, December 2020
17. Lyu, K., Chen, H., Liu, Z., Zhang, B., Wang, R.: 3D human motion prediction: a survey. Neurocomputing **489**, 345–365 (2022)
18. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the 17th IEEE/CVF International Conference on Computer Vision, pp. 9489–9497, November 2019
19. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2891–2900, July 2017
20. Puchert, P., Ropinski, T.: A3gc-ip: attention-oriented adjacency adaptive recurrent graph convolutions for human pose estimation from sparse inertial measurements. Comput. Graph. **117**, 96–104 (2023)
21. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1227–1236, June 2019
22. Sofianos, T., Sampieri, A., Franco, L., Galasso, F.: Space-time-separable graph convolutional network for pose forecasting. In: Proceedings of the 18th IEEE/CVF International Conference on Computer Vision, pp. 11209–11218, October 2021
23. Sovrasov, V.: ptflops: a flops counting tool for neural networks in pytorch framework (2018-2024). https://github.com/sovrasov/flops-counter.pytorch
24. Taylor, G.W., Hinton, G.E., Roweis, S.: Modeling human motion using binary latent variables. Adv. Neural. Inf. Process. Syst. **19**, 1345–1352 (2005)
25. Wang, B., Adeli, E., Chiu, H.k., Huang, D.A., Niebles, J.C.: Imitation learning for human pose prediction. In: Proceedings of the 17th IEEE/CVF International Conference on Computer Vision, pp. 7124–7133, November 2019

26. Wang, J., Hertzmann, A., Fleet, D.J.: Gaussian process dynamical models. Adv. Neural. Inf. Process. Syst. **18**, 1441–1448 (2005)
27. Xu, C., Tan, R.T., Tan, Y., Chen, S., Wang, X., Wang, Y.: Auxiliary tasks benefit 3D skeleton-based human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9509–9520, October 2023
28. Xu, C., et al.: Eqmotion: equivariant multi-agent motion prediction with invariant interaction reasoning. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1410–1420, June 2023

Conversation Activity Recognition Using Interaction Video Sequences in Pedestrian Groups

Wataru Ganaha(✉), Takumi Ozaki, Michiko Inoue, and Masashi Nishiyama

Graduate School of Engineering, Tottori University, 101 Minami 4-chome, Koyama-cho, Tottori 680-8550, Japan
nishiyama@tottori-u.ac.jp

Abstract. We introduce a method for recognizing conversation activity in a group of people walking outdoors using a color video sequence acquired from a camera. Many methods have been developed to recognize whether people are walking together or talking together in a color video sequence. However, a method has yet to be proposed to recognize conversation activity in a pedestrian group walking outdoors. In this paper, we design a feature extraction approach for conversation activity recognition using physical body interactions caused by pedestrians' conversations. Our method generates an interaction video sequence in a virtual space using a temporal posture signal and a temporal walking position signal that represent pedestrians' body interactions. Our method uses the interaction video sequence as an informative and visible feature to determine a conversation activity label. The experimental results showed that our interaction video sequence recognized conversation activity more accurately than alternative techniques that use the appearance of the body regions of a pedestrian group or time-series changes of the posture and walking position among pedestrians.

Keywords: Conversation activity recognition · Pedestrian groups · Human body interaction

1 Introduction

A demand exists for technology that can automatically recognize human interactions within a group of people walking outdoors. In this study, we focus on conversation activity as one form of human interaction in a pedestrian group. We define conversation activity as whether a conversation is occurring within a pedestrian group and whether the conversation is active or inactive. One possible application of conversation activity recognition is marketing in a scenario in which many pedestrian groups are walking in the aisles of a shopping mall. Figure 1 shows an example of the application. By comparing the number of pedestrian groups engaged in active conversation between visitors that are arriving

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 359–374, 2025.
https://doi.org/10.1007/978-3-031-78341-8_23

Fig. 1. We assume that an application for conversation activity recognition exists. This application can determine whether pedestrian groups are satisfied with their visit by comparing the number of pedestrian groups engaged in active conversation between visitors that are arriving and leaving.

and leaving, it may be possible to determine whether visitors are satisfied with their visit.

We consider what feature can be used to recognize conversation activity in a pedestrian group. A possible feature is the chronological change of speech sounds, such as the timing of pedestrians' utterances and the inflection of pedestrians' voices. However, because we target a group of pedestrians walking outdoors, it is difficult to use a microphone for voice sensing for each pedestrian. Instead, we consider using a color video sequence acquired from a surveillance camera as a feature that represents a human body interaction performed in a pedestrian group. We assume that the time-series changes in gestures performed by each pedestrian, and the time-series changes in pedestrians' body orientation and walking position, provide a visible and informative feature for conversation activity recognition. When analyzing speech among people [12], it is well known that gestures, that is, movements produced by the body in response to speech, are helpful. Regarding the analysis of pedestrian group behavior [23] and the development of the group detection method [2], it is well known that body orientation and the walking position, which are interrelated among pedestrians that belong to one group, are helpful. In this study, our definition of physical body interaction consists of gestures, pedestrians' body orientation, and pedestrians' walking position.

We consider how to design a method to recognize conversation activity using the body interaction feature in a video sequence. To the best of our knowledge, a method has yet to be proposed to recognize conversation activity in a pedestrian group. Instead, we survey existing methods for recognizing the presence or absence of body interaction in a pedestrian group, such as whether the pedestrians are walking together or talking together in a video sequence. These existing methods can be divided into two main categories. The first category contains methods [2,5,18,19,24] that detect the presence or absence of a pedestrian group. The second category contains methods [9,13,17] that recognize whether

people in a group are talking together, given that a pedestrian group has been detected. More recently, methods [3,6,15] have emerged that detect the presence or absence of a pedestrian group and simultaneously recognize the presence or absence of conversations within that group. However, even when these existing methods are applied, it is impossible to recognize whether the conversation is active or inactive in a pedestrian group.

In this paper, we propose a novel method for recognizing conversation activity in a pedestrian group by extracting an interaction video sequence as a feature, which has high recognition accuracy and can be visually confirmed by human observers. Our method generates an interaction video sequence in a pedestrian group using a temporal posture signal and temporal walking position signal estimated from a color video sequence. By applying this interaction video sequence to the class classification network, our method determines a conversation activity label: active conversation, inactive conversation, or no conversation. The active conversation label indicates the state in which the pedestrian group is having a lively conversation on topics of mutual interest. The inactive conversation label indicates the state in which the group is not having a lively conversation on topics of no interest. The no conversation label indicates the state in which no conversation is occurring.

The salient contributions of this paper are as follows:

- We extract an informative feature using an interaction video sequence rendered in a virtual space by fixing the viewpoint position of the virtual camera in front of a pedestrian group.
- We design a visible feature that allows human observers to directly see physical body interaction performed in a pedestrian group.
- On an originally collected outdoor pedestrian dataset of 624 video sequences in 52 groups, we demonstrated that our interaction video sequences achieved high accuracy in conversation activity recognition.

From the experimental results, we confirmed that our method using an interaction video sequence recognized conversation activity more accurately than using color video sequences of pedestrian body regions or using a temporal posture and walking position signal.

2 Method for Recognizing Conversation Activity

2.1 Overview

In this paper, we assume that body interaction arising from conversation activity among pedestrians is represented explicitly by time-series signals of the posture and walking position. Figure 2 shows an overview of our method. In the following, we describe the procedure in our method.

P1. Body region estimation:
We estimate the region that represents the body of a pedestrian at each time point in a color video sequence acquired from a surveillance camera.

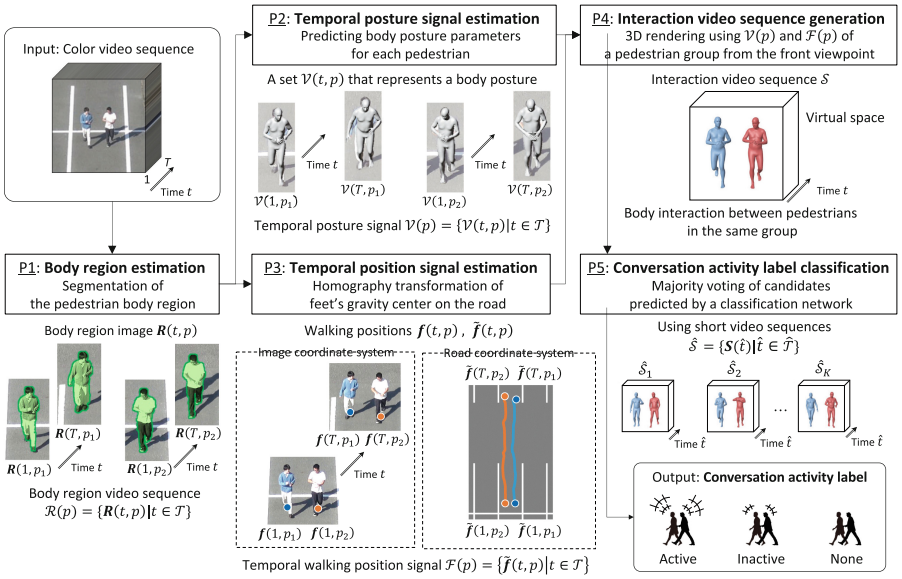


Fig. 2. Overview of our method for conversation activity recognition. We estimate body region images from a color video sequence acquired from a camera in P1. Our method obtains a temporal posture signal and temporal walking position signal that represents the body interaction caused by the conversation in P2 and P3. We generate an interaction video sequence in a virtual space for extracting an informative and visible feature in P4 and determine a conversation activity label using the interaction video sequence in P5.

P2. Temporal posture signal estimation:
We estimate a temporal posture signal from the appearance of a pedestrian body region at each time point. Specifically, we use the three-dimensional (3D) human body model to extract a time-series signal that represents only the posture change of each pedestrian. Using this signal, we represent posture changes in gestures and body orientation while pedestrians engage in conversation in a group.

P3. Temporal walking position signal estimation:
We estimate a temporal walking position signal by calculating the feet's center of gravity from the pedestrian's body region at each time point. Specifically, we estimate the center of gravity of the feet's contour for each pedestrian and determine the walking position on the road surface by applying a homography transformation. Using this signal, we extract temporal changes that represent positional relationships in a conversation in a group.

P4. Interaction video sequence generation:
Our method generates an interaction video sequence of a pedestrian group using 3D rendering with a temporal posture signal of P2 and temporal walking position signal of P3. By always fixing the virtual camera viewpoint in

front of the pedestrian group, we extract a feature that can capture the body interaction that effectively recognizes the conversation activity label. We also design a feature that allows human observers to visually and temporally confirm physical body interactions in a group.

P5. Conversation activity label classification:
We determine the conversation activity label using a classification network for an interaction video sequence of P4. We use three conversation activity labels: active conversation, inactive conversation, and no conversation. We explain the details of these labels in Sect. 3.2. We generate multiple short video sequences from a single interaction video sequence and output multiple candidate labels from the classification network using these short video sequences. A majority vote among these candidate labels determines the final conversation activity label.

In the following sections, we describe each procedure in detail.

2.2 Body Region Estimation

In procedure P1, we estimate the pedestrian body region from a video sequence acquired from a camera. The body region video sequence $\mathcal{R}(p)$ that consists of pedestrian body pixels and surrounding background pixels is expressed as

$$\mathcal{R}(p) = \{\boldsymbol{R}(t,p) \mid t \in \mathcal{T}\}, \tag{1}$$

where $\boldsymbol{R}(t,p)$ is the body region image of each pedestrian p at time point t, $\mathcal{T}$ is a set that consists of the times when the images were acquired, and T is the total number of times that belong to the set $\mathcal{T}$. T also represents the length of time from when a pedestrian enters the camera's field of view until the pedestrian leaves. Note that $\boldsymbol{R}(t,p)$ consists of a pedestrian body region and the background region surrounding it. $\boldsymbol{R}(t,p)$ stores a mask, whether each pixel belongs to the body or background region, and the RGB value of each pixel. We use Mask R-CNN [7], which is internally called from within PHALP [16], at each time point to estimate the pedestrian body region. PHALP is a body posture and shape estimation method, as described in the next section. This method also performs pedestrian tracking and determines each pedestrian p of $\boldsymbol{R}(t,p)$.

2.3 Temporal Posture Signal Estimation

In procedure P2, we estimate a temporal posture signal from the body region video sequence $\mathcal{R}(p)$ to represent the changes of gestures and body orientation in a conversation among pedestrians. First, we estimate the pedestrian's posture from the body region image $\boldsymbol{R}(t,p) \in \mathcal{R}(p)$. The posture is denoted by $\mathcal{V}(t,p)$, a set of 3D vertices $\boldsymbol{v}(t,p)$ on the pedestrian's body surface, and their adjacent vertices. A temporal posture signal $\mathcal{V}(p)$ is expressed as

$$\mathcal{V}(p) = \{\mathcal{V}(t,p) \mid t \in \mathcal{T}\}. \tag{2}$$

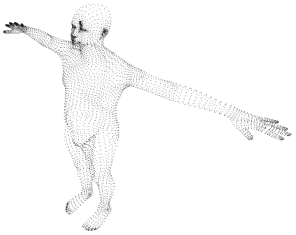


Fig. 3. Examples of vertices on the body surface.

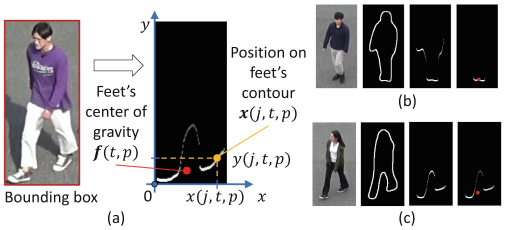


Fig. 4. Parameters used to calculate position $\boldsymbol{f}(t,p)$ of the feet's center of gravity.

In this study, to estimate $\mathcal{V}(t,p)$, which represents the posture changes, we apply PHALP [16] described in the previous section. PHALP is a method for tracking people in monocular movies by predicting their future 3D representations. This method involves estimating temporal models for the 3D pose, position, and appearance and using these models for probabilistic matching and updating tracklets. PHALP uses SMPL [11], which is a 3D human body model, to represent the posture and body shape parameters. The posture parameters are specifically expressed as a rotation matrix at the 23 joint points of the human body and a rotation matrix over the whole body. Using the estimated posture parameters and the standard body shape parameters, we generate a set $\mathcal{V}(t,p)$ that consists of 6,890 vertices $\boldsymbol{v}(t,p)$ on the body surface and their adjacent vertices. Figure 3 shows examples of vertices on the body surface.

When estimating a temporal posture signal, outliers in the time direction often occur suddenly. We detect outliers by applying a Hampel filter to the time-series signal of the 3D vertex $\boldsymbol{v}(t,p) \in \mathcal{V}(t,p)$. Then we interpolate the posture parameters at the time of the outlier using the nearest neighbor technique from the values at the surrounding time.

2.4 Temporal Walking Position Signal Estimation

In P3, we estimate a temporal walking position signal on a road surface to represent the pedestrian's positional relationship caused by the conversation. First, in the body region image $\boldsymbol{R}(t,p) \in \mathcal{R}(p)$, our method estimates the position of the feet's center of gravity $\boldsymbol{f}(t,p)$ in the image coordinate system. Next, by applying a homography transformation to convert the image coordinate system to the road surface coordinate system, our method obtains the walking position $\tilde{\boldsymbol{f}}(t,p)$. The temporal walking position signal $\mathcal{F}(p)$ is expressed as

$$\mathcal{F}(p) = \{\tilde{\boldsymbol{f}}(t,p) \mid t \in \mathcal{T}\}. \tag{3}$$

In the following, we describe how to calculate the position $\boldsymbol{f}(t,p)$ of the feet's center of gravity in the image coordinate system. Figure 4(a) shows the parameters used to calculate this position. In the body region image $\boldsymbol{R}(t,p)$ of pedestrian p at time t, our method obtains the image position $\boldsymbol{x}(j,t,p) = (x(j,t,p), y(j,t,p))$

of the point on the feet's contour. Let $\mathcal{J} = \{j\} : j$ be a natural number and $\forall j, k \in \mathcal{J} : j < k \Rightarrow x(j,t,p) < x(k,t,p)$. The origin is the lower left corner of the bounding rectangle of the pedestrian region. Using the component $y(j,t,p)$, which is the distance from the bottom $(x(j,t,p),0)$ of the bounding rectangle to the feet's contour, we calculate weight $w(j,t,p)$ as

$$w(j,t,p) \sim \mathcal{N}(y(j,t,p)|0,\sigma^2), \tag{4}$$

where $\mathcal{N}()$ is a normal distribution with mean 0 and standard deviation σ. Note that $w(j,t,p)$ satisfies $\sum_{j\in\mathcal{J}} w(j,t,p) = 1$. We obtain the position $\boldsymbol{f}(t,p)$ of the feet's center of gravity in the image coordinate system as follows:

$$\boldsymbol{f}(t,p) = \sum_{j\in\mathcal{J}} w(j,t,p)\boldsymbol{x}(j,t,p). \tag{5}$$

By applying a homography transformation and setting the height on the road surface to 0, we obtain the 3D walking position $\tilde{\boldsymbol{f}}(t,p)$ in the road surface coordinate system.

In the following, we explain why weight $w(j,t,p)$ is assigned to point $\boldsymbol{x}(j,t,p)$ on the feet's contour. Figure 4(b) shows an example when the legs are closed during walking, and (c) shows an example when the legs are open. In the case of closed legs, the candidate contour points mainly appear on the feet, and partially on the hands and other body parts, as shown in the middle part of Fig. 4(b). In the case of open legs, the candidate contour points mainly appear on the feet, and partially on the crotch and other body parts, as shown in the middle part of Fig. 4(c). To suppress the influence of candidate points that do not belong to the feet, we assign small weights to these points in Eq. (5).

The temporal walking position signal $\mathcal{F}(p)$ sometimes contains outliers when the feet's contour is not estimated correctly because of the shadow of a pedestrian on the road surface or markings, such as white lines. Our method detects outliers by applying a Hampel filter and performs a linear interpolation.

2.5 Interaction Video Sequence Generation

In procedure P4, we extract a feature that allows human observers to confirm the body interaction visually. Specifically, we place pedestrians in the same group in a virtual space using a temporal posture signal $\mathcal{V}(p)$ and temporal walking position signal $\mathcal{F}(p)$, and generate an interaction video sequence $\mathcal{S}$ using 3D rendering. In this virtual space, we visualize the temporal posture signal and temporal walking position signal of each pedestrian using the standard body shape parameters, which is the average person's body shape prepared in SMPL, as described in Sect. 2.3. When we render an interaction video sequence in a virtual space, we always set the virtual camera viewpoint at a fixed position in front of the pedestrian group to capture the physical body interaction which increases the accuracy of conversation activity recognition.

In the following, we explain how to generate an interaction video sequence $\mathcal{S}$. Our method places the 3D vertex $\boldsymbol{v}(t,p) \in \mathcal{V}(t,p) \in \mathcal{V}(p)$ on the body surface

obtained in Sect. 2.3 at the walking position $\tilde{\boldsymbol{f}}(t,p) \in \mathcal{F}(p)$ obtained in Sect. 2.4. The 3D vertex $\tilde{\boldsymbol{v}}(t,p)$ in the virtual space is converted as follows:

$$\tilde{\boldsymbol{v}}(t,p) = \boldsymbol{v}(t,p) + \tilde{\boldsymbol{f}}(t,p). \tag{6}$$

All vertices $\boldsymbol{v}(t,p)$ in a set $\mathcal{V}(t,p)$ are converted to $\tilde{\boldsymbol{v}}(t,p)$. Suppose that a converted set $\tilde{\mathcal{V}}(t,p)$ consists of $\tilde{\boldsymbol{v}}(t,p)$ and their adjacent vertices. A temporal posture and walking position signal $\tilde{\mathcal{V}}(p)$ in the virtual space is expressed as

$$\tilde{\mathcal{V}}(p) = \{\tilde{\mathcal{V}}(t,p) \mid t \in \mathcal{T}\}. \tag{7}$$

Note that our method determines pedestrian p that belongs to the same group using the distance between the walking positions $\tilde{\boldsymbol{f}}(t,p)$ of pedestrians. After obtaining $\tilde{\mathcal{V}}(p)$ for a pedestrian group, we place each pedestrian that belongs to the same group and perform 3D rendering to generate an image $\boldsymbol{S}(t)$. An interaction video sequence $\mathcal{S}$ for each pedestrian group is expressed as

$$\mathcal{S} = \{\boldsymbol{S}(t) \mid t \in \mathcal{T}\}. \tag{8}$$

The posture parameters are sometimes estimated with an unnaturally large tilt of the human body if a temporal posture and walking position signal $\tilde{\mathcal{V}}(p)$ is directly used for rendering an interactive video sequence. Our method corrects the human body's inclination relative to the road surface by always setting the rotation angle to 0 degrees.

2.6 Conversation Activity Label Classification

In P5, we apply an existing classification network developed in action recognition to determine conversation activity labels using an interaction video sequence $\mathcal{S}$. We use the C3D network [20] that consists of 3D convolution layers designed for action recognition. Our method divides an interaction video sequence into multiple short video sequences, which are input into the C3D network to predict candidate labels that represent conversation activity for each short video sequence. A majority vote among these candidates determines the final label.

In the following, we explain the details of our method for determining the conversation activity label. Our method generates K short video sequences with different initial times from a single interaction video sequence $\mathcal{S}$ during the C3D network training and prediction process. Short video sequence $\hat{\mathcal{S}}$ is expressed as

$$\hat{\mathcal{S}} = \{\boldsymbol{S}(\hat{t}) \mid \hat{t} \in \hat{\mathcal{T}}\}, \tag{9}$$

where $\hat{\mathcal{T}}$ is a set of time points $\hat{t}$ of the image $\boldsymbol{S}(\hat{t})$ that belong to the short video sequence. Our method randomly determines the initial time point $\hat{t}_1$. We generate a short movie sequence $\hat{\mathcal{S}}$ when $\hat{T}(< T)$ images are collected by progressing time at equal intervals I from $\hat{t}_1$. $\hat{T}$ also represents the total number of time points in the short video sequence. During the training process, we train the C3D network using LK short video sequences generated from L interaction video sequences prepared in advance. During the prediction process, we calculate K candidates for the conversation activity label using the input short video sequences generated from an interaction video sequence, and finally determine the output label using majority voting among candidates.

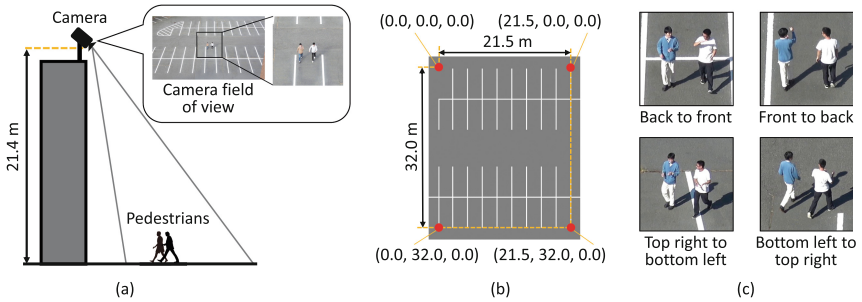


Fig. 5. Camera setting for collecting color video sequences of pedestrian groups while they were walking outdoors and conversing.

3 Experiments

3.1 Dataset

To investigate the effectiveness of our method, we collected color video sequences of pedestrian groups while they were walking outdoors and conversing. Figure 5(a) shows the camera setting. We set the height from the road surface to the camera (SONY, FDR-AX55) to 21.4 m to obtain an overhead view of an outdoor parking lot. The camera resolution was 3840×2160 pixels and the frame rate was 30 fps. Figure 5(b) shows the road surface coordinate system described in Sect. 2.4. We pre-computed the homography matrix from four white line intersections on the road surface. The camera position in the road surface coordinate system was (10.7, 59.3, 21.4).

We recruited 20 participants (19 men, one woman, 22.6±1.3 years old, university students, Japanese ethnicity). When recruiting the participants, we required that they be somewhat acquainted with each other to avoid a lack of conversation when they first met each other. We controlled the number of pedestrians in a group to a minimum of two participants with whom a conversation could occur. We randomly selected two pedestrians from the 20 participants without duplicates to form a single pedestrian group. We prepared 52 pedestrian groups. We controlled each pedestrian group so that the two participants walked side by side, which is considered to occur most frequently in real scenarios.

We acquired color video sequences of pedestrian groups walking outdoors for each conversation activity label (active conversation, inactive conversation, and no conversation). In one color video sequence, a pedestrian group appeared in the camera's field of view from the start to the end, when it disappeared. To confirm the robustness of the virtual camera viewpoint used in our method, we set four walking paths on the road surface: back to front, front to back, top right to bottom left, and bottom left to top right, as shown in Fig. 5(c). We randomized the order in which the participants walked along each path and the order in which the two participants lined up next to each other. In total, we collected 52 (groups) × 3 (labels) × 4 (walking paths) = 624 color video sequences. Figure 6 shows examples of the pedestrian group video sequences

Fig. 6. Examples of the pedestrian group video sequences $\mathcal{R}'$ generated from the color video sequences.

$\mathcal{R}'$ generated from the collected color video sequences. To generate $\mathcal{R}'$, we set a region of interest for the color video sequence so that two pedestrians that belonged to the same group were within the same field of view using the body region image $\boldsymbol{R}(t, p) \in \mathcal{R}(p)$ estimated in procedure P1.

3.2 Conversation Activity Labels

When collecting color video sequences, we only instructed the participants on the topic of the conversation and did not give any explanation or instructions regarding the physical body interaction. We set the following conditions for collecting color video sequences for each conversation activity label.

Active conversation:
As a topic of conversation, we instructed the participants to introduce their hobbies while walking. We collected color video sequences while a pedestrian talked about a hobby, the other pedestrian responded to it, and started a new conversation about a hobby.

Inactive conversation:
As a topic of conversation, we instructed the participants to talk about topics of little interest to each other while walking. The topic was chosen by the participants from several candidate topics prepared in advance (e.g., economic situation and political situation in a country that the participants had never visited and had almost no knowledge of).

No conversation:
We instructed the participants not to engage in any conversation while walking.

Fig. 7. Examples of interaction video sequences $\mathcal{S}$.

We randomized which pedestrians in the group initiated the conversation when collecting active and inactive labels.

3.3 Experimental Conditions

In the following, we describe the experimental conditions for procedures P1 through P3. We used the default parameters provided for PHALP in P1 and P2. The window size of the Hampel filter in P2 and P3 was 5. We set the body shape parameters of SMPL to the default parameters provided by PHALP. We automatically determined the σ of Eq. (4) in P3 according to the height of the pedestrian's bounding rectangle. Specifically, σ increased as the height increased and σ decreased as the height decreased.

Next, we describe how to determine the virtual camera viewpoint for generating interaction video sequences in procedure P4. We determined the direction in which a pedestrian group walks on a road surface by fitting a straight line using the group's center positions at all time points. We always kept the virtual camera viewpoint at a distance of 4.25 m from the center position in the direction of the pedestrian group. The height of the virtual camera viewpoint was 0.85 m from the road surface. Figure 7 shows examples of the interaction video sequences $\mathcal{S}$. The color scheme for each pedestrian was either light red or light blue and was determined randomly without duplication. We believe that human observers can visually confirm the posture among pedestrians, such as arm bending and face orientation, and the positioning of the pedestrians in each group, from the interaction video sequences in the figure.

The C3D network [20] in procedure P5 consisted of four convolution layers, four pooling layers, and two affine layers. The filter size for 3D convolution was $3 \times 3 \times 3$. Time length $\hat{T}$ of a short video sequence $\hat{\mathcal{S}}$ was 16. We set the array

size of the short video sequences to 100 (pixels) × 100 (pixels) × 3 (colors) × 16 (time points). We set $I = 18$ and $K = 50$ for the parameters described in Sect. 2.6. We used RMSprop as the optimizer when training the C3D network, with a learning rate of 0.0001 and mini-batch size of 16. We trained the C3D network from scratch.

We applied leave-one-group-out when evaluating the accuracy of conversation activity recognition. Specifically, we used 12 interaction video sequences generated from one pedestrian group for the prediction process and $L = 612$ interaction video sequences generated from the remaining 51 pedestrian groups for the training process. We repeated the training and prediction processes for all 52 pedestrian groups. We prepared 3 (labels) × 4 (walking paths) = 12 interaction video sequences per pedestrian group.

We evaluated the computational cost of our method on a PC equipped with a GPU (RTX 2080 Ti) and CPU (i9-9940X). The processing time was 0.29 s for P1, 0.66 s for P2, 0.05 s for P3, and 0.58 s for P4 per video sequence frame. The processing time for P5 was 0.01 s per short video sequence during prediction. The total GPU memory usage was 4.7 GB.

3.4 Basic Performance

We evaluated the effectiveness of our method using interaction video sequences as features. For comparison, we used the following features to calculate the accuracy of conversation activity recognition.

M1: Interaction video sequence $\mathcal{S}$
We used $\mathcal{S}$ generated in procedure P4 of our method as the feature. Specifically, we generated short video sequences $\hat{\mathcal{S}}$ in procedure P5 from $\mathcal{S}$. The array size of the short video sequence was 100 (pixels) × 100 (pixels) × 3 (colors) × 16 (time points).

M2: Pedestrian group video sequence $\mathcal{R}'$
We used $\mathcal{R}'$, which represents the appearance of the pedestrian group, as the feature. Examples of $\mathcal{R}'$ were already shown in Fig. 6. We directly passed the pedestrian group video $\mathcal{R}'$ to procedure P5 and generated short video sequences from $\mathcal{R}'$. The array size of the short video sequence was 100 (pixels) × 100 (pixels) × 3 (colors) × 16 (time points).

M3: Temporal posture signal $\mathcal{V}(p)$
We used $\mathcal{V}(p)$ estimated from each pedestrian that belonged to the same group as the feature. Specifically, we directly passed $\mathcal{V}(p)$ estimated in procedure P2 to procedure P5 and then generated short temporal signals. The array size of the short temporal signal was 6890 (vertices) × 2 (pedestrians) × 3 (components) × 16 (time points).

M4: Temporal posture and walking position signal $\tilde{\mathcal{V}}(p)$
We used $\tilde{\mathcal{V}}(p)$, combining a temporal posture signal $\mathcal{V}(p)$ with the temporal walking position signal $\mathcal{F}(p)$ estimated from each pedestrian that belonged to the same group as the feature. Specifically, we directly passed $\tilde{\mathcal{V}}(p)$ generated in procedure P4 to procedure P5 and then generated short temporal

Table 1. Comparison of the accuracy of conversation activity recognition using each feature.

Feature for conversation activity recognition	Accuracy (%)
M1: Interaction video sequence $\mathcal{S}$	**76.2±0.7**
M2: Pedestrian group video sequence $\mathcal{R}'$	57.3±1.3
M3: Temporal posture signal $\mathcal{V}(p)$	72.9±0.9
M4: Temporal posture and walking position signal $\tilde{\mathcal{V}}(p)$	74.1±0.7

signals. The array size of the short temporal signal was 6890 (vertices) × 2 (pedestrians) × 3 (components) × 16 (time points).

We input each feature into the C3D network to predict the conversation activity label in P5. We calculated accuracy using the number of correctly predicted conversation activity labels. Because there was random sampling when we extracted each feature, we set the number of trials used to calculate recognition accuracy to 10. In M3 and M4, to align the dimensionality with other features, we randomly sampled 5000 vertices and then transformed the array size from 5000 × 2 × 3 × 16 to 100 × 100 × 3 × 16. In each accuracy evaluation trial, we assumed that the vertices sampled in all short temporal signals were the same. The other experimental conditions were the same as those described in Sect. 3.3.

Table 1 shows the accuracy of using each feature in conversation activity recognition. Recognition accuracy was 76.2±0.7% for interaction video sequence $\mathcal{S}$ of M1, 57.3±1.3% for pedestrian group video sequence $\mathcal{R}'$ of M2, 72.9±0.9% for temporal posture signal $\mathcal{V}(p)$ of M3, and 74.1±0.7% for temporal posture and walking position signal $\tilde{\mathcal{V}}(p)$ of M4. In all cases, we confirmed that our method M1 was more accurate than M2, M3, and M4. These results indicate that using a feature of an interaction video sequence generated by our method was more effective in recognizing conversation activity than using a feature of a pedestrian group video sequence, a temporal posture signal, or a temporal posture and walking position signal.

Instead of C3D, we applied TimeSformer [1] as a video action recognition method and LSTM [8] as a time series analysis method. TimeSformer performed fine-tuning on a model pre-trained with Kinetics-400, whereas LSTM trained a model from scratch. The recognition accuracies were 71.8±0.7% for TimeSformer and 67.1 ± 1.1% for LSTM. Our method obtained higher recognition accuracy (76.2±0.7%) than the existing methods. The GPU memory usage was 1.3 GB for C3D used in our method, 6.1 GB for TimeSformer, and 0.7 GB for LSTM. We believe that our method is reasonable in terms of the trade-off between accuracy and memory usage.

We evaluated the recognition accuracy of our method for the case of several groups walking simultaneously. The number of groups in each frame ranged from 0 to 3. We used a total of 120 groups. The accuracy of our method was 67.4 ± 0.2%. Although our method performed well in this case with minimal occlusion, it is important to note that real-world scenarios often involve heavy

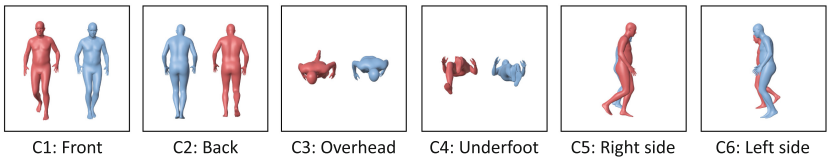


Fig. 8. Examples of interaction video sequences generated from different virtual camera viewpoints in procedure P4 of our method.

occlusion caused by people overlapping. This presents a significant limitation that we need to address in future work. For practical applications, we must develop methods for various scenarios, such as heavy occlusion and interaction with objects such as shopping trolleys.

3.5 Evaluation of Different Virtual Camera Viewpoints

We evaluated the accuracy of conversation activity recognition for different virtual camera viewpoints when generating an interaction video sequence in procedure P4. We set the positions of the virtual camera viewpoints on C1 front, C2 back, C3 overhead, C4 underfoot, C5 right side, and C6 left side. Figure 8 shows interaction video sequences generated using these virtual camera viewpoints. We changed only the position of the virtual camera viewpoint; the other experimental conditions were the same as those described in Sect. 3.4.

Table 2 shows the accuracy for each virtual camera viewpoint when generating interaction video sequences. We confirmed that C1, in which the virtual camera viewpoint was the front of the pedestrian group, had higher recognition accuracy than C2, C3, C4, C5, and C6, in which the virtual camera viewpoint was not the front of the pedestrian group. Furthermore, we checked the recognition accuracy of our method C1 for each walking path in Fig. 5(c). We achieved the same level of accuracy for all walking paths. Based on these results, when generating interaction video sequences in procedure P4, placing the virtual camera viewpoint in a position that always captured a pedestrian group from the front led effectively to the recognition of conversation activity.

4 Conclusions

We proposed a method for recognizing conversation activity in a group of pedestrians walking outdoors using interaction video sequences that represent human body interactions. The experimental results demonstrated that our method is superior to the alternative techniques using pedestrian body region video sequences or temporal posture and walking position signals in conversation activity recognition. We believe that our method can be implemented in a variety of potential applications in addition to the marketing applications described in Sect. 1. For example, we considered medical applications for dementia checking,

Table 2. Accuracy of conversation activity recognition when generating interaction video sequences from different virtual camera viewpoints in P4.

Virtual camera viewpoint	Accuracy (%)
C1: Front	**76.2 ± 0.7**
C2: Back	74.8 ± 0.3
C3: Overhead	70.2 ± 0.8
C4: Underfoot	72.0 ± 0.9
C5: Right side	40.0 ± 1.5
C6: Left side	48.2 ± 2.6

office applications for mental health checking, and educational applications for bullying detection. In future work, we intend to develop a method to recognize conversation activity at multiple levels and a robust method for occlusion. We will expand evaluations by increasing the number of pedestrians in the same group and changing the positional relationship of pedestrians within a group. We will perform a performance comparison with group activity recognition methods, for example, ARG [21], Actor-Transformers [4], GroupFormer [10], DIN [22], and KRGFormer [14]. We appreciate Professor Yoshio Iwai's valuable advice and suggestions during this study. We would like to thank Mr. Norihiko Torii, Mr. Tomohiro Miyake, and Mr. Osamu Yoshimura of SEIRYO ELECTRIC Corporation for their helpful advice on this paper.

References

1. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the International Conference on Machine Learning (2021)
2. Chamveha, I., Sugano, Y., Sato, Y., Sugimoto, A.: Social group discovery from surveillance videos: a data-driven approach with attention-based cues. In: Proceedings of the British Machine Vision Conference, pp. 1–12 (2013)
3. Ehsanpour, M., Saleh, F., Savarese, S., Reid, I., Rezatofighi, H.: JRDB-Act: a large-scale dataset for spatio-temporal action, social group and activity detection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 20951–20960 (2022)
4. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.M.: Actor-transformers for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 836–845 (2020)
5. Ge, W., Collins, R.T., Ruback, R.B.: Vision-based analysis of small groups in pedestrian crowds. IEEE Trans. Pattern Anal. Mach. Intell. **34**(5), 1003–1016 (2012)
6. Han, R., Yan, H., Li, J., Wang, S., Feng, W., Wang, S.: Panoramic human activity recognition. In: Proceedings of the European Conference on Computer Vision, pp. 224–261 (2022)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)

8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
9. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: discriminative models for contextual group activities. In: Advances in Neural Information Processing Systems, vol. 1, p. 1216–1224 (2010)
10. Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., Yi, S.: Groupformer: group activity recognition with clustered spatial-temporal transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13648–13657 (2021)
11. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. **34**(6), 1–16 (2015)
12. McNeill, D.: Hand and Mind: What Gestures Reveal about thought. University of Chicago Press, Chicago (1992)
13. Odashima, S., Shimosaka, M., Kaneko, T., Fukui, R., Sato, T.: Collective activity localization with contextual spatial pyramid. In: Proceedings of the European Conference on Computer Vision, pp. 243–252 (2012)
14. Pei, D., Huang, D., Kong, L., Wang, Y.: Key role guided transformer for group activity recognition. IEEE Trans. Circuits Syst. Video Technol. **33**(12), 7803–7818 (2023)
15. Qing, L., et al.: Public life in public space (PLPS): a multi-task, multi-group video dataset for public life research. In: Proceedings of the International Conference on Computer Vision Workshops, pp. 3611–3620 (2021)
16. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people by predicting 3D appearance, location and pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2740–2749 (2022)
17. Rota, P., Conci, N., Sebe, N.: Real time detection of social interactions in surveillance video. In: Proceedings of the European Conference on Computer Vision, pp. 111–120 (2012)
18. Solera, F., Calderara, S., Cucchiara, R.: Socially constrained structural learning for groups detection in crowd. IEEE Trans. Pattern Anal. Mach. Intell. **38**(5), 995–1008 (2016)
19. Su, J., Huang, J., Qing, L., He, X., Chen, H.: A new approach for social group detection based on spatio-temporal interpersonal distance measurement. Heliyon **8**(10), e11038 (2022)
20. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
21. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9956–9966 (2019)
22. Yuan, H., Ni, D., Wang, M.: Spatio-temporal dynamic inference network for group activity recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7456–7465 (2021)
23. Zanlungo, F., Brcic, D., Kanda, T.: Pedestrian group behaviour analysis under different density conditions. Transp. Res. Procedia **2**, 149–158 (2014)
24. Zanotto, M., Bazzani, L., Cristani, M., Murino, V.: Online bayesian nonparametrics for group detection. In: Proceedings of the British Machine Vision Conference, pp. 1–12 (2012)

Efficient Multi-Band Temporal Video Filter for Reducing Human-Robot Interaction

Lawrence O'Gorman(✉)

Nokia Bell Labs, Murray Hill, NJ, USA
larry.o_gorman@nokia-bell-labs.com

Abstract. Although mobile robots have on-board sensors to perform navigation, their efficiency in completing paths can be enhanced by planning to avoid human interaction. Infrastructure cameras can capture human activity continuously for the purpose of compiling activity analytics to choose efficient times and routes. We describe a cascade temporal filtering method to efficiently extract short- and long-term activity in two time dimensions, isochronal and chronological, for use in global path planning and local navigation respectively. The temporal filter has application either independently, or, if object recognition is also required, it can be used as a pre-filter to perform activity-gating of the more computationally expensive neural network processing. For a testbed 32-camera network, we show how this hybrid approach can achieve over 8 times improvement in frames per second throughput and 6.5 times reduction of system power use. We also show how the cost map of static objects in the ROS robot software development framework is augmented with dynamic regions determined from the temporal filter.

Keywords: Human-robot interaction (HRI) · Video analytics · Mobile robots · Robot navigation · Activity filter · Pedestrian dynamics

1 Introduction

Robots must navigate with respect to both their static world (walls and fixed objects) and dynamic (people and other robots). The dynamic world can be classified in terms of short- and long-term time frames. Robots capture short-term events by using their on-board sensors; for instance, a person steps in front of the robot and the robot should stop. But there is also activity that repeats in predictable, longer-term periodic cycles. Repetition over a regular time frame is termed *isochronal*. Examples of isochronal time periods include factory shifts, scheduled deliveries, and employee breaks. The effects of these activities on navigation are just as real as for static objects except that their occurrence is time dependent. In this paper, we determine both long- and short-term activity by temporal video filtering for use in robot navigation and path planning.

Video analysis of human activity can be performed using convolutional neural networks or vision transformers to detect and track people. This neural network

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 375–390, 2025.
https://doi.org/10.1007/978-3-031-78341-8_24

processing can achieve a high level of recognition, but at high cost of computation. If only activity detection is required – not individually segmented persons – then motion flow [13,20,26,29] is a less expensive alternative. Instead of GPU-processing needed for real-time neural network detection, motion flow can be performed by IoT-level processors, typified by low cost, low power, small memory, and narrow bandwidth. This low level of processing is a practical cost- and power-usage alternative for installations that might have tens or hundreds of cameras. A hybrid solution including a temporal filter and neural network object detection is also shown to realize cost and power efficiencies.

The application goal of this paper is to use fixed cameras to detect human activity such that it can be avoided for the purpose of efficient robot path planning and navigation. We consider both off-line global path planning, where the goal is to schedule the robot for regular (daily, etc.) tasks on selected paths and at times that are efficient and safe with respect to human activity; and real-time local navigation where the goal is to choose the best of current path options for immediate robot navigation.

The technology goal of this paper is to offer an efficient multi-band temporal video filter for extracting short- and long-term activity bands from both chronological and isochronal time. Extraction of these bands requires multiple low, high, and bandpass filters. These could be implemented separately, however we show how a cascade filter architecture can extract all these bands efficiently from a single video stream. We show how use of the cascade filter both reduces video processing and video storage. Although temporal video filters are common and activity detection to avoid human-robot interaction is often used, we believe the design and use of a single cascade filter to efficiently extract multiple bands over chronological and isochronal time is novel.

The main contributions of this work are:

1. A cascade filter that extracts temporal video information of long- and short-term human activity more efficiently than through separate filters.
2. Use of a single, efficient cascade filter to identify long- and short-term activity to aid global and local robot navigation.
3. An efficiency analysis of using pixel- and feature-based activity analytics either independently or as a hybrid combination of pre-filter and neural network object detection.
4. Practical implementation on the ROS robot operating system.

In Sect. 2, we review related literature. In Sect. 3, we describe the system architecture to extract long- and short-term activity. Section 4 shows costs of computation of activity detection, object detection, and a hybrid of both.

2 Background

Early work in robot navigation dealt with a static environment of building walls and fixed-placed objects, e.g., [11]. Inclusion of moving objects (other robots) in

dynamic environments followed, e.g., [3]. While navigation with respect to inanimate objects is a challenge, human presence adds the trade-off of safety versus efficiency. In [28] the procedure of SLAM (Simultaneous Localization and Mapping) is augmented to include humans. For prolonged observation, the SLAM robot can both observe and extrapolate human trajectories to create human motion maps. However, there is a degree of unpredictability when dealing with humans [6] that makes the success of trajectory prediction variable.

Because full trajectories are often difficult to track, many approaches represent floor space as an occupancy grid and determine statistics independently within each grid cell [16]. In [25], this is done with 2-state Markov probabilities of entry and exit to a cell. Direction is added in [30], in which a 9-state Hidden Markov Model describes motion direction from each grid cell, and a 9th state for staying in the same cell. In [10], grid flow is extended to be either observed (statistical) or spatially extrapolated from cell directions to predict continuing trajectories. Mobile robots cannot be in all places at all times so it is understandable that predicted flow is a valuable complement to observed flow. Finally, work such as [31] combine methods discussed here to yield a multi-layer representation (static layer from SLAM and object layer from YOLO [23]).

Besides avoiding human-occupied areas, advantage can be gained by observing the paths humans travel and to follow these. Imitation learning, or inverse reinforcement learning, is a machine learning approach that doesn't require training with labeled samples. Instead, an agent observes how experts behave (humans in our case), learns a reward function that the experts are unconsciously acting upon, and seeks to maximize that reward [7]. In [32] human trajectories are observed to learn their normal paths with respect to objects. With this prediction, efficient human-aware robot paths can be planned. In [14], inverse reinforcement learning is used with particular emphasis on socially normative navigation in dense and complex scenes such as meeting places and hallway intersections. This can be extended beyond just navigation to where robots can learn more complex human movements for the purpose of human-robot collaboration [17].

Besides static location of objects and prediction of forward path, affordance is another relevant factor for robot navigation among objects and humans. Affordance describes how an object is used, and for navigation purposes this relates to spatial interaction between human and object [12,18,27].

For the previously described work and for many robot navigation systems, sensors on the robot are used for navigation. But many situations limit robots to indoors and on paths traveled repeatedly. In these cases, fixed cameras can augment onboard sensors to aid navigation. In [22], fixed cameras are used to create a heat map-based path planner. Motion pixels are found and accumulated into "heat values". Resulting cost values at regular grid locations are associated with their closest path edges. This reduction from grid points to many fewer path edges reduces storage and subsequent communication of cost values to the robot. A relatively new fixed-camera alternative is an event camera, which contains bio-inspired vision sensors to capture scene changes [9]. Although these capture activity, as is our goal, they do not also capture traditional frames for video processing as is also our goal, so are outside the domain of this paper.

This paper has similarities and differences with respect to the literature described. Unlike work that combines SLAM and person detection [16], we detect only people, but do so by their activity rather than their identity. Unlike work that categorizes objects by their affordances [12], we deal only indirectly by learning observed human activity and creating a model similar to cost maps and social force models [14,21,22,27]. A difference in our work from cost maps and social force models, which directs a robot *away from* obstacles is that our model directs it *toward* higher probability paths. In contrast to work using on-board robot sensors and cameras to aid navigation [3,10,16,21,27,28,31], we use fixed cameras as do [22,30,32]. There is much work in learning and avoiding humans in close-up human-robot interaction with static robots [2,15], which has both similarities and differences to mobile robot interaction investigated here.

Our work is closest in purpose and methods to those proposing dynamic occupancy grids [10,22,25,30]. These are created by unsupervised learning of human activity in a grid-space over time. Our work also has similarities to the inverse reinforcement learning approaches used for predicting trajectories in [32] and socially normative behaviors in [14,21]. Whereas these seek to generalize beyond specific objects and locations, our approach imitates what humans do with emphasis on fixed locations and times.

3 Method

3.1 Definitions

Our methods distinguish different types of human activity related to different navigation tasks as shown in Table 1. *Long-term activity* refers to human motion that is statistically stationary in time and place. We also use the term isochronal, meaning that this activity happens on a cyclic basis in some time frame. For simplicity in this paper, our time cycle is one day, so long-term activity refers to the activity that is statistically determined over many days at each chosen time of day. We designate isochronal time as t^*, so an example of an isochronal sequence is ${t_1}^* = 18:23$ Monday, ${t_2}^* = 18:23$ Tuesday, ...

Short-term activity refers to human motion at the current time, of which we distinguish two types. *In-place activity* is static in location. In-place activity may include people who are stationary in location such as waiting in line or dwelling at a shop window. In-place activity also includes people who are not stationary in location, but who create a location that is active by, for instance, passing through a crowded bottleneck such as an entranceway. Opposite to in-place activity is *moving activity*. This refers to people movement with changing location, such as people walking.

We distinguish two types of global path planning. *Off-line global path planning* pertains to the task of choosing a robots full path for a future time. If we are arranging a planned daily trip of a robot delivery cart for example, we would seek to choose the times and paths that are statistically of least activity. *Real-time global path planning* pertains to planning a full path that is to be

Table 1. Activity types corresponding to path planning and navigation.

Activity	Global Planning Off-line	Global Planning Real-time	Local	Navigation Planning
Long-term	1	2	–	–
Short-term, in-place	–	1	1	1
Short-term, moving	–	1	1	1

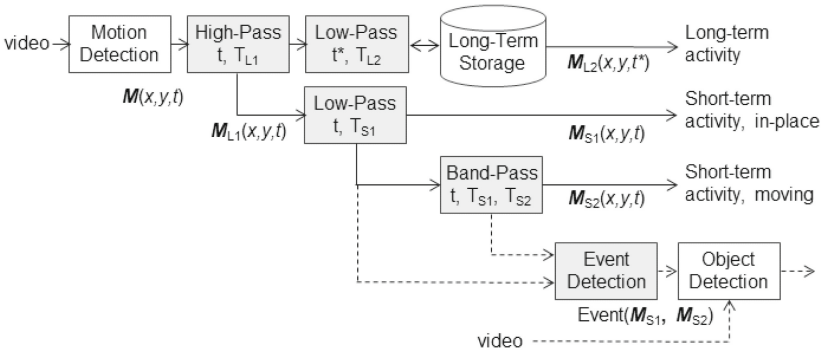


Fig. 1. Functional diagram shows temporal filter cascade and event detection.

begun at the current time. *Local planning* pertains to altering the global path with information local to (i.e., a short distance from) the robot at that time.

In Table 1, off-line global path planning can only be performed with respect to long-term activity because short-term activity is not known off-line. However, for real-time global path planning, in-place short-term activity can be used because both are happening at the current time. We designate in-place short-term activity as first choice "1" for this column and second choice "2" for long-term activity. This is because, when a path is altered due to short-term activity, the real-time global planning may also use long-term activity information.

3.2 Architectural Overview

A functional diagram of the cascade filter is shown in Fig. 1. Motion detection is performed on each video frame, then a cascade of temporal video filters extracts long- and short-term activity. We describe each component in sections below.

3.3 Motion Detection

Motion detection is performed on each frame to obtain a motion image of K blocks subsampled from the full frame, each block k containing 2 motion features, $\mathbf{f}_k$ = (density, direction), corresponding to an (x, y) location,

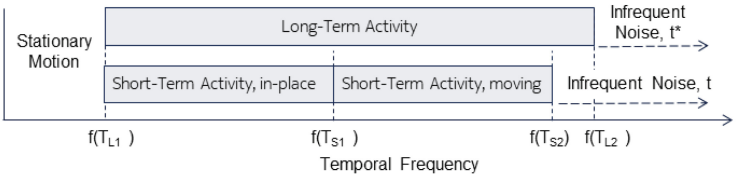


Fig. 2. Temporal filter bands associated with activity types.

$$\mathbf{M}(x, y, t) = \{f_k\}_t, 0 < k < K \tag{1}$$

Density is a measure of the motion in a block, a function of the number of motion pixels and their gradient values. Direction is quantized to 8 angles. The motion features are found from motion flow, specifically by [20], but similarly by [13,29] or optical flow [8] methods. For brevity below we write a single block as $b(t)$, where $\mathbf{b}(t) = \mathbf{M}(x_i, y_j, t)$, and x_i, y_i are top-left x, y block coordinates, which increment with the block size as described in Sect. 4.2.

3.4 Temporal Filtering

After motion detection, long- and short-term activity are found using a cascade of temporal video filters shown in Fig. 1. Figure 2 shows the temporal filter bands of frequencies corresponding to the filter time constants TL1, TL2, TS1, and TS2, which are described below.

For all filtering (with one exception noted below), we use a first-order IIR filter, also called an exponential moving average filter, to give more weight to the most recent block $\mathbf{b}(t)$ than past blocks $\mathbf{b}'(t-1)$, and obtain the block result $\mathbf{b}'(t)$,

$$\mathbf{b}'(t) = \alpha\mathbf{b}'(t-1) + (1-\alpha)\mathbf{b}(t), \alpha \in [0, 1] \tag{2}$$

We choose the filter parameter value α through a more intuitive parameter, which we call the 10%-decay duration, T. This is the amount of time during which a filtered signal will decay to 10% of original with zero input. In Eq. 2, if input $\mathbf{b}(t) = 0$ for n samples, then $\mathbf{b}'(t = n)/\mathbf{b}'(t = 0) = \alpha^n = 0.1$. So, we can obtain α with chosen T as follows,

$$\alpha = 0.1^{(1/n)}, n = rT, \tag{3}$$

where the number of samples is equal to the video frame rate r in frames per second times the 10%-decay duration T [sec]. The temporal video filters, along with rational for their parameters and resultant image results, are described in more depth in [19].

Filtering begins in Fig. 1 with a high-pass filter F_{L1} applied to the frame-rate stream of motion vectors,

$$\mathbf{M}(x, y, t) * F_{L1} \rightarrow \mathbf{M}_{L1}(x, y, t) \tag{4}$$

The filter time constant T_{L1} is chosen to reduce low frequency "stationary motion noise" as described in Sect. 4.2.

The result of Eq. 4 is combined with the long-term, isochronal motion vector from storage at corresponding time $t = t^*$ using a low-pass filter F_{L2}, and the resultant $\mathbf{M}_{L2}$ is stored,

$$\mathbf{M}_{L1}(x, y, t^*) * F_{L2} \rightarrow \mathbf{M}_{L2}(x, y, t^*) \tag{5}$$

Long-term activity is updated in isochronal time, in our case 1 sample per day for each t^*, $0 < t^* < 1440$, where 1440 is the number of minutes in a day. Because of this long sample period, there is a tradeoff between the duration of samples needed to obtain a good measure of long-term activity at any t^* and the delay within which the measure adapts to changes in long-term activity. We choose a low-pass filter value to reduce infrequent (shot) noise as described in Sect. 4.2.

Short-term, in-place activity can be identified by applying a low-pass filter F_{S1} to $\mathbf{M}_{L1}$,

$$\mathbf{M}_{L1}(x, y, t) * F_{S1} \rightarrow \mathbf{M}_{S1}(x, y, t) \tag{6}$$

The time constant T_{S1} is set to capture people activity in the same location, as described in Sect. 4.2.

Short-term, moving activity is identified using a band-pass filter. The low end of the filter is T_{S1} and the high end T_{S2}. The time constant T_{S1} separates the signal from in-place activity and T_{S2} removes high-frequency, infrequent noise. Short-term, moving activity is found by subtracting $\mathbf{M}_{S1}$ from $\mathbf{M}_{L1}$ to rid the in-place activity (effectively a high-pass filter) and then applying a low-pass filter to rid infrequent noise. This combination results in band-pass filtering,

$$(\mathbf{M}_{L1}(x, y, t) - \mathbf{M}_{S1}(x, y, t)) * F_{S2} \rightarrow \mathbf{M}_{S2}(x, y, t) \tag{7}$$

Note that the two filters at the high frequency end of Fig. 2 are not redundant because one reduces noise in isochronal time t^* at T_{L2}, and the other in chronological time t at T_{S2}.

Finally, event detection is performed on short-term, in-place and moving activity to act as a gate on more computationally expensive processing such as object detection,

$$\begin{aligned} \text{Event}(\mathbf{M}_{S1}, \mathbf{M}_{S2}) &= 1, \text{do object detection} \\ &= 0, \text{do nothing.} \end{aligned} \tag{8}$$

3.5 Off-Line Global Path Planning

Most commonly, a path is planned that avoids human activity in time and space. It is less common to choose path segments of higher activity, but we do this in the following way. In Fig. 3, the long-term storage contains motion statistics for each (isochronal) minute of the day $\mathbf{M}(p_i, t^*)$, where we have replaced x, y of Eq. 1 with p_i to indicate this is for an x, y space corresponding to path segment p_i. We time-collapse and binarize this as follows,

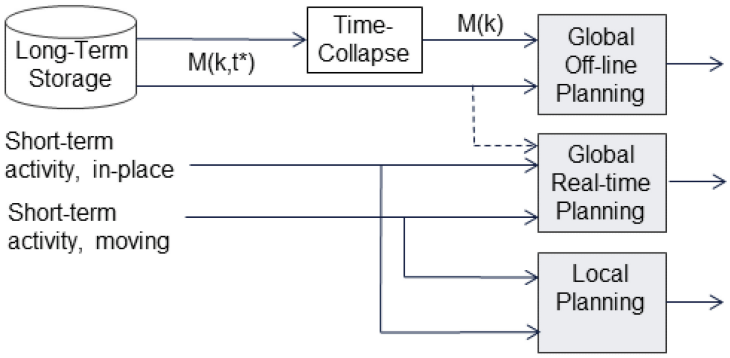


Fig. 3. Different types of activity information used for different path planning tasks.

$$\begin{aligned} \mathbf{M}'(p_i) &= 1, \text{ if } \mathbf{M}(p_i, t^*) \neq 0 \text{ for any } t^*, 0 < t^* < 1440 \\ &= 0, \text{ otherwise.} \end{aligned} \tag{9}$$

The result $\mathbf{M}'(p_i)$ can be thought of as a path attribute learned from people activity. If $\mathbf{M}'(p_i)$ is 0, then no people travel through this path segment, so its reasonable to assume it is not passable (for whatever reason) and for a robot to avoid this path as well.

Off-line path planning through path segments at isochronal time t^* is now a function of two activity-related values, $\mathbf{M}(p_i, t^*)$ and $\mathbf{M}'(p_i)$. Therefore, the activity-related cost for off-line planning over a complete path containing segments p_i can be written,

$$\begin{aligned} \text{Cost}_1(\{p_i\}, t^*) = &\sum_i \text{Cost}(\mathbf{M}(p_i, t^*), \text{ if all } \mathbf{M}'(p_i) = 1 \\ &= \infty, \text{ if any } \mathbf{M}'(p_i) = 0. \end{aligned} \tag{10}$$

3.6 Real-Time Global Path Planning

Since real-time global path planning is performed just before the robot begins a path, there is current short-term activity information available as well as long-term information as shown in Table 1. It may make sense to weight the short-term information higher than the long-term information $\{w1, w2\}$, although we do not explore that further here. For simplicity, we do not repeat the second line of Eq. 10, leaving it implicit that any locations of $\mathbf{M}'_p = 0$ are not included in a path. The activity-related cost for a path through segments $\{p_i\}$ from off-line planning is,

$$\begin{aligned} \text{Cost}_2(\{p_i\}, t)) = w_1 \text{Cost}_1(\{p_i\}, t^* = t) \\ + w_2 \sum_{i=1}^{N} \text{Cost}(\mathbf{M}_{p_i}(t)). \end{aligned} \tag{11}$$

The top line of Eq. 11 is the long-term activity cost at isochronal time $t^* = t$, which for real-time planning is the current time of path planning. The cost in the bottom line includes both in-place and static short-term activity. Short-term moving activity in Table 1, which is captured from the on-board robot sensors for local planning, is outside the focus of this paper, but if captured it would be added to Eq. 11.

4 Experiments and Results

4.1 Scope of Experiments

The focus of this paper is on efficient design of an activity filter and application to human-robot interaction. It is important to state what is outside the scope of this paper. Experiments showing the effectiveness of activity filtering with the same filters but not the same efficient architecture have already been described in [20]. Other references describe the performance of activity filtering on a variety of datasets and applications [4]. We do not repeat these. This paper is also not a comparison between pixel-based activity filtering and neural network object detection. The former only detects activity; the latter can detect activity as well but in addition detect higher level features. However, we do show how their hybrid combination can yield both levels of information in an efficient manner.

4.2 Filter Parameters

Filter parameter values are determined by balancing the signal-to-noise ratio for noise conditions learned statistically for each particular deployment. The values described in this paper are for our deployment of robot path planning in a factory setting described in Sect. 4.7. Typical of indoor surveillance cameras, the factory cameras are mounted to view activity from ceiling height and at a perspective such that the activity objects vary in size from about 10×10 pixels to a quarter of the size of a 640×480 size image. Activity detection has close to 100% reliability for this range in our indoor office and factory conditions. While the size range attests to scale robustness of the filter parameter values described in this section, the scope of our experiments here did not extend to poor lighting and outdoor conditions. We refer interested readers to earlier motion detection experiments that included outdoor and low-lighting conditions [20].

The filter F_{L1} in Eq. 4 is designed to reduce "stationary motion noise". This is motion that occurs in-place and continuously such as from rustling tree leaves or a flashing light. Choice of the value has a wide tolerance, the main consid-

eration being that it should not be too short to reduce activity of interest. We choose to remove motion of duration 30 min and longer, so at 30 frames per second, $r = 30$, $T_{L1} = 30 \times 60$, and Eq. 3 yields $\alpha = 0.794$.

For filter F_{L2} in Eq. 5, we choose a low-pass filter to reduce infrequent (shot) noise. To accomplish this, we choose a filtering duration of 10 d (this is 10 samples in isochronal time), so at 1 frame per day, $r = 1$, $T_{L2} = 1 \times 10$, and Eq. 3 yields $\alpha = 0.999957$.

For filter F_{S1} in Eq. 6, we choose a low-pass filter to capture people activity in the same location and eliminate people moving across locations. The discrimination between static and moving activity is somewhat arbitrary, so the filter value choice also has tolerance. We choose a low-pass filter with time constant greater than or equal to 20 s to define this activity, and this activity is updated not at frame rate but at 1/sec, so $r = 1$, $T_{S1} = 20$, and $\alpha = 0.89$.

The band-pass filter of Eq. 7 uses F_{S1} on the low end, which is already specified. On the high end, F_{S2}, it is set to remove high-frequency, infrequent noise. This is the most intolerant of the filter parameters, since this noise has variable periodicity. We use a FIR filter to average activity values over 1 s.

Finally, event detection is based upon the filtered results exceeding the activity average (Fig. 4), which is statistically learned, plus a chosen standard deviation. In contrast to neural network methods where activity detection would automatically learn hundreds or thousands of network parameter values, the use of two intuitive features (density and direction in Eq. 1) incurs far less computation while incorporating statistical learning for robustness. Since for the hybrid approach, the activity filter is positioned as a pre-filter to more expensive neural network object detection, we can choose a conservative standard deviation to allow more activity events, from which the subsequent neural network object detection can eliminate false positives for higher accuracy. The computation savings of this hybrid approach is shown in Sect. 4.6.

4.3 Isochronal Activity

Figure 4 shows two examples of isochronal activity with 1-day periodicity. The left plot is from an office hallway. The data was collected and averaged over 2 years. The same filter parameter values detailed in Sect. 4.2 are used for all applications. It shows an increase of activity starting at 6 am, a lull in mid-day, increase to 4:30 pm, and activity decreasing to 9 pm. The right plot is data from a university hallway showing activity collected and averaged over 1 month of the school term. It shows rises and falls coinciding with hourly class changes. The red vertical lines show times where robot navigation might best be planned (within the work or school day) to avoid high activity periods.

4.4 Cascade and Non-cascade Filter

For our application of using fixed cameras to monitor human and robot spaces, an industrial or business installation may use hundreds of cameras. It is important to limit costs of hardware and computation. We compare the computation

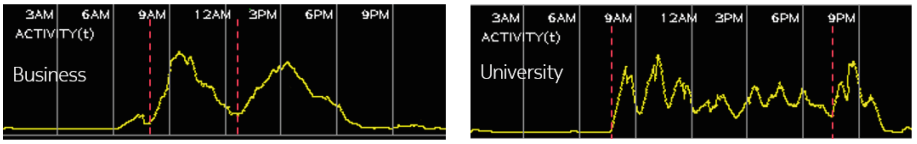


Fig. 4. Isochronal activity plots showing magnitude of hallway activity in a business place (left) and a university (right).

Table 2. Computational costs of non-cascade versus cascade filter.

Cost	Non-Cascade	Cascade	Reduction
Multiplies	5×	4×	20%
Memory	3×	2×	33%

and memory cost of the cascade temporal filter described in Sect. 3.4 and Fig. 1 against a non-cascaded filter that accomplishes the same task.

In Fig. 1, there are 5 filters, where the bandpass filter counts as 2 filters, a combined low- and high-pass filter. So a non-cascaded filter requires 5 filter operations. A cascade filter economizes by using the low-pass filter (F_{S1}) for both the short-term in-place and moving activity operations. Therefore, the cascade filter has an advantage of 4 versus 5 filtering operations.

For memory, both cascade and non-cascade filters require the long-term storage. For sequential filtering, the cascade filter needs only one storage for both short-term in-place and moving activity. However, the non-cascade filter needs storage for both. Therefore, the cascade filter has an advantage of 2 versus 3 motion feature frames. Results of this comparison are shown in Table 2, in which *Multiplies* is a multiple of motion frame filter operations and *Memory* is a multiple of motion frame size.

4.5 Cost of Computation of Activity and Object Detection

Due to the high accuracy of neural network object detection (which we subsequently shorten to *object detection*), this is likely to be the first choice of many practitioners for detecting humans to reduce human-robot interaction. This will indeed perform the task well, but at a relatively high computational cost. In this section, we show the computational cost of activity and object detection separately, and in the following section of a hybrid of both.

Activity detection and object detection are different operations, the latter being much more versatile than the former. By extracting information on number of people, their pose, etc., an object detector can extract much more reliable information than a temporal filter. Our comparison in this section is strictly computational of combinations of solutions. Where low cost of computation is important, the more lightweight activity filter may be all that is needed. Where higher-level information is needed, the activity filter can act as pre-filter to an

Table 3. Comparing methods for detecting activity.

	Single Camera			32-Camera Network			
	GPU	FPS	Power [watts]	Number CPUs	Number GPUs	FPS/ Camera	Power [watts]
F. R-CNN	yes	6.78	135	–	32	6.78	4320
YOLOv3	yes	14.79	153	–	32	14.79	4895
OpenPose	yes	6.2	175	–	32	6.2	5600
tiny YOLO	no	17.78	102	4	0	2.9	520
Activity	**no**	**30+**	**50**	**1**	**0**	**25**	**80**

object detector to form a hybrid solution to reduce overall computation. And where the reliability or additional information of an object detector is always needed, we compare these costs as well.

Table 3 shows computation results of comparing different methods for activity detection. Object detection methods are included in publication order, Faster R-CNN [24], YOLOv3, and Tiny Yolo [23]. We have added OpenPose [5], which finds people as well as their poses, because pose can be useful when working with affordances (as described in Sect. 2). The computational requirements were measured from a testbed 32-camera network viewing hallways and public areas of a building containing offices and laboratories. The computing specifications for processing the video streams are, CPU: AMD Ryzen 5 Pro 2600, 6-core, 8 GB RAM; and GPU: NVIDIA GeForce GTX 1060, 6 GB.

Table 3 shows in general that the pixel-based activity detector is much more computationally efficient than the object detection approaches. These are both based upon a processing rate of each frame, either filtering each frame or object detection on each frame. For a single camera, the activity detector can run at (and above) the rate of a 30 frames per second video feed, whereas the neural methods run at half or less rate. CPU power required for the activity detector is about half of Tiny Yolo on a CPU and about a third of the other object detectors running on a GPU.

In an industrial application, for instance, where there are multiple cameras, the difference is more compelling as seen on the right side of Table 3. On our test machine, we can perform activity detection on up to 32 cameras. Using this as a baseline, we compare for a network of 32 cameras at which activity detection drops to 25 fps. For this workload, Tiny Yolo requires 4 CPUs and the frame rate drops to 6.2 fps. Frame rate for the other methods stays the same as for 1 camera because each of these uses a full GPU per single camera feed. Besides cost of GPUs, energy usage is an important system consideration for real applications. Power consumption is about 5× greater for Tiny Yolo than activity detection and over 50× greater for the other methods. These results support using activity detection alone or as a pre-filter for less frequent object detection as will be discussed in Sect. 4.6.

Table 4. Activity and object detection with an average of 300 events in a workday.

Detection	Single Camera			32-Camera Network		
	Number CPUs	Number GPUs	Energy [w-h]	Number CPUs	Number GPUs	Energy [w-h]
Activity	1	0	500	1	0	800
Hybrid	**1**	**1**	**500.9**	**1**	**1**	**828**
Object	1	1	2030	1	32	49460

4.6 Hybrid Activity Filter and Object Detector

The experimental results of Sect. 4.5 show that performing activity detection is much more efficient than object detection on every frame. But, what if we want more detailed information than the presence or absence of activity? When this is the case, we can employ activity detection as a pre-filter (or gate) to perform or not perform object detection. The argument for a hybrid approach such as this is dependent upon the application and the activity density. For instance, if activity is continuous, we might just as well perform object detection on all frames. If there are periods of inactivity, then use of the activity pre-filter can be more efficient.

We have an example of real data where activity in a business hallway was monitored for 2 years (activity plot shown in Fig. 4). There were, on average, 300 activity events per camera per workday. If an event duration is 10 s, only 8.3% of camera time contains an event.

Using the 32-camera numbers from Table 3, if we performed YOLOv3 on one frame of each event detected by the activity detector, then the extra cost above activity detection is 1 GPU and 28w-h (watt-hours). This is the hybrid approach shown in Table 4. If we were to run continuous YOLOv3 neural network processing to do object detection without a pre-filter, this incurs an extra cost of 32 GPUs and 60× the energy.

4.7 Incorporation into ROS

In practice, we manage our robots on the Robot Operating System (ROS) [1]. A preliminary task in using ROS is to populate a cost map with a floorplan of walls and other static objects. By assigning cost values to (x, y) locations, walls can be designated impenetrable, regions can be marked forbidden, and buffers zones can be placed around objects to help guide robots along safe and efficient paths. Figure 5 shows a ROS cost map of our robot test area. Walls are marked in pink with cyan buffer zones, red marks forbidden zones, and blue marks low-cost areas preferable for robot travel.

We augment the static cost map with dynamic human activity cost determined by temporal filtering. The yellow circle in Fig. 5 indicates a region of human activity as shown by cyan dots. Just as the robot avoids cyan buffer

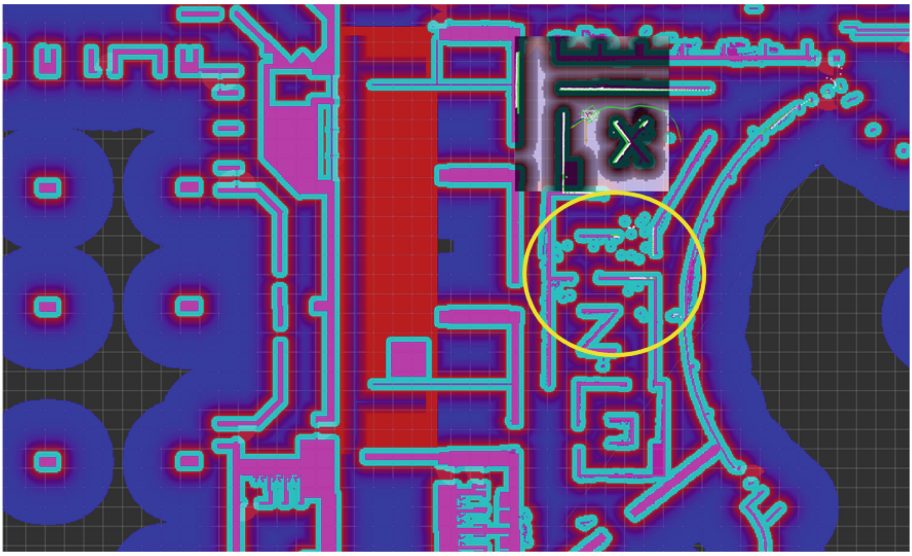


Fig. 5. ROS cost map showing people activity with cyan dots inside yellow circle. (Color figure online)

zones, it will also avoid the cyan activity locations – the difference being that the activity locations can move as they are detected in different locations. We currently assign the same cost to human activity as for static objects. However, it is reasonable, since humans can move, that a lower cost could be assigned to human activity, with the balance between safety and efficiency being a factor in choosing that value.

5 Conclusions

Although mobile robots carry sensors to aid navigation, there are complementary benefits from fixed cameras that view the paths that robots travel. A major benefit is continuous view of an area from which a cycle of activity can be determined. Through knowledge of daily activity patterns, long-term path planning can be performed to avoid areas and times that are crowded, and instead choose paths at off-peak times. We have shown that a cascade filter applied to activity captured in both isochronal and chronological time can efficiently provide activity information for detecting long- and short-term activity. Furthermore, we have shown that a hybrid solution of temporal filtering for event detection, followed by object detection can yield power and cost efficiencies.

References

1. ROS (robot operating system), https://www.ROS.org
2. Arzani, M.M., Fathy, M., Azirani, A.A., Adeli, E.: Switching structured prediction for simple and complex human activity recognition. IEEE Trans. Cybern. **51**(12), 5859–5870 (2021)
3. Belkhouche, F.: Reactive path planning in a dynamic environment. IEEE Trans. Robot. **25**(4), 902–911 (2009)
4. Bendali-Braham, M., Weber, J., Forestier, G., Idoumghar, L., Muller, P.A.: Recent trends in crowd analysis: a review. Mach. Learn. Appl. **4** (2021)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
6. Charalampous, K., Kostavelis, I., Gasteratos, A.: Recent trends in social aware robot navigation: a survey. Robot. Auton. Syst. **93**, 85–104 (2017)
7. El-Shamouty, M., Wu, X., Yang, S., Albus, M., Huber, M.F.: Towards safe human-robot collaboration using deep reinforcement learning. In: 2020 IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 4899–4905 (2020)
8. Fortun, D., Bouthemy, P., Kervrann, C.: Optical flow modeling and computation: a survey. Comput. Vis. Image Underst. **134**, 1–21 (2015)
9. Iaboni, C., Patel, H., Lobo, D., Choi, J.W., Abichandani, P.: Event camera based real-time detection and tracking of indoor ground robots. IEEE Access **9** (2021)
10. Jumel, F., Saraydaryan, J., Simonin, O.: Mapping likelihood of encountering humans: application to path planning in crowded environment. In: 2017 European Conference on Mobile Robots (ECMR), pp. 1–7 (2017)
11. Kavraki, L., Svestka, P., Latombe, J.C., Overmars, M.: Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Trans. Robot. Autom. **12**(4), 566–580 (1996)
12. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 14–29 (2016)
13. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: 2009 IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1446–1453 (2009)
14. Kuderer, M., Kretzschmar, H., Burgard, W.: Teaching mobile robots to cooperatively navigate in populated environments. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3138–3143 (2013)
15. Kumar, S., Savur, C., Sahin, F.: Survey of human-robot collaboration in industrial settings: awareness, intelligence, and compliance. IEEE Trans. Syst. Man Cybern. Syste.**51**(1), 280–297 (2021)
16. Moravec, H., Elfes, A.: High resolution maps from wide angle sonar. In: Proceedings of 1985 IEEE Int. Conference on Robotics and Automation, vol. 2, pp. 116–121 (1985)
17. Mukherjee, D., Gupta, K., Chang, L., Najjaran, H.: A survey of robot learning strategies for human-robot collaboration in industrial settings. Robot. Comput.-Integr. Manuf. **73** (2022)
18. Munguia-Galeano, F., Veeramani, S., et al.: Affordance-based human-robot interaction with reinforcement learning. IEEE Access **11**, 31282–31292 (2023)
19. O'Gorman, L.: Temporal filter parameters for motion pattern maps. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 2612–2617 (2018)

20. O'Gorman, L., Yin, Y., Ho, T.K.: Motion feature filtering for event detection in crowded scenes. Pattern Recogn. Lett. **44**, 80–87 (2014)
21. Okal, B., Arras, K.O.: Learning socially normative robot navigation behaviors with bayesian inverse reinforcement learning. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 2889–2895 (2016)
22. Ravankar, A., Ravankar, A., Hoshino, Y., et al.: Safe mobile robot navigation in human-centered environments using a heat map-based path planner. Artif. Life Robot. **25**, 264–272 (2020)
23. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc. (2015)
25. Saarinen, J., Andreasson, H., Lilienthal, A.J.: Independent markov chain occupancy grid maps for representation of dynamic environment. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3489–3495 (2012)
26. Saleemi, I., Hartung, L., Shah, M.: Scene understanding by statistical modeling of motion patterns. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2069–2076 (2010)
27. Truong, X.T., Ngo, T.D.: Toward socially aware robot navigation in dynamic and crowded environments: a proactive social motion model. IEEE Trans. Autom. Sci. Eng. **14**(4), 1743–1760 (2017)
28. Wada, T., Wang, Z., et al.: Building human motion map for mobile robot in the indoor dynamic environment. In: 2010 IEEE International Conference on Robotics and Biomimetics, pp. 543–548 (2010)
29. Wang, X., Ma, X., Grimson, W.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 539–555 (2009)
30. Wang, Z., Jensfelt, P., Folkesson, J.: Building a human behavior map from local observations. In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 64–70 (2016)
31. Wu, P.T., Yu, C.A., Chan, S.H., Chiang, M.L., Fu, L.C.: Multi-layer environmental affordance map for robust indoor localization, event detection and social friendly navigation. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2945–2950 (2019)
32. Ziebart, B.D., Ratliff, N., et al.: Planning-based prediction for pedestrians. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3931–3936 (2009)

ForceGNN: A Force-Based Hypergraph Neural Network for Multi-agent Pedestrian Trajectory Forecasting

Jiaqian Zhou(✉), Jichao Jiao, and Ning Li

Beijing University of Posts and Telecommunications, Beijing 100876, China
{zhoujiaqian,jichaojiao,Inmmdsy}@bupt.edu.cn

Abstract. Multi-agent trajectory prediction is crucial for many real-world applications. This task faces challenges in effectively capturing individual temporal patterns and complex interactions between intelligent agents. Existing models either solely focus on single agent dynamics or neglect multi-agent collective interactions. To address these difficulties, we propose a graph neural network-based trajectory forecasting approach, named ForceGNN. Specifically, it combines Transfomer to extract temporal patterns for each agent and utilizes a multi-scale hypergraph neural network to simulate complex crowd interactions. Meanwhile, we incorporate a social force model with strong inductive biases, abstracting pedestrian interactions as social forces, thereby achieving more efficient trajectory prediction. We compare ForceGNN with state-of-the-art deep learning methods on the UCY/ETH and SDD datasets, achieving optimal performance on both.

Keywords: Pedestrian Trajectory Forecasting · Hypergraph Neural Network · Social Force Model

1 Introduction

Accurate human trajectory prediction is indispensable for autonomous systems operating in dynamic real-world environments. Applications such as self-driving vehicles, service robots, and intelligent surveillance all rely on the ability to forecast pedestrian behavior and movements [1]. As human spaces become more crowded and complex, modeling the intricate interactions between people and their surroundings remains an open challenge [2]. Recent trajectory prediction methods have modeled pedestrian paths based on individual goals, static obstacles, and interactions among people [3–7]. These approaches include: early rule-based social force models that represent interactions using differential equations [3,4]; data-driven models where recurrent neural networks capture pedestrian sequence dynamics and graph neural networks or attention mechanisms extract complex inter-personal relationships [6,7]. More recently, goal-driven models have been explored that first predict long-term targets to aid future behavior

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 391–405, 2025.
https://doi.org/10.1007/978-3-031-78341-8_25

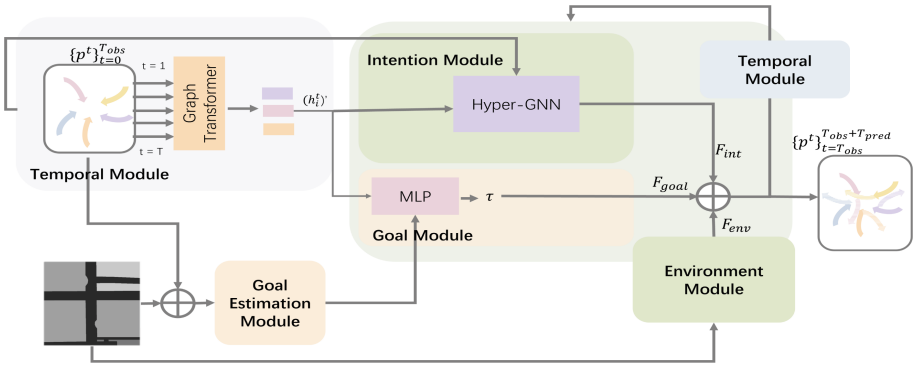


Fig. 1. Model Architecture. Our prediction method consists of three main components: Temporal module, Group Interaction module, and environment module. In addition, there is a auxiliary modules: goal estimation module.

forecasting [8]. However, these methods either disregard individual movement patterns or complex interaction intents. To this end, we propose a goal-driven graph neural network based on social force models called ForceGNN.

Specially, Sequential learning with neural networks tends to forget earlier concepts [9]. To overcome this forgetting, we apply Transformer to selectively learn important model parameters, capturing temporal patterns, named Graph Transformer, as shown in Fig. 1. Graph-based modeling and attention mechanisms have greatly advanced trajectory forecasting by modeling pairwise interaction intensities [6,10]. However, they are limited in reasoning about the complex interaction intensities between agents in dense scenes. Pairwise modeling fails to efficiently capture such intricate relationships in crowded spaces. To enable more comprehensive modeling of interactions, we construct a multi-scale hypergraph to simulate the complex interaction intensities. The hypergraph encodes hierarchical interactions at different scales, from individual goals to group behaviors.

Meanwhile, deep learning methods face challenges in interpretability and efficiency. Therefore, we integrate a rule-based social force model [11], which uses sequential motion information as input. Then a goal module is modeled as the attractive force of the goal on the agent. Meanwhile, we learn the group interactions force between agents adaptively through a hypergraph neural network, named Hyper-GNN, as shown in Fig. 1. Furthermore, we use enviornment module to model the repulsive force of static obstacles in the scene on the agent, avoiding collisions with scenes. Through this decomposition of attractive and repulsive forces at individual, group, and scene levels, our model can realistically simulate agent behavior in complex crowded scenes. Additionally, To estimate the goal on the agent within a reasonable inference time, we apply a goal estimation module that connects historical trajectories with semantic scene information. Then the information is fed into a U-Net structure [12] to predict potential goals. The U-Net structure leverages an encoder-decoder architecture to fuse spatial scene and temporal motion clues for goal prediction. By com-

bining bottom-up perceptual features and top-down contextual reasoning, the model can learn robust representations of potential goals. Overall, this work has three main contributions:

(1) We propose a novel hypergraph neural network prediction framework based on social force model, named ForceGNN.
(2) Our model includes not only individual temporal patterns, but also the interaction between agents and static obstacles, as well as the group-level interaction between agents.
(3) ForceGNN achieves state-of-the-art performance on multiple public human trajectory dataset.

2 Related Work

2.1 Social Interaction Analysis

Essentially, studies on crowd interaction models can be divided into either physics-based methods and learning-based methods. The first group of methods such as the social force model uses external forces to model the social Interaction of pedestrians [11]. The model and its extensions [13,14] perform well in certain cases while poorly on complex interactions in crowded scenarios. Learning-based models enable more flexibility and capacity to capture underlying interactive patterns [7]. However, these methods usually ignore physical feasibility constraints. Several studies attempt to combine physics with deep learning for trajectory prediction [3]. However, these methods ignore the group behavior's influence on pedestrians. In this work we combine graph neural network and social force model to model social interactions.

2.2 Graph Neural Network

Graph Neural Networks have been extensively used in tasks such as traffic prediction and trajectory prediction [7,10,15–17]. For simple graphs, prior trajectory prediction methods based on undirected graphs assign the same weight for each pair of nodes [15,16]. Meanwhile, EvolveGraph [18] and HEAT [10], can flexibly handle edge and heterogeneity features for interaction modeling in multi-agent trajectory prediction. However, these graph neural networks can only model the pair-wise interaction in a graph but ignoring the group behavior's influence. The multiscale hypergraphs can capture group behaviors at different sizes, thereby modeling agent interactions more comprehensively [7]. In this paper, we use a hypergraph neural network to adaptively capture the group interaction between pedestrians.

2.3 Trajectory Prediction Methods

The problem of predicting pedestrian trajectories has been extensively explored in prior research [19]. Traditional approaches relied on handcrafted rules and

energy potentials, which struggled to accommodate intricate interactions [11]. To overcome these limitations, deep learning techniques have been harnessed for trajectory prediction [3,20]. Recurrent Neural Networks have been employed to model temporal dependencies [6], and subsequently, other advanced neural network architectures have been integrated into trajectory prediction. Notable examples include the use of Generative Adversarial Networks [21], conditional Variational Autoencoders [5,22], Convolutional Neural Networks [4], Transformers [23], and diffusion models [24]. In the pursuit of accurately capturing spatial features and interactions among pedestrians, graph neural networks have been introduced to reason and predict future trajectories [16,17,25]. In contrast to prevailing deep learning methodologies, our method not only incorporates pedestrian prior knowledge, but also achieves superior overall performance.

3 Method

In this section, we will introduce our proposed network ForceGNN in detail. We first formalize the multi-agent trajectory prediction problem as follows: Spatial coordinates of N pedestrians is denoted as $p^t = \{(x_i^t, y_i^t)\}$, $i \in \{1, 2, \ldots, N\}$. Given the segmentation map S of each scenario I and spatial coordinates of N pedestrians observed within a time period T_{obs}, $\{p^t\}_{t=0}^{T_{obs}}$, ForceGNN aims to predict the most likely trajectories of these pedestrians for the next T_{pred} time steps, $\{p^t\}_{t=T_{obs}+1}^{T_{obs}+T_{pred}}$.

As shown in Fig. 1, our framework is comprised of five modules. In particular, we first employ a temporal module by Graph-Transformer to capture the temporal patterns of pedestrians. Then, the temporal patterns along with estimation goals produced by a goal estimation model are input into a goal module to capture the influence of the goals. Simultaneously, we merge the temporal patterns with the spatial coordinates of pedestrians and feed it into a group interaction module. Additionally, an environment module is utilized to perceive interactions between pedestrians and their surroundings. Eventually, the network merges insights from the environment, individual goals, and social interactions to facilitate accurate trajectory prediction.

3.1 Temporal Module

As shown in Fig. 2, we first construct a graph G to model the dependencies in the time steps. For each agent i, $G_i = \left(V_i, {E_i}^{(t)}\right)$ contains a set of nodes $V_i = \{v_i^{(1)}, v_i^{(2)}, \cdots, v_i^{(T)}\}$, where $v_i^{(t)}$ is represented by $[p^t, (p^t)']^T$ which denote the position and velocity. Generally speaking, we first uses each node v_i^t as input and output a new embedding feature h_i^t, which is formulated as:

$$h_i^t = f_v(v_i^t) \tag{1}$$

where f_v denotes a multi-layer perceptron(MLP). Then we adopt Transformer to capture the temporal patterns of each agent, which helps overcome the issue

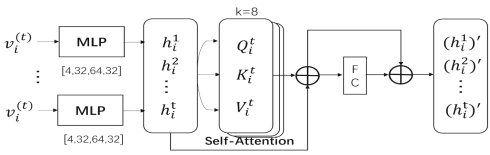


Fig. 2. Graph Transformer. The figures enclosed within square brackets represent the count and dimensionality of layers within each constituent element. k denotes the number of attention heads.

of network forgetting. For i-th pedestrian, The self-attention block first learns the query matrices Q_i^t, key matrix K_i^t and the value matrix V_i^t, where f_Q, f_K, f_V are learnable functions by three MLPs.

$$Q_i^t = f_Q\left(\left\{h_i^j\right\}_{j=1}^{t}\right), \quad K_i^t = f_K\left(\left\{h_i^j\right\}_{j=1}^{t}\right), \quad V_i^t = f_V\left(\left\{h_i^j\right\}_{j=1}^{t}\right) \tag{2}$$

For i-th pedestrian, the temporal pattern at current time $\left(h_i^t\right)'$ is formulated as:

$$\begin{aligned} \left(h_i^t\right)' &= f_h\left(\left[\text{ head }_j^t\right]_{j=1}^{k}\right), \\ \text{where } head_j^t &= \text{Att}_j\left(Q_i^t, K_i^t, V_i^t\right) \end{aligned} \tag{3}$$

$$\text{Att}_j\left(Q_i^t, K_i^t, V_i^t\right) = \frac{\text{Softmax}\left(Q_i^t K_i^{t\ \mathrm{T}}\right)}{\sqrt{d_k}} V_i^t \tag{4}$$

The module uses a MLP f_h to merge information from k attention heads. Each head Att_j focuses on different parts of the inputs. The final embedding combines two skip connections and a last fully connected layer (FC layer).

3.2 Goal Module

A pedestrian's trajectory is directly related to their expected destination. Specifically, we concatenate a processed trajectory heatmap and scene segmentation map S with dimension of $H \cdot W \cdot K_c$ to obtain an input of dimension $H \cdot W \cdot (K_c + M + 1)$ into the U-Net structure. H and W are the height and width of S, K_c is the number of classes for segmentation, and M is the length of the time steps. The network finally outputs a probability distribution map over possible destinations. We train the goal estimation model by minimizing the sum of binary cross-entropy loss and the Kullback-Leibler divergence between the predicted destination and ground truth. During testing, we employ the test-time sampling technique introduced in [24] to sample destinations $X_{i_{dest}}^t$ for better performance, rather than selecting the maximum probability location.

The effect of the destinations on the pedestrian can be abstracted as a target attraction force. Then, by applying an improved social force model [5], the influence of the destination on the pedestrian is abstracted as a target attraction force, which is formulated as:

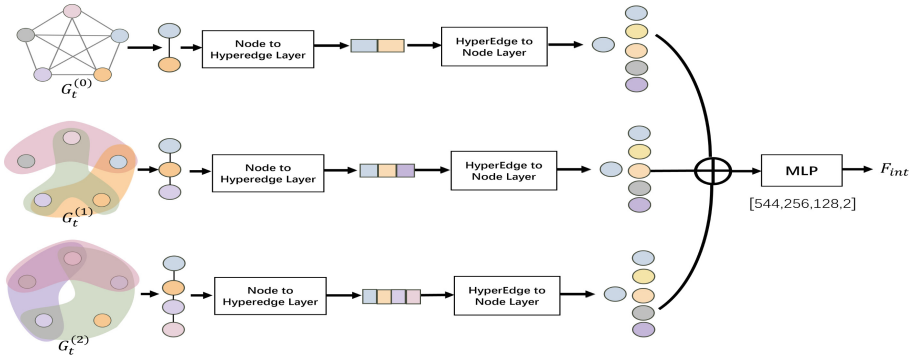


Fig. 3. Group Interaction Module. The graph interaction module is composed of the construction of multi-scale hypergraph $G_t^{(s)}$ and hypergraph neural message passing. Hypergraph neural message passing continuously gets the embedding of agents and interactions through node to hyperedge layer and hyperedge to node layer.

$$F_{goal} = \frac{1}{\tau}(\nu_0^t \boldsymbol{e} - \boldsymbol{\nu}_i^t). \tag{5}$$

$$\tau = f_{goal}(f_{dest}(X^t_{i_{dest}}), (h_i^t)') \tag{6}$$

where $\nu_0^t = \frac{||X^t_{i_{dest}} - v_i^t||}{(T-t)\Delta t}$ and $\boldsymbol{e} = \frac{X^t_{i_{dest}} - v_i^t}{||X^t_{i_{dest}} - v_i^t||}$represent the magnitude and direction of the desired velocity respectively. $\boldsymbol{\nu}_i^t$ denotes the current velocity, and τ indicates the expected time to reach the ideal velocity. f_{goal}, f_{dest} are two functions by MLPs.

3.3 Group Interaction Module

Pedestrians often exist in dynamic scenes and inevitably have complex interactions with surrounding pedestrians. For example, pedestrians always try to maintain a certain distance from others to avoid collisions. Such interactions can be modeled in the form of repulsive forces. However, pedestrian interactions are complex, with many pedestrians walking in groups or having common destinations. Therefore, we incorporate a hypergraph neural network to model interactions between pedestrians, enabling adaptive modeling of interactions within pedestrian crowds. As shown in Fig. 3, we first construct a multiscale hypergraph G to reflect the interactions between different pedestrian groups. In other words, the multiscale hypergraph $G_t = \{G_t^{(0)}, G_t^{(1)}, ...G_t^{(s)}\}$ represents multiple hypergraphs at different scales s at time t. At each scale s, the hypergraph $G_t^{(s)} = \left(V_t, E_t^{(s)}\right)$ contains a set of hyperedges $E_t^{(s)} = \{e_1^{(s)}, e_2^{(s)}, \cdots, e_{M_s}^{(s)}\}$, $V_t = \{v_1^t, v_2^t, \cdots, v_N^t\}$, where $v_i^t = (x_i^t, y_i^t)$. Each hyperedge connects a group of agents, capturing their common relations. Meanwhile, The elements of the

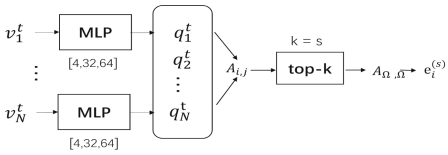


Fig. 4. Spatial Hypergraph Construction. Top-k denotes that the top k values in $A_{i,j}$ are taken out each time to form a superedge, and the rest is set to 0.

incidence matrix $H^{(s)}$ is given by: $H^{(s)}_{i,j} = 1$ if the ith node is included in the jth hyperedge, otherwise $H^{(s)}_{i,j} = 0$.

Spatial Hypergraph Construction. Specifically, we map the trajectories into high-dimensional feature vectors to measure the correlations between spatial edges. Agents with high correlations are grouped together to construct the hyperedges. As shown in Fig. 4, we use the following notation to describe the affinity matrix of the hypergraph:

$$A_{i,j} = q_i^\top q_j / (||q_i||_2 ||q_j||_2) \tag{7}$$

$$q_i = f_q(v_i) \tag{8}$$

which denotes the relation weight between the ith agent and the jth agent, reflects the correlation between the two agents. We construct hyperedges at different scales based on the affinity matrix.

$$e_i^{(s)} = \underset{\Omega \subseteq V}{argmax} \, \|A_{\Omega,\Omega}\|_{1,1} \,, \tag{9}$$

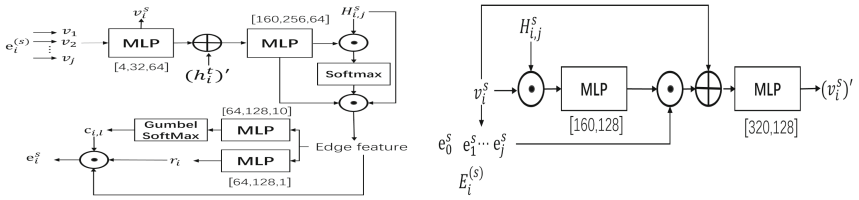
$$s.t. |\Omega| = s, v_i \in \Omega, i = 1, ..., N, s \neq 0 \tag{10}$$

where $\|\cdot\|_{1,1}$ denotes the sum of the absolute values of all elements.

Hypergraph Neural Message Passing. For each scale s, we iteratively pass information from nodes to hyperedges and from hyperedges back to nodes. As shown in Fig. 5(a), in the node-to-hyperedge stage, for the ith hyperedge $e_i^{(s)}$, its interaction embedding is computed as:

$$e_i^s = r_i \sum_{\ell=1}^{L} c_{i,l} f_{agg\_v}\Big(\sum_{v_j \in e_i^{(s)}} v_j \Big),$$

where r_i represents the intensity of node interaction and $c_{i,l}$ falls within the range [0, 1], signifying the probability of the i-th node's interaction being of category l. Furthermore, for each possible category, a learnable interaction intensity function implemented by a MLP is assigned, named f_{agg_v}. Each of these elements is designed to be trainable within an end-to-end framework. As



(a) Node to Hyperedge Layer. (b) Hyperedge to Node Layer.

Fig. 5. Hypergraph neural message passing. Hypergraph neural message passing continuously gets the embedding of agents and interactions through node to hyperedge layer and hyperedge to node layer.

shown in Fig. 5(b), in the hyperedge-to-node stage, let $E_i^{(s)} = \{e_j^{(s)} | v_i \in e_j^{(s)}\}$ be the set of hyperedges associated with the ith node v_i^s. The embedding of $(v_i^s)'$ is updated as:

$$(v_i^s)' = f_{agg\_e}\big(\big[v_i^s, \sum_{e_j \in E_i} e_j^s\big]\big)$$

where the notation $[\cdot, \cdot]$ signifies the amalgamation of embeddings from a single node and its connected hyperedges, f_{agg_e} denotes a MLP. Then the agent embeddings from all scales v_i^s are ultimately merged to form a cohesive representation of the agents' embedding as follows:

$$(v_i)' = [(v_i^0)', (v_i^1)', \cdots, (v_i^s)']$$

Group Interaction Force. The multi-scale hypergraph features are then input into the decoder to obtain the final output predictions F_{int}, where f_{dec} denotes by MLP:

$$F_{int} = f_{dec}((v_i)') \tag{11}$$

3.4 Environment Module

In addition to their pedestrians' motion and other pedestrians, pedestrians are also influenced by the environment. We model the influence F_{env} as:

$$F_{env} = \frac{r_{env}}{||v_i^t - v_{obs}||}\Big(\frac{v_n^t - v_{obs}}{||v_i^t - v_{obs}||}\Big) \tag{12}$$

where v_{obs} are the center points of static obstacles within the visible scope r_{scope} and r_{env} is a learnable parameter.

3.5 Loss Functions and Training

This final output aggregates the outputs from the three modules:

$$\ddot{v}_i^t = F_{goal} + F_{int} + F_{env} \tag{13}$$

where $\ddot{v}_i^t$ represents the second order derivative of v_i^t with respect to t. The final loss is computed as:

$$L_{traj} = \frac{1}{N(T-M)} \sum_{n=1}^{N} \sum_{t=M+1}^{T} ||v_n^t - \bar{v}_n^t||_2^2 \tag{14}$$

where N is the total number of samples, M is the history length, and T is the total trajectory length. L_{traj} minimizes the difference between the predicted positions and the ground truth.

4 Experiments

In this section, we discuss the datasets used to evaluate ForceGNN performance and the hyperparameters used in the experiments. We also discuss the metrics for evaluating trajectory predictions from the models and baseline methods.

4.1 Datasets

We conduct experiments on publicly available SDD [26], ETH [27] and UCY [28] datasets. The ETH subset includes ETH and HOTEL scenes, while the UCY subset consists of ZARA1, ZARA2 and UNIV scenes. We adopt a leave-one-out testing strategy, where models are trained on four subsets and tested on the left out subset. Trajectories are sampled at 2.5 frames per second, with 8 time steps (equivalent to 3.2 s) observed and next 12 time steps (4.8 s) predicted.

4.2 Implementation Details

The entire network was optimized using the Adam optimizer [29] with a learning rate of 1×10^{-3} and a batch size of 32. During training, mean squared error (MSE) loss function was employed to facilitate accurate position prediction. Transformer was utilized to capture temporal patterns, with a head size of 8 for the multi-head attention. In the group interaction module, the scale size of the hypergraph was [2, 3]. Regarding the environment module, the SDD scene was categorized into three segments - navigable, non-navigable, and weak obstacle areas. The ETH/UCY scene was divided into navigable and non-navigable regions.

Table 1. Results of State-of-the-Art and Our Proposed Model on ETH/UCY Dataset. Bold/underlined fonts represent the best/second-best result.

Datasets	SFM-NN	Y-net	Grouptron	ForceFormer	Goal-SAR	NSP-SFM	Ours
	ADE/FDE	ADE/FDE	ADE/FDE	ADE/FDE	ADE/FDE	ADE/FDE	ADE/FDE
ETH	0.36/0.82	0.28/0.33	0.7/1.56	0.36/0.52	0.28/0.38	**0.25**/**0.24**	0.28/0.29
HOTEL	0.68/1.63	0.10/0.14	0.21/0.46	**0.09**/0.14	0.12/0.17	**0.09**/0.13	**0.09**/**0.09**
UNIV	0.46/1.12	0.24/0.41	0.38/0.97	**0.21**/0.42	0.25/0.43	**0.21**/**0.38**	0.23/**0.38**
ZARA1	0.35/0.85	0.17/0.27	0.30/0.76	**0.15**/**0.22**	0.17/0.26	0.16/0.27	0.16/0.24
ZARA2	0.38/0.95	0.13/0.22	0.22/0.56	**0.12**/**0.20**	0.15/0.22	**0.12**/**0.20**	0.13/0.22
AVG	0.45/1.07	**0.18**/0.27	0.36/0.86	0.19/0.30	0.19/0.29	**0.17**/**0.24**	**0.17**/**0.24**

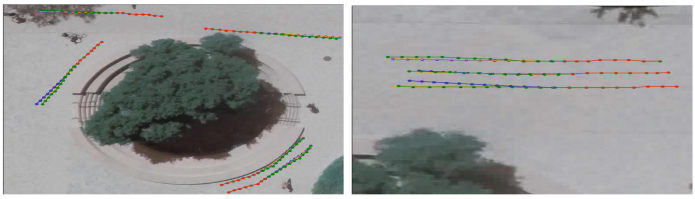
4.3 Metrics

We adopt average displacement error (ADE) and final displacement error (FDE) as evaluation metrics. For multi-modal future predictions, ADE calculates the minimum over K predicted trajectories of the average $L2$ distance from the ground truth trajectory at each time step, while FDE takes the minimum over the final time step. The multi-modal prediction size K is set to 20 in our experiments.

4.4 Baselines

We compare our proposed method ForceGNN with existing approaches based on combined physical models and neural networks, graph neural networks, and goal-driven models. Moreover, to demonstrate the performance of our model, we also make comparisons with several state-of-the-art pedestrian trajectory prediction baselines from recent years.

SFM-NN [3] is a dynamic equation prediction approach integrating social force model and neural networks. Grouptron [32] performs trajectory prediction on spatio-temporal dynamic graphs at individual, group, and scene levels using graph neural networks. GroupNet [7] is a multi-agent prediction method based on multiscale hypergraphs. Y-net [12] utilizes scene semantic information for multimodal modeling of goals and paths for trajectory prediction. ForceFormer [23] incorporates social forces into a stochastic generator backbone based on Transformer, uses destination-driven modeling to simulate goal-driven pedestrian interactions, and models inter-pedestrian collision avoidance as repulsive forces. Goal-SAR [8] proposes a lightweight multi-head attention recurrent backbone, as well as a goal estimation module similar to ForceFormer and ForceGNN. LED [24] is a diffusion model based trajectory prediction approach. TDOR [33] is an end-to-end interpretable trajectory prediction framework based on inverse reinforcement learning. NSP-SFM [20] predicts trajectories by utilizing social scene features and physical dynamics to achieve multi-behavioral group trajectory prediction.



(a) (b)

Fig. 6. Trajectory Prediction in Low crowd density scenes. The observed trajectories are shown in red, the ground truth future trajectories are in green, F_{goal}(full) are in blue, and ForceGNN are in orange. (Color figure online)

Table 2. Results of State-of-the-Art and Our Proposed Model on SDD Dataset. Bold/underlined fonts represent the best/second-best result.

Datasets	Y-net	Goal-SAR	LED	GroupNet	TDOR	NSP-SFM	Ours
	ADE/FDE	ADE/FDE	ADE/FDE	ADE/FDE	ADE/FDE	ADE/FDE	ADE/FDE
SDD	7.85/11.85	7.75/11.83	8.48/11.66	9.31/16.11	6.77/**10.46**	<u>6.52</u>/10.61	**6.49**/<u>10.58</u>

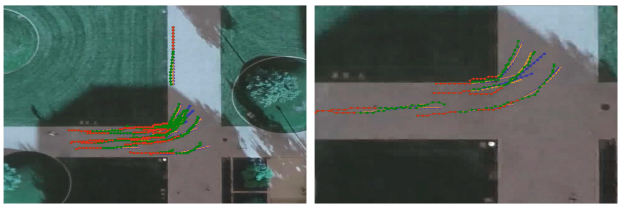
5 Results

5.1 Quantitative Analysis

In Table 1, we compare our method with the current state-of-the-art approaches. Firstly, compared to the current best-performing method, NSP-SFM, ForceGNN demonstrates comparable performance. Additionally, our proposed model, ForceGNN, outperforms the baseline model, ForceFormer, on all subsets of the ETH dataset and achieves comparable results on the UCY dataset.

For the SDD dataset, whose metrics are reported in pixels, we compare our proposed method ForceGNN with current state-of-the-art methods in Table 2. Compared to the Y-net baseline, ForceGNN provides significant improvements, reducing FDE by 17.1% and ADE by 8.5%. On the SDD dataset, ForceGNN outperforms the current best method NSP-SFM in terms of ADE, reducing ADE by 0.5%. The pixel-based metrics show that ForceGNN advances pedestrian path prediction on this challenging dataset.

Our model demonstrates strong performance in both dense crowded scenarios and sparse scenarios in the UCY/ETH dataset and SDD dataset, as shown in Fig. 6 and Fig. 7. This improvement is attributed to the hypergraph neural networks' capability in handling complex interactions and the social force model's ability to better incorporate strong inductive biases. The hypergraph networks can capture intricate relationships between agents, while the physics-based social force formulation encodes useful domain knowledge about pedestrian dynamics. Together, they enable robust trajectory forecasting across varied crowds, whether sparse or congested. Our experiments validate that this combination leads to superior multi-agent modeling compared to prior approaches on the diverse UCY/ETH and SDD benchmark.



(a) (b)

Fig. 7. Trajectory Prediction in Crowded scenes. The observed trajectories are shown in red, the ground truth future trajectories are in green, F_{goal}(full) are in blue, and ForceGNN are in orange. (Color figure online)

Table 3. Results of different modules on SDD dataset.

SDD	F_{goal}(LSTM)	F_{goal}(Transformer)	F_{goal}(with env)	ForceGNN
ADE	6.59	6.55	6.51	**6.49**
FDE	10.68	10.67	10.61	**10.58**

5.2 Ablation Study

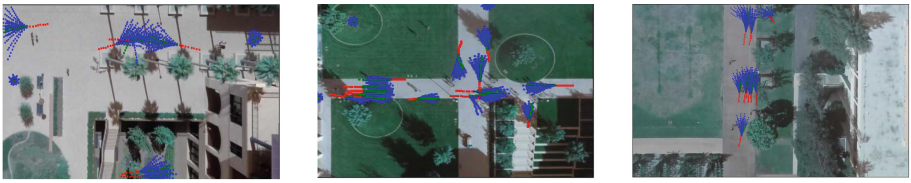
In order to analyze the contributions of different modules in ForceGNN, we conduct an ablation study on three model variants using the SDD dataset: F_{goal} (LSTM) uses only Long Short-Term Memory (LSTM) for modeling temporal patterns; F_{goal} (Transformer) which incorporates temporal patterns into the goal module via Graph-Transformer; F_{goal} (with env) integrates the environment module into the network on the basis of Transformer, as well as the complete ForceGNN model. The results are presented in Table 3.

Notably, F_{goal} (LSTM) alone already achieves reasonable performance by only considering individual dynamics, as illustrated in Fig. 6(a).F_{goal} (Transformer) further enhances the performance by mitigating neural network forgetting and enabling adaptive learning of agent trajectories. Finally, the full model additionally captures complex collective interactions among multiple agents through the integration of group interaction modeling. As observed in Fig. 6(b), the full model demonstrates improved capability in capturing agents' interactions.

Moreover, in crowded complex scenes such as Fig. 7, the output of F_{goal} (Transformer) better approximates the ground truth, validating the efficacy of the group interaction module for multi-agent modeling.

5.3 Qualitative Analysis

In conclusion, our ablation study verifies the role of each component of ForceGNN for trajectory forecasting, including inherent dynamics modeling, temporal pattern extraction, and multi-agent interaction reasoning. As shown in Fig. 7, Fig. 7(a) is the entire pedestrian trajectory map in the complex scene,



(a) A simple scene (b) A very crowded scene (c) A crowded scene

Fig. 8. Qualitative Trajectory Forecasting Results. We visualize the multi-agent trajectory predictions across three distinct scenarios from the SDD test set. The observed trajectories are shown in red, the ground truth future trajectories are in green, and our multimodal trajectory predictions are in blue.

which can be roughly seen that the yellow line (ForceGNN) is closer to the green groundtruth line and Fig. 7 (b) is the result of sampling Fig. 7 (a), which can be clearly seen that the yellow line is closer to the true value. In summary, our ablation study validates the role of each component of ForceGNN in trajectory prediction, including intrinsic dynamics modeling, motion mode extraction, and multi-agent interactive inference. In Fig. 8, we demonstrate the qualitative results of multi-agent trajectory prediction on the SDD dataset. It is observed that ForceGNN is capable of jointly predicting trajectories for all agents within a given scene, achieving relatively good performance in both sparse (Fig. 8(a)) and crowded scenarios (Fig. 8(b) and Fig. 8(c)).

In summary, ForceGNN demonstrates strengths in joint trajectory forecasting for multiple agents, while there is still room for improvement regarding handling abrupt changes. Specifically, when there are more surrounding pedestrians, ForceGNN can adjust the motion direction and speed through the group interaction module. However, we also notice the failure in capturing some sudden changes accurately in the prediction, such as the case shown in Fig. 8a. This could be attributed to the lack of considering sufficient uncertainty for some abrupt cases.

6 Conclusions

In this work, we propose ForceGNN, a novel framework that integrates hypergraph neural networks and social force models for multi-agent trajectory prediction. Compared to existing approaches using either hypergraph neural networks and social force incorporation, our model achieves significant improvements on the UCY/ETH and SDD benchmarks, even reaching state-of-the-art performance. However, there remains ample room for improvement. In the future, we plan to accommodate more extreme cases such as sudden trajectory changes by incorporating additional uncertainty and enriching the group interaction module. Meawhile, we will extend the network to systems with higher crowd density interactions and explore arbitrary observation length trajectory forecasting.

References

1. Rudenko, A., et al.: Human motion trajectory prediction: a survey. Int. J. Robot. Res. **39**(8), 895–935 (2020)
2. Golchoubian, M., et al.: Pedestrian trajectory prediction in pedestrian-vehicle mixed environments: a systematic review. IEEE Trans. Intell. Transp. Syst. (2023)
3. Antonucci, A., Gennaro, M., Rao, S., Abbattista, G.: Generating reliable and efficient predictions of human motion: A promising encounter between physics and neural networks. arXiv preprint arXiv:2006.08429 (2020)
4. Xu, C., et al.: EqMotion: equivariant multi-agent motion prediction with invariant interaction reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
5. Xu, P., Hayet, J.-B., Karamouzas, I.: SocialVAE: human trajectory prediction using timewise latents. In: European Conference on Computer Vision. Springer, Cham (2022)
6. Bisagno, N., Zhang, B., Conci, N.: Group LSTM: group trajectory prediction in crowded scenarios. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
7. Xu, C., et al.: roupNet: multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
8. Chiara, L.F., et al.: Goal-driven self-attentive recurrent networks for trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
9. Lesort, T., et al.: Continual learning for robotics: definition, framework, learning strategies, opportunities and challenges. Inf. Fusion **58**, 52–68 (2020)
10. Mo, X., et al.: Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network. IEEE Trans. Intell. Transp. Syst. **23**(7), 9554–9567 (2022)
11. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**, 4282 (1995)
12. Mangalam, K., et al.: From goals, waypoints & paths to long term human trajectory forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
13. Pascucci, F., Rinke, N., Schiermeyer, C., Berkhahn, V., Friedrich, B.: Should i stay or should i go? A discrete choice model for pedestrian-vehicle conflicts in shared space. Technical report (2018)
14. Ahmed, S., Johora, F.T., Müller, J.P.: Investigating the role of pedestrian groups in shared spaces through simulation modeling. In: International Workshop on Simulation Science, pp. 52–69. Springer (2019)
15. Shi, L., et al.: SGCN: sparse graph convolution network for pedestrian trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
16. Mohamed, A., et al: Social-STGCNN: a social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
17. Alinezhad Noghre, G., et al.: Pishgu: universal path prediction network architecture for real-time cyber-physical edge systems. In: Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023) (2023)

18. Li, J., et al.: EvolveGraph: multi-agent trajectory prediction with dynamic relational reasoning. Adv. Neural. Inf. Process. Syst. **33**, 19783–19794 (2020)
19. Huang, R., et al.: Multimodal trajectory prediction: a survey. arXiv preprint arXiv:2302.10463 (2023)
20. Yue, J., Manocha, D., Wang, H.: Human trajectory prediction via neural social physics. In: European Conference on Computer Vision. Springer, Cham (2022)
21. Mangalam, K., et al.: It is not the journey but the destination: endpoint conditioned trajectory prediction. In: ECCV 2020, Part II, vol. 16, Springer, Cham (2020)
22. Ivanovic, B., Pavone, M.: The trajectron: probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
23. Zhang, W., et al.: ForceFormer: exploring social force and transformer for pedestrian trajectory prediction. arXiv preprint arXiv:2302.07583 (2023)
24. Mao, W., et al.: Leapfrog diffusion model for stochastic trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
25. Yang, B., et al.: A novel graph-based trajectory predictor with pseudo-oracle. IEEE Trans. Neural Netw. Learn. Syst. **33**(12), 7064–7078 (2021)
26. Robicquet, A., et al.: Learning social etiquette: human trajectory understanding in crowded scenes. In: ECCV 2016, Part VIII, vol. 14, Springer, Cham (2016)
27. Pellegrini, S., Ess, A., Van Gool, L.: Improving data association by joint modeling of pedestrian trajectories and groupings. In: ECCV 2010, Part I, vol. 11, Springer, Heidelberg (2010)
28. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Computer Graphics Forum, vol. 26, no. 3. Blackwell Publishing Ltd., Oxford (2007)
29. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
30. Graves, A., Graves, A.: Long short-term memory. In: Supervised Sequence Labelling with Recurrent Neural Networks, pp. 37–45 (2012)
31. Zhang, Y., et al.: D2-TPred: discontinuous dependency for trajectory prediction under traffic lights. In: European Conference on Computer Vision. Springer, Cham (2022). In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
32. Zhou, R., et al.: Grouptron: dynamic multi-scale graph convolutional networks for group-aware dense crowd trajectory forecasting. In: 2022 International Conference on Robotics and Automation (ICRA). IEEE (2022)
33. Guo, K., Liu, W., Pan, J.: End-to-end trajectory distribution prediction based on occupancy grid maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

NeuralMultiling: A Novel Neural Architecture Search for Smartphone Based Multilingual Speaker Verification

P. N. Aravinda Reddy[1(✉)], Raghavendra Ramachandra[2], K. Sreenivasa Rao[1], and Pabitra Mitra[1]

[1] Indian Institute of Technology Kharagpur, Kharagpur, India
aravindareddy.27@iitkgp.ac.in
[2] Norwegian University of Science and Technology (NTNU), Gjøvik, Norway

Abstract. Multilingual speaker verification introduces the challenge of verifying a speaker in multiple languages. Existing systems were built using i-vector/x-vector approaches along with Bi-LSTMs, which were trained to discriminate speakers, irrespective of the language. Instead of exploring the design space manually, we propose a neural architecture search for multilingual speaker verification suitable for mobile devices, called **NeuralMultiling**. First, our algorithm searches for an optimal operational combination of neural cells with different architectures for normal cells and reduction cells and then derives a CNN model by stacking neural cells. Using the derived architecture, we performed two different studies:1) language agnostic condition and 2) interoperability between languages and devices on the publicly available Multilingual Audio-Visual Smartphone (MAVS) dataset. The experimental results suggest that the derived architecture significantly outperforms the existing Autospeech method by a 5–6% reduction in the Equal Error Rate (EER) with fewer model parameters.

Keywords: Biometrics · Multilingual speaker verification · Neural architecture search · Mobile devices · Light weight models

1 Introduction

Biometric-based secure verification is widely deployed in many applications, such as door locks, security devices, home automation, IoT, smart speakers, game consoles, border control, smartphone unlocking, banking, and financial transactions. Over the years, the evolution of smartphones has enabled the biometric-based secure verification of several financial applications, including banking transactions. Biometric verification of smartphones can be achieved using physiological

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78341-8_26.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 406–423, 2025.
https://doi.org/10.1007/978-3-031-78341-8_26

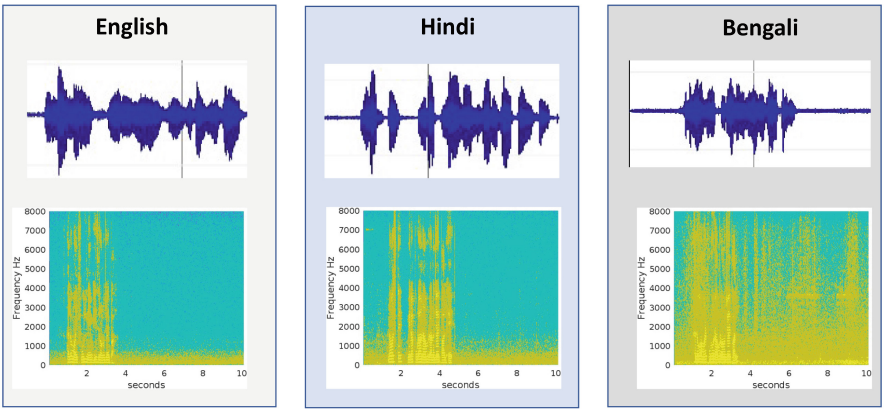


Fig. 1. Illustration of speech signal and corresponding spectrogram of the different languages uttered by the same subject

and behavioral biometrics. The most commonly used biometric characteristics in smartphone verification include faces [1], irises or eyes [2], fingerphotos [3] and voice [4]. Each biometric characteristic has its own advantages and disadvantages in terms of usability, accuracy, and user experience.

Voice-based biometric verification is widely employed in various smartphone applications including banking [5]. The main advantages of using voice biometrics in smartphone applications are accuracy, scalability, and usability. Conventional voice biometric systems enrol speakers in one language by using short sentences. During the verification, the speaker will utter the same sentence (in the case of text-dependent) or different sentences (in the case of text independence) in the same language used during the enrolment, which will be compared with the enrolled sample to make the verification decision. However, the use of the same language limits both service providers (or vendors) and users, in terms of scalability and usability. Because users can speak more than one language at a time, it is more convenient for speakers to use multilingual verification than a single language. From a vendor's perspective, it is important to build language-independent models that can achieve scalability. These factors motivated multilingual speaker verification, which allows the user to enrol in one language and verify it with another language. Therefore, multilingual speaker verification aims to verify speaker identity based on speech utterances from one or more languages and ensure that the voice-based security system is robust and generalizable to various applications [4].

Figure 1 illustrates example time signals and corresponding spectrograms for different languages uttered by the same subject. The different characteristics of the language, especially the sequence of phonemes, and the language-specific spectral characteristics of utterances from the same subject introduce challenges for reliable multilingual verification. However, as the same speaker is speaking

multiple languages, even though the sequence of phonemes may differ in languages, the individual phoneme characteristics may remain somewhat the same across languages because of the geometry offered by the same vocal tract, which has motivated researchers to develop multilingual speaker verifications. Earlier studies have explored multilingual speaker verification systems by extracting i-vectors or x-vectors from speech utterances [6,7], which are trained to discriminate speakers and map utterances to fixed dimensional embeddings. Recently, the 2D CNN model [8] was trained on English data and evaluated in the Chinese and Uyghur languages. It significantly outperforms the baseline i-vector model by a large margin [8].

The deployment of a multilingual speaker-verification model for smartphones is challenging because of the requirement for a lightweight text-independent model that can be generalized across different languages. In this study, we propose a novel CNN architecture for multilingual speaker verification based on Neural Architecture Search (NAS) methods [9] to derive the best CNN architecture for text-independent multilingual speaker verification. The proposed method has two novel features 1) An automatic network search that can result in an optimized network architecture for a multilingual speaker verification model. 2) A different architecture for normal and reduction cells to achieve reliable multilingual speaker verification with a lightweight model.

The main contributions of this study are as follows.

1. Novel method for multilingual speaker verification using differentiable neural architecture search to achieve the optimized lightweight model.
2. The proposed method is initialized to have different architecture for normal and reduction cells to better quantify the speaker characteristics.
3. Extensive experiments are presented on the publicly available MAVS dataset with 37,800 utterances representing three different languages. The MAVS dataset was collected in three different sessions using five different smartphones from 103 subjects with unique data.
4. The performance of the proposed method is compared with the Autospeech [10], which derives the architecture in automated way.

The rest of the paper is organized as follows: Sect. 2 discusses the related work on multilingual speaker verification, Sect. 3 presents the proposed method, Sect. 4 discusses the experimental protocols, architecture search, and quantitative results, and Sect. 6 concludes the paper.

2 Related Work

Multilingual speaker verification has attracted significant interest from researchers in the recent decades. Early work began with the introduction of the first Spanish corpus named AHUMADA [11] by NIST [12]. Joint Factor Analysis (JFA) was adopted by [13], in which language factors were captured in training and testing utterances. The evaluation results showed a significant improvement in the performance of the non-English trials.

The NIST Speaker Recognition Challenge in 2016 revealed the importance of score normalization for mismatched data conditions. Therefore, [14] compared several normalization techniques, as well as different cohorts, and analyzed the nature of the files selected for the cohort in adaptive score normalization. Unsupervised speaker verification was conducted using adversarial training [15]. For short utterances, hard prototype mining as a computationally efficient hard negative mining strategy to fine-tune the x-vectors was adopted by [16]. A large-scale study of 46 languages was conducted by [17], in which a hybrid novel triage mechanism was introduced for both text-dependent and text-independent methods. Lately [18] proposed disentangled representation learning to disentangle speaker module and language module. Both modules have a speaker feature extractor, embedding layer, and classifier to achieve reliable speaker verification.

It is worth noting that all existing studies are mainly focused on a non-smartphone environment, where the requirement of lightweight models is of paramount importance. Recently, [4] benchmarked a smartphone-based SWAN dataset consisting of four different languages by performing a cross-lingual speaker verification using the x-vector method. However, the development of lightweight models is important, particularly in smartphone environments. The first neural architecture search-based speaker recognition autospeech was recently proposed by [10]. The experimental results indicate a lightweight model with robust performance in English, which motivated us to propose a neural search method for multilingual speaker verification. *We hypothesize that ameliorating network architecture design matters for deriving a lightweight model for multilingual speaker verification for mobile devices.* Therefore, we consid-

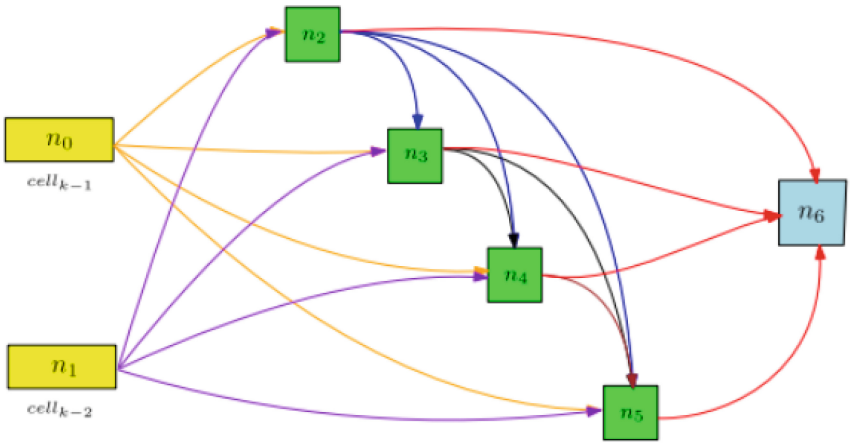


Fig. 2. Depiction of a neural cell. The transitional nodes(x_2 to x_5) are thickly connected during the search process. Only two operations with the highest softmax probabilities are retained during architecture derivation for the transitional nodes.

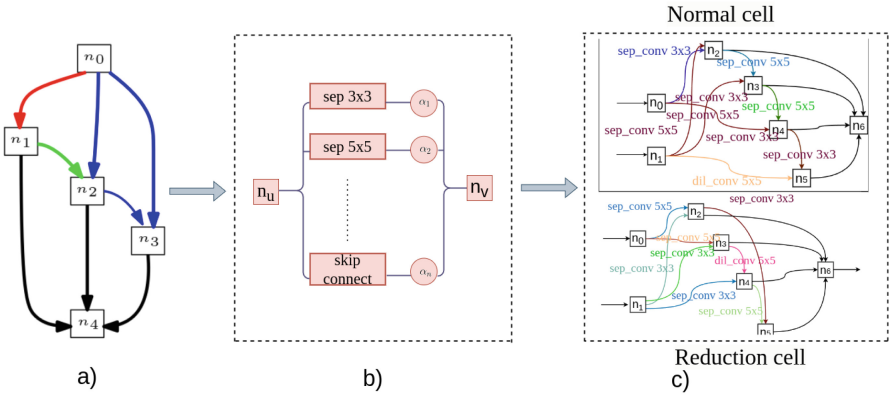


Fig. 3. a) illustration of neural architecture search, b) illustration of search space between node u, v the d) obtained different architecture for normal and reduction cell

ered Autospeech [10] as a baseline model and proposed a modified architecture that considers normal cells and reduction cells to have different architectures to increase the search space within the specified space, by which better speaker characteristics are captured.

3 Proposed Method: *NeuralMultiling*

The inspiration for a CNN-based search space is the cognizance that architecture engineering with a CNN often pinpoints repeated patterns consisting of convolutional filter banks, nonlinearities, and a judicious selection of connections to accomplish state-of-the-art results. In this section, we introduce the modified NAS, which is automated to find par excellent architectures. First, we introduce the neural cell in Sect. 3.1, and in Sect. 3.2 we define the candidate operations, in Sect. 3.3 and in Sect. 3.4, we modify the basic cell architectural parameters of [9] in and finally derive the discrete architecture.

3.1 Neural Cell

A block diagram of the proposed method is shown in Fig. 3. First, the search space is composed of nodes, and candidate operations exist between nodes. The operation with the highest weight is selected as the connection operation. After all connected edges and corresponding operations are selected, the final structure is determined. This section describes the process of building this architecture. To build a scalable architecture, we need 2-types of convolutional cells to deliver 2-main functions when taking a feature map as input:1) Normal cell: A convolutional cell that returns a feature map of the same dimensions and maintains the same number of channels. The normal cell is responsible for capturing features

and patterns from the input data and passing them to the next stage of the network. 2) The reduction cell is a convolutional cell that reduces the input feature map by a factor of two using a stride of two for all its operations, in contrast to the normal cell. In addition, this increases the number of channels, resulting in a reduction in the computational cost and complexity of the network. Reduction cells are used to downsample feature maps and compress the information before passing it to the next stage of the network.

A cell is a directed acyclic graph consisting of an ordered sequence of n nodes; in our case, n=7. Each node $n^{(i)}$ is a latent representation (e.g., a feature map on a convolutional neural network) and has directed edges (u, v) associated with some operation $o^{(u,v)}$ that transforms $n^{(i)}$. The structure of each cell was fixed, with each cell having two input nodes, four transitional cells, and one output node. A neural cell consists of two types of parameters:1) architectural parameters, which specify the structure of a neural cell in terms of the edges (or transformation tensors) connecting the source and target nodes and the operation being performed on them within the cell. 2) Weight parameters: The weight parameters are optimized while keeping the architectural parameters of the cell fixed.

Input Node: The input to the k^{th} cell is the output of the last two cells, the first input node n_0 is the output of the $(k-1)^{th}$ cell, and the second input node is the output of the $(k-2)^{th}$ cell. Both inputs to the 1^{st} cell are the same speech spectrogram. For the 2^{nd} cell, the first input n_0 is the output from the first cell and the second input is the speech spectrogram.

Intermediate Node: All intermediate nodes $(n_2 \rightarrow n_5)$ are densely connected, and each intermediate node n_i is computed as the summation of operations based on all its predecessors:

$$n^v = \sum_{u<v} o^{(u,v)}(n^{(u)}) \tag{1}$$

Because there were four intermediate nodes in a cell for our experiment, the number of edges within a cell was 14.

Output Node: The output from all the intermediate nodes is concatenated to form the output node. The architecture of a cell with all of its edges is shown in Fig. 2.

3.2 Candidate Operations

Each edge, connecting from the input node to the intermediate node and from the intermediate node to another intermediate node is associated with one of the operations from the set of candidate operations. These operations must be well-defined and is capable of capturing the variability and generality of data. The set of candidate operations utilized to derive the architecture are $'maxpool\,3\times3'$, $'avgpool\,3\times3'$, $'skip\,connect'$, $'sepconv\,3\times3'$,$'dilconv\,3\times3'$, $'sepconv\,5\times5'$ and $'dilconv\,5\times5'$

Each of the above operations has a constrained filter size, which we refer to as a channel that can be fine-tuned to obtain the optimal channel size for a given dataset. These operations are common in modern CNN architectures. Finally, each edge was associated with one of these operations. The best combination of operations with these edges is obtained at the end of the search process. It is noteworthy that the operations chosen for each cell were independent of each other. These operations constitute search space (O). Our convolutional cell consisted of N=8 neural cells and an initial number of channels C=16; the network was formed by stacking them together. Following previous studies [9] reduction cells are located at $\frac{1}{3}$ and $\frac{2}{3}$ positions of the total depth of the network, and the rest are normal cells.

3.3 Continuous Relaxation over the Cells and Bi-Level Optimization

Let O be a set of candidate operations (e.g., convolutions, max pooling, and zero), where each operation refers to a function $o(.)$ to be applied to n^u. We use normal and reduction cells to have different architectures, which is contrary to [9] and formulate the continuous search space relaxing the categorical choice of operations to be a softmax over all possible operations:

$$\bar{o}^{(k,u,v)} = \sum_{o \in O} \frac{exp(\alpha_O^{(k,u,v)})}{\sum_{o' \in O} exp(\alpha_{o'}^{(k,u,v)})} o(n) \tag{2}$$

The goal of the architecture search is then reduced to learning in a continuous variable $\alpha = \{\alpha^{(k,u,v)}\}$ where k is the cell index as illustrated in Fig. 4. After searching among candidate architectures, a discrete architecture is obtained by jointly optimizing the mixing probabilities and network weights by solving the bi-level optimization problem for each normal and reduction cell with the most likely operations, that is, $o_{normal}^{(k,u,v)} = argmax_{o \in O}\ \alpha_o^{(k,u,v)}$ for normal cells and $o_{reduction}^{(k,u,v)} = argmax_{o \in O}\ \alpha_o^{(k,u,v)}$ for the reduction cell, contrary to [9]. Subsequently, we aim to jointly learn the candidate architectures and weight parameters. After soothing out, we aim to jointly learn the candidate architectures and weight parameters.

In each iteration of the Algorithm 1 two steps are being carried out for each cell(k):

- **Weight parameter (ω) update:** During this step, the weight parameters are optimized while keeping the architectural parameters of the cell fixed.
- **Architectural parameter (α^k) update:** In this step, the architectural parameters of the cell are updated based on the architectural loss while fixing the weight parameters. The update of both the weight and architecture parameters is achieved through the minimization of the respective cross-entropy loss equation 4.

3.4 Re-Defining the Architecture Parameters of Normal and Reduction Cell

According to [9], normal and reduction cells have the same architecture parameters, that is, 14(*edges*) × 8(*operations*) for both types of cells. In contrast,

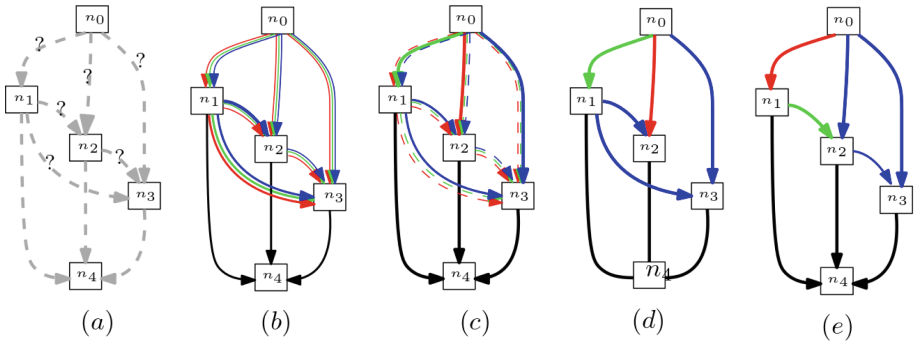


Fig. 4. An overview of Continuous relaxation: a) Initial architecture with unknown operations. b) Continuous relaxation of the searched space on each of the edges by setting up candidate operations. c) Two-way optimization of network weights and probabilities of each node. d) & e) Spawning the final architecture from the learned probabilities for normal cell and reduction cell.

Algorithm 1 Search Algorithm:

Input: ← Training data $\mathcal{D}_{train}$ and validation data $\mathcal{D}_{val}$
Output: Searched Architecture
procedure ENTROPY CALCULATION

$$E = \sum_{k \in C} \sum_{(u,v)} \sum_{o \in \mathcal{O}} \alpha^{o}_{kuv} \log \alpha^{o}_{kuv} \tag{3}$$

end procedure

procedure NAS
  while *entropy decreases* **do** ▷ \ ∗ . . . entropy of the cells have decreased \
    for each cell k **do**
      Fix the Architectural Parameters for a cell (α_k)
        $\mathcal{L}_{train}$ ← training loss $\mathcal{D}_{train}$
        $\nabla_{\omega}\mathcal{L}_{train}$ ← gradient on $\mathcal{D}_{train}$
        Update the weight parameters (ω)
      Fix the weight parameters(ω)
        $\mathcal{L}_{val}$ ← Validation loss $\mathcal{D}_{val}$
        $\nabla_{\alpha_k}\mathcal{L}_{val}$ ← gradient on $\mathcal{D}_{val}$
        Update the architectural parameters (α_k)
    end for
  end while
end procedure

we assume that the normal and reduction cells have different architectures by modifying the parameter dimensionality of the normal cell to (number of cells−2) × 14(*edges*) × 8(*operations*) and 2 × 14(*edges*) × 8(*operations*) for the reduction cell. By doing so, we increase the search within the specified search space, which better captures the speaker variability across various speakers. More details about the modified architecture is given in the supplementary material.

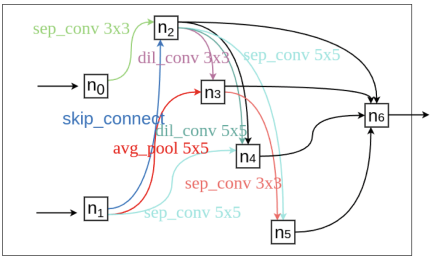


Fig. 5. Normal cell: Architecture derived from our proposed search algorithm

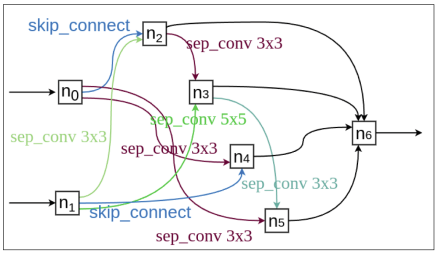


Fig. 6. Reduction cell: Architecture derived from our proposed search algorithm

The main objective of architecture search is to produce an excellent architecture α^* that minimizes the validation loss $L_{val} = (\omega^*, \alpha^*)$ where ω^* is the weight parameter obtained by minimizing the training loss $\omega^* = argmin_\omega \ L_{train}(\omega, \alpha^*)$. The architecture parameters of the normal and reduction cells are considered to be a 3-D tuple (x, y, z) where x = cell index, y = number of edges, z = number of operations are jointly optimized by passing through the Adam optimizer. The outcome of the optimizer is to produce low entropy, and the cross-entropy losses for L_{train} and L_{val} are described as follows:

$$L_{CE} = -\sum_{i=1}^{N} t_i log(p_i) \tag{4}$$

where t_i: ground truth speaker, N: Number of speakers, $logp_i$: softmax probability of speaker i.

Deriving Discrete Architectures. To construct each node in the architecture, we keep the top-2 highest softmax probabilities among all non-zero operations accumulated from previous nodes. The softmax probability of an operation O between the nodes (u, v) is defined as:

$$p^o_{(k,u,v)} = \frac{exp(\alpha^o_{(k,u,v)})}{\sum_{o' \in O} exp(\alpha^{o'}_{(k,u,v)})} \tag{5}$$

The visualization of the architectures of the searched normal cell and reduction cell is as shown in Fig. 5 and 6.

Table 1. Verification performance of the proposed and existing methods for multilingual speaker verification for language agnostic scenario

EER (%)								
Trained on ↓	Tested on							
	Proposed				Autospeech [10]			
	English	Hindi	Bengali	Parameters	English	Hindi	Bengali	Parameters
English	**20.99**	21.33	23.74	362383	**27.04**	25.72	27.44	418079
Hindi	22.68	**17.73**	19.75	362383	26.02	**22.21**	24.67	418079
Bengali	21.95	19.59	**18.95**	362383	25.90	25.48	**23.18**	418079

4 Experiments and Results

In this section, we discuss the quantitative results of the proposed and existing methods for multilingual speaker verification. First, we present the multilingual dataset employed in this study, followed by the implementation details and discussion of the results.

4.1 Multilingual Voice Dataset

We conducted our experiments on the MAVS database [4], which consists of 37,810 utterances in three languages recorded in three different sessions using five different mobile phones: iPhone 6 s, iPhone 10 s, iPhone 11, Samsung S7, and Samsung S8 spoken by 103 speakers (70 male and 33 female). The three different sessions include session-1 with no noise, session-2 with controlled noise, and session-3 with uncontrolled noise from the natural background. The dataset was divided into $32,250(80\%)$ utterances for training and $5560(20\%)$ utterances for testing in both stages, that is, for the architecture search and training of the searched architecture for multilingual speaker verification. In the first stage, we search for each cell architecture using our search procedure explained in Algorithm 1. If the entropy calculated using (3) remains the same for dozens of epochs, we can conclude that the algorithm has converged. In the second stage, the searched architecture was trained from scratch, and its performance on the test set was reported. We used a single-searched architecture for all the experiments reported in Sect. 4.3.

4.2 Implementation Details

For each utterance, we excerpted a 257-dimensional spectrogram with a 25ms window and 10ms overlap. We implemented the proposed architecture search using Pytorch and trained it on a paramshakti supercomputer which has 22 nodes, each of the node has two GPUs of 16 GB named V100 Tesla, and we used one node for training. The NAS search process model described in Sect. 3 was trained for 50 epochs, with a batch size of 8. We utilized the Adam optimizer to optimize both the weight ω and the architecture parameters α by setting the

initial learning rate to 10^{-1} and the weight decay of the optimizer to 3×10^{-4}. The entire search process took five days to converge. In the second stage, the searched architecture was trained from scratch for 200 epochs, with a batch size of 48. The optimizer learning rate was set to 0.15 the weight decay to 3×10^{-4}, and the verification process took less than a day.

4.3 Results and Discussion

In this section, we present quantitative results of the proposed method for multilingual speaker verification. The performance of the proposed method was compared with that of Autospeech [10], which is based on neural searching. Autospeech was trained using the MAVS [4] dataset under similar training conditions as described in Autospeech for a fair comparison. There was another model based on Bi-LSTMs [17] where they presented a lighweight speaker verification models operated on 46 languages. But to do comparison neither the code nor the dataset is available in open source. We present two different experiments: (1) Language agnostic, in which the speaker is enroled with one language and probed with another language. (2) Interoperability across smartphones (or devices) and languages in which the speaker is enrolled with one device and one language and probed with other devices and languages. The performance of the proposed method is presented using an Equal Error Rate (EER(%)), which corresponds to the False Match Rate (FMR), and is equal to the False Non-Match Rate (FNMR). Table 1 shows the quantitative performance of the proposed and existing methods in language-agnostic experiments. Here we consider language

Table 2. Interoperability results in EER(%) for the proposed method. E-refers to English, H-refers to Hindi, B-refers to Bengali and each entry in table shows the EER.

	Devices															
Trained on ↓	iPhone 6s			iPhone 10s			iphone 11			Samsung S7			Samsung S8			
iPhone 6s	E	H	B	E	H	B	E	H	B	E	H	B	E	H	B	
E	**16.71**	18.96	22.21	**23.25**	22.35	23.35	**23.45**	21.26	21.62	**22.45**	20.16	22.92	**20.19**	19.60	22.10	
H	17.96	**15.72**	21.63	26.25	**21.13**	24.45	25.62	**20.18**	20.62	23.62	**19.62**	21.68	21.36	**18.72**	21.01	
B	20.22	19.72	**17.13**	25.26	24.45	**20.13**	24.15	22.62	**19.72**	24.56	21.27	**20.16**	22.57	20.61	**19.63**	
iphone 10s																
E	**26.79**	24.20	26.90	**19.25**	15.36	17.35	**19.71**	17.42	20.21	**21.84**	26.78	26.79	**21.78**	22.72	23.88	
H	27.38	**24.97**	24.85	19.94	**13.99**	15.45	18.14	**15.61**	21.42	21.84	**24.12**	25.31	25.03	**23.84**	22.50	
B	27.85	24.56	**24.27**	17.24	15.53	**16.74**	21.18	16.44	**17.52**	23.95	22.53	**24.54**	24.60	27.78	**21.94**	
iPhone 11																
E	**26.53**	23.24	26.19	**18.49**	14.91	17.21	**15.55**	14.96	17.49	**23.40**	24.77	23.15	**22.61**	21.88	19.03	
H	26.20	**26.14**	26.22	19.94	**15.95**	17.03	14.82	**13.77**	14.81	21.72	**22.71**	20.87	20.89	**20.49**	21.06	
B	27.44	23.42	**27.46**	19.50	13.43	**16.21**	16.94	13.24	**15.19**	23.71	20.79	**23.52**	24.81	20.34	**20.66**	
Samsung S7																
E	**25.51**	26.19	26.32	**24.25**	25.57	25.98	**24.57**	25.17	25.86	**18.19**	19.22	20.77	**18.96**	19.41	20.63	
H	26.41	**24.44**	25.73	26.13	**25.22**	26.41	24.64	**22.12**	24.73	15.35	**15.91**	15.02	20.15	**19.23**	20.11	
B	25.23	25.11	**24.92**	25.52	24.78	**24.63**	25.12	24.06	**23.21**	18.16	17.17	**14.63**	19.42	19.71	**18.63**	
Samsung S8																
E	**22.12**	27.18	21.91	**26.54**	20.57	20.71	**22.88**	28.59	24.90	**20.06**	18.19	17.13	**16.94**	18.60	16.09	
H	20.35	**23.61**	20.26	24.67	**28.41**	23.06	20.90	**27.58**	26.94	17.64	**17.14**	17.35	18.32	**16.19**	15.11	
B	27.44	23.42	**27.46**	19.50	13.43	**16.21**	16.94	13.24	**15.19**	23.71	20.79	**23.52**	24.81	20.34	**20.66**	

Table 3. Interoperability results in EER(%) for Autospeech [10]. E- refers to English, H-refers to Hindi, B-Bengali language. Each entry in the table indicates EER.

	Devices														
Trained on ↓	iPhone 6s			iPhone 10s			iphone 11			Samsung S7			Samsung S8		
iPhone 6s	E	H	B	E	H	B	E	H	B	E	H	B	E	H	B
E	**18.71**	20.72	22.72	**25.49**	26.18	25.23	**25.36**	26.18	25.68	**24.45**	21.16	23.62	**22.16**	21.05	20.18
H	21.42	**18.62**	23.31	25.69	**23.39**	24.35	23.14	**22.29**	21.26	24.45	**21.79**	22.55	24.46	**20.60**	23.30
B	20.62	22.92	**17.69**	24.11	23.09	**22.05**	21.04	22.56	**20.62**	23.03	22.72	**22.50**	22.11	23.62	**21.56**
iphone 10															
E	**24.40**	25.10	26.90	**20.25**	19.25	18.19	**20.17**	18.62	22.12	**23.48**	25.78	25.56	**22.06**	23.27	24.81
H	24.07	**23.60**	23.16	16.26	**15.99**	17.32	20.54	**16.51**	22.50	22.19	**26.12**	24.17	26.03	**25.48**	23.17
B	23.92	23.05	**22.16**	19.60	20.11	**18.64**	22.12	18.06	**18.72**	24.56	24.35	**25.46**	25.07	26.90	**23.49**
iPhone 11															
E	**24.43**	25.24	25.14	**20.94**	16.81	18.12	**17.45**	15.19	16.69	**24.40**	25.68	24.17	**24.16**	22.98	20.11
H	24.24	**24.44**	25.22	20.14	**17.59**	16.17	16.92	**15.71**	15.19	22.06	**23.81**	21.78	22.78	**22.98**	22.60
B	25.44	23.42	**24.19**	21.60	14.34	**18.08**	18.49	15.42	**17.11**	24.68	21.55	**23.60**	25.18	21.43	**22.66**
Samsung S7															
E	**24.41**	25.91	25.32	**25.52**	25.17	25.68	**25.75**	24.17	26.68	**20.91**	20.23	21.76	**20.68**	20.14	22.40
H	24.61	**25.44**	25.63	26.31	**25.24**	27.72	25.46	**23.21**	24.37	16.53	**16.76**	17.20	22.61	**21.64**	21.07
B	25.13	24.11	**24.62**	24.22	23.87	**24.36**	25.42	25.60	**24.21**	20.80	19.76	**15.40**	21.36	22.61	**20.11**
Samsung S8															
E	**23.21**	27.81	22.60	**27.13**	22.62	21.18	**23.66**	26.95	25.68	**18.16**	20.22	19.46	**18.19**	20.60	17.11
H	21.55	**24.09**	21.08	25.69	**27.51**	24.60	21.62	**26.65**	27.18	18.56	**19.14**	20.46	20.22	**18.22**	16.44
B	26.44	25.32	**26.64**	20.66	15.56	**15.76**	18.11	15.98	**16.12**	22.66	25.25	**25.18**	25.62	21.24	**21.88**

based speech files from all devices and perform cross language testing. Based on the results in Table 1, the following can be observed.

- The verification performance of the proposed and the existing method indicates the improved performance when trained and tested with the same language. The best performance was observed when trained and tested using Hindi.
- The verification performance degradation is noted with the proposed and the existing method during the cross language test. It can also be observed that the training language can influence the verification performance of the proposed and existing methods. For example, training with the English language indicated less verification performance degradation when tested with other languages such as Hindi and Bengali. Furthermore, it is interesting to note that the cross-language verification performances of the proposed and existing methods are less degraded between Hindi and Bengali. This can be attributed to similarities in language characteristics.
- The proposed method indicates the best performance compared to the existing method on cross and same language experiments.

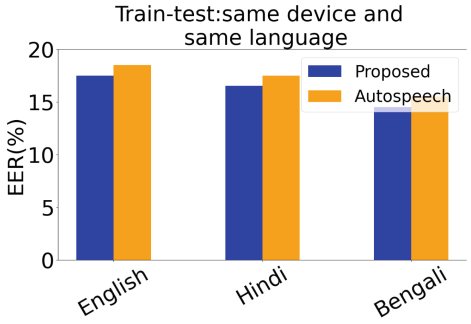


Fig. 7. Case-1: Histogram for same device and same language

– Achieving a language agnostic condition in multilingual speaker verification involves addressing challenges such as variations in phonetic structures, acoustic characteristics and linguistic patterns across different languages. The Autospeech [10] which follows same architecture for normal cell and reduction cell do not capture the above said characteristics hence a degradation of EER is observed whereas for our proposed method the EER is reduced when cross-language testing is performed. This accounts for robust and a generalizable muiltilingual verification model that can adapt to inherent diversity in languages while maintaining lesser number of parameters.

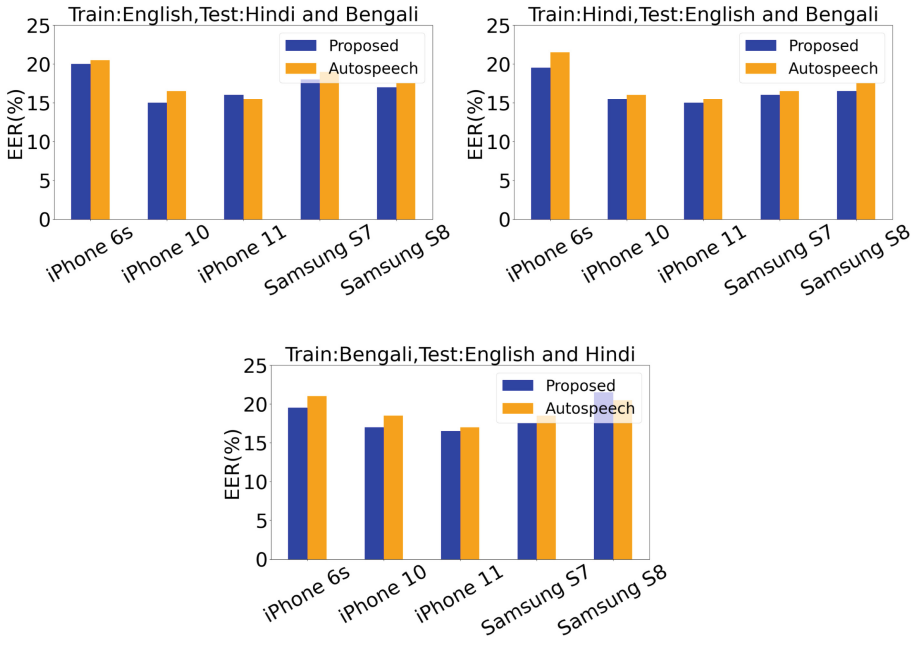


Fig. 8. Case-2: Histogram for cross language and same device

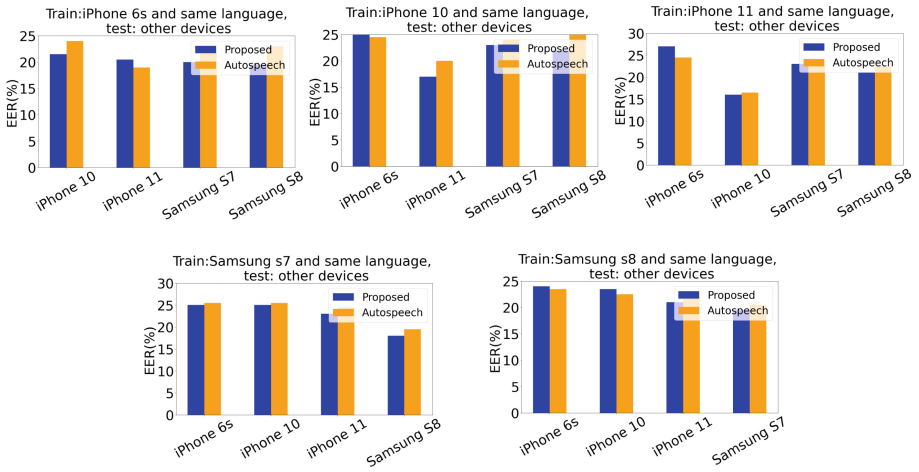


Fig. 9. Case 3: Histogram for same language and cross device

– The proposed method also results in the less number of parameters (362k) and model size of *6.28Mb* compared to the existing method with (418k) parameters with model size of *8.28Mb*. Thus, the proposed method not only outperforms the existing method but also results in a lightweight model suitable for deployment in a smartphone environment.

Tables 2 and 3 show the quantitative performance of the proposed method and Autospeech for interoperability across devices and languages, respectively. The interoperability experimental results were interpreted based on the four cases discussed below.

– **Case-I: Same device and same language:** Here, we analyze the verification performance of the proposed method when same language is trained and tested by the same device. This analysis provides insight into the verification performed on independent languages. Figure 7 shows the average EER(%) with respect to different devices, which is independent of language. As shown in Fig. 7, Hindi had the lowest EER(%), and English had the highest EER(%). The best performance with the Hindi language can be attributed to the fact that the majority of speakers in the MAVS dataset were native Hindi speakers.
– **Case II: Cross language and same device:** Here, we analyze the verification performance of the proposed method when individual devices are trained in one language and tested in another language. Figure 8 shows the verification performance of the proposed method with cross-language and the same device scenario. This experiment allowed us to analyze the interoperability of language across devices. The obtained results indicate that (a) the verification performance is influenced by the language of the individual devices. (b) The iPhone6S has the highest EER(%) across all three languages. (3) The

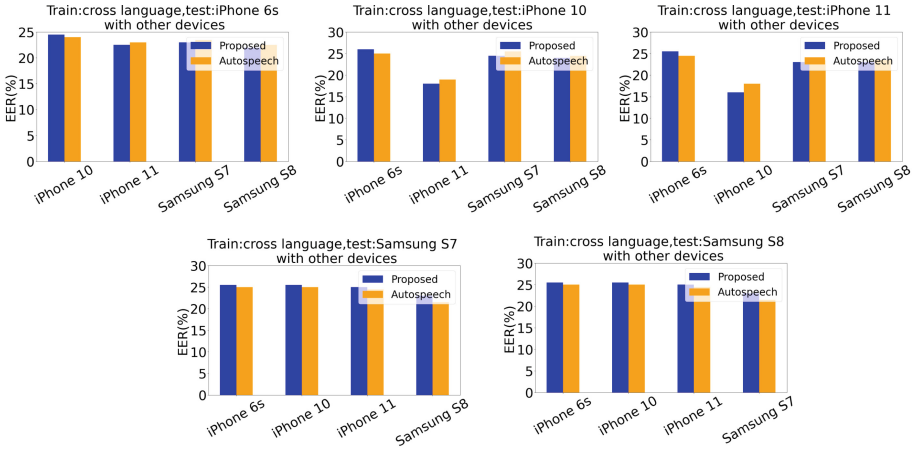


Fig. 10. Case-4: Histogram for cross language and cross device

iPhone11 indicated the best performance across all three languages and thus emerged as the best language-agnostic device with the proposed method.

– **Case III: Same language and cross device:** Here, we analyze the verification performance of the proposed method when the same language is used for training and testing, while cross devices are used for verification. Figure 9 shows the verification performance of the proposed method when the voice data (irrespective of the language) from one type of device are used for training, and testing is performed using the voice data (same language as that of training) collected from another device. This experiment allowed the interoperability of the devices to be analyzed when the same language was used for training and testing. The obtained results indicated the influence of the device data on the verification performance. In some cases, the interoperability of the devices indicated improved performance (for example, when trained with the iPhone6S and tested with Samsung S8). However, the verification performance across other devices was less influenced, particularly when Samsung devices were used.

Table 4. The ablation study of the proposed method and Autospeech [10] with varying number of nodes and channels such that N indicated number of nodes and C indicates number of channels.

Method	No of Parameters	Search cost No of GPU days
Autospeech [10]		
N=8, C=16	418k	7
N=8, C=64	617k	9
N=30, C=64	986k	10
N=8, C=128	1160k	11
Proposed		
N=8,C=16	362k	5
N=8, C=64	418k	6
N=30, C=64	568k	8
N=8, C=128	625k	8.2

– **Case IV: Cross language and cross device:** Here, we analyze the verification performance of the proposed method with cross language (training and testing with different languages) and cross device (enrolment using one device and probe with other devices). Figure 10 shows the verification performance of the proposed method in cross-language and cross-device scenarios. This experiment allowed us to analyze the performance of the proposed system with interoperability for both language and device. Note that (a) the interoperability of the devices indicates higher error rates with cross-language. (b) Verification performance degrades across all devices. For all four cases, our proposed method outperforms Autospeech [10] because our proposed model can capture the speaker characteristics better than Autospeech [10].

5 Ablation Studies

To verify the effectiveness of the proposed method, we use original Autospeech to search on our MAVS dataset. It is obvious that the search cost is greatly reduced through our proposed method. We also varied the number of nodes and number channels during the search process and for each model our proposed method outperforms Autospeech both in terms of number of parameters and also in terms of search cost. The experimental results are as shown in Table 4.

6 Conclusions

In this study, we propose an automatic approach to determine the optimal CNN architecture for multilingual speaker verification. We modified the baseline approach by introducing different architectures for normal and reduction cells. With

this modification, we searched for an excellent CNN architecture for neural cells with different edge operations. Subsequently, with the derived architecture we conducted two different experiments: language-agnostic conditions across various smartphone devices, and interoperability by building language models across different devices and languages on MAVS database. For the language-agnostic condition, our proposed method outperformed the baseline model while maintaining lower model complexity. For interoperability, the proposed model also yields better performance when the trained and test mobile phones are from the same manufacturer; however, for cross devices, a slightly higher EER is observed. Overall, we obtained an automatic architecture that is lightweight and performs better than the baseline model, which can be further deployed into mobile devices for multilingual speaker verification.

References

1. Rattani, A., Derakhshani, R.: A survey of mobile face biometrics. Comput. Electr. Eng. **72**, 39–52 (2018)
2. Das, A., Galdi, C., Han, H., Ramachandra, R., Dugelay, J.L., Dantcheva, A.: Recent advances in biometric technology for mobile devices. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2018, pp. 1–11 (2018). https://doi.org/10.1109/BTAS.2018.8698587
3. Ramachandra, R., et al.: Smartphone multi-modal biometric authentication: database and evaluation, arXiv preprint arXiv:1912.02487
4. Mandalapu, H., Reddy, P.A., Ramachandra, R., Rao, K.S., Mitra, P., Prasanna, S.M., Busch, C.: Multilingual audio-visual smartphone dataset and evaluation. IEEE Access **9**, 153240–153257 (2021)
5. Research, Markets, Voice biometrics market forecast to 2028 - covid-19 impact and global analysis by component, type, authentication process, deployment, vertical, and application, May 2022. https://www.researchandmarkets.com/reports/5623597/voice-biometrics-market-forecast-to-2028-covid
6. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2010)
7. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: robust DNN embeddings for speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE 2018, pp. 5329–5333 (2018)
8. Li, L., Wang, D., Rozi, A., Zheng, T.F.: Cross-lingual speaker verification with deep feature learning. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE 2017, pp. 1040–1044 (2017)
9. Liu, H., Simonyan, K., Yang, Y.: Darts: differentiable architecture search, arXiv preprint arXiv:1806.09055
10. Ding, S., Chen, T., Gong, X., Zha, W., Wang, Z.: Autospeech: neural architecture search for speaker recognition, arXiv preprint arXiv:2005.03215
11. Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V.: Ahumada: a large speech corpus in Spanish for speaker characterization and identification. Speech Commun. **31**(2–3), 255–264 (2000)

12. Greenberg, C.S., Mason, L.P., Sadjadi, S.O., Reynolds, D.A.: Two decades of speaker recognition evaluation at the national institute of standards and technology. Comput. Speech Lang. **60**, 101032 (2020)
13. Lu, L., Dong, Y., Zhao, X., Liu, J., Wang, H.: The effect of language factors for robust speaker recognition. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2009, pp. 4217–4220 (2009)
14. Matejka, P., Novotnỳ, O., Plchot, O., Burget, L., Sánchez, M.D., Cernockỳ, J.: Analysis of score normalization in multilingual speaker recognition. In: Interspeech, pp. 1567–1571 (2017)
15. Xia, W., Huang, J., Hansen, J.H.: Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 5816–5820 (2019)
16. Thienpondt, J., Desplanques, B., Demuynck, K.: Cross-lingual speaker verification with domain-balanced hard prototype mining and language-dependent score normalization, arXiv preprint arXiv:2007.07689
17. Chojnacka, R., Pelecanos, J., Wang, Q., Moreno, I.L.: Speakerstew: scaling to many languages with a triaged multilingual text-dependent and text-independent speaker verification system, arXiv preprint arXiv:2104.02125
18. Nam, K., Kim, Y., Huh, J., Heo, H.S., Jung, J.W., Chung, J.S.: Disentangled representation learning for multilingual speaker recognition, arXiv preprint arXiv:2211.00437

MTSA-SNN: A Multi-modal Time Series Analysis Model Based on Spiking Neural Network

Chengzhi Liu[1(✉)], Zihong Luo[1], Zheng Tao[1], Chenghao Liu[2], Yitao Xu[1], and Zile Huang[1]

[1] Xi'an Jiaotong-Liverpool University, Suzhou, China
chengzhi.liu21@student.xjtlu.edu.cn
[2] Shanghai Normal University, Shanghai, China

Abstract. Time series analysis and modelling constitute a crucial research area. Traditional artificial neural networks struggle with complex, non-stationary time series data due to high computational complexity, limited ability to capture temporal information, and difficulty in handling event-driven data. To address these challenges, we propose a Multi-modal Time Series Analysis Model Based on Spiking Neural Network (MTSA-SNN). The Pulse Encoder unifies the encoding of temporal images and sequential information in a common pulse-based representation. The Joint Learning Module employs a joint learning function and weight allocation mechanism to fuse information from multi-modal pulse signals complementary. Additionally, we incorporate wavelet transform operations to enhance the model's ability to analyze and evaluate temporal information. Experimental results demonstrate that our method achieved superior performance on three complex time-series tasks. This work provides an effective event-driven approach to overcome the challenges associated with analyzing intricate temporal information. Access to the source code is available at https://github.com/Chenngzz/MTSA-SNN.

Keywords: Multi-Modal · Time series analysis · Spiking neural network · Joint learning · Pulse encoder · Wavelet transform

1 Introduction

Traditional artificial neural networks (ANNs) have found extensive applications in time series analysis. They serve as a non-parametric, non-linear model capable of effectively capturing complex non-linear relationships within time series data. This is particularly valuable for addressing numerous time series problems since relationships within such data are typically non-linear. Deep neural networks (DNNs), as an extension of ANNs, exhibit a multi-layer structure that automatically learns features and hierarchical information from data. This characteristic

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 424–438, 2025.
https://doi.org/10.1007/978-3-031-78341-8_27

enhances the capability of DNNs to analyze complex time series data by capturing patterns at various abstraction levels. For instance, deep learning models like Long Short-Term Memory (LSTM) networks have been widely employed to predict future values or sequences using past time steps [1]. ANNs have also been widely applied across a range of applications traditionally addressed by statistical methods, including classification, pattern recognition, prediction, and process control [2].

However, for complex and volatile time series information, traditional ANNs often face challenges in capturing temporal features accurately. Consequently, Spiking Neural Networks (SNNs), as an alternative approach, have garnered considerable attention. Currently, SNNs have been successfully applied in various time series prediction scenarios, including financial time series forecasting, time series classification [3], and real-time online time series prediction [4].

SNNs rely on discrete signals in continuous time to effectively capture complex time patterns. Nonetheless, current SNN models encounter several challenges. First, the transformation of time series data into a suitable spiking representation poses a significant challenge. Second, the firing times of spiking neurons play a crucial role in model performance, necessitating higher demands for stability and accuracy. Moreover, integrating information from different sources into a single spiking network framework for decision-making involves complex issues related to cross-modal time synchronization and information mapping.

To address these challenges, we propose a Multi-Modal Time Series Analysis model based on Spiking Neural Networks (MTSA-SNN). This model consists of three key components: a single-modal spiking encoder, a spiking joint learning module, and an output layer. The spiking encoder is responsible for transforming time-series information from different modalities into spike signals. It includes alternating layers of feature extraction and neuron layers to selectively process input data from each modality. In the spiking joint learning module, we design a joint learning function and weight allocation mechanism to balance and fuse the complex spike information from multiple modalities. The output layer optimally adjusts the fused spike information to adapt to complex time series analysis tasks. The main contributions are as follows:

- A novel Multi-modal Time Series Analysis Model Based on a Spiking Neural Network proposed by us. This model introduces an efficient event-driven approach that overcomes the limitations of traditional time series analysis methods.
- We design SNN joint learning functions and a weight allocation mechanism, effectively addressing the balance and fusion of pulsed information.
- We synergize wavelet transform with pulse networks to bolster the model's capability in analyzing complex and non-stationary temporal data.
- Extensive experiments demonstrate the outstanding performance of our approach across multiple complex time series datasets.

2 Related Work

2.1 Time Series Forecasting

Modelling and forecasting time series data is a valuable task in various domains. It has evolved significantly, transitioning from traditional methods to deep learning techniques, resulting in improved prediction accuracy and relevance over time.

Initially, time series forecasting relied on traditional approaches such as the ARIMA model [5] and Fourier analysis [6]. ARIMA, which includes autoregressive (AR) and moving average (MA) components with differencing (I) to address non-stationarity, had challenges related to parameter selection and model identification. Fourier analysis was used for frequency domain analysis to identify periodic and seasonal patterns in the data.

Later, deep learning methods such as RNN and LSTM emerged to handle temporal dependencies [7]. LSTM, an improved version of RNN, performed better with long sequences due to its enhanced memory and forgetting mechanisms, becoming the preferred model for many time series problems. Nonetheless, they encountered challenges related to gradient vanishing and exploding when handling extended sequences, which restricted their practicality.

In contrast to single-modal time series forecasting, multi-modal time series forecasting leverages multiple data sources, such as text, images, and sensor data, to capture a broader perspective, enabling a wider range of pattern and trend recognition. This approach offers benefits like information synthesis, complementarity of different data types, model robustness, and improved generalization. Multi-modal deep learning models use CNN and BiLSTM to extract features from multi-modal time series data. Ensemble models, including probabilistic time series prediction based on Hidden Markov Models [8] and stacked ensembles, have been used to enhance accuracy and reduce overfitting.

Specific algorithms, including interpretable ML models and multi-modal meta-learning techniques [9], have been applied in diverse use cases, ranging from early Parkinson's disease detection to time series regression tasks. These applications highlight their potential in various domains, reflecting the diversity and complexity of time series modelling and forecasting. They underscore the evolving methods and technologies that offer robust tools for a broad spectrum of application scenarios (Fig. 1).

2.2 Spiking Neural Network

Multi-modal time series models struggle with complex, irregularly and non-uniformly sampled data due to their continuous computations, difficulty in handling event-driven data patterns, and high computational complexity. However, Spiking Neural Networks (SNNs) hold promise in mitigating these challenges. SNNs, a unique class of neural networks that communicate using discrete spike signals in a continuous-time framework [10], are capable of emulating biological neural systems' sparsity and encoding temporal information [11]. SNNs find

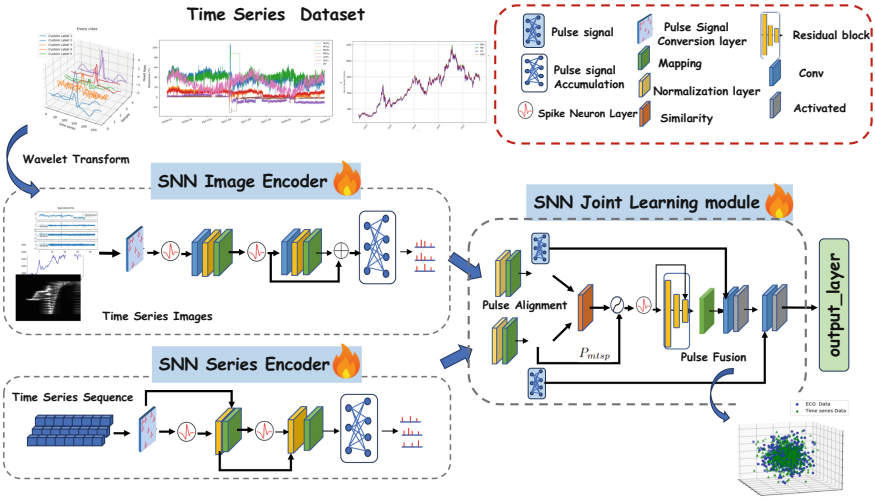


Fig. 1. The MTSA-SNN structure consists of three main components: Single-mode pulse encoding modules used to extract features from time series data; SNN Joint Learning Module utilizes joint learning and probability distribution methods to map multi-modal signals to a shared joint learning space, enabling the fusion of pulse signals. Output layer used to generate predictions and classification results for multi-modal time series data.

practical application in various time series prediction scenarios, including financial time series forecasting, time series classification [3], and real-time online time series prediction [4].

SNNs pose challenges due to their complex neurons and non-differentiable pulse-based operations. Choosing a multi-modal time series model depends on the problem and data characteristics. The multi-spike network SNN variant is useful for financial time series prediction. Therefore, you should select the most appropriate model based on the problem and data characteristics.

SNNs (Spiking Neural Networks) present challenges due to the complexity of their neurons and the non-differentiable nature of pulse-based operations, making training complex. The choice of a multi-modal time series model should depend on the problem and data characteristics. For instance, a variant like multi-spike networks has proven valuable in time series prediction, especially for non-stationary data [12]. Thus, selecting the right model should align with the problem and data intricacies.

In summary, the proposed MTSA-SNN model efficiently encodes multimodal information into spikes. It utilizes a spike-based cooperative learning module to effectively map and integrate complex spike information. This method provides an accurate and practical event-driven approach that addresses the analysis of complex and non-stationary temporal information, demonstrating strong performance across multiple time series datasets.

3 Methodology

The MTSA-SNN structure consists of three main components: SNN Encoder Module used to extract features from time series data; SNN Joint Learning Module utilizes joint learning and probability distribution methods to map multimodal signals to a shared joint learning space, enabling the fusion of pulse signals. Output layer used to generate predictions and classification results for multimodal time series data. The entire workflow is shown in Algorithm 1.

Algorithm 1. MTSA-SNN Overall Model

Require: Data of different models
Input: S_{image}, S_{series}
Single-Modal Pulse Encoding Module:
$\hat{S}_1 = \text{Encoder}_i(S_{\text{image}})$, $\hat{S}_1 \in \mathbb{R}^{T\times B\times C\times H\times W}$
$\hat{S}_2 = \text{Encoder}_t(S_{\text{series}})$, $\hat{S}_2 \in \mathbb{R}^{T\times B\times C\times T}$
Pulse Joint Learning Module:
$J_{\text{align}} = \Psi(FT(\hat{S}_1, \hat{S}_2))$, $J_{\text{align}} \in \mathbb{R}^{T\times B\times C\times T}$
$J_{\text{fusion}} = \textbf{JWAM}(\hat{S}_{1/2})$, $J_{\text{fusion}} \in \mathbb{R}^{T\times B\times C\times T}$
$J_{MTSA} = \textbf{Outputlayer}(J_{\text{fusion}})$, $J_{MTSA} \in \mathbb{R}^{T\times B\times N}$
return J_{MTSA}

3.1 Single-Modal Pulse Encoding Module

The visual pulse encoder is a component that processes time-series image information into pulse representations and extracts features. This encoder alternates between the Feature Extraction (FM) module and the Leaky Integrate-and-Fire (LIF) SNN module. Visual information initially passes through the SNN layer to be transformed into a unified and compatible pulse signal format, making it suitable for subsequent network operations. The FM module further performs feature extraction on the visual information converted into pulse signals, including operations such as convolution and pooling. After feature extraction, the pulse signal $\hat{S}_1$ is then passed to the pulse co-learning module.

The sequence pulse encoder is another modality encoder used for pulse-coding and feature extraction of temporal data sequences. These sequence data initially pass through the SNN layer and are then transformed into pulse signals. The network employs alternating operations between mapping layers and neurons. Neurons receive pulse information from the previous layer and membrane potential from the preceding time step in the sequence. By introducing this self-feedback mechanism, the pulse network can utilize membrane potential information from the previous time step to influence the calculations at the current time step. Consequently, the encoder is better equipped to capture the temporal correlations and dynamic changes in time-series data. The pulse information encoded through sequence encoding is $\hat{S}_2$. The SNN encoder entire workflow is shown in Algorithm 2.

Algorithm 2. SNN Encoder Algorithm

Require: $S_{img/series}, factor, threshold$
SNN Encoder Layer:
$L \leftarrow length(S)$
$\Delta \leftarrow [0]$
for $t = 1$ to $L - 1$ **do**
  $\Delta[t] \leftarrow S[t+1] - S[t]$
end for
$threshold \leftarrow mean(\Delta) + factor \times std(\Delta)$
$output \leftarrow [0]$
for $t = 1$ to L **do**
  if $\Delta[t] \geq threshold$ **then**
    $output[t] \leftarrow 1$
  else
    $output[t] \leftarrow 0$
  end if
end for
FM/Mapping Layer:
$\hat{S}_1 \leftarrow FM(output) \oplus output$
$\hat{S}_2 \leftarrow Mapping(output) \oplus output$
return $\hat{S}_1, \hat{S}_2$

Due to the strong temporal information processing capabilities of SNN, we employ the Leaky Integrate-and-Fire (LIF) model to describe the neural dynamics of multi-modal information. The following formula can represent the dynamic equation for the LIF model under continuous-time sequences:

$$\tau_m \frac{dV(t)}{dt} = -(V(t) - V_{\text{rest}}) + R \cdot I(t) \tag{1}$$

$$V(t) = V(t-1) + \frac{1}{\tau}\left(I(t) - (V(t-1) - V_{\text{rest}})\right) \tag{2}$$

$V(t)$ is a membrane potential function concerning time t. V_{rest} represents the resting membrane potential of the neuron. τ_m is a constant that characterizes the charging and discharging rate of the neuron's membrane potential. $I(t)$ is the synaptic pulse input function. R denotes the membrane's responsiveness to input currents.

When the membrane potential $V(t)$ exceeds the threshold potential V_{th}, the neuron is activated and triggers a spike, denoted as $H(t)$. $\Theta(x)$ is the Heaviside step function, which is 1 when $x \geq 0$ and 0 otherwise. V_{th} represents the threshold potential. V_{reset} is the reset potential, to which the membrane potential is reset when the neuron is activated.

$$\begin{cases} H(t) = \Theta(V(t) - V_{\text{th}}) \\ V(t) = V_{\text{reset}} \end{cases} \tag{3}$$

A neuron receives multiple pulse signals. Their effects are not independent but accumulate within the neuron, leading to a sustained change in membrane potential. By controlling the pulse frequency and timing, neurons can integrate and encode input information over time. Assuming that N neurons generate multiple pulses at different time points, these pulse timings can be represented by a series of time sequences $\{t_1^{(i)}, t_2^{(i)}, ..., t_j^{(i)}\}$. The cumulative effect of multiple pulses can be expressed as $P(t) = \sum_{i=1}^{N} \sum_{j=1}^{j} f(t - t_j^{(i)})$.

$f(t)$ represents the Dirac Delta function, signifying the generation of a pulse at the firing time. $P(t)$ is the output of the cumulative effect of multiple pulses, which corresponds to the pulse output of the encoder $\hat{S}_1$ & $\hat{S}_2$. Algorithm 2 is the workflow of the SNN encoder.

3.2 Multi-modal Pulse Joint Learning Module

The pulse signals extracted from different encoders are first subjected to normalization and mapping operations before input into a unified pulse co-learning module. The pulse signals $\hat{S}_1$ and $\hat{S}_2$ obtained from two heterogeneous spaces are then transformed from the time domain to the frequency domain through Fourier transformation. Fourier transformation $FT(s)$ decomposes the signal into different frequency components, which aids in analyzing the frequency domain characteristics of different modal signals.

$$\hat{S}_1 = \text{Encoder}_i(S_{\text{image}}) \in \mathbb{R}^{N \times D_i} \tag{4}$$

$$\hat{S}_2 = \text{Encoder}_t(S_{\text{series}}) \in \mathbb{R}^{N \times D_t} \tag{5}$$

To better integrate and align the information from two different signal spaces, we introduce a joint learning function denoted as Ψ. This function aims to adjust the feature representations of the signals, mapping the signals from space D_i and space D_t to a common frequency domain space. During the training process, this function is continuously adjusted to make the pulse information in different modalities more consistent, achieving effective fusion and alignment of heterogeneous signals. J_{align} denotes the fusion of pulse information in the joint learning space. D_j is the dimension of joint learning space, where data from different modalities coexist in a shared representation.

$$FT(s) = \int_{-\infty}^{\infty} \hat{S} \odot e^{-iwT} \in \mathbb{R}^{N \times D_j} \tag{6}$$

$$J_{align} = \Psi\left(FT(\hat{S}_1), FT(\hat{S}_2)\right) \tag{7}$$

We introduce a more effective pulse-based joint weight allocation mechanism (JWAM). This mechanism involves mapping the similarity results in *sim* of multi-modal pulse signals into different spatial dimensions of the probability distribution matrix (P_{mtsa}). The similarity probability distribution is adaptively

adjusted based on the features of each modality and their relative importance to achieve information fusion. P_{mtsa} integrates information from various modalities, providing a quantitative method for scoring cross-modal information representation. sim is a metric function used to measure the similarity between two pulse information representations in heterogeneous spaces. This function employs the Euclidean distance calculation method to assess the similarity between different modality representations. σ^2 is used to adjust the sensitivity of the similarity measurement function. It is worth noting that it can dynamically adapt based on the distribution information of different modality features, enhancing the robustness and adaptability of similarity measurements.

$$sim = \exp\left(-\frac{|\sum_{dim=1}^{i,t}(\hat{S}_1 - \hat{S}_2)|}{2\sigma^2}\right) \tag{8}$$

$$P_{mtsa} = \frac{\exp\left(sim_i, sim_t\right)}{\sum_{dim=1}^{j} \exp\left(Sim_i, Sim_t\right)} \tag{9}$$

Furthermore, matrix transformations of the information in the joint space are utilized to adjust pulse signals. This operation aims to optimize the feature space while taking into consideration information from different modalities in order to better accommodate the characteristics of pulse sequences from other modalities. Additionally, we interact this process with cross-modal probability distributions to obtain the pulse fusion representation denoted as J_{fusion}. This can be expressed as:

$$J_{fusion} = \textbf{Softmax}\left(\frac{\hat{S}_{1/2} \odot J_{align}}{\sqrt{D_j}}\right) \odot P_{mtsa} \tag{10}$$

The Output layer is responsible for two major tasks: predicting and classifying information from multi-modal time series pulse fusion data. It employs network layer techniques such as residual connections and ReLU to transform the fused information into a common format, making it available for various downstream tasks.

4 Experiment

4.1 Datasets

We conduct experimental evaluations for classification and regression tasks on two traditional time series datasets, MIT-BIH Arrhythmia (MIT-BIH) [13] and Electricity Transformer Temperature (ETT) [14]. Additionally, we perform a market forecasting analysis on relevant stock indices of the Chinese stock market from June 6, 2013, to June 6, 2023, covering a ten-year period, focusing on the opening and closing prices.

4.2 Pulse Signal Processing Based on Wavelet Transform

To effectively address the non-stationary, non-linear characteristics and constraints in multi-scale feature analysis of time-series data, we employ the wavelet transform analysis method. Wavelet transform possesses exceptional time-frequency locality and multi-scale analysis capabilities, making it more suitable for capturing local features of signals at different time and frequency scales. The MTSA-SNN network based on wavelet transform can capture richer feature representations, endowing it with a significant advantage in handling non-stationary signals, extracting critical signal features, and analyzing signals across multiple scales.

MTSA-SNN employs wavelet transform to decompose input signals into four subbands: LL, LH, HH and HL, which represent distinct signal characteristics in terms of different frequencies and spatial scales. This multi-scale and multi-frequency analysis approach equips the MTSA-SNN model with a comprehensive understanding of multimodal data, enhancing its learning capabilities. As illustrated in Fig. 2 and Fig. 3, the temporal visualizations of these four subbands in the ETT and stock prediction datasets demonstrate the effectiveness of this multi-scale analysis.

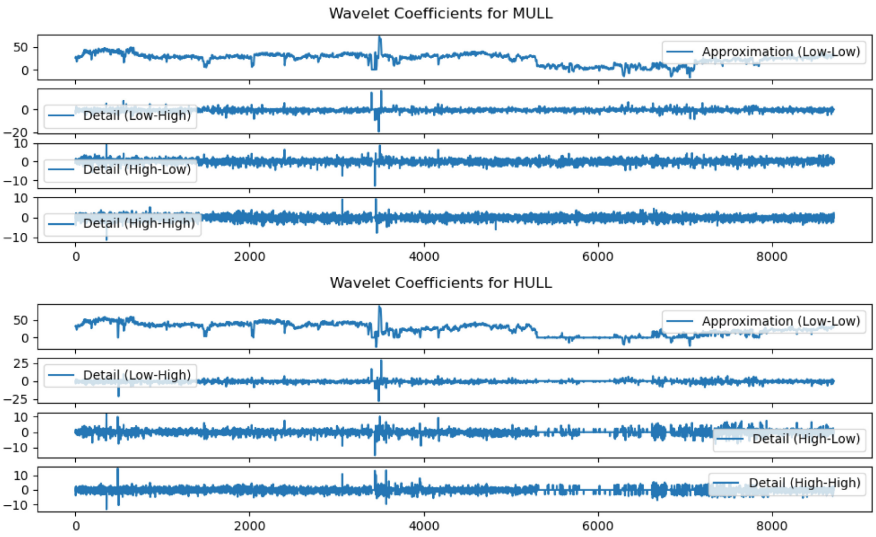


Fig. 2. ETT dataset signal features across different frequency and spatial scales. (LL captures low-frequency signal components. LH and HH capture high-frequency components in both low and high-frequency signals. HL contains low-frequency components of high-frequency signals.)

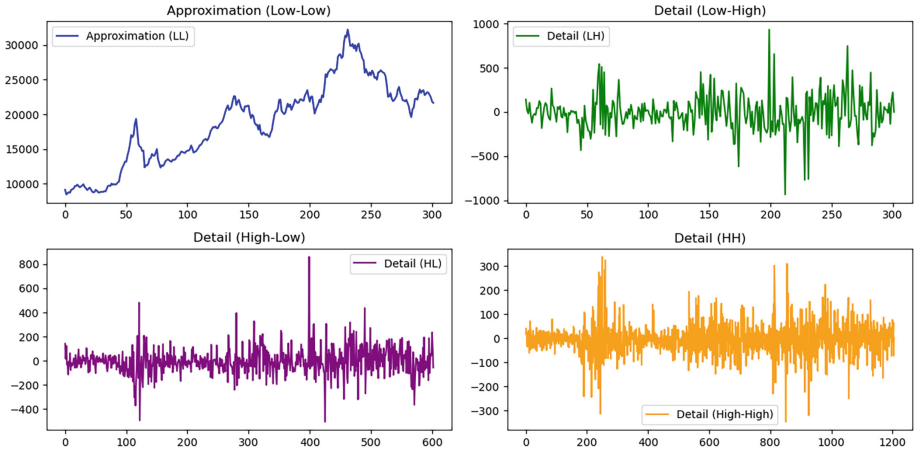


Fig. 3. Stock prediction dataset signal features across different frequency and spatial scales

Figure 4 depicts the pulse network outputs based on the MIT-BIH dataset with different processing methods. It is evident that the pulse output subjected to wavelet transform more accurately captures the features of multimodal signals, resulting in a more stable and effective neural activation.

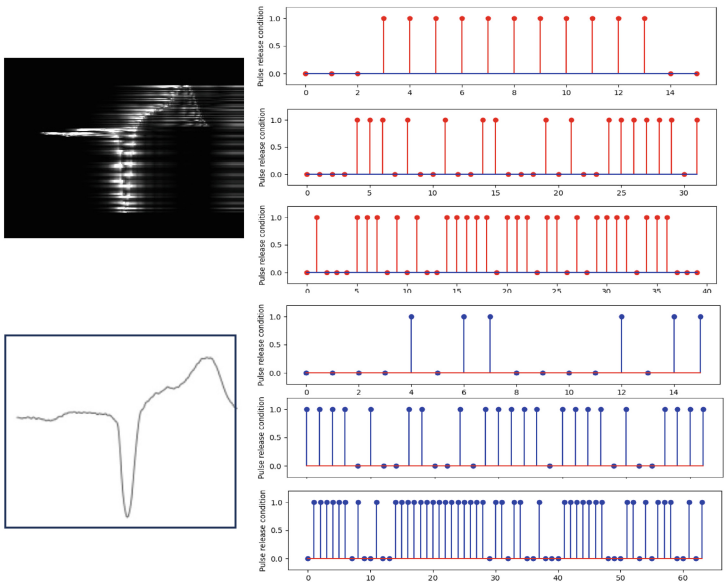


Fig. 4. Data (wavelet transform) converted into pulse signals by MTSA-SNN (above) & Original data converted into pulse signals by MTSA-SNN

4.3 Comparison with Other Methods

MTSA-SNN demonstrates remarkable performance advantages in the field of biological time-series data analysis. The experimental results in Table 4 demonstrates that our model has achieved advanced performance in the detection of cardiac arrhythmias in multimodal electrocardiogram data. With a dataset classification accuracy of 98.75%, MTSA-SNN markedly outperforms previous leading algorithms. This is attributable to the effective simulation of the neural signal conduction process in biological systems through MTSA-SNN's pulse-based fusion approach, resulting in performance advantages. Although the improvements in accuracy and precision seem incremental, these advancements are significant in biological time-series data analysis. Each percentage point increase can substantially reduce hundreds or even thousands of misdiagnoses, which is crucial for the medical field (Table 1).

Table 1. Comparison of MTSA-SNN with other methods on the MIT-BIH dataset

Network	Accuracy (%)↑	F1(%) ↑	Precision(%)↑
Mousavi et al. [12]	97.62	85.82	91.46
Yang et al. [15]	97.76	88.28	94.34
Hammad et al. [16]	98.00	89.70	86.55
Xing et al. [17]	98.26	89.09	–
Vision Transformer + Autoformer	98.38	91.09	88.43
Vision Transformer + Informer	98.54	93.50	92.47
MTSA-SNN (ours)	**98.75**	**94.31**	**94.62**

In addition, our method exhibits outstanding performance in various prediction tasks, including transformer temperature monitoring and stock market forecasting. Analyzing the results presented in Table 2, our model demonstrates the lowest MAE and MSE across four different time steps in the ETT dataset. Furthermore, in Table 3, MTSA-SNN achieves remarkably low errors of 0.96 and 1.15 in the stock market price prediction task compared to traditional time-series prediction models such as LSTM and XGBoost. MTSA-SNN, by converting complex and diverse multimodal time series data into a pulse-based representation, significantly enhances the model's predictive and analytical capabilities regarding time-series information.

Table 2. Comparison of MTSA-SNN with other methods on ETT dataset

Methods		NLinear [18]		DLinear [18]		Autoformer [19]		Informer [14]		MTSA-SNN (ours)	
Metric		MSE↓	MAE ↓	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT	96	0.374	0.394	0.375	0.399	0.449	0.459	0.865	0.713	**0.235**	**0.247**
	192	0.408	0.415	0.405	0.416	0.500	0.482	1.008	0.792	**0.345**	**0.371**
	336	0.429	0.427	0.439	0.443	0.521	0.496	1.107	0.809	**0.358**	**0.362**
	720	0.440	0.453	0.472	0.490	0.514	0.512	1.181	0.865	**0.396**	**0.439**

Table 3. Comparison of MTSA-SNN with other methods on Stock market price prediction dataset

Network	LSTM	XGBoost	LSTM-XGBoost	MTSA-SNN (ours)
MAE ↓	2.465	2.317	1.394	**0.961**
MSE ↓	2.839	2.285	1.461	**1.152**

4.4 Computational Costs and Speed Performance

Under identical operating conditions with a GPU RTX 3090 24GB, we conduct a comprehensive comparison between existing multimodal networks and our MTSA-SNN on the MIT-BIH dataset, including GPU memory usage, training speed, and inference speed. Training speed refers to the time required for forward and backward propagation of a batch of data, while inference speed indicates the time required for forward propagation of a batch of data, measured in milliseconds (ms). To reduce variance, we calculate the average time over 50 batches. The results are shown in Table 4.

Our MTSA-SNN demonstrates exceptional performance across all metrics. Its memory usage is significantly reduced to just 11.89 GB, while other methods range from 16.53 GB to 23.78 GB. In terms of training speed, MTSA-SNN achieves a notable improvement with 142 ms per batch, showcasing superior efficiency. For inference speed, our method also leads with 23 ms per batch, indicating high processing performance. These results suggest that MTSA-SNN has significant advantages in resource utilization and processing speed, proving its superiority in handling the MIT-BIH dataset.

Table 4. Memory usage and speed performance comparison of our method with existing methods on MIT-BIH dataset

Network	Memory(GB)	Training Speed (ms/batch)	Inference Speed (ms/batch)
Mousavi et al. [12]	22.13	210	50
Yang et al. [15]	19.54	234	53
Hammad et al. [16]	21.46	175	41
Xing et al. [17]	16.53	158	36
Vision Transformer + Informer	23.78	255	63
MTSA-SNN (ours)	**11.89**	**142**	**23**

4.5 Ablation Study

We conduct a comprehensive ablation study to evaluate different components of the MTSA-SNN model. As shown in Fig. 5, we present pulse signal output heatmaps for different components at the same time step using the MIT-BIH dataset. The brightness of the colours in the figure represents the activation levels of neurons.

In comparison to the activation patterns from single-modal encoders, the joint learning module of MTSA-SNN activates more neurons, thus enriching the representation of temporal information. Furthermore, the application of wavelet transform enhances the representation of temporal information within the MTSA-SNN. This suggests that joint learning of pulses effectively balances multi-modal pulse signals and fuses them together. Simultaneously, wavelet transformation contributes to enhancing the representation of temporal information in the pulse network.

In addition, we analyze the spectral information of the waveform plots during the training process of the single-modal encoder and the joint learning module. In Fig. 6, the horizontal axis represents the time steps, while the vertical axis represents the amplitude. This indicates that the MTSA-SNN model effectively integrates and analyzes multi-modal signals while enhancing the overall robustness of the model.

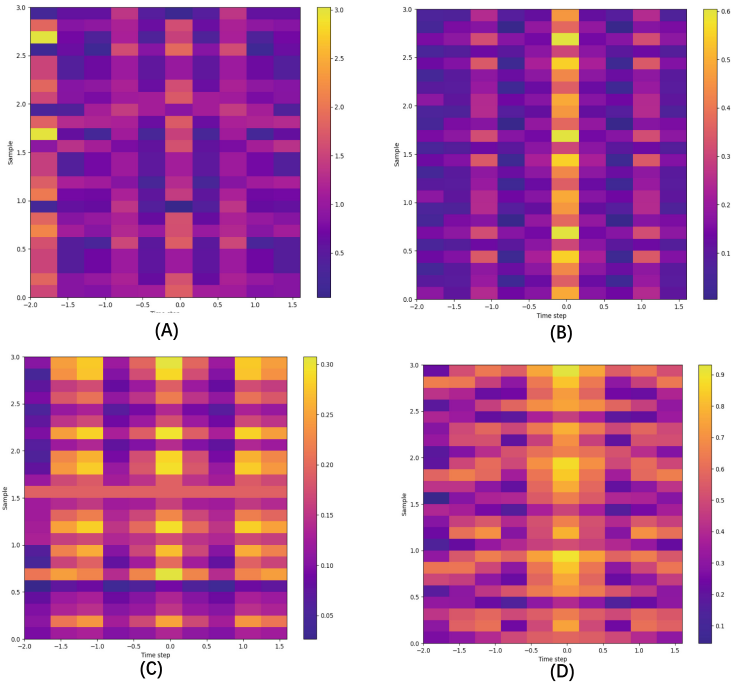


Fig. 5. The heatmap of MTSA-SNN's various component neuron activations. Specifically, (A) and (B) represent the neuron activation patterns after the time series information passes through the image encoder and sequence encoder of MTSA-SNN. (C) demonstrates the fused output after the joint learning process for the original temporal information. (D) represents the pulse fusion after applying wavelet transform in MTSA-SNN.

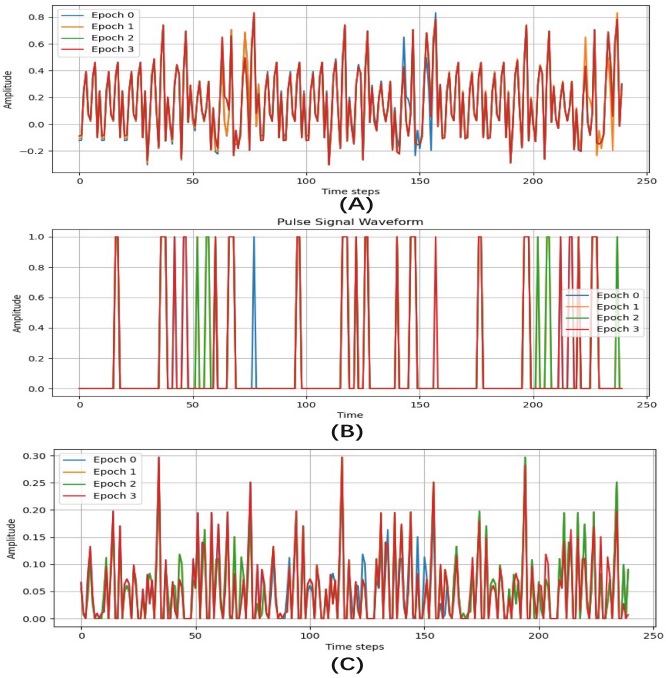


Fig. 6. The spectral analysis of waveform plots during training (The first four epochs). (A) and (B) show the waveforms generated by the single-modal encoder, revealing unstable characteristics of the pulse signals and relatively weak robustness in the individual modality. (C) shows the output of the MTSA-SNN model, exhibiting significant frequency domain stability as it consistently remains within a defined range of amplitudes.

5 Conclusion

In this paper, we introduce an innovative Multi-modal Time Series Analysis Model based on the Spiking Neural Network. The model's pulse encoder is designed to uniformly pulse-code multi-modal information. The pulse joint learning module is employed to effectively integrate complex pulse-encoded data. Additionally, we incorporate wavelet transform operations to enhance the model's capability to analyze and evaluate time series data. Experimental results on three distinct time series datasets demonstrate the outstanding performance of our proposed approach across multiple tasks.

References

1. Hua, Y., Zhao, Z., Li, R., Chen, X., Liu, Z., Zhang, H.: Deep learning with long short-term memory for time series prediction. IEEE Commun. Mag. **57**(6), 114–119 (2019)

2. Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., Arshad, H.: State-of-the-art in artificial neural network applications: a survey. Heliyon **4**(11) (2018)
3. Fang, H., Shrestha, A., Qiu, Q.: Multivariate time series classification using spiking neural networks. In: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7 (2020)
4. George, A.M., Dey, S., Banerjee, D., Mukherjee, A., Suri, M.: Online time-series forecasting using spiking reservoir. Neurocomputing **518**, 82–94 (2023)
5. Kong, L., et al.: Time-aware missing healthcare data prediction based on arima model. IEEE/ACM Trans. Comput. Biol. Bioinform. (2022)
6. Stein, E.M., Shakarchi, R.: Fourier Analysis: An Introduction, vol. 1. Princeton University Press, Princeton (2011)
7. Xian, Q., Liang, W.: A multi-modal time series intelligent prediction model. In: Qian, Z., Jabbar, M., Li, X. (eds.) Proceeding of 2021 International Conference on Wireless Communications, Networking and Applications. WCNA 2021. LNEE, pp. 1150–1157. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-2456-9_115
8. Zhang, M., Jiang, X., Fang, Z., Zeng, Y., Xu, K.: High-order hidden Markov model for trend prediction in financial time series. Phys. A **517**, 1–12 (2019)
9. Chen, Z., Wang, D.: Multi-initialization meta-learning with domain adaptation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 1390–1394 (2021)
10. Yamazaki, K., Vo-Ho, V.-K., Bulsara, D., Le, N.: Spiking neural networks and their applications: a review. Brain Sci. **12**(7), 863 (2022)
11. Li, S.C.-X., Marlin, B.: Learning from irregularly-sampled time series: a missing data perspectiv. In: International Conference on Machine Learning, PMLR, 2020, pp. 5937–5946 (2020)
12. Liu, Q., Long, L., Yang, Q., Peng, H., Wang, J., Luo, X.: LSTM-SNP: a long short-term memory model inspired from spiking neural p systems. Knowl.-Based Syst. **235**, 107656 (2022)
13. Moody, G.B., Mark, R.G.: The impact of the MIT-BIH arrhythmia database. IEEE Eng. Med. Biol. Mag. **20**(3), 45–50 (2001)
14. Zhou, H., et al.: Informer: beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 11106–11115 (2021)
15. Yang, W., Si, Y., Wang, D., Guo, B.: Automatic recognition of arrhythmia based on principal component analysis network and linear support vector machine. Comput. Biol. Med. **101**, 22–32 (2018)
16. Hammad, M., Iliyasu, A.M., Subasi, A., Ho, E.S., Abd El-Latif, A.A.: A multitier deep learning model for arrhythmia detection. IEEE Trans. Instrum. Meas. **70**, 1–9 (2020)
17. Xing, Y., Zhang, L., Hou, Z., Li, X., Shi, Y., Yuan, Y., Zhang, F., Liang, S., Li, Z., Yan, L.: Accurate ECG classification based on spiking neural network and attentional mechanism for real-time implementation on personal portable devices. Electronics **11**(12), 1889 (2022)
18. Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 9, pp. 11121–11128 (2023)
19. Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. Adv. Neural Inf. Process. Syst. (NIPS) **34**, 22419–22430 (2021)

PDET: Progressive Diversity Expansion Transformer for Cross-Modality Visible-Infrared Person Re-identification

Mingfu Xiong[1], Jingbang Liang[1], Yifei Guo[2], Ik Hyun Lee[3], Sambit Bakshi[4], and Khan Muhammad[5(✉)]

[1] School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, Hubei, China
xmf2013@whu.edu.cn

[2] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China
yifeiguo@hust.edu.cn

[3] Department of Mechatronics Engineering, Tech University of Korea, and IKLAB Inc., Seoul, South Korea
ihlee@tukorea.ac.kr

[4] Department of Computer Science and Engineering, National Institute of Technology Rourkela, Rourkela, India
bakshisambit@nitrkl.ac.in

[5] School of Convergence, Sungkyunkwan University, Seoul 03063, South Korea
khanmuhammad@g.skku.edu

Abstract. Visible-Infrared Person Re-identification (VI-ReID) would effectively improve the recognition performance in weak-lighting and nighttime scenes, which is an important research direction in pattern recognition and computer vision. However, existing methods usually focus on reducing the image differences between modalities (visible and infrared) to extract more reliable features, while neglecting the ability to discriminate the different identities with similar appearances. To address this problem, we propose a framework called "Progressive Diversity Expansion Transformer (PDET)", which includes a Diversity Distinguishing Vision Transformer Module (DDViTM) and a Cross-Modality Similarity Matching (CMSM) module for VI-ReID in this study. The DDViTM is proposed to implement the multiple embedded output vectors for a single input, learning feature representations of individual pedestrians in different modalities. The second module (CMSM) is used to improve the feature similarity between visible and infrared images, and dynamically adjust the image sequence weights of the two modalities to complete the training and optimization efficiency for the entire network. We conducted extensive experiments on the SYSU-MM01 and RegDB datasets, widely recognized public datasets for VR-ReID. The results demonstrate that the algorithm presented in this work has achieved promising performance compared to state-of-the-art methods. The code is available at https://github.com/jxsiaj/PEDT.git.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 439–454, 2025.
https://doi.org/10.1007/978-3-031-78341-8_28

Keywords: Transformer · Visible-Infrared Person Re-identification · Cross-modality Retrieval · Progressive Diversity Expansion

1 Introduction

Person Re-identification (ReID), aiming at discriminating whether pedestrians captured from various views of the cameras are the same one or not by analyzing (comparing) his/her appearance features, is a hotspot issue in the communities of pattern recognition and social security [1,35]. It is widely used in video surveillance [20], public safety [29], intelligent transportation [37], and other fields. Traditional ReID methods mainly rely on the pedestrians' appearance feature extraction (manually designed or deep features) and similarity measure manners in visible environments, which is not applied effectively in weak-lighting and nighttime conditions [2,20]. Therefore, the Visible-Infrared Re-identification (VI-ReID) has been proposed to combine both the visible and infrared images for personal appearance representation, which can provide clearer visual features in weak-lighting, nighttime, or complex environments, thus achieving better performance in ReID tasks [2]. Compared to the traditional single-modality-based ReID approaches, the VI-ReID methods aim to address this issue in complex scenes effectively, especially when presenting weak-lighting and nighttime conditions. Naturally, it also faces some new challenges, which include: (1) The large differences between modalities, (2) Visible images are more vulnerable during cross-modality feature matching, and (3) Publicly available datasets for VI-ReID are limited [23,42]. Consequently, this often results in unstable ReID performance in real-world scenarios, presenting an ongoing and formidable challenge yet to be fully addressed [3].

Existing VI-ReID approaches are mainly classified into two categories: (i) Non-generative-based methods [4]; and (ii) Generative-based methods [42]. The former is similar to the traditional approaches, i.e., feature extraction and distance learning manner to finish the VR-ReID task. The dual-stream network structure is usually designed to perform on the visible and infrared images separately [4,12,35]. While these methods effectively extract modality-specific features and leverage the complementary information from visible and infrared images to minimize modality differences, achieving shared features across modalities remains challenging. The generative-based methods mainly integrate the modality differences at the data-level, which generate cross-modality pedestrian samples via the modality shift strategies [6,21,31]. However, these manners are easy to increase the computational complexity and other noise image traits for real VI-ReID applications. Furthermore, current VI-ReID methods primarily minimize modality differences between visible and infrared images, overlooking the ability to distinguish different individuals with similar appearances. In addition, they usually exploit the dilated convolution with multiple branches in ResNet-50 to achieve a single input to produce multiple different outputs, which disrupt the performance of the model [14,44].

To address the above issues, we have proposed a novel framework called PDET: Progressive Diversity Expansion Transformer for Cross-Modality Visible-

Infrared Person Re-identification task in this study. The proposed framework includes a Diversity Distinguishing Vision Transformer Module (DDViTM) and a Cross-Modality Similarity Matching (CMSM) module. Specifically, to enhance the appearance discriminative ability between modalities, the DDViTM is proposed to implement the multiple embedded output vectors for a single input, learning features for the individual pedestrians' cross-modality-sharing representation. Furthermore, the CMSM module is used to improve the feature similarity between visible and infrared images for the same person, and dynamically adjust the image sequence weights of the two modalities to complete the training and optimization efficiency for the entire network. Our investigation extends across the VI-ReID public datasets, including (SYSU-MM01 [33] and RegDB [25]), serving as the litmus test for evaluating the efficacy of our approach. The outcomes underscore the promising strides made by our proposed method, outperforming contemporary algorithms. In essence, the key contributions of this work can be encapsulated as follows:

- We present a new PDET framework: Progressive Diversity Expansion Transformer for VI-ReID, which aims to increase the amount of data and reduce inter-modal differences implicitly.
- The proposed framework includes DDViTM and CMSM modules to enhance the ability of individual appearance discrimination between modalities and to adjust the image sequence weights of the two modalities for the entire network, respectively.
- Extensive experiments have been conducted on public VI-ReID datasets, namely, SYSU-MM01 and RegDB from a variety of aspects. Results show that the presented algorithm in this work has achieved a promising performance against SOTA methods.

2 Related Work

In this section, we have reviewed the relevant work for VI-ReID. Firstly, we review the overall research status of VR-ReID. After that, the transformer-based methods for person ReID are described. The most noteworthy works are given as follows.

2.1 Visible-Infrared-Based Methods for Person ReID

The purpose of VI-ReID is to retrieve the same person as the target sample in both modalities. Visible ReID is only for visible environments and is more sensitive to light in practical applications. To solve the problem that RGB images do not work at night, Wu et al. [33] provided the first in-depth study of VI-ReID and compared several feasible network structures. Also, they released a large-scale dataset called SYSU-MM01, with cross-modality, which is an important contribution to further research in this field. The cross-modality-sharing feature transfer module, namely, cm-SSFT is presented by Lu et al. [24] to explore the

shared information between modalities. They have employed a complementary feature learning strategy to acquire discriminative and complementary shared and modality-specific features. Gao et al. [12] used a joint optimization network with edge feature enhancement templates and edge fusion in multi-feature space to enhance modality-shared features within each modality. In [26], Park et al. introduced a Learning by Aligning (LBA) method to exploit the correspondences between visible and infrared images, which address the modality differences in the pixel-level for VR-ReID. Zhu et al. [44] learnt inter-modal invariant information between two heterogeneous modalities by controlling the intra-class centroid distance between the two modalities. Ye et al. [36] presented HAT (a homogeneous augmented three-modal learning method), which generates homogeneous grayscales approximating infrared maps using RGB images for three-modal learning. Fu et al. [11] thought that the batch normalization (BN) layer can improve the performance of cross-modality retrieval with proper separation. In addition, a modal obfuscation learning network is presented by Hao et al. [13] that no longer distinguishes between samples of different modalities during training to minimize the differences between modalities. Josef et al. [10] used semantic alignment to proofread the same features between modalities. Meanwhile, there are other methods to mitigate the inter-modal differences by directly processing the images.

2.2 Transformer-Based Methods for Person ReID

Transformer [30] was originally proposed to solve the problems of computationally difficult and long-term dependency of RNN networks in natural language. For example, Das et al. [7] improved the accuracy of language text recognition by using a transformer. Transformers have also been used in recent years in computer vision due to their excellent modeling capabilities as well as their ability to have better parallelism, faster training and reasoning, and can deal with longer sequence fields while achieving better results. For example, in single-modal Visible ReID, He et al. [15] improved the Vision Transformer (ViT) [9] using the camera viewpoint information, which is embedding the patch module to learn the global-local features for each person. Meanwhile, Zhu et al. [43] quickly localized image body parts and other parts in the patch, adding partial labeling vectors to learn partial features that can be directly discriminated. Zhang et al. [38] proposed a multilevel feature aggregation network combining CNN and transformer to achieve multiscale feature learning. Li et al. [22] proposed a more compact transformer that learns more robust and differentiated embeddings to recognize similar classes. Although the previously mentioned transformer-based approach achieves excellent performance in the ReID task, it falls short in adding diverse outputs and extracting modality-sharing features in VI-ReID.

3 The Proposed PDET Framework

This part introduces the details of the proposed PDET framework, which is described as Fig. 1. The detailed content is described as follows.

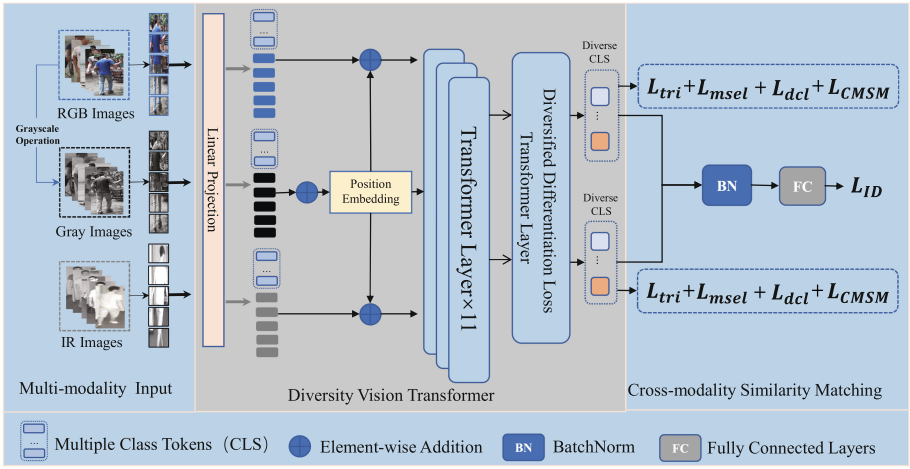


Fig. 1. Overall network structure of the proposed PDET. The left part is the multi-modality images input, which includes RGB, Grayscale, and IR (Infrared) images. The middle one is the Diversity Vision Transformer Module with the Multiple Class Tokens for the corresponding modality of the previous part and visual transformation. The loss functions are following at the right part for the framework's optimization.

3.1 Diversity Distinguishing Vision Transformer

Generally, increasing the amount of data can provide more samples for the model, allowing it to better adapt to various changes during the training process, which would learn more comprehensive and accurate feature representations to make the model more robust. Meanwhile, the model can recognize more details for the trained samples when increasing the amount of similar data, thereby reducing the overfitting of the model for noise and abnormal samples. Therefore, to achieve different output representations for the single input, we have proposed a Diversity Distinguishing Vision Transformer Module (DDViTM) with Multiple Class Tokens (CLS) to achieve this goal and get the same effect as increasing the amount of similar data as described previously. In [40], multiple inflated convolutions with mutually independent branches are used to produce multiple distinct outputs, but more branches also consume a lot of computational resources. Inspired by [15], we just need to improve the last stage of the ViT [9] model to achieve the effect of embedding a single input and outputting multiple embedding vectors in parallel, without the need for additional convolution calculations. In addition, the ViT model, being pre-trained on ImageNet [8], is used as our backbone.

Without changing the pre-trained ViT weights, we set three different modality images as the input. Each of them can be represented as $X \in R^{W \times H}$ as the import image, where W and H denote the width and height of the input image, respectively. Then P patches are set as the same size, which is denoted as $\{x^i | i = 1, .., i..., P\}$. Sequentially, the divided patches are all mapped into the d-dimensional vector via the linear mapping layer of the pre-trained ViT model,

which is denoted as $F(x^i)$. Then we have exploited the combination strategy to concatenate the CLS labels and patch embeddings to form a series of feature vectors $\{x^{cls1}, ..., x^{clsN}, F\left(x^1\right), ..., F(x^P)\}$, and this process is described in the middle part of Fig. 1. N is the number of the CLS. At last, the added learnable position embedding sequences are used to obtain the diversified representation.

In addition, due to the multi-attention mechanism, our class tokens are transparent to each other. They not only collect information from different patches but from other CLS among themselves, thereby improving the efficiency of information collection. After that, we believe that each CLS does not overlap with each other, which would learn different features. Maximizing the differences between CLS can result in boosting the model's discriminative ability to distinguish similar classes and reduce the amount of computation. Therefore, we have proposed a Diversity Discrimination Loss (DDL) with dynamic adjustment in the last layer of the ViT model to push the distance between CLS. Specifically, we have used the paired CLSs to perform an orthogonal operation for each other to achieve the de-correlation process, i.e., $\cos(\mathrm{x}^{\mathrm{cls1}}, \mathrm{x}^{\mathrm{cls2}}) = 0$. However, as the number of CLS increases, each one needs to achieve cosine similarity minimization, which leads to the complexity being high, making the model's optimization process difficult due to the high cosine similarity of some CLS. To solve this problem, we propose a dynamic adjustment strategy, which recalculates the weights of the softmax function for each pair of CLS. The CLS markers with higher similarity gain higher weights, making the model pay more attention to them. The DDL and dynamic weight adjustment strategy are described as Eq. (1) and Eq. (2).

$$\mathrm{L}_{DDL} = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} |\cos(f^i, f^j)|, i < j, \tag{1}$$

$$w_{ij} = \frac{\exp(|\cos(f^i, f^j)|)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \exp(|\cos(f^i, f^j)|)}, \tag{2}$$

where f^i and f^j denote the i-th and j-th CLS, respectively. w_{ij} indicates the dynamic adjustment of the weights of each pair of CLS. N represents the number of the CLS.

3.2 Cross-Modality Similarity Matching Module

Although the proposed DDViTM would provide diverse expressions of individual appearance features, it also faces an important challenge, which involves the cross-modality pedestrian matching in VI-ReID. Inspired by PMT [23], we also exploit a two-stage strategy to achieve the cross-modality similarity matching for VI-ReID. Specifically, a cross-modality similarity matching loss is proposed which is utilized to calculate the similarity distribution between the cross-modality (visible and infrared) images, correlating representations with different modalities. It is a more comprehensive and richer sample representation to combine the cross-modality information. The details are described as follows.

Generally, the visible and infrared images are represented as x^{vis} and x^{ir} respectively. The corresponding grayscale image of the visible image is denoted as x^{gray}. The sample images $\left\{x^{vis}, x^{ir}, x^{gray}\right\}$ are input into the diversified expansion Transformer $F(\cdot)$ to obtain the corresponding feature representation, respectively. The process is described as Eq. (3)–(5).

$$\{f^{v1}, f^{v2}, \cdots, f^{vi}, \cdots, f^{vN}\} = F(x^{vis}), \tag{3}$$

$$\{f^{g1}, f^{g2}, \cdots, f^{gi}, \cdots, f^{gN}\} = F(x^{gray}), \tag{4}$$

$$\{f^{ir1}, f^{ir2}, \cdots, f^{iri}, \cdots, f^{irN}\} = F(x^{ir}), \tag{5}$$

where f^{vi}, f^{gi}, and f^{iri} are the corresponding feature vectors of visible, grayscale, and infrared images, respectively. The size of CLS is represented as N. Different from existing methods that just consider the representations within modalities, we have introduced a novel cross-modality similarity matching loss. Firstly, a label-based similarity matrix is calculated, and then the visible image vector is projected onto the infrared image space, meanwhile the infrared image vector is also projected into the visible one. Next, according to the projected vector and similarity matrix, we calculated the loss values between the two processes (denoted as L_{v2ir} and L_{ir2v}), which are added together to get the final L_{CMSM} loss. The above process is described as Eq. (6) and (7).

$$L_{cmsm} = L_{v2ir} + L_{ir2v}, \tag{6}$$

$$L_{CMSM} = \frac{1}{N}\sum_{i=1}^{N} {L^i}_{cmsm}, \tag{7}$$

where L_{v2ir} and L_{ir2v} denote the mutual projection loss between visible and infrared images. Minimizing the loss would make similar visible and infrared image vectors closer during cross-modality retrieval. Inspired by [18], we have selected M visible-infrared image pairs, which are represented as $\{(f_i^v, f_j^{ir}), y_{ij}\}_{j=1}^{M}$, where y_{ij} are the true matching labels, $y_{ij} = 1$ indicates that (f_i^v, f_j^{ir}) is the same identity, and $y_{ij} = 0$ indicates a different one. The probability of cross-modality matching (p_{ij}) is then calculated by softmax function as Eq. (8).

$$p_{ij} = \frac{\exp(\cos({f_i}^v, {f_j}^{ir})/\gamma)}{\sum_{k=1}^{M}\exp(\cos({f_i}^v, {f_k}^{ir})/\gamma)}, \tag{8}$$

where γ is a hyper-parameter that controls the probability distribution peaks. Subsequently, the CMSM loss is computed from the visible image to the infrared image within a mini-batch as Eq. (9).

$$L_{v2ir} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}+\sigma}\right), \tag{9}$$

where σ is introduced for avoid numerical problems, and $q_{ij} = \frac{y_{ij}}{\sum_{k=1}^{M} y_{ik}}$ represents the true matching probability. Symmetrically, the CMSM loss from infrared to visible L_{ir2v} can be formulated by exchanging f^v and f^{ir}.

3.3 Model Optimization

The staged loss function is proposed to optimize the whole model. Firstly, the identity loss L_{ID} [41] and triplet loss L_{tri} [23] are utilized to learn modality-independent features. L_{DDL} constrained diversity representation and L_{CMSM} are also utilized to increase the correct probability of cross-modality matching. The process can be described as Eq. (10) and Eq. (11).

$$\mathrm{L}_1 = \frac{1}{N}\sum_{i=1}^{N}\left(L^i{}_{ID} + L^i_{tri}\right) + \lambda_1 L_{CMSM} + \lambda_2 L_{DDL}. \tag{10}$$

In the second stage, we further use diversity L_{msel} [23] to extract reliable modal-sharing features and enhance the discrimination with diversity L_{dcl} [23].

$$\mathrm{L}_2 = \frac{1}{N}\sum_{i=1}^{N}\left(L^i{}_{ID} + L^i_{tri}\right) + \lambda_2 L_{DDL} + \frac{1}{N}\sum_{i=1}^{N}\lambda_3(L^i_{msel} + L^i_{dcl}). \tag{11}$$

4 Experimental Results and Analysis

In this section, we provide the experimental results and relevant analysis. Firstly, we introduce the datasets, evaluation protocols, implementation details, and parameter settings in Subsect. 4.1. Subsequently, the ablation study is detailed in Subsect. 4.2. At last, the comparison with the latest SOTA methods is described in Subsect. 4.3. The details are described as follows.

4.1 Datasets and Evaluation Protocols

Datasets: To validate the performance of our method, extensive experiments were conducted on widely-used public VR-ReID datasets, including, **SYSU-MM01** [33] and **RegDB** [25].

- The **SYSU-MM01** [33] includes 491 person identities with 286, 628 visible and 15, 792 infrared images, respectively. Out of this data, 395 identities with 22, 258 and 11, 909 for visible and infrared images, respectively, are used to train the model. The remaining data is used as the test set with 3803 infrared, and 301 visible images for querying and gallery, respectively. In addition, the dataset exploits all search images and indoor ones for testing.
- The **RegDB** [25] dataset includes 412 pedestrian identities. There are both 10 visible and infrared images for each person. Without loss of generality, we divided the total pedestrian identities in this dataset evenly into two halves, with half used as the training set and the other half as the testing set.

Evaluation Protocols: As usual, we also exploited the Cumulative Matching Characteristics (CMC) [20] and mean Average Precision (mAP) [16] as the performance evaluation metrics. In addition, the Mean Inverse Negative Penalty (mINP) [35] strategy is also used in this study for performance evaluation.

Implementation Details: The proposed algorithm runs on the NVIDIA RTX-2080Ti GPUs, which are set as the hardware platform. The software environment is based on Python 3.8.10 and the Ubuntu 20.04 operating system. ImageNet is used to pre-train the ViT-B/16 model and the overlap step set as 12. The input images are resized as 256×128. There are 4 different identities with 64 batch sizes for each one. In addition, both 4 visible and infrared images have been selected for one person's identity.

Parameter Settings: Following [23], the AdamW optimizer of the learning rating (lr) is set as 3×10^{-4}. Weight decay was set as 1×10^{-4}. The training epochs were set as 20 (SYSU-MM01) and 58 (RegDB), respectively. Besides, the training epoch was set at 6 on both the datasets in the first stage. The parameters λ_1 and λ_3 are set as 0.7. λ_2 is set to 1, and the margin parameter m is set to 0.1. Hyper-parameter γ and σ are set as 50 and 1×10^{-8}, respectively. At the testing phase, all the output features are spliced into 3084 dimensions.

4.2 Ablation Studies

To verify the effectiveness of each component of our proposed PDET, we have conducted a series of ablation studies on the above two public datasets described below.

Effectiveness of the CLS: To verify the efficiency of the CLS, we have added multiple extra CLS tokens (denoted as "cls") and exploited the loss function L_{DDL} to constrain the CLS tokens to learn different features in Sect. 3.1. From Table 1, it can be seen that CLS indeed have better performance, and the result of the method can be further improved by adding the L_{DDL} loss function. In addition, to further verify how many extra "cls" are more effective, we also conducted the corresponding experiments. It can be seen that adding 4 "cls" is more effective in Table 1.

Table 1. Ablation experiments with CLS on the RegDB dataset. 'cls' denotes the extra added CLS token and L_{DDL} is the loss function that constrains CLS to learn different features.

Method	ViT	cls	cls	cls	cls	L_{DDL}	mAP ↑	mINP↑	rank-1↑
Baseline	√	√					76.55	62.90	84.83
DDViTM	√	√					78.76	64.07	85.19
	√	√	√	√			80.72	66.44	87.23
	√	√	√	√	√		79.76	65.11	86.75
	√	√	√			√	79.56	63.92	87.14
	√	√	√	√		√	81.20	66.85	86.89
	√	√	√	√	√	√	**81.58**	**67.69**	**88.16**

Impact of Loss Functions: To evaluate the impact of different loss functions on VI-ReID performance, we have also conducted the ablation experiments on the RegDB dataset, which is shown in Table 2. We have assessed the following

loss functions: Identity Loss (L_{ID}), Triplet Loss (L_{tri}), Modal Sharing Enhancement Loss (L_{msel}), Discriminant Center Loss(L_{dcl}) and Cross-modality similarity Matching Loss(L_{CMSM}). As can be seen from the final result, our unification of all the above loss functions has an essential impact on the model's performance.

Table 2. Performance of different loss functions on the RegDB dataset.

L_{ID}	L_{tri}	L_{msel}	L_{dcl}	L_{CMSM}	mAP↑	mINP↑	rank-1↑
√	√	√	√		76.55	62.90	84.83
√	√				71.18	57.27	75.63
√	√			√	75.44	60.53	81.50
√	√	√	√	√	**82.90**	**68.16**	**90.10**

Performance of DDViTM and CMSM: To verify the effect of DDViTM and CMSM included in the proposed framework in this study, we have also conducted the corresponding ablation experiments for each of them. The specific results are shown in Table 3. From the table, the two models are stacked together, where the performance of mAP and mINP slightly decreases due to an increase in feature duplication on the SYSU dataset. It may be due to inconsistent optimization objectives between the two modules, resulting in an increase in model complexity. Further investigation is needed in this regard, but due to time constraints, it is beyond the scope of current research. We will further address this issue in future research.

Table 3. Performance of the DDViTM and CMSM on the SYSU and RegDB datasets.

Baseline	DDViTM	CMSM	SYSU			RegDB		
			mAP↑	mINP↑	rank-1↑	mAP↑	mINP↑	rank-1↑
√			64.98	51.86	67.53	76.55	62.90	84.83
√	√		**66.43**	**54.07**	67.85	81.58	67.69	88.16
√		√	65.84	53.86	67.25	82.05	66.41	88.93
√	√	√	66.42	53.80	**68.36**	**82.90**	**68.16**	**90.10**

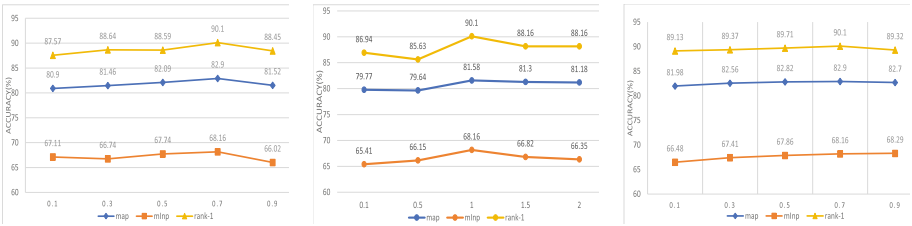
Computational Complexity: To demonstrate the computational complexity of our method, we also conducted a series of experiments to verify it. We mainly validate the model in terms of several dimensions such as the size of parameters, FLOPs, Throughput, and Inference Time. The specific results are shown in Table 4. From the table, it can be seen that although our method has an increase in the number of parameters compared to the baseline, this does not significantly impact computational efficiency. By adding more CLS, we capture richer feature representations, which require more parameters and time to learn these shared features. Despite the increase in computational complexity in some

Table 4. Computational complexity of PEDT. 'Throughput' is the amount of data processed per second. 'Inference time' refers to the time required to make predictions for a single batch of input data.

Model	Parameters (M)↓	FLOPs (G)↓	Throughput(B/S)↑	Inference Time(S)↓
Baseline	85.97	18.08	0.84	1.18
Ours	93.25	18.34	0.72	1.41

areas, these additional computational resources significantly enhance the model's performance as verified from Table 5 and Table 6.

Parameters Analysis:. In this part, we discuss the effect of parameters for the proposed model. It mainly includes three hyper-parameters: λ_1 (controlling the CMSM), λ_2 (constraining the CLS), and λ_3 (constraining the loss of Modal Sharing Enhancement), which are shown in Fig. 2. When λ_1 is small, its gradient may be too weak and the model will not fully optimize this loss term during training, which reduces the performance of the model. When $\lambda_1 = 0.7$, the correct rate of cross-modality matching increases significantly, which is illustrated in Fig. 2a. When $\lambda_2 = 0.1$, it does not play a role. The main reason is that the weights are too small, and the features learned by the CLS are too similar, resulting in a lack of differences. With the value increased, each CLS learns more and more diverse features, and its performance is optimal when $\lambda_2 = 1$. If its value continues to increase, too large weight makes it difficult for CLS to converge on the optimization, failing to learn the optimal features. Finally, we have discussed the effect of λ_3 for model performance. Due to the extra CLS in our model, it requires larger weights compared to the baseline to learn shared features between the two modalities. When it is set to 0.7, the entire model achieves the best accuracy.



(a) Impact of λ_1 (b) Impact of λ_2 (c) Impact of λ_3

Fig. 2. Impacts of λ_1, λ_2, and λ_3

Visual Retrieval Results: To visually verify the effectiveness of the algorithm, we have visualised the results. The top 10 retrieval results are displayed in Fig. 3 to verify the visual effect for the proposed model. It can be seen that we can quickly recognize the same person in the first few cases.

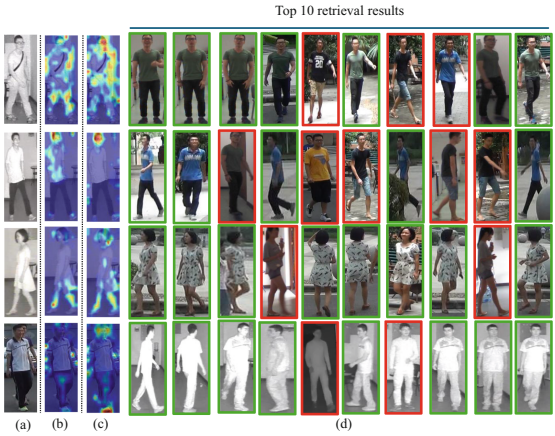


Fig. 3. Visualization representations. (a) Query images. (b) The map of PMT [23] with the attention operation. (c) Attention map of the PDET. (d) Top 10 retrieval results.

4.3 Comparison with the Latest Methods

In this part, we have discussed the corresponding experimental results on the above two datasets and compared them with the latest SOTA methods.

Comparisons on SYSU-MM01 Dataset: In Table 5, we have summarized the performance of our method and that of the SOTA competitors on the SYSU-MM01 dataset. For the benchmark dataset, 13 latest methods are proposed, including Zero-Pad [33], Hi-CMD [5], HAT [36], AGW [35], NFS [4], CM-NAS [11], LBA [26], MID [17], SPOT [2], FMCNet [39], PMT [23], DARD [32], and CMRViT [28]. It can be seen that our approach achieved relatively competitive results. However, our model is trained on an "all" scenario, where the outdoor background is more complex and the indoor background is relatively simple. The model is better at handling complex backgrounds, but not as good as the DARD method [32] for simple backgrounds. The reason for our analysis is that we have added "CLS" in the "all" scenario. The model is trained on more diverse data, and each "CLS" can learn different feature representations. In the "indoor" scenario, it contains only indoor data, the model does not have enough diverse data to learn. The singularity of the data in the "indoor" scenario leads to the fact that there is still room for further improvement of our algorithms, and this will be the next problem we need to solve.

Comparisons on RegDB Dataset: Correspondingly, we have also compared our method with the 12 SOTA methods (namely, Zero-Pad [33], DDAG [34], HAT [36], AGW [35], NFS [4], MSO [12], CM-NAS [11], LBA [26], SPOT [2], PMT [23], PartMix [19], and PAPG [27]) on RegDB dataset, which is depicted in Table 6. The method proposed in this study exhibits notable accuracy advantages, achieving a Rank-1 accuracy of 90.10% and an mAP of 82.90%.

Table 5. Comparison results on the SYSU-MM01 dataset.

SYSU-MM01		All search					Indoor search				
Method	Venue	r=1	r=10	r=20↑	mAP↑	mINP↑	r=1	r=10	r=20↑	mAP↑	mINP↑
Zero-Pad [33]	ICCV'17	14.80	54.12	71.33	15.95	–	20.58	68.38	85.79	26.62	–
Hi-CMD [5]	CVPR'20	34.94	77.58	–	35.94	–	–	–	–	–	–
HAT [36]	TIFS'20	55.29	92.14	97.36	53.89	–	62.10	95.75	99.20	69.37	–
AGW [35]	TPAMI'21	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
NFS [4]	CVPR'21	56.91	91.34	96.52	55.45	–	62.79	96.53	99.07	69.79	–
CM-NAS [11]	ICCV'21	61.99	92.87	97.25	60.02	–	67.01	97.02	99.32	72.95	–
LBA [26]	ICCV'21	55.41	91.12	–	54.14	–	58.46	94.13	–	66.33	–
MID [17]	AAAI'22	60.27	92.90	–	59.40	–	64.86	96.12	–	70.12	–
SPOT [2]	TIP'22	65.34	92.73	97.04	62.25	48.86	69.42	96.12	99.12	74.63	70.48
FMCNet [39]	CVPR'22	66.34	–	–	62.51	–	68.15	–	–	74.09	–
PMT [23]	AAAI'23	67.53	95.36	98.64	64.98	51.86	71.66	96.73	99.25	76.52	72.74
DARD [32]	TIFS'23	68.33	94.32	97.52	65.65	–	**77.21**	**98.32**	99.18	**81.91**	–
CMRViT [28]	PR'24	68.05	**97.12**	96.87	65.17	53.54	72.43	97.16	98.89	77.58	72.19
PDET	**Ours**	**68.36**	96.01	**98.81**	**66.42**	**53.80**	74.45	97.80	**99.48**	79.36	**75.99**

Table 6. Comparison Results on the RegDB Dataset.

RegDB		Visible to infrared					Infrared to visible				
Method	Venue	r=1	r=10	r=20↑	mAP↑	mINP↑	r=1	r=10	r=20↑	mAP↑	mINP↑
Zero-Pad [33]	ICCV'17	17.74	34.21	44.35	18.90	–	16.63	34.68	44.25	17.82	–
DDAG [34]	ECCV'20	69.34	86.19	91.49	63.46	49.29	68.06	85.15	90.31	61.80	48.62
HAT [36]	TIFS'20	71.83	87.16	92.16	67.56	–	70.02	86.45	91.61	66.30	–
AGW [35]	TPAMI'21	70.05	86.21	91.55	66.37	50.19	70.49	87.12	91.84	65.90	51.24
NFS [4]	CVPR'21	80.54	91.96	–	72.10	–	77.95	90.45	–	69.79	–
MSO [12]	ACM '21	73.6	88.6	–	66.9	–	74.6	88.7	–	67.5	–
CM-NAS [11]	ICCV'21	84.54	95.18	97.85	80.32	–	82.57	94.51	97.37	78.31	–
LBA [26]	ICCV'21	74.17	–	–	67.64	–	72.43	–	–	65.46	–
SPOT [2]	TIP'22	80.35	93.48	96.44	72.46	56.19	79.37	92.79	96.01	72.26	56.06
PMT [23]	AAAI'23	84.83	–	–	76.55	–	84.16	–	–	75.13	–
PartMix [19]	CVPR'23	85.66	–	–	82.27	–	84.93	–	–	82.52	–
PAPG [27]	SPL'24	88.35	–	–	83.18	–	86.46	–	–	80.08	–
PDET	**Ours**	**89.95**	**97.14**	**98.88**	**83.69**	**70.86**	**90.10**	**97.48**	**99.17**	**82.90**	**68.16**

5 Conclusion

This paper proposed a framework called Progressive Diversity Expansion Transformer (PDET), which includes a diversity distinguishing vision transformer module (DDViTM) and a cross-modality similarity matching (CMSM) module for VI-ReID. The DDViTM was used to enhance the ability of individual appearance discrimination between modalities, thereby learning feature representations of individual pedestrians in different modalities. Subsequently, the CMSM was introduced to mitigate inter-modality disparities, resulting in enhanced accuracy for cross-modality retrieval. Comprehensive evaluations across various metrics demonstrate that our approach has delivered promising performance compared to state-of-the-art methods, thus affirming its efficacy. In future, our aim is to

develop a lighter yet more efficient model to identify modality-shared features, effectively addressing the cross-modality differences in VR-ReID.

Acknowledgments. This work was supported by the Science Foundation of Hubei Province (No. 2021CFB568) and the National Natural Science Foundation of China (No.62371350).

References

1. Behera, N.K.S., Sa, P.K., Muhammad, K., Bakshi, S.: Large-scale person re-identification for crowd monitoring in emergency. IEEE Trans. Autom. Sci. Eng. 1–9 (2023)
2. Chen, C., Ye, M., Qi, M., Wu, J., Jiang, J., Lin, C.W.: Structure-aware positional transformer for visible-infrared person re-identification. IEEE Trans. Image Process. **31**, 2352–2364 (2022)
3. Chen, X., et al.: Oh-former: omni-relational high-order transformer for person re-identification. arXiv preprint arXiv:2109.11159 (2021)
4. Chen, Y., Wan, L., Li, Z., Jing, Q., Sun, Z.: Neural feature search for RGB-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 587–597 (2021)
5. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: Hi-cmd: hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10257–10266 (2020)
6. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: International Joint Conference on Artificial Intelligence, vol. 1, p. 6 (2018)
7. Das, A., Biswas, S., Banerjee, A., Lladós, J., Pal, U., Bhattacharya, S.: Harnessing the power of multi-lingual datasets for pre-training: towards enhancing text spotting performance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 718–728 (2024)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
9. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
10. Farooq, A., Awais, M., Kittler, J., Khalid, S.S.: Axm-net: implicit cross-modal feature alignment for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 4477–4485 (2022)
11. Fu, C., Hu, Y., Wu, X., Shi, H., Mei, T., He, R.: Cm-nas: cross-modality neural architecture search for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11823–11832 (2021)
12. Gao, Y., et al.: MSO: multi-feature space joint optimization network for RGB-infrared person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 5257–5265 (2021)
13. Hao, X., Zhao, S., Ye, M., Shen, J.: Cross-modality person re-identification via modality confusion and center aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16403–16412 (2021)

14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
15. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15013–15022 (2021)
16. Henderson, P., Ferrari, V.: End-to-end training of object class detectors for mean average precision. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10115, pp. 198–213. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54193-8_13
17. Huang, Z., Liu, J., Li, L., Zheng, K., Zha, Z.J.: Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 1034–1042 (2022)
18. Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2787–2797 (2023)
19. Kim, M., Kim, S., Park, J., Park, S., Sohn, K.: Partmix: regularization strategy to learn part discovery for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18621–18632 (2023)
20. Leng, Q., Ye, M., Tian, Q.: A survey of open-world person re-identification. IEEE Trans. Circuits Syst. Video Technol. **30**(4), 1092–1108 (2019)
21. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 4610–4617 (2020)
22. Li, W., et al.: DC-Former: diverse and compact transformer for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1415–1423 (2023)
23. Lu, H., Zou, X., Zhang, P.: Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1835–1843 (2023)
24. Lu, Y., et al.: Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13379–13389 (2020)
25. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)
26. Park, H., Lee, S., Lee, J., Ham, B.: Learning by aligning: visible-infrared person re-identification using cross-modal correspondences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12046–12055. IEEE (October 2021)
27. Qian, Y., Tang, S.K.: Pose attention-guided paired-images generation for visible-infrared person re-identification. IEEE Signal Process. Lett. **31**, 346–350 (2024)
28. Sarker, P.K., Zhao, Q.: Enhanced visible-infrared person re-identification based on cross-attention multiscale residual vision transformer. Pattern Recogn. **149**, 110288 (2024)
29. Tahboub, K.: Person Re-identification and Intelligent Crowdsourcing with Applications in Public Safety. Ph.D. thesis, Purdue University (2017)
30. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)

31. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3623–3632 (2019)
32. Wei, Z., Yang, X., Wang, N., Gao, X.: Dual-adversarial representation disentanglement for visible infrared person re-identification. IEEE Trans. Inf. Forensics Secur. (2023)
33. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: RGB-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380–5389 (2017)
34. Ye, M., Shen, J., J. Crandall, D., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12362, pp. 229–247. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58520-4_14
35. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2872–2893 (2021)
36. Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Trans. Inf. Forensics Secur. **16**, 728–739 (2020)
37. Yu, Z., Huang, Z., Pei, J., Tahsin, L., Sun, D.: Semantic-oriented feature coupling transformer for vehicle re-identification in intelligent transportation system. IEEE Trans. Intell. Transp. Syst. (2023)
38. Zhang, G., Zhang, P., Qi, J., Lu, H.: Hat: hierarchical aggregation transformers for person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 516–525 (2021)
39. Zhang, Q., Lai, C., Liu, J., Huang, N., Han, J.: FMCNet: feature-level modality compensation for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7349–7358 (2022)
40. Zhang, Y., Wang, H.: Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2153–2162 (2023)
41. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned cnn embedding for person reidentification. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **14**(1), 1–20 (2017)
42. Zhong, X., Lu, T., Huang, W., Ye, M., Jia, X., Lin, C.W.: Grayscale enhancement colorization network for visible-infrared person re-identification. IEEE Trans. Circuits Syst. Video Technol. **32**(3), 1418–1430 (2021)
43. Zhu, K., et al.: AAformer: auto-aligned transformer for person re-identification. IEEE Trans. Neural Netw. Learn. Syst. (2023)
44. Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X., Tao, D.: Hetero-center loss for cross-modality person re-identification. Neurocomputing **386**, 97–109 (2020)

DTMIReID: Person Re-identification Based on Deformable Transformer to Incorporate Mutual Information Between Images

Han Yang, Haodi Feng(✉), and Xuefeng Cui

Shandong University, Jinan 250100, Shandong, People's Republic of China
fenghaodi@sdu.edu.cn

Abstract. Person Re-identification (ReID) aims to retrieve a target pedestrian from an image gallery captured by cameras in varied scenarios. It is crucial for ReID to extract extensive discriminative feature representations from images for achieving desirable performance. The majority of current methods focus on mining data that can identify a pedestrian from a single image by investigating different dimensions of the image. However, a single image is sometimes insufficient to precisely characterize all the necessary features for identifying a pedestrian especially when the data quality is not guaranteed. Since a pedestrian tends to be caught in numerous images, information missed in a single image is expected to be supplemented from other images. Therefore, we consider extracting more robust feature representations benefiting from relationships between multiple pedestrian images and propose a new method DTMIReID. Firstly, we suggest a Dual Branch Attention Module (DBAM) based on Transformer to extract global and local features from single images. Then we combine the extracted features of multiple images together and input them into our proposed Deformable Transformer Module (DTM) to simultaneously fuse the global and local features from these multiple images by a Sample-Points-Based Attention (SPBA) mechanism. To the best of our knowledge, our method is the first ReID model that uses the Deformable Transformer to establish relationships between multiple features. Experimental results on four large ReID datasets show that the new method outperforms state-of-the-art published works by a large margin. DTMIReID is available at https://github.com/Titaniumyh/DTMIReID.git.

Keywords: Person Re-identification · Deformable transformer · Sample-points-based attention

1 Introduction

Person re-identification(ReID) aims to retrieve the target pedestrian from the image gallery. The pedestrian images in the gallery are captured by a set of cameras with non-overlapping shooting ranges. ReID is extensively applied in suspects searching, target individuals identification, and other scenarios. However, lighting changes, pedestrian posture shifting, filming angle varying, blurring, occlusion, and other unexpected issues make ReID a great challenge. The majority of current research focuses on

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 455–472, 2025.
https://doi.org/10.1007/978-3-031-78341-8_29

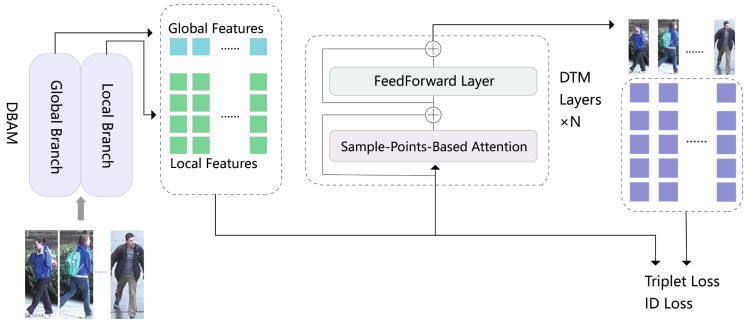


Fig. 1. The pipeline of the proposed method. A batch of images are input into the Dual Branch Attention Module (DBAM) to obtain global and local features for each image. Then, these features are feeded into the Deformable Transformer Module (DTM) for further fusion and reconstruction. Triplet loss and ID loss are calculated respectively for the features output by the two modules.

how to extract more discriminative feature representations of pedestrians. Example works include methods concerning data augmentation [9–12], methods using local features [13, 14, 16–22], and approaches based on attention mechanisms [4–8].

Despite that quite a lot of works have been published, most methods pay their attention on extracting features from individual images of a pedestrian. However, a single image is prone to losing some information due to occlusion, blurring, and inappropriate shooting angles. In genuine surveillance circumstances, a pedestrian is usually captured in multiple images. The missing information from a single image is expected to be supplemented from other images. Therefore, a few methods, such as NFormer [31] and HLGAT [34], turn to consider constructing relationships between multiple pedestrian images to enhance the identification performance. Inspired by their ideas, in this work we use a Deformable Transformer Module(DTM) to automatically construct relationships between multiple pedestrian images.

Alexey Dosovitskiy et al. proposed ViT [30], a universal solution to computer vision problems using the Transformer structure. Based on the structrue of ViT, TransReID [26] became the first Transformer-based ReID model. Experiments have shown that TransReID achieves better performance than traditional CNN methods. Our work is also based on ViT. Actually, we propose a Dual Branch Attention Module(DBAM) based on ViT and TransReID: A global branch and a local branch for extracting the global and local features of individual pedestrian images respectively. Afterwards, the feature vectors of multiple pedestrian images are combined and input into the Deformable Transformer Module(DTM) for further learning and reconstruction. Experiment results verify the efficiency of our proposed method which outperforms the state-of-the-art published works by a large margin.

2 Related Works

Usually, given a target pedestrian image set Query(Q) and an image Gallery(G), the goal of Person ReID is to retrieve the target pedestrian in Q from G. The ReID

algorithms initially relied mainly on metric learning [1–3]. At present, the popular ReID methods are primarily based on representation learning. This kind of method commonly starts by training a deep learning model to extract the feature representations of input images, and then achieves image matching and retrieval through similarity calculation of feature vectors. Therefore, obtaining more discriminative pedestrian feature representations is the key to the success of the person ReID algorithm.

Certain methods [9–12] improve the model's performance from the perspective of data augmentation. For example, utilizing a generative model like GAN can provide more supplementary sample data. In addition, numerous studies have demonstrated that local features are particularly helpful for improving the ReID algorithm's recognition accuracy. These are the major means of gathering local features:

- Some methods are based on strip segmentation, such as PCB [13], AlignedReID++ [14], and MGN [15].
- Some methods [16–22] align different parts by locating human key points. Typically, parsing or key point estimation models [23–25] are employed to assist in obtaining local features.
- Other methods [26–28] are based on Transformer to automatically obtain local features.

Some methods [4–8] add attention mechanisms to the model. The attention mechanism can effectively strengthen the expression of valuable information while suppressing irrelevant information expression.

Transformer [29], as a flexible structure for processing different types of data, has also shown its strong discriminative ability for fulfilling the task of ReID.

ViT [30] is the earliest proposed solution to computer vision problems based on the Transformer structure. Through the Patch Embedding procedure, ViT divides the image into numerous fixed-sized small patches, and then encodes each patch into a feature vector. Numerous experiments have demonstrated that ViT outperforms conventional CNN models when given a large amount of training data. Later, more Transformer-based models emerged in the CV field. Detr [35] utilizes Transformer to solve the problem of object detection, and achieves complete end-to-end detection in contrast with traditional object detection methods based on CNN and prior boxes. Deformable Detr [36] introduces a deformable attention mechanism based on sampling points. It effectively increases the model's training pace and outperforms Detr greatly in terms of detection performance.

Many Transformer-based ReID models have gradually emerged as a result of Transformers' broad use in the CV area. In order to extract pedestrian image features, CNN-based algorithms typically start with ResNet [37] or other models as baseline and then add attention mechanisms, local features, and other techniques. TransReID [26] is the first pure Transformer-based ReID model. It takes ViT as the baseline and incorporates two training techniques for ReID problems: Side Information Embedding(SIE) and Jiasaw Patch Module(JPM). TransReID has achieved better performance than CNN-based methods on several vehicle and pedestrian ReID datasets. Zhu et al. [27] proposed AAformer, which can automatically partition image parts based on Transformer. Zhu et al. [28] proposed a Dual Cross-Attention Learning(DCAL) algorithm that includes Global-Local Cross-Attention (GLCA) and Pair-Wise Cross-Attention (PWCA). In

addition, there are some methods that combine CNN and Transformer, such as HAT [32], PAT [33], etc.

Most of the methods mentioned above only consider how to extract discriminative features from a single image, while ignoring the potential relationships between different images. Noticing this point, Wang et al. [31] modified the Transformer and proposed NFormer which can effectively construct relationships between all input images based on the observation that the images of the same person have closer distances in the feature space. NFormer [31] has improved the multi-head self-attention mechanism and proposed the Land-Mark Agent Attention and the Reciprocal Neighbor SoftMax. Another tool, HLGAT [34], utilizes the Graph Attention model to simultaneously construct the inter-local relation and the intra-local relation to improve the identification accuracy. The outstanding performance of NFormer and HLGAT demonstrates the power of building relationships between multiple images. Inspired by their success, we aim to develop a model that can automatically combine the characteristic of multiple images of one person efficiently. Notice that local features are proved particularly helpful in improving ReID accuracy, we introduce a Dual Branch Attention Module based on ViT [30] and TransReID [26] for processing single images, incorporating a mask attention component as in NFormer [31] for extracting discriminative local features. Impressed by the success of introducing the sampling points [36], we design a Deformable Transformer Module for fulfilling the task of automatically merging the features extracted from multiple images.

3 Method

We'll give our proposed approach a thorough introduction in this section. The model consists of two modules: The first is a Dual Branch Attention Module (DBAM) designed to extract global and local features of individual images; The other is a Deformable Transformer Module (DTM), which is applied to construct relationships between multiple pedestrian images.

3.1 Pipeline

Assuming that N pedestrian images are utilized as batch input, and the i-th input image is recorded as $x_i \in \mathbb{R}^{C\times H\times W}$. Firstly, we input a batch of images into the Dual Branch Attention Module(DBAM) for preliminary feature extraction. After the DBAM, for each input image x_i, a global feature representation $g_i \in \mathbb{R}^D$ and several local features $l_i \in \mathbb{R}^{m\times D}$ can be obtained. Next, we will combine all the feature vectors of N input images to form a feature map $F \in \mathbb{R}^{D\times N\times(m+1)}$. Then we input F into the Deformable Transformer Module(DTM) to simultaneously fuse the global and local features of multiple images. After DTM, we can get the reconstructed $m+1$ feature vectors for each input image x_i. Finally, we concatenate these $m+1$ vectors together as the feature representation of this image. Figure 1 displays the pipeline of our method.

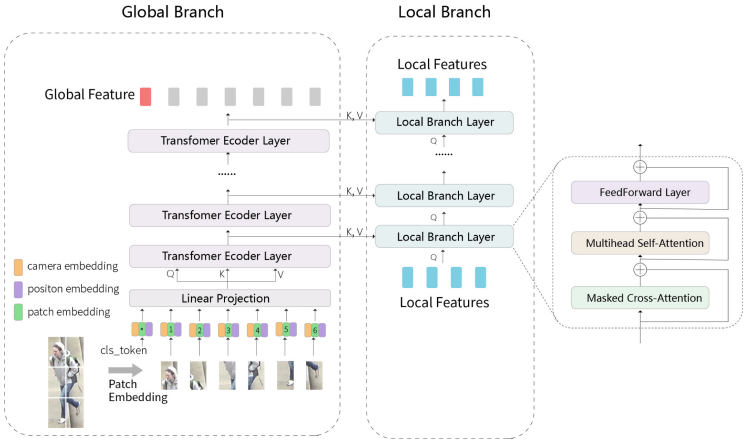


Fig. 2. The structure of the Dual Branch Attention Module (DBAM). The left side represents the global branch, and the right side represents the local branch. After the DBAM, one global feature (shown in the red small square) and m local features (shown in the blue small squares) can be obtained for each input image. (Color figure online)

3.2 Dual Branch Attention Module

Figure 2 depicts the structure of the Dual Branch Attention Module(DBAM). DBAM includes a global branch for obtaining global feature representation of the image and a local branch for extracting a number of local features.

The global branch is designed based on ViT [30] and incorporates the Side Information Embedding (SIE) module proposed in TransReID [26]. First, we partition the input image x_i into several patches using Patch Embedding to convert x_i into a sequence $y_i \in \mathbb{R}^{n \times D}$. Similar to ViT, we add a CLS-Token as the image's global feature representation at the start of sequence y_i. Then we incorporate camera embedding and position embedding. We currently get the final input sequence $M_0 \in \mathbb{R}^{(n+1) \times D}$. It is introduced into a multiple-layer Transformer Encoder, and the output of the i-th layer is recorded as $M_i \in \mathbb{R}^{(n+1) \times D}$. We take out the CLS-Token output from the last layer as the global feature of the current image, denoted as $g_i \in \mathbb{R}^D$.

The local branch has a multi-layer structure to match the global branch layer by layer. Firstly, we randomly initialize m vectors as local features, denoted as $L_0 \in \mathbb{R}^{m \times D}$. We then input them into the local branch and record the output of i-th layer as $L_i \in \mathbb{R}^{m \times D}$. Three components make up each layer of the local branch: a Masked Cross-Attention component, which we shall discuss in more details in the following section, a multi-head self-attention mechanism as the second component, and a simple feed-forward neural network as the last component. The input of each layer in the local branch is the output of the previous layer and the output of the corresponding layer in the global branch. For example, for the i-th layer, its inputs are L_{i-1} and M_i. We ultimately use the output of the last layer as local features of the current image, denoted as $l_i \in \mathbb{R}^{m \times D}$.

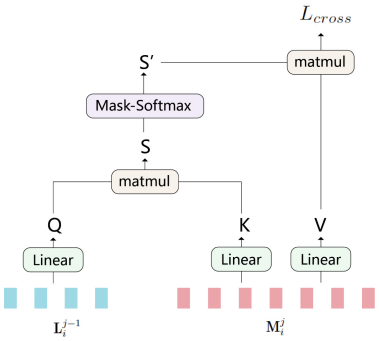


Fig. 3. The calculating mechanism of the Masked Cross-Attention.

Through the Dual Branch Attention Module, for each input image x_i, a global feature g_i and several local features l_i can be extracted.

3.3 Masked Cross-Attention

In the multi-head self-attention mechanism, the Q, K, and V vectors usually come from the same input sequence. However, in our cross-attention mechanism, the Q vectors come from the output of the previous layer, and the K and V vectors come from the output of the corresponding layer of the global branch. For example, in the i-th layer, Q, K, and V vectors can be represented as:

$$\begin{aligned} Q &= L_i^{j-1} \cdot W_q \ , \\ K &= M_i^j \cdot W_k \ , \\ V &= M_i^j \cdot W_v \ . \end{aligned} \tag{1}$$

Figure 3 illustrates the calculating mechanism for the Masked Cross-Attention. Firstly, we obtain a relationship matrix S:

$$S = \frac{Q \cdot K^\top}{\sqrt{D}}. \tag{2}$$

Then we introduce a Mask mechanism in order to ensure that each local feature will only concentrate on a specific area of the image. We initialize a matrix $\mathcal{M}$ as a Mask. The i-th row of matrix S records the relationship between the i-th local feature and all patches of the image. We find the top k maximum values of $S_{i,;}$ and set the corresponding position in $\mathcal{M}$ to 1, indicating the patches that the i-th local feature needs to be focused on. Simultaneously set the values of other positions in $\mathcal{M}$ to 0:

$$\mathcal{M}_{ij} = \begin{cases} 1, & j \in \text{topk}(S_{i,:}), \\ 0, & else. \end{cases} \tag{3}$$

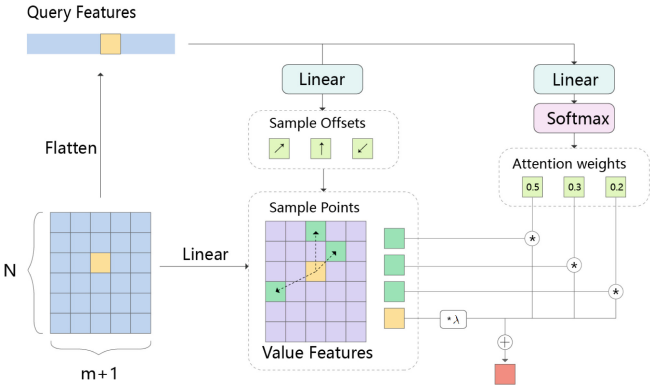


Fig. 4. The calculating mechanism of the Sample-Points-Based Attention.

Drawing inspiration from the design of the Reciprocal Neighbor Softmax in NFormer [31], we add the Mask into the Softmax function and further process S to obtain S':

$$S'_{ij} = \frac{\mathcal{M}_{ij} \exp\left(S_{ij}\right)}{\sum_k \mathcal{M}_{ik} \exp\left(S_{ik}\right)}. \tag{4}$$

Finally, S' is applied to V to obtain the output $L_{cross} \in \mathbb{R}^{m \times D}$:

$$L_{cross} = S' \cdot V \ . \tag{5}$$

3.4 Deformable Transformer Module

Through the Dual Branch Attention Module, we can obtain the global and local features of each input image. Then we combine these features together and input them into the Deformable Transformer Module for simultaneously fusing global and local features from multiple images.

After DBAM, for any input image, a global feature and m local features can be obtained. We will create a two-dimensional feature map $F \in \mathbb{R}^{D \times N \times (m+1)}$ with all the features from a batch of input images. Then we will input it into the DTM. DTM is a multi-layer structure with each layer consisting of two parts: One is the Sample-Points-Based Attention mechanism, which is designed from the inspiration by deformable Detr [36]; And the other is a simple feed-forward neural network.

Figure 4 displays the calculating mechanism of the Sample-Points-Based Attention. Firstly, flatten F to obtain query features $Q \in \mathbb{R}^{N(m+1) \times D}$ and index them with q. For any vector Q_q, it corresponds to a position $P_q(p_{qx}, p_{qy})$ in the feature map F. At the same time, we conduct a linear map on the features in F to get the value features V. Next, we need to predict a set of offsets and attention weights for each feature. Specifically, for each feature Q_q, K offsets are predicted through a linear layer. Each offset contains offset values in both X and Y directions. The i-th offset is denoted by

Table 1. Contributions from Algorithmic Components. The bold numbers indicates that the complete model with both DBAM and DTM outperforms dramatically, which suggests the necessity of both components.

Experimental Settings	DukeMTMC		Market1501	
	mAP	Rank-1	mAP	Rank-1
Base	81.6	89.8	88.6	95.2
Base+DBAM	82.0	90.4	89.0	95.5
Base+DTM	90.3	95.1	96.3	98.6
Base+DBAM+DTM	**93.2**	**95.8**	**97.5**	**99.2**

$\Delta P_{q,i}(\Delta p_{qx,i},\ \Delta p_{qy,i}), i \in [0, K)$. Based on the original position P_q of Q_q, K sampling points can be obtained by offsetting according to K offsets. The original coordinates and offsets are added to determine the coordinates of each sampling point:

$$\begin{aligned} P'_{q,i} &= P_q + \Delta P_{q,i} \\ &= (p_{qx} + \Delta p_{qx,i},\ p_{qy} + \Delta p_{qy,i}) \ . \end{aligned} \tag{6}$$

It should be noted that the coordinates may not be integers, therefore, bilinear interpolation method needs to be used for further processing. In addition, another linear layer is also applied to Q_q in order to forecast K weights, which correspond to K sample points one by one. Then the calculation mechanism of sampling-points-based attention can be expressed as:

$$\boldsymbol{Q'_q} = W_q \sum_{i=0}^{K} (\alpha_{q,i} \cdot Sample(V, P'_{q,i}) + \lambda \boldsymbol{V_q}) \ . \tag{7}$$

$Sample(\cdot)$ represents a sampling operation, i.e., based on the coordinates of the sampling points, to extract the corresponding feature vectors from the value features. α is the attention weights. λ is a hyper-parameter used to explicitly retain a portion of the current features during computation. The above operations of the single-head attention mechanism, like Transformer, have been extended to the multi-head mechanism.

4 Experiments

4.1 Experiments Setting

Table 2. Statistics of person ReID datasets.

dataset	identities	images	cameras
Market1501	1501	32668	6
DukeMTMC-ReID	1404	36411	8
MSMT17	4101	126441	15
Occluded-Duke	1404	35489	8

Datasets and Evaluation Protocols. To evaluate the efficiency of our method, we conducted experiments on four large person ReID datasets: Market1501 [38], DukeMTMC-ReID [11], MSMT17 [39], and Occluded-Duke [40]. Occluded-Duke is an occluded ReID dataset, with images selected from DukeMTMC, and all images in the query set are pedestrian images with occlusions. Each identity is captured by multiple images from various cameras or settings in all of the above datasets. The detailed information of the above datasets is shown in Table 2.

We employ the conventional ReID algorithm's setting for experiments and use the cumulative matching characteristic (CMC) curve and mean Average Precision (mAP) as evaluation indicators. CMC shows the precision of the top k rankings by calculating the true positive results among the top k individuals in the sorted query results. When k is taken as 1, top-1 represents the first hit rate, which is one of the most important evaluation criteria for the ReID method. The mAP measures the gallery set's overall re-identification accuracy.

Implementation. We choose the pre-trained ViT-Base [30] model on the ImageNet [41] dataset as the baseline of our method. This model has 12 Transformer encoding layers, 12 heads for multi-head attention, and feature vector size is set to 768. Every input pedestrian image is resized to 256×128. We use random horizontal flipping, padding with 10 pixels, random cropping, and random erasing [10] for data augmentation. We set the batch size to 64, indicating each batch includes 16 pedestrian IDs and 4 images for each ID. Stochastic Gradient Descent (SGD) optimizer is applied for training the model. The initial Learning rate is set to 0.01, and the cosine Learning rate decay strategy is applied. Our method was implemented using Pytorch and all experiments were conducted on one GeForce RTX 3090.

4.2 Ablation Study

Contributions from Algorithmic Components. To demonstrate the necessity of each component of our proposed method, we conducted ablation experiments on the DukeM TMC-ReID and Martket1501 datasets. Experiment 1 applies the ViT model to extract pedestrian features and incorporates the overlapping patches and SIE training tips suggested in TransReID. Experiment 2 adds the local branch on the basis of Experiment 1, i.e., to use our Dual Branch Attention Module (DBAM) to extract global and local features of a pedestrian, and concatenate the features together as the final feature representation of the pedestrian. Experiment 3 adds our proposed Deformable Transformer Module (DTM) to that of Experiment 1, i.e., to conduct preliminary feature extraction using the basic model from Experiment 1, and then combine the extracted features together to input into DTM. Experiment 4 is designed to illustrate the performance of our complete method, which includes our proposed DBAM and DTM modules. Table 1 presents the results of the above groups of experiments.

When comparing the results of Experiment 1 and Experiment 2, it can be observed that our proposed DBAM performs by a small margin better than Baseline. On the DukeMTMC-ReID dataset, the Rank-1 and mAP of the DBAM increase in comparison to Baseline by 0.6% and 0.4%, respectively. On the Market1501 dataset, Rank-1 and

mAP values increase by 0.3% and 0.4% respectively. The above results indicate that our proposed DBAM method for extracting global and local feature representations of pedestrians is effective. Comparing the results of Experiment 3 and Experiment 1, the performance of the algorithm is shown to be greatly improved by adding our proposed DTM. On the DukeMTMC-ReID dataset, the model incorporating DTM improves on Rank-1 and mAP by 5.3% and 8.7%, respectively compared to Baseline. On the Market1501 dataset, Rank-1 and mAP increase by 3.4% and 7.7%. The above experimental results fully demonstrate the effectiveness of our idea of using DTM to extract more discriminative information from multiple pedestrian images. Compared to Baseline, in experiment 4, Rank-1 and mAP increase by 6.0% and 11.6% on the DukeMTMC-ReID dataset, and by 4% and 8.9% on the Market1501 dataset, respectively. Notice that experiment 4 corresponds to our complete framework which includes both the DBAM and DTM. Compared to experiments 2,3 where DBAM or DTM are added separately, the complete framework achieves significantly better performance in experiment 4, which illustrates the effectiveness of our overall approach.

Table 3. Ablation Study of λ. Bold numbers show that the model performs better when λ is set to middle-sized numbers. We set λ to 0.5 accordingly.

λ	DukeMTMC		Market1501	
	mAP	Rank-1	mAP	Rank-1
0	90.8	92.3	96.8	98.1
0.3	90.8	92.5	**97.4**	**99.4**
0.5	**91.8**	**95.3**	97.1	99.3
0.8	89.9	94.8	96.0	98.5
1	89.1	94.4	95.7	98.3

Ablation Study of λ. In the Sample-Points-Based Attention mechanism of the DTM, λ is a hyper-parameter, which is used to explicitly specify the proportion of preserving the current feature itself. We compare the impact of different λ on the performance of the model to determine a reasonable λ for our method. For this purpose, we conducted five sets of experiments on the DukeMTMC-ReID and Market1501 datasets, with λ set as 0, 0.3, 0.5, 0.8, and 1, respectively. The experimental results are shown in Table 3.

According to the experimental data shown in Table 3, different values of λ have a certain impact on the model's final performance. The model performs best on the DukeMTMC-ReID and Market1501 datasets when λ is adjusted to 0.5 and 0.3, respectively. The Rank-1 and mAP values of the algorithm will decrease to some extent with a smaller or bigger λ. This might be the case that if λ is too tiny, information on the feature itself would be lost, while if λ is too large, learning from other features adequately will be hard. Therefore, a reasonable value of λ is necessary. We ultimately set λ to 0.5 based on the experimental findings.

Amount of Sample Points. The number of sampling points K is an important parameter in DTM, which affects the range of mutual learning and fusion between features.

Table 4. Amount of Sample Points. Bold numbers show that the model has achieved best performance with the amount of sample points set to 4. This is most likely having too many sampling points can result in the blending of irrelevant information from other features. We set the number of sampling points K to 4 in our method.

K	DukeMTMC		Market1501	
	mAP	Rank-1	mAP	Rank-1
4	**93.2**	95.8	**97.5**	**99.2**
8	93.0	**96.5**	96.5	98.8
12	91.7	95.8	96.6	98.9

Therefore, we determined a reasonable number of sampling points through a set of ablation experiments on K. We conducted three sets of experiments on the DukeMTMC-ReID and Market1501 datasets, with the number of sampling points K set to 4, 8, and 12, respectively. The experimental results are shown in Table 4.

When K is set to 4, the model has achieved ideal performance on both the DukeMTMC-ReID and Market1501 datasets. When the value of K increases, the model's Rank-1 and mAP do not significantly improve, while even decrease. This is most likely because having too many sampling points can result in the blending of irrelevant information from other features. Therefore, we set the number of sampling points K to 4 in our method.

Table 5. Amount of Deformable Transformer Layers. Bold numbers show that the model performs optimally on both datasets with fewer layers. The features will become amalgamated with excessive irrelevant information when there are too many layers. In our approach, we set the number of DTM layers to 3.

Layers	DukeMTMC		Market1501	
	mAP	Rank-1	mAP	Rank-1
3	**93.2**	**95.8**	**97.5**	99.2
6	91.8	95.3	97.1	**99.3**
9	90.5	94.8	96.5	98.9

Amount of Deformable Transformer Layers. Like the traditional Transformer, the DTM is also a multi-layer structure. We will determine a reasonable number of DTM layers through experiments. With the DTM layers adjusted to 3, 6, and 9, we carried out three sets of experiments on the DukeMTMC-ReID and Market1501 datasets. The experimental results are shown in Table 5.

According to the experimental findings, the DTM's layer count significantly affects the model's overall performance. We discovered that the model performed optimally on both datasets with fewer layers, such as three layers. When the number of layers of the DTM increases, the Rank-1 and mAP of the model do not increase but decrease instead. This is attributed to the fact that the receptive field expands as the number of

DTM layers increases. Consequently, if there are too many layers, the features will become amalgamated with excessive irrelevant information, which is not desirable. As a result, reducing the number of layers is sensible. Our approach limits the number of DTM layers to three.

Table 6. Amount of Local Features. Bold numbers show that increasing the amount of local features does not significantly enhance the algorithm's performance but increase the cost of time and space. Therefore, we set up 4 local features in our model.

m	DukeMTMC		Market1501	
	mAP	Rank-1	mAP	Rank-1
4	93.2	95.8	97.5	99.2
6	**93.4**	**96.4**	97.6	98.9
12	92.7	95.5	**97.7**	**99.3**

Amount of Local Features. In this part, we conducted experiments to examine the effects of various local feature amounts on algorithm performance. On the DukeMTMC-ReID and Market1501 datasets, we ran three sets of experiments with 4, 6, and 12 local features. Table 6 presents the experimental outcomes. We discovered that increasing the amount of local features does not significantly enhance the algorithm's performance, but rather may slightly worsen the results. Additionally, adding more local features will cost the algorithm more in terms of time and space. Based on the above considerations, we ultimately set up four local features in our model.

4.3 Comparison with SOTA Methods

Table 7 shows the comparison results of our method DTMIReID with state-of-the-art published ReID methods on four datasets: DukeMTMC-ReID, Market1501, MSMT17, and Occluded-Duke. Note that these methods are commonly cited for comparison in other works and they cover CNN-based and ViT-based methods as well as the methods that also apply multi-image retrieval. Overall, our method achieved overwhelming results on all these four datasets and outperformed other compared methods by a large margin. Notice that the new method demonstrates excellent performance even when dealing the occluded images. MP-ReID [51], based on prompt learning and language models, follows our method and achieves the second best performance on Market1501 and DukeMTMC-ReID. FPC [50], designed mainly for circumventing occluded person ReID, leads the others on the Occluded-Duke data but still falls behind ours. SOLIDER [55], as a general self-supervised learning framework, achieves quite leading performance with the second position on MSMT17 and the third on Market1501. Notice that method HLGAT [34], which also takes mutual information of multiple images into consideration, performs equally well on almost all the data-sets, which validates the significance of this strategy.

Table 7. Comparison with SOTA methods. Bold numbers show that our method of merging features of multiple images using DTM outperforms other compared methods dramatically. The prompt-learning-based method MP-ReID, the method FPC designed mainly for circumventing occluded person ReID, and the self-supervised learning framework SOLIDER perform well against the others.

Method	references	Market1501		DukeMTMC		MSMT17		Occluded-Duke	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PCB [13]	ECCV(2018)	81.6	93.8	69.2	83.3	40.4	68.2	33.7	42.6
MGN [15]	MM(2018)	86.9	95.7	78.4	88.7	52.1	76.9	–	–
ABDNet [42]	ICCV(2019)	88.3	95.6	78.6	90.0	60.8	82.3	–	–
OSNet [43]	ICCV(2019)	84.9	94.8	73.5	88.6	52.9	78.7	–	–
IANet [44]	CVPR(2019)	83.1	94.4	73.4	87.1	46.8	75.5	–	–
CBN [45]	ECCV(2020)	83.6	94.3	70.1	84.8	–	–	–	–
ISP [46]	ECCV(2020)	88.6	95.3	80.0	89.0	–	–	–	–
HOReID [47]	CVPR(2020)	84.9	94.2	75.6	89.6	–	–	43.8	55.1
PGFA [19]	ICME(2018)	76.8	91.2	65.5	82.6	–	–	37.3	51.4
CDNet [48]	CVPR(2021)	86.0	95.1	76.8	88.6	54.7	78.9	37.3	51.4
PAT [33]	CVPR(2021)	88.0	95.4	78.2	88.8	–	–	53.6	64.5
TransReID [26]	ICCV(2021)	88.9	95.2	82.0	90.7	67.4	85.3	59.2	66.4
AAFormer [27]	arxiv(2021)	87.7	94.4	80.0	90.1	63.2	83.6	–	–
HAT [32]	MM(2021)	89.5	95.6	81.4	90.4	61.2	82.3	–	–
DCAL [28]	CVPR(2022)	87.5	94.7	80.1	89.0	64.0	83.1	–	–
FED [49]	CVPR(2022)	86.3	95.0	78.0	89.4	–	–	56.4	68.1
NFormer [31]	CVPR(2022)	93.0	95.7	85.7	90.6	62.2	80.8	–	–
HLGAT [34]	CVPR(2021)	93.4	97.5	87.3	92.7	73.2	87.2	–	–
FPC [50]	AAAI(2024)	91.4	95.1	–	–	–	–	**72.8**	**76.7**
MP-ReID [51]	AAAI(2024)	**95.5**	**97.7**	**88.9**	**95.7**	–	–	–	–
ISR [52]	ICCV(2023)	92.3	96.9	–	–	71.5	88.4	–	–
DC-Former$^{384\uparrow}$ [53]	AAAI(2023)	90.6	96.0	–	–	70.7	86.9	–	–
CLIP-ReID+SIE+OLP [54]	AAAI(2023)	90.5	95.4	83.1	90.8	75.8	89.7	60.3	67.2
SOLIDER [55]	CVPR(2023)	93.9	96.9	–	–	**77.1**	**90.7**	–	–
DTMIReID (Ours)		**97.5**	**99.2**	**93.2**	**95.8**	**86.9**	**93.6**	**74.5**	**81.4**

5 Conclusion

In this paper, we propose a person ReID model DTMIReID based on Deformable Transformer, which consists of two main modules: the Dual Branch Attention Module (DBAM) and the Deformable Transformer Module (DTM). Specifically, we first perform preliminary feature extraction on individual images using the DBAM to obtain global and local features of the images. Then we assemble the features from all of the images into a two-dimensional feature map and input it into the DTM. The Sampling-Points-Based Attention mechanism in DTM can help us automatically construct the relationships between all features of multiple images and reconstruct them. The features output from the DTM are more robust and discriminative. The experimental results on four ReID benchmarks demonstrate that our method achieves state-of-the-art performance and outperforms other compared methods substantially. Despite the high identification precision, the proposed method needs more time for model training than meth-

ods using single image for retrieval. Besides, supervised training requires all the training data annotated which is very labor consuming. How to combine the features of multiple images without extra resources consumption and how to migrate the architecture to semi-supervised or unsupervised situations will be of interest.

Acknowledgments. This work is supported by the National Natural Science Foundation of China under No. 61672325. We sincerely thank the anonymous reviewers for their valuable comments and suggestions.

References

1. Zheng, W., Gong, S., and Xiang, T.: Reidentification by relative distance comparison. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 653–668 (2013). https://doi.org/10.1109/TPAMI.2012.138
2. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P., and Bischof, H.: Large scale metric learning from equivalence constraints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 2288–2295. IEEE Computer Society (2012). https://doi.org/10.1109/CVPR.2012.6247939
3. Liao, S., and Li, Z.: Efficient PSD constrained asymmetric metric learning for person re-Identification. In: 2015 IEEE International Conference on Computer Vision(ICCV), pp. 3685–3693. IEEE Computer Society (2015). https://doi.org/10.1109/ICCV.2015.420
4. Li, W., Zhu, X., and Gong, S.: Harmonious attention network for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 2285–2294. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00243
5. Wang, C., Zhang, Q., Huang, C., Liu, W., and Wang X.: Mancs: a multi-task attentional network with curriculum sampling for person re-identificatione. In: Proceedings of the 15th European Conference on Computer Vision(ECCV), pp. 356–381. Springer (2018)
6. Wang, Y., Chen, Z., Wu, F., Wang, G.: Person re-identification with cascaded pairwise convolutions. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 1470–1478. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00159
7. Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z.: Relation-aware global attention for person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3183–3192. IEEE Computer Society (2020). https://doi.org/10.1109/CVPR42600.2020.00325
8. Song, C., Huang, Y., Ou Y., Wan L., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1179–1188. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00129
9. Huang, H., Li, D., Zhang, Z., Chen, X., Huang, K.: Adversarially occluded samples for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 5098–5107. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00535
10. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 7, pp. 13001–13008. Association for the Advancement of Artifcial Intelligence (2020). https://doi.org/10.1609/aaai.v34i07.7000

11. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3774–3782. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.405
12. Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4099–4108. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00431
13. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling (and A Strong Convolutional Baseline). In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 501–518. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_30
14. Luo, H., Jiang, W., Zhang, X., Fan, X., Qian, J., Zhang, C.: AlignedReID++: dynamically matching local information for person re-identification. Pattern Recogn. **94**, 53–61 (2019). https://doi.org/10.1016/j.patcog.2019.05.028
15. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia(MM), pp. 274–282. Association for Computing Machineray (2018). https://doi.org/10.1145/3240508.3240552
16. Suh, Y., Wang, J., Tang, S., Mei, T., Lee, K.M.: Part-aligned bilinear representations for person re-identification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 418–437. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_25
17. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3239–3248. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.349
18. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: GLAD: global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 25th ACM International Conference on Multimedia(MM), pp. 420–428. Association for Computing Machinery (2017). https://doi.org/10.1145/3123266.3123279
19. Zhuo, J., Chen, Z., Lai, J., Wang, G.: Occluded person re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE Computer Society (2018). https://doi.org/10.1109/ICME.2018.8486568
20. Guo, J., Yuan, Y., Huang, L., Zhang, C., Yao, J., Han, K.: Beyond human parts: dual part-aligned representations for person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision(ICCV), pp. 3641–3650. IEEE Computer Society (2019). https://doi.org/10.1109/ICCV.2019.00374
21. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose-invariant embedding for deep person re-identification. IEEE Trans. Image Process. **28**(9), 4500–4509 (2019). https://doi.org/10.1109/TIP.2019.2910414
22. Kalayeh, M., Basaran, E., Gokmen, M., Kamasak, M., Shah, M.: Human semantic parsing for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 1062–1071. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00535
23. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5686–5696. IEEE Computer Society (2019). https://doi.org/10.1109/CVPR.2019.00584
24. Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y.: OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. In: IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(1), pp. 172–186 (2021)

25. Güler, R., Neverova, N., Kokkinos, I.: DensePose: dense human pose estimation in the wild. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 7297–7306. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00762
26. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: TransReID: transformer-based object re-identification. In: 2021 IEEE/CVF International Conference on Computer Vision(ICCV), pp. 14993–15002. IEEE Computer Society (2021). https://doi.org/10.1109/ICCV48922.2021.01474
27. Zhu, K., et al.: AAformer: auto-aligned transformer for person re-identification. In: arXiv preprint arXiv:2104.00921. (2021)
28. Zhu, H., Ke, W., Li, D., Liu, J., Tian, L., Shan, Y.: Dual cross-attention learning for fine-grained visual categorization and object re-identification. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 4692–4702. IEEE Computer Society (2022)
29. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010. Association for Computing Machineray (2017)
30. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: 2021 International Conference on Learning Representations (ICLR), pp. 1–22. OpenReview.net (2021)
31. Wang, H., Shen, J., Liu, Y., Gao, Y., Gavves, E.: NFormer: robust person re-identification with neighbor transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7297–7307. IEEE Computer Society (2022)
32. Zhang, G., Zhang, P., Qi, J., Lu, H.: HAT: hierarchical aggregation transformers for person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia(MM), pp. 516–525. Association for Computing Machineray (2021). https://doi.org/10.1145/3474085.3475202
33. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: occluded person re-identification with part-aware transformer. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 2897–2906. IEEE Computer Society (2021). https://doi.org/10.1109/CVPR46437.2021.00292
34. Zhang, Z., Zhang, H., Liu, S.: Person re-identification using heterogeneous local graph attention networks. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 12136–12145. IEEE Computer Society (2021)
35. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
36. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: 2021 International Conference on Learning Representations(ICLR), pp. 1–16. OpenReview.net (2021)
37. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
38. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1116–1124. IEEE Computer Society (2015). https://doi.org/10.1109/ICCV.2015.133
39. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 79–88. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00016

40. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 542–551. IEEE Computer Society (2019). https://doi.org/10.1109/ICCV.2019.00063
41. Deng, J., Dong, W., Socher, R., Li, L., Kai L., Li, F.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. IEEE Computer Society (2009). https://doi.org/10.1109/CVPR.2009.5206848
42. Chen, T., et al.: ABD-Net: attentive but diverse person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision(ICCV), pp. 8350–8360. IEEE Computer Society (2019). https://doi.org/10.1109/ICCV.2019.00844
43. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3701–3711. IEEE Computer Society (2019). https://doi.org/10.1109/ICCV.2019.00380
44. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9309–9318. IEEE Computer Society (2019). https://doi.org/10.1109/CVPR.2019.00954
45. Zhuang, Z., et al.: Rethinking the distribution gap of person re-identification with camera-based batch normalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 140–157. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_9
46. Zhu, K., Guo, H., Liu, Z., Tang, M., Wang, J.: Identity-guided human semantic parsing for person re-identification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12348, pp. 346–363. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58580-8_21
47. Wang, G., et al.: High-order information matters: learning relation and topology for occluded person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 6448–6457. IEEE Computer Society (2020). https://doi.org/10.1109/CVPR42600.2020.00648
48. Li, H., Wu, G., Zheng, W.: Combined depth space based architecture search for person re-identification. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 6725–6734. IEEE Computer Society (2021). https://doi.org/10.1109/CVPR46437.2021.00666
49. Wang, Z., Zhu, F., Tang, S., Zhao, R., He, L., Song, J.: Feature erasing and diffusion network for occluded person re-identification. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 4744–4753. IEEE Computer Society (2022). https://doi.org/10.1109/CVPR52688.2022.00471
50. Ye, Y., et al.: Dynamic feature pruning and consolidation for occluded person re-identification. In: Proceedings of the 2024 AAAI Conference on Artificial Intelligence, vol. 38, no. 7, pp. 6684–6692. Association for the Advancement of Artifcial Intelligence (2024). https://doi.org/10.1609/aaai.v38i7.28491
51. Zhai, Y., Zeng, Y., Huang, Z. ., Qin, Z., Jin, X., Cao, D.: Multi-prompts learning with cross-modal alignment for attribute-based person re-identification. In: Proceedings of the 2024 AAAI Conference on Artificial Intelligence, vol. 38, no. 7, pp. 6979–6987. Association for the Advancement of Artifcial Intelligence (2024). https://doi.org/10.1609/aaai.v38i7.28524
52. Dou Z., Wang Z., Li Y., Wang S.: Identity-seeking self-supervised representation learning for generalizable person re-identiffcation. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision(ICCV), pp. 15847–15858. IEEE Computer Society (2023). arXiv:2308.08887

53. Li W., et al.: DC-Former: diverse and compact transformer for person re-identification. In: Proceedings of the 2023 AAAI Conference on Artificial Intelligence, vol. 37, no. 2, pp. 1415–1423. Association for the Advancement of Artifcial Intelligence (2023). https://doi.org/10.1609/aaai.v37i2.25226
54. Li S., Sun L., Li Q.: CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels. In: Proceedings of the 2023 AAAI Conference on Artificial Intelligence, vol. 37, no. 1, pp. 1405–1413. Association for the Advancement of Artifcial Intelligence (2023). https://doi.org/10.1609/aaai.v37i1.25225
55. Chen W., et al.: Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 15050–15061. IEEE Computer Society (2023)

Integral Probability Metrics for Perceptual Learning in Generative Cross-Modal Person Re-Identification

Nirmala Murali(✉) and Deepak Mishra

Indian Institute of Space Science and Technology, Thiruvananthapuram, Kerala, India
nirmalamurali891@gmail.com, deepak.mishra@iist.ac.in

Abstract. Person Re-Identification (Person-ReID) is the problem of recognizing an identity in various instances across various cameras. Cross-modal Person-ReID extends this task to match images of different modalities, posing a significant challenge due to the considerable gap between modalities. The availability of the exact RGB-IR pairs for each identity and pose will help the system understand the feature space better. Thus, a generative model leveraging Optimal Transport Theory is proposed to synthesize IR images corresponding to available RGB images, enhancing the training data for the Person-ReID model. These images can be considered as distributions, and finding out how each distribution differs will eventually tell the model how each identity varies. Comparing the feature vectors using the conventional distance metrics might only work for some cases. Hence, this is done using the Integral Probability metrics, which finds the difference between two probability distributions by bringing in perceptual similarity while also aligning the inter-modality images. Additionally, a part feature attention module is proposed to learn the essential features in every RGB-IR pair. This method combines various loss functions based on Integral Probability metrics, including Wasserstein distance and Maximum Mean Discrepancy. The proposed method showed significant improvements in the cross-modal Person-ReID result.

Keywords: Person Re-Identification · integral probability metrics · Wasserstein distance · generative adversarial networks · cross-modality

1 Introduction

Person Re-Identification (Person-ReID) [1] is an image retrieval problem where a person's identity is recognized by a model in another instance. This problem plays a significant role in many other tasks like tracking and image retrieval. The applications of Person-ReID are in security, real-time surveillance, crowd

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78341-8_30.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 473–488, 2025.
https://doi.org/10.1007/978-3-031-78341-8_30

management etc. Any Person-ReID model has to overcome certain challenges, like low image resolution, incomplete annotations, illumination variations, occlusions, and viewpoint variations.

Fig. 1. Sample images from SYSU-MM01 dataset.

This study addresses heterogeneous modality [2], where the images are taken during morning and nighttimes. The modality of the images changes when the nighttime images are captured using IR cameras and the daylight images are captured using standard RGB cameras, as shown in Fig. 1. The model needs to distinguish between different identities while also aligning the changes in modalities. This means that the model must be trained to understand how features change when modality changes for the same identity.

Some of the recent works like [3] and [4] solve the issue of cross-modality Person-ReID by providing a baseline and by introducing camera invariant learning. However, we try to address the problem by enhancing the feature space and bringing in perceptual understanding to learn the features. We propose to solve the cross-modality problem in two steps, first, identity discrimination has to be learnt, and as the next step, the modality gap is reduced. In order for the model to understand RGB and IR feature space and to match the identities across modalities, exact RGB-IR pairs are essential. Therefore, we propose a generative model to generate synthetic IR images for the available RGB images by limiting the CycleGAN network using the Lipschitz constraint. Person-ReID faces a high inter-class similarity, and to solve this, the model needs to understand the features across domains perceptually. To induce perceptual understanding and to help with the cross-modality alignment, we propose a loss function that makes use of two integral probability metrics, namely, Wasserstein distance and maximum mean discrepancy. Also, to map the corresponding RGB-IR pairs, an attention module is proposed. The proposed system consists of two parts. The first part is to generate the corresponding IR pairs for the RGB images, and the next part uses the generated image pairs for the re-identification task. The generative Person-ReID has been studied in a few works like [5]. Also, [6] learns identity features using Wasserstein distance. Inspired by the above works, we propose a cross-modal person re-identification system and experiment in different settings. The primary contributions of this study are,

– Proposed a perceptual learning-based generative model - Wasserstein Cycle Generative Adversarial Network (WCGAN) that can generate IR images for

the given RGB images. This can improve the perceptual understanding of the model.
- Proposed a new model for cross-modal Person-ReID - Generative Cross-Modal Person-ReID using Integral Probability Metrics (GCM-IPM)
- Proposed a part feature attention module that focuses on most discriminative features and aligns the RGB-IR feature attention maps.

This study is structured as follows: Sect. 2 discusses the works related to this study in literature; Sect. 3 provides an overview of the OTT basics and explains the proposed methodology; Sect. 4 explains the implementation details and analyzes the results obtained from this proposed work. In Sect. 5, the future work is discussed.

2 Related Work

Person-ReID is an important problem in computer vision and has received a lot of attention in the recent literature. There are two main classifications in Person-ReID systems [2], namely open-world and closed-world re-identification. An open-world re-identification system works with a lot of assumptions, such as all query images will be from the gallery, all the annotations are correct, and only the same modality images are used in Re-ID systems. In contrast, closed-world re-identification does not assume any conditions and is designed to handle various kinds of data, making it a more challenging task to design such models.

2.1 Feature Based Person Re-ID

A comprehensive analysis of diverse Re-ID approaches is presented in [1]. [7] proposed a model that uses local features to solve the occlusion problem, and the modality difference is solved by strengthening the global features. The authors have also designed a batch-normalized global feature enhancement method and a method to fuse multi-granularity features, thus making use of global and local features. [8] proposed a siamese model that shares the layers so that feature fusion can be done. Modality-based and modality-invariant representations(MSR) are learned separately in this work. In DDAG [9], a dual attention learning-based approach is proposed. [10] proposed an end-to-end system for cross-modality-based re-identification that shares features across modalities. This technique makes use of modality information and modality-specific information as well.

To effectively address the heterogeneous feature space, we need to provide an exact RGB-IR pair for every instance of the image. This will enrich the feature space and help the model to learn and align the identities across different modalities effectively. Hence a generation module WCGAN is proposed to generate exact RGB-IR pairs and use the generated images for further training.

2.2 Generative Person-ReID Models

An approach based on knowledge distillation [5] is introduced where adversarial learning is employed to align the source data with the target data distribution. A larger model named the teacher, is used to learn an extensive dataset, while soft labels are used to train the small model which will be deployed. [11] proposed a thermal GAN that converts RGB to IR images and then uses those images for training the re-identification model. [12] proposed a model that tries to solve the misalignment problem. This method explicitly removes the modality-specific features and keeps all the modality-invariaxnt features. Sparse GAN uses skip connections; thus, it reduces the number of parameters and makes fast retrieval possible. [13] proposes a contrastive learning-based image generation network, which fuses the generated images. [14] proposed a model that handles the colour difference by converting both modalities to grayscale and then generating paired images.

On the contrary, this study introduces a generative model that employs a perceptual similarity metric incorporating Wasserstein distance and cycle consistency constraints. The objective of this approach is to improve the model's capacity to perceive images accurately and produce images of better quality.

2.3 Loss Based Re-Identification Models

Loss-based models define a new loss function that improves network learning. The dual alignment learning method by [15] uses semantic and skeletal graphs and optimal transport-based graph matching techniques. Earth mover's distance/Wasserstein distance has been used for image retrieval tasks [16]. Distribution alignment using Wasserstein distance [17] exploits the local features of images and aligns the features using Wasserstein distance which shows significant improvement in ReID results. Wasserstein distance is used in [6] for cross-modality image alignment. This work reduces the intra-identity distance by using work done as a metric. [18] proposed the MMD-ReID model, one of the first kernel methods in ReID that does margin-based modification, which helps in reducing over-fitting. A Wasserstein distance-based approach is proposed in [19] which finds workdone between the RGB and IR modalities.

Inspired by the above works, we extract features from the RGB image and the corresponding generated IR image and then train the twin model using basic identity loss and for distribution comparison, integral probability metrics are used.

3 Proposed Method

The overview of the system: a generative structure WCGAN is proposed to handle the RGB to IR conversion. Then, the generated RGB-IR images are used to train the proposed Person-ReID model GCM-IPM along with the already available RGB images. Then the extracted features are used for discrimination learning and modality alignment. The high-level architecture of the proposed model is depicted in Fig. 2.

3.1 Paired Image Generation Module

In the benchmark cross-modality datasets [2,20], there are many images for each identity for RGB and IR; however, the specific matching pairs of RGB and IR images for the same individual are not provided in the SYSU-MM01 dataset. Therefore, this is not a direct image-to-image mapping and we need to map the RGB and IR domains without a definite pair. In order to perform unpaired cross-modality translation, we impose two constraints, namely, cycle consistency and Lipschitz constraint.

Critic in CycleGAN: Inspired by the cycleGAN, we impose the cycle consistency constraint, which forces the generator to learn important features of the target domain and to learn the workdone to transport the source domain and target domain. The discriminator is replaced with the critic as in the original CycleGAN, and the discriminator will output the score for every generated image. In order to compare between the generated images and the real images, we use the Wasserstein Distance [21]. Details of the Wasserstein Distance are discussed in the next section. To train the GAN with Wasserstein distance, the second constraint, 1-Lipschitz continuity, needs to be ensured. For all real values $\mathbf{x_1}$ and $\mathbf{x_2}$, a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is Lipschitz continuous if,

$$\frac{|f(\mathbf{x_1}) - f(\mathbf{x_2})|}{|\mathbf{x_1} - \mathbf{x_2}|} \leq k. \tag{1}$$

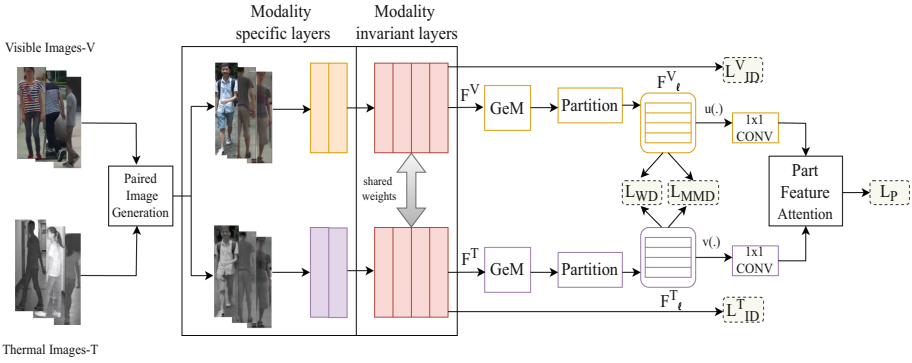


Fig. 2. The proposed model is shown here. The image generation module is used for paired image generation. The paired RGB-IR images are passed to the twin network to learn features. The first three convolutional blocks are modality-specific layers, and the next three layers are modality-invariant layers. Weights are shared among these layers. Then, the proposed IPM loss and part feature attention loss are calculated.

Here, derivatives of the function f are bounded to 1. Therefore, the gradients of the mapping function from the RGB to IR domain need to be less than or equal to 1. This ensures convergence and will not let the gradient be stuck at some local optimum, and the critic will output scalar scores.

Adversarial Learning Loss. For each generator $G_{V\to T}$ and $G_{T\to V}$, the objective is to align the identities by minimizing the Wasserstein distance between the distribution of generated samples and the distribution of real samples in their respective domains. Generator loss can be defined as follows,

$$\mathcal{L}_{\text{gen}} = -\frac{1}{N}\sum_{i=1}^{N}\left(D_{T\to V}(G_{V\to T}(V_i)) + D_{V\to T}(G_{T\to V}(T_i))\right) \tag{2}$$

Discriminator loss can be defined as follows,

$$\begin{aligned}\mathcal{L}_{\text{disc}} =& \frac{1}{N}\sum_{i=1}^{N}\Big(D_{V\to T}(V_i) - D_{V\to T}(G_{T\to V}(T_i)) \\ &+ D_{T\to V}(T_i) - D_{T\to V}(G_{V\to T}(V_i))\Big)\end{aligned} \tag{3}$$

where V_i and T_i are samples from the RGB and IR training set.

3.2 Two Stream Network

The baseline architecture is a two-stream network where both the Convolutional neural networks are trained together for cross-modality image learning. The input to the network is the RGB-IR image pairs generated by the proposed WCGAN in the last section. The initial convolutional blocks are set to learn the modality-specific features. The first three convolution blocks of the shallow layers are not shared. After the three layers, the convolution layers start sharing weights so that the networks learn the representation of the other modality and also so that the feature representation can be made modality invariant. Three such convolutional blocks are added to the networks. Then, the Generalized mean average pooling layer (GeM) is added to filter the extracted feature vectors. Batch normalization is applied to the features after the last three layers. The extracted features $F^V, F^T \in \mathbb{R}^{\mathbb{N}}$ are partitioned into p parts to exploit the local features $F_l^V = \{f_1^V, f_2^V, ...f_p^V\}$, $F_l^T = \{f_1^T, f_2^T, ...f_p^T\} \in \mathbb{R}^{\mathbb{N}\times 1}$. To learn the discriminative features of each identity, the Identification loss L_{ID} is calculated separately at the end of both networks. This is the cross-entropy loss, where entropy is the average of information obtained from one sample.

$$L_{ID} = -\sum_{i=1}^{N}\sum_{j=1}^{p} y_{ij} log(f_{ij}^M), where, M \in \{V, T\} \tag{4}$$

where f_{ij} is the local feature extracted by the proposed model with N samples and p parts per sample.

Part Attention Loss. Now that we have part-wise features for every RGB-IR pair, we propose to calculate the pairwise attention between each feature part. There are p parts for each feature vector of each image in every modality.

Therefore we have $p \times N \times 1$ part features in each modality. Because we have paired inputs, we take each part feature of RGB modality and find attention to the respective part feature of IR modality. To find the similarity between each feature pair, a dot product is first performed,

$$f(f_i^V, f_j^T) = exp(u(f_i^V)^T v(f_j^T)) \tag{5}$$

where u(.) and v(.) represents the 1×1 convolution layer. Now to form the attention map,

$$\beta_{i,j}^p = \frac{f(f_i^V, f_j^T)}{\sum_i^p f(f_i^V, f_j^T)} \tag{6}$$

β is the attention map that holds the importance of each part feature modality-wise. Now to enhance the feature space, the original feature vectors are multiplied with the attention maps as follows,

$$f_i^{'p} = \beta_i * f_i^p \tag{7}$$

The part feature attention loss for N samples is defined as follows,

$$L_P = -\frac{1}{N} \sum_i^N y_i log(P(y_i / f_i^{'p})) \tag{8}$$

3.3 IPM Based Cross-Modality Learning

The integral probability metric refers to finding the distance between two probability distributions. Each image is considered as a distribution of pixels/features, and the distance between the distributions can be calculated to find the similarity. Two loss metrics namely, Wasserstein distance and Maxmimum mean discrepancy are used in training the neural network to discriminate between identities and modalities.

Learning via Perceptual Similarity. Perceptual similarity refers to how we humans perceive any object and the difference between various objects. Wasserstein distance is the measure of work done to move a mass from one place to another. The distribution of one image can be compared with the distribution of another image by using KL Divergence and EMD. KL divergence and total variation find the distance only as a point-to-point match [16]. Also, by trying to convert one distribution to another, the model essentially learns how much effort it has to put in for the task. Therefore, Wasserstein distance can bring in perceptual learning. An optimal transportation plan is to be learnt from source distribution to the target distribution with the least possible cost. The Wasserstein loss between RGB and IR feature vectors is defined as follows,

$$WD(\mathbf{F^V}, \mathbf{F^T}) = min \sum_{\mathbf{i}} \sum_{\mathbf{j}} \mathbf{T}_{ij} . D(f_i^V, f_j^T). \tag{9}$$

where $\mathbf{T}_{ij}$ is the weight to transport points from i to j. D is the point-to-point distance between source point i and destination point j. Wasserstein distance does two stages of distribution comparison. The first step is to find the point-to-point cost and then use the cost to find the work done to transport one distribution to another. The similarity is determined by the work done cost, where if the work done is less, the query image and the gallery image are similar, and they belong to the same class.

$$L_{WD} = WD(\mathbf{F^V}, \mathbf{F^T}), \tag{10}$$

where WD is the work done to transport $\mathbf{F^V}$ to $\mathbf{F^T}$. In this work, we fit the model's output to the ground truth and Wasserstein loss is calculated by finding the cumulative distribution of the model features and the ground truth feature vectors. The difference between these two measures will tell the actual work done, as using only the distributions at that point will give only the point-to-point mapping. By learning the perceptual similarity along with point-to-point similarity, the model can differentiate between features effectively.

Modality Alignment Using Maximum Mean Discrepancy. Maximum Mean Discrepancy is one of the kernel methods [22] used in machine learning. A kernel is generally considered to be the dot product between features. For MMD, the kernels are considered to be in Hilbert space. If the dot product between the points is large, then the points are similar. If the dot product is 0, the points are perpendicular to each other. In other words, they are aligned completely in opposite directions, and there is no similarity.

$$K(\mathbf{x_1}, \mathbf{x_2}) =< \phi(\mathbf{x_1}), \phi(\mathbf{x_2}) > . \tag{11}$$

In the above equation, $\phi(\mathbf{x_1})$ and $\phi(\mathbf{x_2})$ are the feature vectors of the source and target distributions, respectively. Instead of doing this point-to-point, MMD finds the difference between the moments of the distributions.

Let $\mathbf{V}$, and $\mathbf{T}$ be RGB and IR distributions. Then MMD between the feature distributions can be expressed as follows,

$$L_{MMD}(\mathbf{V}, \mathbf{T}) = ||E_{\mathbf{v}}\mathbf{F^V} - E_{\mathbf{T}}\mathbf{F^T}||^2 \tag{12}$$

Here, $F^V = \{f_1^V, f_2^V, ..., f_n^V\}$ and $F^T = \{f_1^T, f_2^T, ..., f_n^T\}$. In the above equation, MMD is the distance between the means of both distributions. L_MMD loss can be defined as follows,

$$\begin{aligned} L_{MMD}(\mathbf{V}, \mathbf{T}) &= (E_{\mathbf{v}}\mathbf{F^V})^2 + (E_{\mathbf{T}}\mathbf{F^T})^2 - 2(E_{\mathbf{v}}\mathbf{F^V})(E_{\mathbf{T}}\mathbf{F^T}) \\ &= E_{\mathbf{V}}[k(f^V, f^V)] + E_{\mathbf{T}}[k(f^T, f^T)] - 2E_{\mathbf{VT}}[k(f^V, f^T)] \end{aligned} \tag{13}$$

Here k is the Gaussian kernel used for the MMD computation. The first two terms represent the within-modality comparison. The last term is the inter-modality alignment term. MMD shows the proximity between these two distributions. By computing MMD for every sample, the model learns how far the distributions are from each other and aligns the modalities to learn a modality invariant feature representation.

Overall Loss. Identity loss L_{ID} is calculated for each identity with respect to the modality. Hence, this is done at the end of the fully connected layer to make it modality-specific. Using the part features, the part feature attention loss L_P is calculated. Finally, IPM loss is calculated with the modality invariant features. First, the Wasserstein loss, L_{WD}, is calculated to bring in the perceptual similarity, then the MMD loss, L_{MMD}, is calculated to find a modality invariant feature representation. The sum of all four losses gives the overall loss of the model.

$$Total\ Loss = \alpha_1 L_{ID}^V + \alpha_1 L_{ID}^T + \alpha_2 L_P + \alpha_3 L_{WD} + \alpha_4 L_{MMD}, \quad (14)$$

The hyper-parameters $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ represent the weights assigned to each loss metric, indicating their respective significance.

4 Experimental Results and Evaluation

4.1 Dataset

Two popular cross-modality datasets, SYSU-MM01 [2] and RegDB [20], are used. The SYSU-MM01 dataset is one of the most popular cross-modal image datasets. It has RGB-IR images for 491 identities. Images from 6 cameras are available in the dataset. Of these, four are RGB cameras, and two are IR cameras.

The dataset has a total of 30071 RGB images and 15792 IR images. The images in SYSU-MM01 are of high resolution, but there are no exact RGB-IR pairs. The RegDB dataset contains images of 412 identities. For every identity, ten RGB-IR pairs are available.

4.2 Evaluation Metrics

Two main evaluation metrics are used for comparing the accuracy of the Person-ReID model. The first one is the rank-k accuracy. To find rank-k accuracy, top k matching samples are taken and compared with the query image. If the identity matches, then the Re-ID system correctly identifies the identities. This paper uses R1, R10, and R20 metrics for evaluation. The other metric used for evaluation is the precision - mean Average Precision(mAP).

4.3 Implementation Details

The training set images are resized to the shape $256 \times 256 \times 3$. Adversarial learning is used for RGB to IR translation tasks. The WCGAN is trained for 50 epochs with a learning rate of 0.0001, and then the learning rate is gradually increased to 0.005. SYSU-MM01 images are unpaired and are used to generate IR images using the proposed GAN for every RGB image. Resnet-50 is used as the backbone architecture. Four more convolutional layers are added to both the visible and thermal networks. Three more weight-sharing layers are added to the model. The model is trained using an RMS prop optimizer. Four loss functions,

Table 1. This table presents the accuracy of the proposed model alongside comparisons with state-of-the-art models on the SYSU-MM01 dataset. Accuracy metrics, Rank-1 (R1), Rank-10 (R10), and Rank-20 (R20) and mAP are shown.

Settings Methods	All-search				Indoor-search			
	R1	R10	R20	mAP	R1	R10	R20	mAP
LOMO [23]	3.64	23.18	37.28	4.53	5.75	34.35	54.9	10.19
One-stream [24]	12.04	49.68	66.74	13.67	16.94	63.55	82.1	22.95
Two stream [24]	11.65	47.99	65.5	12.85	15.6	61.18	81.02	21.49
Zero-Pad [24]	14.8	54.12	71.33	15.95	20.58	68.38	85.79	26.92
D-HSME [25]	50.85	73.36	81.66	47	50.15	72.40	81.07	46.16
Align GAN [26]	42.4	85	93.7	40.7	45.9	87.6	94.4	54.3
CMM-CML [27]	51.8	92.72	97.71	51.21	54.98	94.38	99.41	63.7
DDAG [9]	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
MMD-ReID [18]	66.75	94.16	97.38	62.25	71.64	97.75	99.52	75.95
CMCL [28]	69.97	95.26	98.27	67.42	76.48	97.92	99.68	79.94
CM-EMD [6]	73.39	96.24	98.82	68.56	80.53	98.31	**99.91**	82.71
CM-LSP [7]	76.28	94.38	97.08	76.52	82.31	98.12	**99.91**	85.16
AGMNet [14]	69.63	96.27	98.82	66.11	74.68	97.51	99.14	78.30
DMA [3]	74.57	–	–	70.41	82.85	–	–	85.10
DEN [29]	76.36	–	–	71.3	83.56	–	–	84.65
DGFFN [30]	74.76	–	–	70.13	81.66	-	-	83.70
BEMSSNet [4]	72.0	–	–	68.8	80.5	–	–	83.2
PSFLNet [31]	74.0	96.5	99.0	70.51	79.5	97.5	99.24	82.1
G^2DA [15]	57.07	90.99	96.28	55.05	63.70	94.06	98.35	69.83
GC-IFS [13]	74.83	97.32	99.10	71.54	78.70	98.68	99.63	82.29
GCM-IPM(proposed model)	**77.68**	**98.18**	**98.85**	**76.90**	**82.54**	**98.66**	98.92	**86.03**
GCM-IPM*(with re-ranking)	**88.95**	**98.95**	**99.52**	**86.67**	**91.77**	**98.92**	99.67	**92.44**

Fig. 3. Sample output images obtained from WCGAN(proposed). The RGB image on the left is given as input to the GAN. The IR image on the right is the generated output of the WCGAN.

namely, L_{ID}, L_P, L_{WD}, and L_{MMD} are used, hence loss weights need to be set. While training the SYSU-MM01 dataset, loss weights $[\alpha_1, \alpha_2, \alpha_3, \alpha_4]$ are set to [1, 0.5, 0.5, 0.5]. For the RegDB dataset, loss weights $[\alpha_1, \alpha_2, \alpha_3, \alpha_4]$ are set to [1, 0.3, 0.3, 0.3]. The batch size used for training is 256 and the initial learning rate is set to 0.0001 and later increased to 0.001 after 40 epochs. The model is implemented in TensorFlow and trained for a total of 150 epochs. Training is done on NVIDIA Tesla V100 with the GAN requiring 26 h for training, while the Re-ID model requires 5 h to complete its training.

Table 2. This table presents the accuracy of the proposed model alongside comparisons with state-of-the-art models on the RegDB dataset. Accuracy metrics, Rank-1 (R1), Rank-10 (R10), and Rank-20 (R20)and mAP are shown.

Settings Methods	Visible to Thermal				Thermal to Visible			
	R1	R10	R20	mAP	R1	R10	R20	mAP
LOMO [23]	0.85	2.47	4.1	2.28	–	–	–	–
One-stream [24]	13.11	32.98	42.51	14.01	–	–	–	–
Two stream [24]	12.43	30.36	40.96	13.42	–	–	–	–
Zero-Pad [24]	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
D-HSME [25]	50.85	73.36	81.66	47	50.15	72.40	81,07	46.16
Align GAN [26]	57.9	–	–	53.6	56.3	–	–	53.4
CMM-CML [27]	59.81	80.39	88.69	60.86	–	–	–	–
DDAG [9]	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
CMCL [28]	93.40	97.63	98.90	86.77	94.16	97.70	98.69	86.69
MMD-ReID [18]	95.06	99.67	99.31	88.95	93.65	97.55	98.38	87.30
CM-EMD [6]	94.37	98.93	99.42	88.32	92.77	98.50	99.66	86.85
CM-LSP [7]	94.13	–	–	88.86	93.16	–	–	87.26
AGMNet [14]	88.40	95.10	96.94	81.45	85.34	94.56	97.48	81.19
DMA [3]	93.30	–	–	88.34	91.50	–	–	86.80
DEN [29]	95.34	–	–	90.21	94.98	–	–	90.24
DGFFN [30]	94.53	–	–	91.61	93.93	–	–	90.94
BEMSSNet [4]	94.3	–	–	92.0	94.1	–	–	91.7
PSFLNet [31]	95.87	98.63	99.23	91.08	92.32	97.45	98.53	88.28
G^2DA [15]	71.72	87.13	91.92	65.90	69.50	84.87	89.85	63.88
GC-IFS [13]	94.40	99.89	100.00	92.19	92.87	99.80	99.95	91.00
GCM-IPM(proposed model)	**95.82**	**98.97**	**99.67**	**89.34**	**94.08**	**98.97**	**99.78**	**87.90**
GCM-IPM*(with re-ranking)	**97.50**	**99.95**	**99.99**	**97.80**	**96.62**	**99.93**	**100.0**	**97.08**

4.4 WCGAN Results

The resulting images from the proposed WCGAN are shown in Fig. 3. RGB images indicate the input given to the model, and corresponding IR images are the images generated by WCGAN. The image quality has considerably improved when compared with the vanilla CycleGAN. The Re-ID results also proved that IR images can aid with the re-identification process when RGB images are taken outdoors during the night or at places where proper illumination is not available.

4.5 Re-ID Results

An extensive evaluation was conducted on the SYSU-MM01 and RegDB datasets across various settings. For SYSU-MM01, experiments were performed in all-search and indoor-search settings. Table 1 presents the rank-k and mAP accuracy of the model under two architectures: CM-IPM and the proposed GCM-IPM. GCM-IPM, which incorporates IPM loss, achieved a rank-1 accuracy of 77.68% and a rank-10 accuracy of 98.18%, marking a 3.8% improvement over the state-of-the-art. The results demonstrate that integrating IPM loss enhances the model's discriminative ability, as seen in the superior performance of GCM-IPM across all metrics in all-search settings.

Table 3. Ablation study on SYSU-MM01 dataset. Impact of different components on GCM-IPM model

Settings Methods	Components				Evaluation Metrics	
	WCGAN	Part features	Part attention	Re-ranking	R1	mAP
Baseline		✓	✓		72.36	74.91
	✓	✓	✓		77.68	76.90
	✓	✓	✓	✓	83.91	81.02

For the RegDB dataset, analyzed in Visible-to-Thermal and Thermal-to-Visible settings, only the IPM module was applied due to the paired nature of the images, omitting the generation module. As shown in Table 2, the model achieved a rank-1 accuracy of 95.82

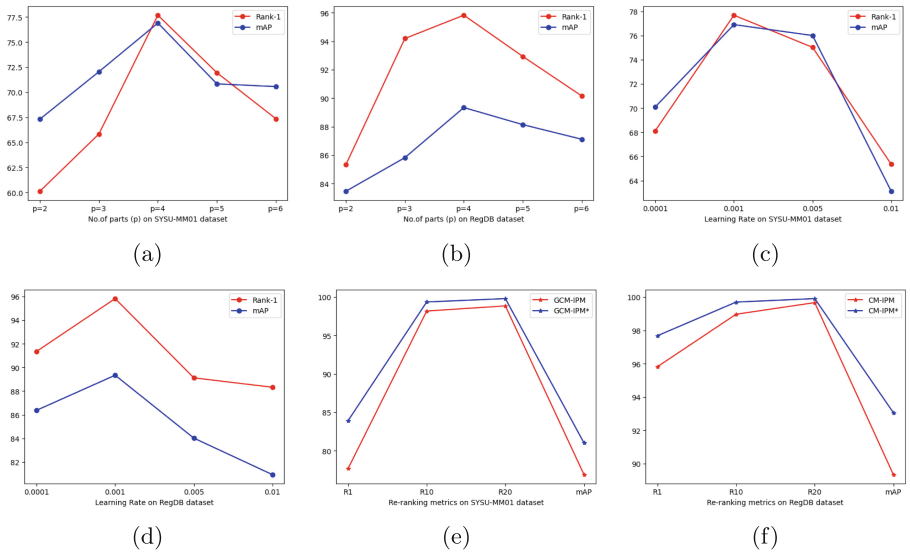


Fig. 4. Hyperparameter analysis on the proposed GCM-IPM model

Table 4. Ablation study on SYSU-MM01 and RegDB datasets. Impact of different loss functions on GCM-IPM model

Settings Methods	Components				SYSU-MM01		RegDB	
	L_{ID}	L_{WD}	L_{MMD}	L_P	R1	mAP	R1	mAP
Baseline	✓				53.80	61.34	70.14	69.85
	✓	✓			62.45	70.09	87.93	82.40
	✓	✓	✓		69.84	73.52	93.16	85.06
	✓	✓	✓	✓	77.68	76.90	95.82	89.34

Table 5. Estimation of loss weights α on SYSU-MM01 dataset.

Loss weights				Evaluation Metrics	
α_1	α_2	α_3	α_4	R1	mAP
2	2	0.5	0.5	65.97	63.42
0.5	0.5	2	2	72.33	69.13
1	1	1	1	73.19	70.25
1	0.5	0.5	0.5	77.68	76.90

Table 6. Estimation of loss weights α on RegDB dataset.

Loss weights				Evaluation Metrics	
α_1	α_2	α_3	α_4	R1	mAP
2	2	0.5	0.5	82.78	77.74
1	1	1	1	88.02	80.69
0.5	0.5	2	2	92.13	84.41
1	0.3	0.3	0.3	95.82	89.34

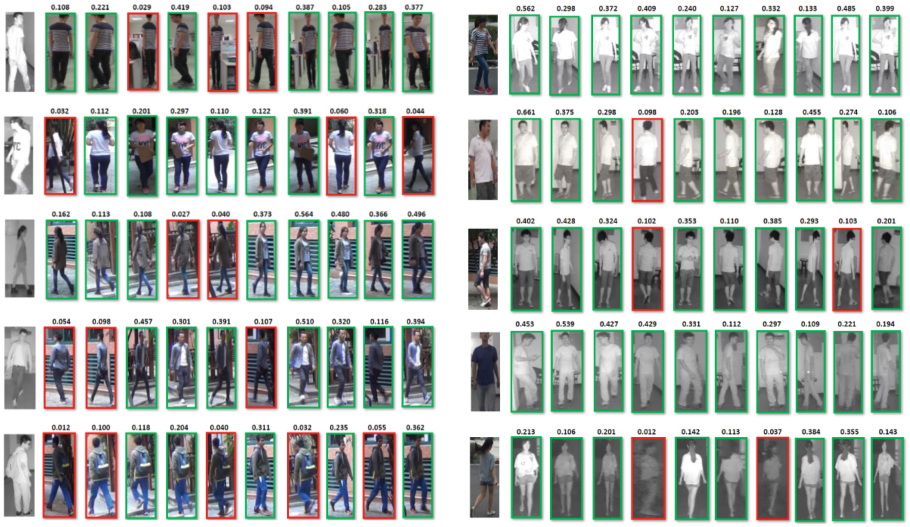


Fig. 5. Image similarity check is done by randomly sampling ten identities from the gallery and trying to match images in two settings: (a) visible to thermal, (b) thermal to visible. The values on the top of each image are the image similarity values obtained from the final layer of the proposed model. Green boxes indicate a correct match, and red boxes indicate a wrong match. (Color figure online)

Parameter Analysis. Inorder to analyse the hyper-parameters used while training, we perform the parameter analysis which includes parameters namely, loss weights Table 5 and 6, no. of parts Fig. 4a and Fig. 4b, learning rate Fig. 4c and Fig. 4d and re-ranking results Fig. 4e and Fig. 4f.

Ablation Study. To understand how effective each component is, we perform ablation study on the proposed GCM-IPM model as shown in Table 3. It can be observed that once WCGAN module results are added to the training, the performance of the model improves. Inorder to test the loss function's impact on the training, we perform ablation study on different loss functions as shown

in Table 4. The model's R1 improves to 77.68% when the proposed combination of loss functions are used.

Visualisation of Re-ID Results. The image retrieval results from the proposed GCM-IPM model are shown in Fig. 5. Evaluation is done in two settings, Visible-to-Thermal and Thermal-to-Visible. Ten identities were randomly chosen for the evaluation. The first five identities are tested under the Visible-to-Thermal setting. A single RGB image serves as the query image, while ten IR images are drawn from the test set to assess re-identification accuracy. This is repeated for all ten identities and for both modalities. The green boxes represent the samples the model has correctly re-identified, and the red boxes represent the samples that the model did not re-identify correctly. The values on top of the sample are the output similarity scores obtained from the proposed system.

5 Conclusion

This work proposes a framework that can translate RGB images to IR images and use those images for Person-ReID. To solve the unavailability of RGB-IR pairs, the WCGAN is proposed. Wasserstein distance-based generative model converges and can be trained till optimality. Furthermore, a re-identification system utilizing integral probability metrics is introduced. This system is integral in evaluating the similarity between images across different modalities. The proposed loss function based on the integral probability metric aids in learning the perceptual similarities between different image modalities. It helps align the features extracted from different modalities within the model, enabling a more coherent representation. To aid in modality alignment, part-based feature attention loss is proposed. Extensive experiments and analysis on two major datasets prove that the addition of the generation module has enhanced the feature space by providing paired images to the re-identification module. The analysis also delves into how the combined use of part attention loss, Wasserstein distance and Maximum Mean Discrepancy loss contributes to reducing the variance between different modalities for the same individual. This process leads to a more robust and aligned feature space, critical for accurate person re-identification across different image types. As a future direction, the integration of additional data modalities like text and sketches is proposed to enrich the feature space further.

6 Code Availability

The code used for the proposed model is given as supplementary material.

Acknowledgements. The authors would like to thank the Indian Institute of Space Science and Technology for funding this research. We would also like to thank the Indian Institute of Technology, Palakkad, for supporting this work under Technology IHub Foundation, DST, Government of India, Grant IPTIF/HRD/DF/025/SEP36.

References

1. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future. *arXiv preprint*arXiv:1610.02984, 2016
2. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2872–2893 (2021)
3. Cui, Z., Zhou, J., Peng, Y.: DMA: dual modality-aware alignment for visible-infrared person re-identification. IEEE Trans. Inf. Forensics Secur. (2024)
4. Gwon, S., Kim, S., Seo, K.: Balanced and essential modality-specific and modality-shared representations for visible-infrared person re-identification. IEEE Signal Process. Lett. (2024)
5. Zhang, Z., Jiang, S., Huang, C., Li, Y., Xu, R.Y.D.: RGB-IR cross-modality person ReID based on teacher-student GAN model. Pattern Recogn. Lett. **150**, 155–161 (2021)
6. Ling, Y., et al.: Cross-modality earth mover's distance for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1631–1639 (2023)
7. Wang, X., Li, C., Ma, X.: Cross-modal local shortest path and global enhancement for visible-thermal person re-identification. arXiv preprint arXiv:2206.04401, 2022
8. Feng, Z., Lai, J., Xie, X.: Learning modality-specific representations for visible-infrared person re-identification. IEEE Trans. Image Process. **29**, 579–590 (2019)
9. Ye, M., Shen, J., J. Crandall, D., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. ECCV 2020. LNCS, vol. 12362, pp. 229–247. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58520-4_14
10. Lu, Y., et al.: Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13379–13389 (2020)
11. Kniaz, V.V., Knyaz, V.A., Hladuvka, J., Kropatsch, W.G., Mizginov, V.: Thermalgan: multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
12. Wang, G.-A., et al.: Cross-modality paired-images generation for RGB-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12144–12151 (2020)
13. Qi, J., Liang, T., Liu, W., Li, Y., Jin, Y.: A generative-based image fusion strategy for visible-infrared person re-identification. IEEE Trans. Circuits Syst. Video Technol. (2023)
14. Liu, H., Xia, D., Jiang, W.: Towards homogeneous modality learning and multi-granularity information exploration for visible-infrared person re-identification. IEEE J. Sel. Top. Signal Process. (2023)
15. Wan, L., Sun, Z., Jing, Q., Chen, Y., Lijing, L., Li, Z.: G2DA: geometry-guided dual-alignment learning for RGB-infrared person re-identification. Pattern Recogn. **135**, 109150 (2023)
16. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**, 99–121 (2000)
17. Zhang, Z., Xie, Y., Li, D., Zhang, W., Tian, Q.: Learning to align via wasserstein for person re-identification. IEEE Trans. Image Process. **29**, 7104–7116 (2020)

18. Jambigi, C., Rawal, R., Chakraborty, A.: Mmd-reid: a simple but effective solution for visible-thermal person reid (2021)
19. Murali, N., Mishra, D.: Wasserstein distance for attention based cross modality person re-identification. In: 2022 IEEE 19th India Council International Conference (INDICON), pp. 1–6. IEEE (2022)
20. Wu, A., Zheng, W.S., Gong, S., Lai, J.: RGB-IR person re-identification by cross-modality similarity preservation. Int. J. Comput. Vis. (2020)
21. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223. PMLR, 2017
22. Gretton, A., Borgwardt, K.M., Rasch, M.J., Scholkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. **13**(1), 723–773 (2012)
23. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206, 2015
24. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: RGB-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380–5389 (2017)
25. Hao, Y., Wang, N., Li, J., Gao, X.: HSME: hypersphere manifold embedding for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8385–8392 (2019)
26. Wang, G.A., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3623–3632 (2019)
27. Ling, Y., Zhong, Z., Luo, Z., Rota, P., Li, S., Sebe, N.: Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 889–897 (2020)
28. Fang, P., Zhang, Y., Lan, Z.: Beyond a strong baseline: cross-modality contrastive learning for visible-infrared person re-identification. Mach. Vis. Appl. **34**(6), 105 (2023)
29. Kim, S., Gwon, S., Seo, K.: Enhancing diverse intra-identity representation for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2513–2522 (2024)
30. Cai, S., Yang, S., Jing, H., Xi, W.: Dual-granularity feature fusion in visible-infrared person re-identification. IET Image Proc. **18**(4), 972–980 (2024)
31. Chan, S., Feng, D., Tang, T., Zhang, G., Jiang, X., Guan, Q.: Parameter sharing and multi-granularity feature learning for cross-modality person re-identification. Complex Intell. Syst. **10**(1), 949–962 (2024)

Author Index

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15314, pp. 489–490, 2025.
https://doi.org/10.1007/978-3-031-78341-8