

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

LNCS 15328

Pattern Recognition

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

28 Part XXVIII



Lecture Notes in Computer Science

15328

Founding Editors

Gerhard Goos
Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*

Wen Gao, *Peking University, Beijing, China*

Bernhard Steffen , *TU Dortmund University, Dortmund, Germany*

Moti Yung , *Columbia University, New York, NY, USA*

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 XXVIII

Editors

Apostolos Antonacopoulos 
University of Salford
Salford, UK

Rama Chellappa 
Johns Hopkins University
Baltimore, MD, USA

Saumik Bhattacharya 
IIT Kharagpur
Kharagpur, India

Subhasis Chaudhuri 
Indian Institute of Technology Bombay
Mumbai, India

Cheng-Lin Liu 
Chinese Academy of Sciences
Beijing, China

Umapada Pal 
Indian Statistical Institute Kolkata
Kolkata, India

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-031-78103-2

ISBN 978-3-031-78104-9 (eBook)

<https://doi.org/10.1007/978-3-031-78104-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

If disposing of this product, please recycle the paper.

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 bio-signal 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 IJCPD 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
Josef Kittler
Anil Jain

Indian Statistical Institute, Kolkata, India
University of Surrey, UK
Michigan State University, USA

Program Chairs

Apostolos Antonacopoulos
Subhasis Chaudhuri
Rama Chellappa
Cheng-Lin Liu

University of Salford, UK
Indian Institute of Technology, Bombay, India
Johns Hopkins University, USA
Institute of Automation, Chinese Academy of
Sciences, China

Publication Chairs

Ananda S. Chowdhury
Wataru Ohyama

Jadavpur University, India
Tokyo Denki University, Japan

Competition Chairs

Richard Zanibbi
Lianwen Jin
Laurence Likforman-Sulem

Rochester Institute of Technology, USA
South China University of Technology, China
Télécom Paris, France

Workshop Chairs

P. Shivakumara
Stephanie Schuckers
Jean-Marc Ogier
Prabir Bhattacharya

University of Salford, UK
Clarkson University, USA
Université de la Rochelle, France
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 Lladós	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	Jiixin Zhang
Yuliang Liu	Yiwu Zhong
Yongxin Shi	

Reviewers (Conference Papers)

Aakanksha Aakanksha
 Aayush Singla
 Abdul Muqet
 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
 Anwasha Sengupta
 Archisman Adhikary
 Arjan Kuijper
 Arnab Kumar Das

Arnav Bhavsar	Bin-Bin Jia
Arnav Varma	Binbin Yong
Arpita Dutta	Bindita Chaudhuri
Arshad Jamal	Bindu Madhavi Tummala
Artur Jordao	Binh M. Le
Arunkumar Chinnaswamy	Bi-Ru Dai
Aryan Jadon	Bo Huang
Aryaz Baradarani	Bo Jiang
Ashima Anand	Bob Zhang
Ashis Dhara	Bowen Liu
Ashish Phophalia	Bowen Zhang
Ashok K. Bhateja	Boyang Zhang
Ashutosh Vaish	Boyu Diao
Ashwani Kumar	Boyun Li
Asifuzzaman Lasker	Brian M. Sadler
Atefeh Khoshkhahtinat	Bruce A. Maxwell
Athira Nambiar	Bryan Bo Cao
Attilio Fiandrotti	Buddhika L. Semage
Avandra S. Hemachandra	Bushra Jalil
Avik Hati	Byeong-Seok Shin
Avinash Sharma	Byung-Gyu Kim
B. H. Shekar	Caihua Liu
B. Uma Shankar	Cairong Zhao
Bala Krishna Thunakala	Camille Kurtz
Balaji Tk	Carlos A. Caetano
Balázs Pálffy	Carlos D. Martá-Nez-Hinarejos
Banafsheh Adami	Ce Wang
Bang-Dang Pham	Cevahir Cigla
Baochang Zhang	Chakravarthy Bhagvati
Baodi Liu	Chandrakanth Vipparla
Bashirul Azam Biswas	Changchun Zhang
Beiduo Chen	Changde Du
Benedikt Kottler	Changkun Ye
Beomseok Oh	Changxu Cheng
Berkay Aydin	Chao Fan
Berlin S. Shaheema	Chao Guo
Bertrand Kerautret	Chao Qu
Bettina Finzel	Chao Wen
Bhavana Singh	Chayan Halder
Bibhas C. Dhara	Che-Jui Chang
Bilge Günsel	Chen Feng
Bin Chen	Chenan Wang
Bin Li	Cheng Yu
Bin Liu	Chenghao Qian
Bin Yao	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 Coquenat
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
Dipanjoyoti 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	Galal Binamakhshen
Emna Ghorbel	Ganesh Krishnasamy
Enrique Naredo	Gang Pan
Enyu Cai	Gangyan Zeng
Eric Patterson	Gani Rahmon
Ernest Valveny	Gaurav Harit
Eva Blanco-Mallo	Gennaro Vessio
Eva Breznik	Genoveffa Tortora
Evangelos Sartinas	George Azzopardi
Fabio Solari	Gerard Ortega
Fabiola De Marco	Gerardo E. Altamirano-Gomez
Fan Wang	Gernot A. Fink
Fangda Li	Gibran Benitez-Garcia
Fangyuan Lei	Gil Ben-Artzi
Fangzhou Lin	Gilbert Lim
Fangzhou Luo	Giorgia Minello
Fares Bougourzi	Giorgio Fumera
Farman Ali	Giovanna Castellano
Fatiha Mokdad	Giovanni Puglisi
Fei Shen	Giulia Orrù
Fei Teng	Giuliana Ramella
Fei Zhu	Gökçe Uludoğan
Feiyan Hu	Gopi Ramena
Felipe Gomes Oliveira	Gorthi Rama Krishna Sai Subrahmanyam
Feng Li	Gourav Datta
Fengbei Liu	Gowri Srinivasa
Fenghua Zhu	Gozde Sahin
Fillipe D. M. De Souza	Gregory Randall
Flavio Piccoli	Guanjie Huang
Flavio Prieto	Guanjun Li
Florian Kleber	Guanwen Zhang
Francesc Serratosa	Guanyu Xu
Francesco Bianconi	Guanyu Yang
Francesco Castro	Guanzhou Ke
Francesco Ponzio	Guhnoo Yun
Francisco Javier Hernández López	Guido Borghi
Frédéric Rayar	Guilherme Brandão Martins
Furkan Osman Kar	Guillaume Caron
Fushuo Huo	Guillaume Tochon
Fuxiao Liu	Guocai Du
Fu-Zhao Ou	Guohao Li
Gabriel Turinici	Guoqiang Zhong
Gabrielle Flood	Guorong Li
Gajjala Viswanatha Reddy	Guotao Li
Gaku Nakano	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
Jiyang 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
Liviú-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 Kämpel
Martina Pastorino
Marwan Turki
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	Palash Ghosal
Narayan Vetrekar	Pallav Dutta
Narendra D. Londhe	Paolo Rota
Nathalie Girard	Paramanand Chandramouli
Nati Ofir	Paria Mehrani
Naval Kishore Mehta	Parth Agrawal
Nazmul Shahadat	Partha Basuchowdhuri
Neeti Narayan	Patrick Horain
Neha Bhargava	Pavan Kumar
Nemanja Djuric	Pavan Kumar Anasosalu Vasu
Newlin Shebiah R.	Pedro Castro
Ngo Ba Hung	Peipei Li
Nhat-Tan Bui	Peipei Yang
Niaz Ahmad	Peisong Shen
Nick Theisen	Peiyu Li
Nicolas Passat	Peng Li
Nicolas Ragot	Pengfei He
Nicolas Sidere	Pengrui Quan
Nikolaos Mitianoudis	Pengxin Zeng
Nikolas Ebert	Pengyu Yan
Nilah Ravi Nair	Peter Eisert
Nilesh A. Ahuja	Petra Gomez-Krämer
Nilkanta Sahu	Pierrick Bruneau
Nils Murrugarra-Llerena	Ping Cao
Nina S. T. Hirata	Pingping Zhang
Ninad Aithal	Pintu Kumar
Ning Xu	Pooja Kumari
Ningzhi Wang	Pooja Sahani
Niraj Kumar	Prabhu Prasad Dev
Nirmal S. Punjabi	Pradeep Kumar
Nisha Varghese	Pradeep Singh
Norio Tagawa	Pranjal Sahu
Obaidullah Md Sk	Prasun Roy
Oguzhan Ulucan	Prateek Keserwani
Olfa Mechi	Prateek Mittal
Oliver Tüselmann	Praveen Kumar Chandaliya
Orazio Pontorno	Praveen Tirupattur
Oriol Ramos Terrades	Pravin Nair
Osman Akin	Preeti Gopal
Ouadi Beya	Preety Singh
Ozge Mercanoglu Sincan	Prem Shanker Yadav
Pabitra Mitra	Prerana Mukherjee
Padmanabha Reddy Y. C. A.	Prerna A. Mishra
Palaash Agrawal	Prianka Dey
Palaiahnakote Shivakumara	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
Ruwei 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
Siladitya 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 Napoléon
 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	Wataru Ohyama
Tim Raven	Wee Kheng Leow
Ting kai Liu	Wei Chen
Tingting Yao	Wei Cheng
Tobias Meisen	Wei Hua
Toby P. Breckon	Wei Lu
Tong Chen	Wei Pan
Tonghua Su	Wei Tian
Tran Tuan Anh	Wei Wang
Tri-Cong Pham	Wei Wei
Trishna Saikia	Wei Zhou
Trung Quang Truong	Weidi Liu
Tuan T. Nguyen	Weidong Yang
Tuan Vo Van	Weijun Tan
Tushar Shinde	Weimin Lyu
Ujjwal Karn	Weinan Guan
Ukrit Watchareeruetai	Weining Wang
Uma Mudenagudi	Weiqiang Wang
Umarani Jayaraman	Weiwei Guo
V. S. Malemath	Weixia Zhang
Vallidevi Krishnamurthy	Wei-Xuan Bao
Ved Prakash	Weizhong Jiang
Venkata Krishna Kishore Kolli	Wen Xie
Venkata R. Vavilthota	Wenbin Qian
Venkatesh Thirugnana Sambandham	Wenbin Tian
Verónica Maria Vasconcelos	Wenbin Wang
Véronique Ve Eglin	Wenbo Zheng
Víctor E. Alonso-Pérez	Wenhan Luo
Vinay Palakkode	Wenhao Wang
Vinayak S. Nageli	Wen-Hung Liao
Vincent J. Whannou De Dravo	Wenjie Li
Vincenzo Conti	Wenkui Yang
Vincenzo Gattulli	Wenwen Si
Vineet Padmanabhan	Wenwen Yu
Vishakha Pareek	Wenwen Zhang
Viswanath Gopalakrishnan	Wenwu Yang
Vivek Singh Baghel	Wenxi Li
Vivekraj K.	Wenxi Yue
Vladimir V. Arlazarov	Wenxue Cui
Vu-Hoang Tran	Wenzhuo Liu
W. Sylvia Lilly Jebarani	Widhiyo Sudiyono
Wachirawit Ponghiran	Willem Dijkstra
Wafa Khlif	Wolfgang Fuhl
Wang An-Zhi	Xi Zhang
Wanli Xue	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 XXVIII

Enhanced Classification and Segmentation of Brain Tumors in MRI Images Using Custom CNN and U-Net Models with XAI	1
<i>Pathikreet Chowdhury and Gargi Srivastava</i>	
Deep BI-RADS Network for Improved Cancer Detection from Mammograms	17
<i>Gil Ben-Artzi, Feras Daragma, and Shahar Mahpod</i>	
Histopathological Diagnosis of Meningioma and Solitary Fibrous Tumors Based on a Multi-scale Fusion Approach Utilizing Vision Transformer and Texture Analysis	31
<i>Mohamed T. Azam, Hossam Magdy Balaha, Dibson D. Gondim, Akshikumar Mistry, Mohammed Ghazal, and Ayman El-Baz</i>	
HFENet: High-Frequency Enhanced Network for Shape-Aware Segmentation of Left Ventricle in Pediatric Echocardiograms	46
<i>Tianxiang Chen, Ziyang Wang, and Zi Ye</i>	
Chaos Theory Based Gravitational Search Algorithm For Medical Image Segmentation	58
<i>Sajad Ahmad Rather, Partha Pratim Roy, and Sujit Das</i>	
Integrated Grading Framework for Histopathological Breast Cancer: Multi-level Vision Transformers, Textural Features, and Fusion Probability Network	76
<i>Hossam Magdy Balaha, Khadiga M. Ali, Ali Mahmoud, Mohammed Ghazal, and Ayman El-Baz</i>	
Dual-MambaNet: A Lightweight Dual-Branch Brain Image Segmentation Network Based on Local Attention and Mamba	92
<i>Feifei Zhang, Fei Shi, Dayong Ren, Zhenhong Jia, and Jianyi Wang</i>	
Location Matters: Harnessing Spatial Information to Enhance the Segmentation of the Inferior Alveolar Canal in CBCTs	108
<i>Luca Lumetti, Vittorio Pipoli, Federico Bolelli, Elisa Ficarra, and Costantino Grana</i>	

Adaptive Class Learning to Screen Diabetic Disorders in Fundus Images of Eye	124
<i>Shramana Dey, Pallabi Dutta, Riddhasree Bhattacharyya, Surochita Pal Das, Sushmita Mitra, and Rajiv Raman</i>	
EDB-Net: An Edge-Guided Dual-Branch Neural Network for Skin Cancer Classification	138
<i>Amartya Ray, Soumyajit Gayen, Dmitrii Kaplun, and Ram Sarkar</i>	
PSIVUS: Atherosclerotic Plaque Segmentation in Intravascular Ultrasound Images via Active Learning	154
<i>Anuradha Mahato, Paromita Banerjee, Rutvik Narendrabhai Jethava, Bhanu Duggal, Angshuman Paul, Mayank Vatsa, and Richa Singh</i>	
Generalist Segmentation Algorithm for Photoreceptors Analysis in Adaptive Optics Imaging	168
<i>Mikhail Kulyabin, Aline Sindel, Hilde R. Pedersen, Stuart Gilson, Rigmor Baraas, and Andreas Maier</i>	
UNeXt++: A Serial-Parallel Hybrid UNeXt for Rapid Medical Image Segmentation	183
<i>Yan Li, Juelin Wang, Yunteng Deng, Binyang Li, and Junlin Hu</i>	
Efficient Adapter on Pre-trained Visual Feature Reliance in Medical Visual Question Answering	198
<i>Aakansha Mishra, Prateek Keserwani, Vikram N. Rajendiran, and Ashok K. Senapati</i>	
MUMR: Mask-UnMask Regions Framework for AMD Grades Classification Based on Inter-regional Interactions	213
<i>Ibrahim Abdelhalim, Mohamed Elsharkawy, Namuunaa Nadmid, Mohammed Ghazal, Ali Mahmoud, and Ayman El-Baz</i>	
A New Attention Based UNet and Gated Edge Attention Network for Retinal Vessel Segmentation	224
<i>Ayush Roy, Shivakumara Palaiahnakote, Umapada Pal, and Sukalpa Chanda</i>	
TractoEmbed: Modular Multi-level Embedding Framework for White Matter Tract Segmentation	240
<i>Anoushkrit Goel, Bipanjit Singh, Ankita Joshi, Ranjeet Ranjan Jha, Chirag Ahuja, Aditya Nigam, and Arnav Bhavsar</i>	

Unsupervised Domain Adaptation for Cross-Device Iris Liveness
 Detection Model Transfer 256
*Xiuying Wu, Chenxi Du, Hui Zhang, Jing Liu, Dexin Zhang,
 and Hang Zou*

A Collaborative Approach Using Ridge-Valley Minutiae for More
 Accurate Contactless Fingerprint Matching 273
Ritesh Vyas and Ajay Kumar

A Generative Method for Finger Knuckle Print Recognition 288
Yuqi Wang, Bob Zhang, Shuyi Li, and Hao Yang

Multimodal Drivers of Attention Interruption to Baby Product Video Ads 303
Wen Xie, Lingfei Luan, Yanjun Zhu, Yakov Bart, and Sarah Ostadabbas

Facial Wrinkle Segmentation for Cosmetic Dermatology: Pretraining
 with Texture Map-Based Weak Supervision 319
Junho Moon, Haejun Chung, and Ikbeom Jang

ECMISM: Speech Recognition via Enhancing Conformer Models
 with Innovative Scoring Matrices 335
Jiang Zhang, Liejun Wang, Yinfeng Yu, and Miaomiao Xu

Benchmarking AI in Mental Health: A Critical Examination of LLMs
 Across Key Performance and Ethical Metrics 351
Rui Yuan, Wanting Hao, and Chun Yuan

Sampling Rate Adaptive Speaker Verification from Raw Waveforms 367
*Vinayak Abrol, Anshul Thakur, Akshat Gupta, Xiaomo Liu,
 and Sameena Shah*

Adjustable Gating Prompt Transformer for Facial Attribute Recognition
 with Limited Labeled Data 383
*Qinxian Ye, Si Chen, Da-Han Wang, Nanfeng Jiang, Yanfei Su,
 and Yan Yan*

IPHGaze: Image Pyramid Gaze Estimation with Head Pose Guidance 399
*Hekuangyi Che, Dongchen Zhu, Wenjun Shi, Guanghui Zhang,
 Hang Li, Lei Wang, and Jiamao Li*

BCNet: Binocular Cooperative Network for Gaze Estimation 415
*Dongchen Zhu, Minjin Lin, Hekuangyi Che, Wenjun Shi,
 Guanghui Zhang, Hang Li, Lei Wang, and Jiamao Li*


mmAlphabet: Air Writing Alphabet Recognition System Based
on mmWave FMCW Radar and Convolutional Neural Network 431
Chao-Wang Huang, Chien-Yao Wang, and Jia-Ching Wang

FG-MDM: Towards Zero-Shot Human Motion Generation
via ChatGPT-Refined Descriptions 446
*Xu Shi, Wei Yao, Chuanchen Luo, Junran Peng, Hongwen Zhang,
and Yunlian Sun*

Author Index 463



Enhanced Classification and Segmentation of Brain Tumors in MRI Images Using Custom CNN and U-Net Models with XAI

Pathikreet Chowdhury and Gargi Srivastava^(✉) 

Rajiv Gandhi Institute of Petroleum Technology, Jais 229304, Uttar Pradesh, India
{21cs2026,gsrivastava}@rgipt.ac.in
<http://www.rgipt.ac.in>

Abstract. This study aims to classify brain tumors in MRI images into four categories: glioma, meningioma, absent, and pituitary tumors, as well as segment low-grade gliomas. We evaluate our proposed models on four publicly available datasets to ensure robustness and generalizability. For classification tasks, we compare the performance of our custom CNN model against established models like ResNet and VGG. For segmentation tasks, we compare our custom U-Net model with the original U-Net and ResNet-based encoders. To validate the effectiveness of our models, we employ the Explainable AI (XAI) method LIME, providing insights into why our custom architectures outperform others. Our custom U-Net model achieves a validation accuracy of 99.79% and an Intersection over Union (IoU) score of 0.889 for low-grade glioma segmentation. Additionally, we report a LIME explanation stability score of 0.8169 and a sparsity score of 0.1190. The proposed custom CNN model achieves a validation accuracy of 98.70%, weighted avg precision of 97.63%, recall of 97.64% and weighted F1 - Score of 97.63%. The model achieves a LIME stability score of 0.923 and a sparsity score of 0.203. These results highlight the potential of our custom models to enhance accuracy and interpretability in brain tumor classification and segmentation tasks, offering significant improvements over existing methodologies. The custom U-net model is also an excellent negative classifier achieving a perfect 1.00 IoU score for classifying MRI scans which do not have any tumor.

Keywords: Brain Tumor · Explainable AI · LIME

1 Introduction

Segmentation and classification of brain tumors in MRI images have been extensively researched over the past few years. Researchers have been leveraging deep neural networks to achieve significant improvements in this area [1–5].

However, the advent of Explainable AI (XAI) [6] has introduced new possibilities, allowing us to scrutinize network structures and understand the underlying mechanisms behind their final results. This perspective is crucial for discerning

why certain models perform better than others and how existing neural network architectures can be modified for improved efficiency and accuracy. Consequently, this can lead to smaller networks that save training time and provide faster results when deployed.

Previously, deep learning models operated largely as black boxes. Input images were fed into the network, and based on the output, additional images that led to misclassifications were incorporated into the training set to enhance learning. Developers focused on layer adjustment, feedback incorporation, loss prevention, and hyperparameter tuning, often relying on trial and error to achieve better results. With XAI, the opaque nature of neural networks has transformed into a transparent process, providing clear insights into the inner workings of the network. This transparency empowers developers to exert more control over the network, significantly reducing development time and reliance on trial and error [7].

Moreover, traditional deep learning models were constrained to predicting predefined classes without the ability to indicate uncertainty. XAI now enables us to identify when the network reaches an "I don't know" stage, allowing for the development of custom models tailored to specific needs rather than merely adapting existing network architectures for different domains through transfer learning.

Explainable AI also enhances the accountability, fairness, and transparency of deep learning models. This is particularly important in medical diagnostics, where the consequences of misclassification and segmentation can be severe.

In this paper, we develop a custom segmentation and classification model, providing interpretation with the LIME model [8]. We compare our results with benchmark neural network models such as ResNet [9] and VGG Net [10].

The paper is organized as follows: we detail our methodology in Sect. 2, followed by the experimentation details in Sect. 3 and presentation of our results and their discussion in Sect. 4. The conclusions and implications our findings are presented in the Sect. 5.

2 Methodology

2.1 Proposed Neural Network for Segmentation

The proposed segmentation network is a custom U-Net, specifically designed to enhance the segmentation of brain MRI images. The architecture retains the canonical encoder-decoder structure of the original U-Net, with several enhancements aimed at capturing more complex features and improving segmentation performance. Figure 1 displays the network architecture.

Proposed Enhanced U-Net. Our proposed Enhanced U-Net model introduces several critical modifications to the original U-Net architecture, aimed at augmenting its performance for brain tumor segmentation tasks. These enhancements are strategically designed to improve feature extraction, model robustness,

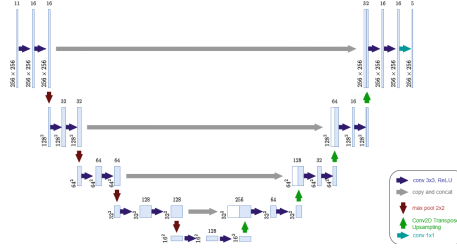


Fig. 1. Proposed Neural Network for Segmentation

and overall segmentation accuracy. Here are the key differences and enhancements compared to the original U-Net:

Residual Connections: Residual connections are integrated within each convolutional block. This strategy helps alleviate the vanishing gradient problem and facilitates the training of deeper networks by allowing gradients to flow more effectively through the network. The residual path is implemented by adding the input to the output of the convolutional layers within the block.

Batch Normalization: Batch normalization layers are employed after each convolutional layer. This technique stabilizes and accelerates the training process by normalizing the activations, thus reducing internal covariate shift. It also allows for higher learning rates and reduces sensitivity to initialization.

Spatial Dropout: Spatial Dropout is incorporated within each convolutional block to improve generalization and prevent overfitting. This form of dropout randomly drops entire feature maps rather than individual elements, which is particularly effective for spatial data, ensuring that the model does not become overly reliant on specific feature maps.

By incorporating these enhancements, our proposed Enhanced U-Net model aims to significantly improve segmentation performance in terms of accuracy, robustness, and generalization capabilities.

2.2 Proposed Neural Network for Classification

The proposed classification network is a custom Convolutional Neural Network (CNN) specifically designed for classifying brain MRI images into benign and malignant tumors. It incorporates advanced convolutional layers with batch normalization, ReLU activations, and strategic pooling operations to enhance feature extraction and classification accuracy. Figure 2 displays the Proposed Neural Network for Classification.

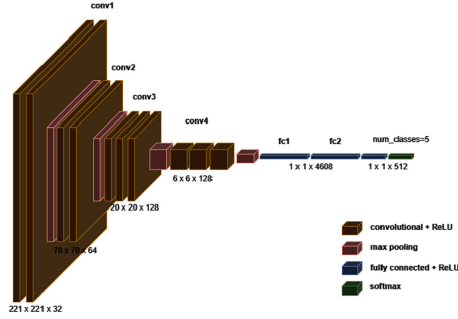


Fig. 2. Proposed Neural Network for Classification

Convolutional Layers: The network comprises four convolutional layers, each contributing to hierarchical feature learning:

- Conv Layer 1: Applies 32 filters of size 4×4 with a stride of 1, utilizing batch normalization and ReLU activation.
- Conv Layer 2: Employs 64 filters of size 4×4 with a stride of 1, followed by batch normalization and ReLU activation.
- Conv Layer 3: Utilizes 128 filters of size 4×4 with a stride of 1, integrated with batch normalization and ReLU activation.
- Conv Layer 4: Applies 128 filters of size 4×4 with a stride of 1, followed by batch normalization and ReLU activation.

Pooling Layers: Pooling operations are strategically placed to reduce spatial dimensions and enhance translational invariance:

- Pooling Layer 1: After the first and second convolutional layers, a max pooling operation with a kernel size of 3×3 and a stride of 3 is applied.
- Pooling Layer 2: Following the third convolutional layer, a max pooling operation with a kernel size of 3×3 and a stride of 2 is employed.

Fully Connected Layers: After feature extraction through convolution and pooling, the network incorporates fully connected layers for final classification:

- Fully Connected (FC) Layer 1: Composed of 512 units with ReLU activation, facilitating complex feature integration.
- Dropout Regularization: Implemented with a dropout rate of 0.5 before the final classification layer to prevent overfitting by randomly deactivating neurons during training.
- Final Classification FC Layer 2: The last layer outputs predictions corresponding to the number of classes, leveraging integrated features for accurate classification.

Enhanced Performance: The strategic use of two different strides in the max pooling layers significantly enhances test accuracy, allowing the network to effectively capture and integrate hierarchical features from varying spatial scales. This design choice optimizes the network's ability to discern between benign and malignant brain tumor images with increased precision and reliability in medical imaging applications.

2.3 Applying LIME on Proposed Models

Application on Classification Models: For the classification task, LIME is applied to understand the reasons behind the classification of brain MRI images as benign or malignant or no tumor. The steps include:

1. **Selecting Instances:** A subset of correctly classified and misclassified images from the test set is selected for explanation.
2. **Generating Explanations:** LIME generates explanations for each selected instance by highlighting the regions of the image that most influenced the model's decision.
3. **Visualizing Explanations:** The explanations are visualized as heatmaps overlaid on the original images, showing which parts of the image contributed most to the classification.

Application on Segmentation Models: For the segmentation task, LIME helps to explain why certain regions of the MRI images were segmented as tumor areas. The steps include:

1. **Selecting Instances:** A set of segmented images, including both successful and failed segmentations, is chosen for explanation.
2. **Generating Perturbations:** Perturbations are applied to the images, and the model's segmentation predictions for these perturbed samples are obtained.
3. **Fitting an Interpretable Model:** A simpler model is fitted to approximate the segmentation decisions of the original model.
4. **Visualizing Explanations:** The important features (image segments) that influenced the segmentation decision are visualized, helping to understand the model's behavior.

2.4 Algorithm for the Proposed Work

The following steps outline the algorithm we employed in our research for preprocessing data, training custom models for classification and segmentation, applying LIME for explainability, and evaluating the models using specific metrics.

Step 1: Data Preparation:

1. **Data Acquisition:** We collected a publicly available dataset of brain MRI images. The dataset includes labels for classification (glioma , pituitary , meningioma ,no tumor) and ground truth masks for segmentation.

2. **Data Preprocessing:** We normalized the images to a standard scale (e.g., $[0, 1]$). The images were resized to a fixed size (e.g., 256×256 pixels) to ensure uniform input dimensions. Used a custom data loading function to load , preprocess MRI Images and load them in a Dataframe. We augmented the dataset using techniques such as cropping out the brain sections only to enhance model generalization and accuracy and also adding data augmentations such as horizontal flips and rotations.

Step 2: Model Training:

1. **Classification Model:** We defined a custom CNN architecture for classification. The model parameters were initialized, and the model was compiled with an appropriate optimizer (Adam) and loss function (cross-entropy loss). The dataset was split into training, validation, and test sets. We trained the model on the training set and validated it on the validation set. The training process was monitored, and early stopping was applied to prevent overfitting.
2. **Segmentation Model:** We defined a custom U-Net architecture for segmentation. The model parameters were initialized, and the model was compiled with an appropriate optimizer (Adam) and loss function (binary cross-entropy with Dice coefficient). The model was trained on the training set and validated on the validation set. The training process was monitored, and early stopping was applied to prevent overfitting.

Step 3: Model Evaluation

1. **Classification Evaluation:** We evaluated the trained classification model on the test set using metrics such as accuracy, precision, recall, and F1 score.
2. **Segmentation Evaluation:** We evaluated the trained segmentation model on the test set using metrics such as Dice coefficient, Intersection over Union (IoU), and pixel-wise accuracy.

Step 4: Applying LIME for Explainability

1. **Instance Selection:** We selected a subset of correctly classified and misclassified images from the test set for classification explainability. We selected a set of successful and failed segmentations from the test set for segmentation explainability.
2. **Generate Explanations:** We applied LIME to generate explanations for the selected instances. For classification, we created perturbed samples and fit an interpretable model to approximate the classifier's behavior locally. For segmentation, we created perturbed samples and fit an interpretable model to approximate the segmenter's behavior locally.
3. **Visualization and Interpretation:** We visualized the explanations as heatmaps to highlight the regions of the images that contributed most to the model's predictions. The visualizations were interpreted to understand the model's decision-making process.

Step 5: Quantitative Evaluation of Explanations

1. **Explanation Stability:** We measured the consistency of the explanations when the input image was slightly perturbed.
2. **Explanation Sparsity:** We evaluated the proportion of the image highlighted in the explanation to assess conciseness.
3. **Explanation Fidelity:** We assessed how well the interpretable model approximated the original model's predictions.

3 Experimental Details

3.1 Dataset Used

For Classification Tasks: For the classification tasks, we employed a combined dataset comprising 7023 images of human brain MRI images. These images are categorized into four distinct classes: glioma, meningioma, no tumor, and pituitary. This dataset amalgamates images from multiple sources, providing a diverse and comprehensive collection for robust classification model training and evaluation.

Sources of Data:

1. **Jun Cheng's Brain Tumor Dataset [11, 12]:** This dataset contains 3064 T1-weighted contrast-enhanced images from 233 patients, distributed across three tumor types: meningioma (708 slices), glioma (1426 slices), and pituitary tumor (930 slices). The dataset is split into four subsets, each archived in a .zip file containing approximately 766 slices. The data is organized in MATLAB (.mat) format, with each file storing the image data and associated annotations.
2. **Br35H Dataset:** The Br35H dataset contains a diverse collection of brain MRI images used for various brain tumor classification tasks. The dataset includes images categorized into the four classes mentioned above, further enhancing the diversity and robustness of our classification model.
3. **SarTaj Dataset:** This dataset includes additional brain MRI images used to complement the classification model's training dataset, ensuring a robust learning process.

The combined dataset from these sources provided a rich variety of images, enhancing the model's ability to generalize across different types of brain tumors. The images in the combined dataset were preprocessed to ensure uniformity in resolution and intensity normalization, followed by augmentation techniques to further increase the dataset size and variability, such as rotation, flipping, and contrast adjustment.

For Segmentation Tasks

Brain MRI Segmentation Dataset: This dataset includes brain MR images along with manual FLAIR abnormality segmentation masks. The images were sourced from The Cancer Imaging Archive (TCIA) and correspond to 110 patients included in The Cancer Genome Atlas (TCGA) lower-grade glioma collection, each having at least one fluid-attenuated inversion recovery (FLAIR) sequence and associated genomic cluster data. FLAIR sequences, known for their high sensitivity to lesions and abnormalities within the brain tissue. Manual segmentation masks were created by expert radiologists, delineating the abnormal regions with high precision. Covers 110 patients, providing a comprehensive and diverse dataset for training and evaluating segmentation models.

Each image has been standardized to a uniform resolution, ensuring consistency across the dataset. Images were preprocessed to correct for intensity inhomogeneities and were normalized to have zero mean and unit variance. Data augmentation techniques, including random rotations, scaling, and elastic deformations, were applied to increase the effective size of the training set and to improve the robustness of the segmentation model. The segmentation masks provide a binary representation of the tumor regions, with 1s indicating the presence of a tumor and 0s representing healthy tissue. These masks are crucial for training supervised learning models for segmentation tasks.

3.2 Evaluation Metrics

Segmentation: For evaluating the performance of our segmentation models, we employed several standard metrics to ensure a comprehensive assessment:

1. Dice Coefficient (Dice Similarity Index, DSC): The Dice coefficient measures the overlap between the predicted segmentation and the ground truth. It ranges from 0 to 1, with 1 indicating perfect overlap.

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

where A is the set of pixels in the predicted segmentation and B is the set of pixels in the ground truth segmentation.

2. Intersection over Union (IoU, Jaccard Index): IoU measures the ratio of the intersection to the union of the predicted and ground truth segmentations. It ranges from 0 to 1, with 1 indicating perfect segmentation.

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}} = \frac{A \cap B}{A \cup B} \quad (2)$$

3. Precision (Positive Predictive Value): Precision indicates the proportion of true positive pixels among all pixels that were predicted as positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

where TP is the number of true positive pixels and FP is the number of false positive pixels.

4. Recall (Sensitivity, True Positive Rate): Recall measures the proportion of true positive pixels among all actual positive pixels.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

where FN is the number of false negative pixels.

5. F1 Score: The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

6. Hausdorff Distance: This metric measures the maximum distance between the predicted segmentation boundary and the ground truth boundary, providing insight into the spatial accuracy of the segmentation.

$$d_H(X, Y) = \max \{ \sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y) \} \quad (6)$$

where sup represents the supremum operator and $d(a, B) = \inf_{b \in B} d(a, b)$. inf is the infimum operator $d(a, B)$ quantifies the distance from a point $a \in X$ to the subset $B \subseteq X$. $d(a, b)$ is the Euclidean distance between points a and b .

Classification: For the classification tasks, the following metrics were utilized to evaluate the model performance:

1. Accuracy: Accuracy is the ratio of correctly predicted instances to the total instances. It provides a straightforward measure of overall performance.
2. Confusion Matrix: The confusion matrix provides a detailed breakdown of the classification performance, displaying the counts of true positives, true negatives, false positives, and false negatives for each class.
3. Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC): The ROC curve plots the true positive rate against the false positive rate at various threshold settings. The AUC provides a single scalar value summarizing the model's performance across all thresholds. An AUC of 1 indicates perfect classification, while an AUC of 0.5 suggests no better performance than random guessing.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (7)$$

Precision, Recall and F1-scores are also used to evaluate the results.

LIME (Local Interpretable Model-Agnostic Explanations): For the interpretability of our models, particularly in understanding their decision-making processes, we employed LIME to generate explanations. The following metrics were used to evaluate the quality of the explanations provided by LIME:

1. **Explanation Stability:** Explanation stability measures the consistency of explanations when slight perturbations are made to the input data. High stability indicates that small changes in the input do not significantly alter the explanation.

$$\text{Stability} = 1 - \frac{1}{n} \sum_{i=1}^n |E(x_i) - E(x'_i)| \quad (8)$$

where $E(x_i)$ is the explanation for instance x_i and $E(x'_i)$ is the explanation for the perturbed instance and n is the number of samples.

2. **Explanation Sparsity:** Explanation sparsity evaluates the proportion of features used in the explanation compared to the total number of features. Sparse explanations are preferred as they are easier to interpret.

$$\text{Sparsity} = 1 - \frac{\text{Number of features in explanation}}{\text{Total number of features}} \quad (9)$$

3. **Explanation Fidelity:** Explanation fidelity measures how well the explanation approximates the original model’s behavior. High fidelity indicates that the surrogate model used for generating explanations closely mimics the original model.

$$\text{Fidelity} = 1 - \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2 \quad (10)$$

where $f(x_i)$ is the prediction of the original model for instance x_i and $g(x_i)$ is the prediction of the surrogate model.

3.3 Setting Up Parameter Values

For both the segmentation and classification tasks, careful selection and tuning of hyperparameters were crucial to optimizing the performance of our neural network models. The parameter values for various components of our proposed work are detailed below:

Segmentation Model (Enhanced U-Net)

1. **Learning Rate:** 1e-4. The learning rate determines the step size at each iteration while moving toward a minimum of the loss function. A smaller learning rate ensures a stable convergence but might require more epochs.
2. **Batch Size:** 16. The batch size indicates the number of training samples used in one forward/backward pass. A moderate batch size balances memory efficiency and gradient stability.
3. **Epochs:** 50 The number of epochs defines how many times the learning algorithm will work through the entire training dataset. More epochs can lead to better convergence but may also increase the risk of overfitting.
4. **Dropout Rate:**
 - Initial layers: 0.1

- Middle layers: 0.2
- Final layer: 0.3

Dropout is used to prevent overfitting by randomly setting a fraction of input units to 0 at each update during training time. Different rates are used for different layers to balance regularization and learning.

5. **Optimizer Adam.** The Adam optimizer is chosen for its adaptive learning rate capabilities and efficient handling of sparse gradients, making it suitable for training deep neural networks.
6. **Loss Function: Binary cross entropy** is used for measuring the performance of a classification model whose output is a probability value between 0 and 1. It is well-suited for segmentation tasks where the output is a binary mask.

$$\text{Binary Cross Entropy} = -\frac{1}{N} \sum_i^N \sum_j^M y_{ij} \log(p_{ij}) \quad (11)$$

where N is the number of samples and M is the number of classes.

Classification Model (Custom CNN)

1. **Learning Rate: 1e-3.** A higher learning rate compared to the segmentation model to ensure faster convergence while maintaining stability.
2. **Batch Size: 32.** A larger batch size to improve gradient estimation accuracy and training speed.
3. **Epochs: 100.** A higher number of epochs to ensure sufficient training for convergence.
4. **Dropout Rate: 0.5.** A higher dropout rate to strongly regularize the model and prevent overfitting, given the smaller dataset size.
5. **Optimizer Adam.** Adam optimizer is chosen for its efficient gradient computation and adaptive learning rates, facilitating robust training.
6. **Loss Function: Cross Entropy Loss.** Cross entropy loss is used for multi-class classification tasks, where it evaluates the performance of a classification model whose output is a probability value between 0 and 1 for each class.

$$\text{Cross Entropy} = -\frac{1}{N} \sum_{j=1}^N [t_j \log(p_j) + (1 - t_j) \log(1 - p_j)] \quad (12)$$

N is the number of data points, t_j is the truth value and p_j is the Softmax probability for taking the truth value.

LIME Parameters

1. **Number of Samples: 1000.** The number of perturbed samples generated to explain each prediction. More samples can improve explanation fidelity but increase computational cost.

2. **Kernel Width: 0.25.** The width of the kernel used for weighting the perturbed samples. A smaller width focuses the explanation on samples closer to the original instance.
3. **Feature Selection Method: Forward Selection.** The method used for selecting the most important features in the perturbed samples. Forward selection iteratively adds features to improve explanation quality.
4. **Regularization: L1 Regularization.** L1 regularization encourages sparsity in the explanation, making it more interpretable by focusing on the most influential features.
5. **Segmenter: Quickshift**
 - **Kernel Size: 4.** Controls the spatial scale of the segmentation. A larger kernel size results in larger segments.
 - **Max Distance: 200.** Limits the distance in the color space between two pixels to be merged.
 - **Ratio: 0.5.** Balances the color proximity and spatial proximity. A higher ratio gives more importance to color similarity.

Quickshift is a mode-seeking segmentation algorithm that clusters pixels based on color similarity and spatial proximity. It is used to generate superpixels, which are smaller segments of the image that preserve local information.

4 Results and Discussion

Table 1 presents the evaluation results of our proposed Custom CNN model compared to ResNet32 and VGG16 models.

The Custom CNN model achieved the highest accuracy (98.70%) compared to ResNet32 (96.35%) and VGG16 (96.2%), indicating superior overall performance in classifying brain tumors. The Custom CNN model also demonstrated the highest precision (97.63%), compared to ResNet32 (95.24%) and VGG16 (95.78%), indicating a lower rate of false positives. Recall: The Custom CNN model achieved a recall of 97.64%, slightly higher than ResNet32 (96.12%) and VGG16 (95.12%), indicating a higher rate of true positives. The Custom CNN model showed the highest F1-Score (97.47%), compared to ResNet32 (96.15%) and VGG16 (95.76%), balancing precision and recall effectively. The Custom CNN model had the highest LIME Explanation Stability Score (0.923), compared to ResNet32 (0.846) and VGG16 (0.687), indicating more stable explanations across similar inputs. The Custom CNN model had a LIME Explanation Sparsity Score of 0.208, compared to ResNet32 (0.196) and VGG16 (0.225), with lower sparsity indicating more concise explanations. ResNet32 had the highest LIME Explanation Fidelity Score (0.556), followed by VGG16 (0.418), and the Custom CNN model (0.310). This measures how well the explanation model approximates the original model.

These results demonstrate that the Custom CNN model generally outperforms both ResNet32 and VGG16 in terms of classification accuracy, precision, recall, and F1-Score, while providing highly stable and reasonably concise explanations as measured by LIME.

Table 1. Evaluation Results of proposed Custom CNN model compared to ResNet and VGG16

Evaluation Metrics	Custom CNN Model	ResNet32	VGG16
Accuracy	98.70%	96.35%	96.2%
Precision	97.63	95.24	95.78
Recall	97.64	96.12	95.12
F1-Score	97.47	96.15	95.76
LIME Explanation Stability Score	0.923	0.846	0.687
LIME Explanation Sparsity Score	0.208	0.196	0.225
LIME Explanation Fidelity Score	0.310	0.556	0.418

Table 2 presents the evaluation results of our proposed Custom U-Net model compared with a U-Net model that uses ResNet as its encoder. The Custom U-Net model achieved a higher validation accuracy (99.8%) compared to the U-Net with ResNet as its encoder (99.11%). The Custom U-Net model demonstrated a lower validation loss (0.0132) than the U-Net with ResNet (0.0425), indicating better performance in minimizing error. The IoU score, which measures the overlap between the predicted and ground truth segments, was higher for the Custom U-Net model (0.889) compared to the ResNet-based U-Net (0.847), suggesting more accurate segmentation. The Custom U-Net model had a higher LIME Explanation Stability Score (0.8169) versus the ResNet-based U-Net (0.7873), indicating more stable explanations across similar inputs. Both models showed similar LIME Explanation Sparsity Scores, with the Custom U-Net model having

Table 2. Evaluation Results for the proposed Custom U-net model compared with ResU-net

Evaluation Metrics	Custom U-net Model	U-net with ResNet as Encoder
Validation Accuracy	99.8%	99.11%
Validation Loss	0.0132	0.0425
Intersection over Union Score (IoU)	0.889	0.847
LIME Explanation Stability Score	0.8169	0.7873
LIME Explanation Sparsity Score	0.1190	0.1221
LIME Explanation Fidelity Score	0.5447	0.6036

a slightly lower score (0.1190) compared to the ResNet-based U-Net (0.1221). Lower sparsity indicates more concise explanations. The ResNet-based U-Net model exhibited a higher LIME Explanation Fidelity Score (0.6036) compared to the Custom U-Net model (0.5447), which measures how well the explanation model approximates the original model. Figure 3 displays the Confusion Matrix of custom CNN. Figure 4 displays the Plot of ground truth scans, masks and predicted masks and labels along with IoU Score. Figure 5 displays the MRI Scan on which LIME visualizations are generated. Figure 6 displays the LIME Visualizations of Custom CNN with feature heatmap, positive and negative influences and top three feature visualization. Figure 7 displays the plot of ground truth scans, masks and predicted masks and labels along with IoU Score Negative Classifier achieving perfect IoU of 1.

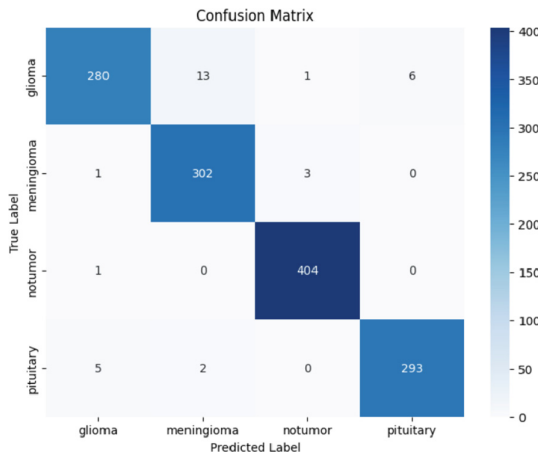


Fig. 3. Confusion Matrix of custom CNN

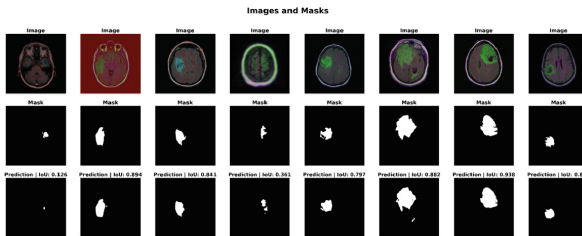


Fig. 4. Plot of ground truth scans , masks and predicted masks and labels along with IoU Score

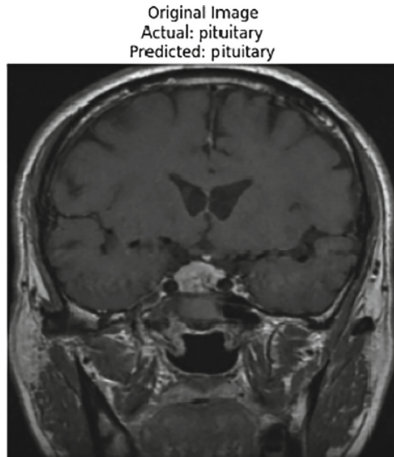


Fig. 5. MRI Scan on which LIME visualizations are generated

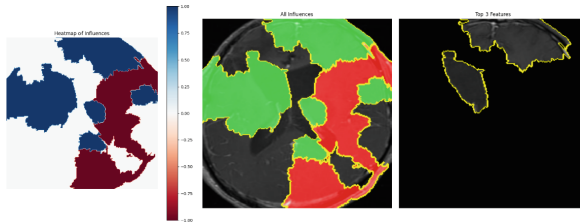


Fig. 6. LIME Visualizations of Custom CNN with feature heatmap, positive and negative influences and top three feature visualization

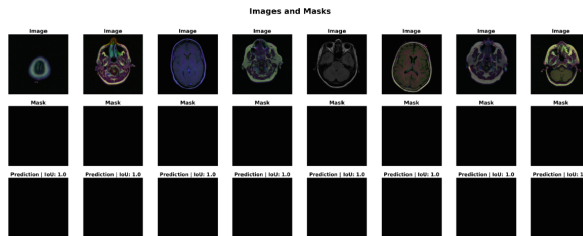


Fig. 7. Plot of ground truth scans, masks and predicted masks and labels along with IoU Score Negative Classifier achieving perfect IoU of 1

5 Conclusion and Future Scope

In this study, we developed and evaluated advanced neural network architectures for brain tumor classification and segmentation using MRI images. Our enhanced U-Net model demonstrated improved segmentation performance, while our custom CNN showed significant accuracy in classification tasks. The integration of

Explainable AI (XAI) techniques, particularly LIME, provided valuable insights into the model’s decision-making processes, enhancing interpretability and trustworthiness.

Combining multiple neural network architectures, such as integrating attention mechanisms or transformers, could improve both classification and segmentation performance. Leveraging larger and more diverse datasets can improve the generalizability of our models across different populations and imaging modalities. Developing models capable of real-time processing can facilitate clinical applications, enabling quicker diagnosis and treatment planning. Incorporating more advanced XAI techniques can provide deeper insights into model behavior, improving interpretability and clinical acceptance. Utilizing pre-trained models on broader datasets and fine-tuning them on specific medical imaging tasks can enhance model performance and reduce training time. Implementing these models in clinical workflows and electronic health records (EHR) systems can aid in automated diagnosis and decision support.

References

1. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* **35**(5), 1240–1251 (2016)
2. Abd-Ellah, M.K., Awad, A.I., Khalaf, A.A.M., Hamed, H.F.A.: A review on brain tumor diagnosis from MRI images: practical implications, key achievements, and lessons learned. *Magnetic Resonance Imaging* **61**, 300–318 (2019)
3. Badran, E.F., Mahmoud, E.G., Hamdy, N.: An algorithm for detecting brain tumors in MRI images. In: *The 2010 International Conference on Computer Engineering & Systems*, pp. 368–373. IEEE (2010)
4. Badža, M.M., Barjaktarović, M.Č.: Classification of brain tumors from MRI images using a convolutional neural network. *Appl. Sci.* **10**(6), 1999 (2020)
5. Prashant, G.S., Singh, V.P.: Ensemble of deep learning approaches for detection of brain. In: *International Journal of Advanced Networking and Applications - IJANA: 1st International Conference on Advancements in Smart Computing and Information Security, ASCIS 2022, Rajkot, India, November 24-26, 2022*, pp. 11–16. Eswar Publications (2022)
6. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fus.* **58**, 82–115 (2020)
7. Van der Velden, B.H.M., Kuijff, H.J., Gilhuijs, K.G.A., Viergever, M.A.: Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* **79**, 102470 (2022)
8. Dieber, J., Kirrane, S.: Why model why? Assessing the strengths and limitations of lime. *arXiv preprint [arXiv:2012.00093](https://arxiv.org/abs/2012.00093)* (2020)
9. Targ, S., Almeida, D., Lyman, K.: Resnet in Resnet: generalizing residual architectures. *arXiv preprint [arXiv:1603.08029](https://arxiv.org/abs/1603.08029)* (2016)
10. Sengupta, A., Ye, Y., Wang, R., Liu, C., Roy, K.: Going deeper in spiking neural networks: VGG and residual architectures. *Front. Neurosci.* **13**, 95 (2019)
11. Cheng, J., et al.: Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS ONE* **10**(10), e0140381 (2015)
12. Cheng, J., et al.: Retrieval of brain tumors by adaptive spatial pooling and fisher vector representation. *PLoS ONE* **11**(6), e0157112 (2016)



Deep BI-RADS Network for Improved Cancer Detection from Mammograms

Gil Ben-Artzi^(✉), Feras Daragma, and Shahar Mahpod

School of Computer Science, Ariel University, Ariel, Israel
{gilba,mahpods}@ariel.ac.il

Abstract. While state-of-the-art models for breast cancer detection leverage multi-view mammograms for enhanced diagnostic accuracy, they often focus solely on visual mammography data. However, radiologists document valuable lesion descriptors that contain additional information that can enhance mammography-based breast cancer screening. A key question is whether deep learning models can benefit from these expert-derived features. To address this question, we introduce a novel multi-modal approach that combines textual BI-RADS lesion descriptors with visual mammogram content. Our method employs iterative attention layers to effectively fuse these different modalities, significantly improving classification performance over image-only models. Experiments on the CBIS-DDSM dataset demonstrate substantial improvements across all metrics, demonstrating the contribution of handcrafted features to end-to-end.

Keywords: Cancer Detection · BI-RADS · Deep Learning · Mammograms · Breast Cancer · Attention · Transformer · Multi-Modal

1 Introduction

In recent years, deep learning techniques have emerged as a powerful tool for breast cancer detection, demonstrating significant potential in enhancing the accuracy of mammography interpretation. State-of-the-art models [6, 17, 29] have achieved impressive results by leveraging information from different mammogram views (craniocaudal (CC) and mediolateral oblique (MLO)) to enhance diagnostic accuracy. However, these approaches often focus solely on end-to-end extracted visual features.

Radiologists use the Breast Imaging Reporting and Data System (BI-RADS) lexicon [1] to document specific lesion descriptors such as size, shape, and margin characteristics during mammogram interpretation. These descriptors can offer crucial insights that aid in distinguishing between benign and malignant lesions. In this paper we investigate whether incorporating BI-RADS descriptors can improve deep learning for cancer detection.

Integrating these descriptors with mammograms poses challenges due to differences in modalities, scales, importance levels, and inconsistencies across radiology reports. To address these challenges and answer our research question, we

propose a multi-modal dual-branch architecture. Each branch, corresponding to CC/MLO views, encodes the mammogram in a multi-resolution manner. We introduce a dedicated iterative attention mechanism [10] that processes input from the previous layer, the current encoded resolution of the mammogram, and processed information from the other branch. By processing information from these sources at each level using the attention mechanism, our model effectively overcomes the differences in modalities and inconsistencies.

We conduct experiments using the CBIS-DDSM dataset, which includes both mammograms and BI-RADS descriptors as metadata. Our results indicate that our multi-modal iterative attention-based approach effectively integrates both visual and textual modalities, outperforming image-only models for benign vs. malignant classification. We achieve performance improvements across all metrics compared to image-only models, with an AUC score of 0.87. Our results demonstrate the significant potential of incorporating handcrafted features with deep learning models, suggesting a promising direction for future research in medical image analysis.

2 Related Work

2.1 Handcrafted Features for Cancer Detection

The Breast Imaging Reporting and Data System (BI-RADS) [1], developed by the American College of Radiology (ACR), acts as a standardized language for describing and classifying breast lesions identified through mammograms, ultrasounds, and MRIs. This system plays a crucial role in improving the consistency, clarity, and accuracy of breast imaging reports. Unlike our suggested features, BI-RADS descriptors are based on the grayscale level of the pixels in the lesions. A similar lexicon, the Thyroid Imaging Reporting and Data System (TI-RADS), has been proposed for thyroid lesions [2].

2.2 Multi-view Cancer Detection

Liu et al. [14] presented a cross-view correspondence reasoning method based on a bipartite graph convolutional network for mammogram mass detection. This approach effectively addresses the challenge of inherent view alignment between different views by learning geometric constraints. Tulder et al. [25] proposed a multi-view analysis method for unregistered medical images using cross-view transformers, addressing the challenge of effectively combining features from unregistered mammogram views (CC/MLO) with perspective differences. Shen et al. [21] presented an interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. This approach effectively addresses the challenge of interpretability in deep learning models for mammogram analysis. Chen et al. [4] proposed a multi-view local co-occurrence and global consistency learning method for mammogram classification generalization, addressing the challenge of effectively combining features

Table 1. The BI-RADS descriptors and descriptors classes in the CBIS-DDSM dataset.

Mass	Calcifications
<u>Margin</u>	<u>Morphology</u>
Circumscribed	Pleomorphic
Ill-defined	Amorphous
Spicular	Linear
Obscured	Punctate
<u>Shape</u>	<u>Distribution</u>
Round	Clustered
Oval	Scattered
Irregular	Diffuse

from unregistered mammogram views (CC/MLO) with perspective differences. While these methods address multi-view analysis, they do not utilize the textual lesion attribute data and cross-view information at each analysis stage - key capabilities of our architecture.

2.3 Incorporating Handcrafted Features

In the field of mammogram-based deep learning for breast cancer detection, current research primarily focuses on predicting BI-RADS descriptors as model outputs. The integration of both these descriptors and visual features in mammogram analysis remains an open research question.

Zhang et al. introduced BI-RADS-NET [30], an explainable deep learning approach for breast cancer diagnosis that outputs BI-RADS descriptors to better explain predictions, although their model was designed for ultrasound images. [16, 23] investigated a deep learning method that utilizes multi-view mammogram images to enhance BI-RADS and breast density assessment, rather than integrating them as in our approach. Liu et al. [13] explored the potential of combining mammography-based deep learning with clinical factors such as age and family history of breast cancer, demonstrating the potential benefits of integrating additional features with visual data in the prediction process.

3 Model Architecture

Our model consists of two branches. Each branch is composed of $N = 6$ stacked identical attention-based layers. An overview of our dual-branch architecture using stacked multi-attention layers (gray background) is presented in Fig. 1.

The attention based layers progressively fuse and process the multi-modal inputs. The input to the first layer is textual attributes with a skip connection to the last layer. In the first layer, learnable query vectors \hat{X} and \hat{Y} are used since no feature queries exist yet. The input to subsequent layers is the extracted

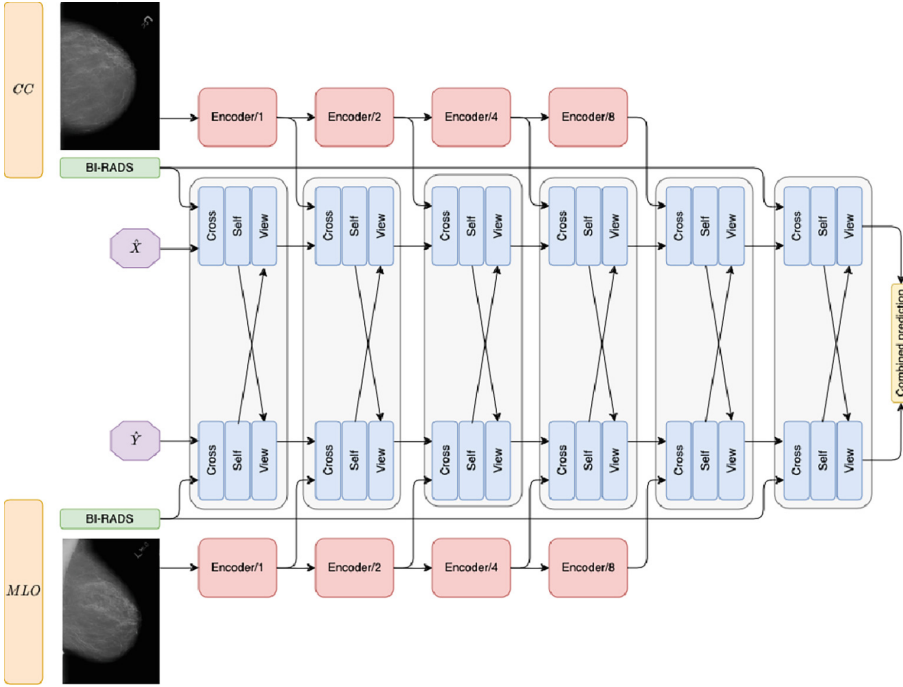


Fig. 1. Our model takes mammograms from the MLO and CC views along with a varying number of textual descriptors classes describing one or more lesions as input. The multi-attention layers (grayed blocks) processes these descriptors along with visual features extracted from mammogram images in different resolutions.

image features at different resolutions using the Big Transfer (BiT) blocks [11], the input from the preceding layer, and the latent features from the other branch.

The output of the final attention layer in each branch is aggregated by averaging to obtain a unified vector $z \in \mathbb{R}^{L \times 1}$. This representation encodes the joint contribution of images and text. The vector is then layer-normalized and reduced into a labeling vector using a fully-connected layer for the benign and malignant classes.

In the following we present a detailed description of each component of our model.

3.1 BI-RADS Descriptors Encoding

Our model utilizes the textual metadata associated with each mammogram, which contains the classes of the lesion and breast descriptors of the Breast Imaging Reporting and Data System (BI-RADS) lexicon [1]. We do not use subjective assessments reflecting radiologist suspicion, like BI-RADS scores, but only the descriptive physical lesion and breast characteristics annotated during routine screening. Table 1 presents examples of the descriptors and descriptors

classes that are incorporated by our model. Both calcifications and mass lesions can have combinations of these descriptors. For instance, a mass lesion can have a ‘‘Circumscribed-Obscured’’ margin or a ‘‘Round-Oval’’ shape. Our approach allows for the integration of a variable number of classes, as well as their combinations.

We assign a unique index $i = 1, \dots, N$ to enumerate the possible values of the descriptors classes across all categories. The input to our model is a binary vector $\phi \in \mathbb{R}^L$ (we use $L = 256$), defined as:

$$\phi(i) = \begin{cases} 1, & \text{if descriptor class } i \text{ exists in the description} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The encoded input vector ϕ represents the BI-RADS descriptors for a single lesion. However, there can be cases with more than one lesion. Our architecture supports a dynamic number of input vectors, so the input is $\Phi = \{\phi_j\}_{j=1}^K$ where K is the number of lesions in the mammogram.

3.2 Feature Extraction

We use BiT layers as our feature extractor, pre-trained on the PatchCamelyon dataset [27]. Our BiT layer \mathbf{F} is based on ResNet50-V2 [7, 8] with modifications made by [11] to Group Normalization [28] instead of Layer Normalization [3], and the use of Weight Standardization [19] for all convolution layers. The output of each of these blocks is the input to our multi-attention layer. Each BiT layer \mathcal{F} reduces the resolution and increases the number of channels using the following formulation:

$$\begin{aligned} \mathbf{F}_0 &= \mathcal{F}_0(I) \\ \mathbf{F}_k &= \mathcal{F}_k(\mathbf{F}_{k-1}), \quad \forall \quad k = 1, \dots, N-1, \end{aligned} \quad (2)$$

$I \in \mathbb{R}^{1 \times H \times W}$ is the input image of height H and width W , $\mathbf{F}_k \in \mathbb{R}^{d_0 \cdot 2^k \times H' \times W'}$ where $H' = \frac{H}{4 \cdot 2^k}$, $W' = \frac{W}{4 \cdot 2^k}$ and $d_0 = 64$.

3.3 Multi-attention Layer

The multi-attention layer has three attention-based [26] sub-layers. The first is a cross-attention mechanism, the second is self-attention and the third is view-attention. They enable the model to establish connections between different resolutions and the attributes, between patches within the same image, and between images from different views.

The utilization of attention enables the exploration of connections between a provided query Q , pre-existing key data K , while representing these relationships using V . It is stated as follows:

$$\mathcal{A}ttn(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (3)$$

where d is the scaling factor corresponding to the dimensionality of the key vectors.

Cross-attention. The first sub-layer, referred to as cross-attention, allows efficient processing of multi-modal inputs including attributes, latent features and images, without relying on domain-specific assumptions. It takes a high-dimensional input and projects it into a lower-dimensional latent bottleneck [10]. It then applies Transformer-style self-attention on this latent space. It combines the preceding latent features with either the attributes or image features at a given resolution.

The cross-attention in layer k , denoted as \mathbf{A}_k^C , is defined as:

$$\mathbf{A}_k^C := \mathcal{A}ttn(\mathbf{C}_{k-1}, \mathbf{F}_{k-1}, \mathbf{F}_{k-1}), \quad (4)$$

where \mathbf{F}_{k-1} is the extracted features from the previous layer (Sect. 3.2) and \mathbf{C}_{k-1} is the output of the previous multi-attention layer.

Positional encoding vectors are employed to encode the feature vector F_k . In the first Cross-attention sub-layer, where no preceding input exists, the query is learnable parameters X :

$$\mathcal{A}ttn(X, \Phi, \Phi). \quad (5)$$

Self-attention. The self-attention sub-layer is placed right after the cross-attention sub-layer. Similar to [10], the goal is to model both short-range and long-range dependencies within the features and capture the global context.

The inputs for the self-attention in layer k , denoted as \mathbf{A}_k^S , are the output of the cross-attention:

$$\mathbf{A}_k^S := \mathcal{A}ttn(\mathbf{A}_k^C, \mathbf{A}_k^C, \mathbf{A}_k^C). \quad (6)$$

View-Attention. The view-attention sub-layer combines the latent features from the current view with the latent features from the other view, enabling the expansion of the context to both MLO and CC. The values V of the view-attention block are the output of the preceding self-attention block, while the query Q and keys K are the output of the view-attention from the other branch at the corresponding level:

$$\mathcal{A}ttn(\hat{\mathbf{A}}_k^S, \hat{\mathbf{A}}_k^S, \mathbf{A}_k^S), \quad (7)$$

where $\hat{\cdot}$ denotes the output of the self-attention sub-layer in the other branch.

Input-Output. To ensure that our multi-attention layer receives input with the same number of channels, we reshape the feature tensor \mathbf{F}_k to have dimensions $\mathbf{F}_k \in \mathbb{R}^{d' \times N_k}$, where $N_k = \frac{H' \cdot W'}{2^{(n-k-1)}}$ and $d' = 4 \cdot d_0$ represents the desired length of the feature vectors inserted into the multi-attention layer.

The output tensors of the multi-attention layer at level k have dimensions $\mathbb{R}^{L \times N_k}$, where L denotes the length of the multi-attention latent vector. The query parameters $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ are learnable parameters with dimensions $\mathbb{R}^{L \times N_Q}$, where N_Q is a hyperparameter set by the user.

3.4 Sub-layer Attention Computation

Given an input sequence $X = (x_1, x_2, \dots, x_N)$, each attention sub-layer computes a weighted sum of the values at all positions in the sequence. This is achieved through the following steps:

Positional Encoding. To provide positional information to the model, we apply a Fourier feature encoding to the input sequences. Similar to [10], we utilize the Fourier feature positional encodings introduced in [22].

Given N input vectors $x_i \in \mathbb{R}^L$ each associated with a position index i , we first normalize the index as:

$$p_i = 2 \cdot \frac{i}{N} - 1 \quad (8)$$

We then define a set of sinusoidal frequency bands:

$$S_b = \{S \mid S = b \cdot \frac{m_{\text{freq}}}{n_{\text{bands}}}, 1 \leq b < n_{\text{bands}}, b \in \mathbb{Z}^+\} \quad (9)$$

where m_{freq} and n_{bands} determine the maximum frequency and number of bands.

The sinusoidal position encoding vectors are then calculated as:

$$PE1_b(p_i) = \sin(p_i \cdot S_b \cdot \pi), \quad (10)$$

$$PE2_b(p_i) = \cos(p_i \cdot S_b \cdot \pi). \quad (11)$$

Finally, we concatenate the normalized index p_i and encoding vectors $PE1_b$, $PE2_b$ to the original input x_i , expanding it to $x_i \in \mathbb{R}^{L+2n_{\text{bands}}+1}$. This injects positional information through sinusoidal functions of different frequencies, allowing the model to utilize the order of the input vectors.

Linear Transformation. The input sequence X is linearly projected into the query (Q), key (K), and value (V) matrices using learnable weight matrices W_Q , W_K , and W_V :

$$Q = XW_Q \quad (12)$$

$$K = XW_K \quad (13)$$

$$V = XW_V \quad (14)$$

where $Q, K, V \in \mathbb{R}^{L \times d_{\text{model}}}$, and d_{model} is the dimensionality of the model.

This transformation projects the input into distinct query, key, and value spaces. The query and key matrices are used to compute attention weights indicating the relevance between inputs. The value matrix holds the input representations that will be aggregated according to the attention weights.

Attention Unit. We compute the attention function (Eq. 3).

Position-wise Feed-Forward Network. After the attention unit, a position-wise feed-forward network is applied to each position independently. The feed-forward network consists of two linear transformations with a ReLU activation function in between:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (15)$$

where x is the input, W_1, W_2 are weight matrices, and b_1, b_2 are bias vectors.

4 Experimental Setup

4.1 Dataset

We use the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) dataset [12] that contains valuable metadata providing additional clinical information about each mammogram and associated lesions. It is a widely used mammography image collection annotated by radiologists, derived from the original DDSM [9] dataset and contains a diverse range of breast abnormalities, including benign and malignant lesions. The images are provided in the Digital Imaging and Communications in Medicine (DICOM) format, along with detailed annotation files. These files specify lesion locations, types, ROI crops, and binary masks across the craniocaudal (CC) and mediolateral oblique (MLO) views. It includes 1566 patients with total of 3,568 abnormalities, 1696 mass and 1872 calcification. In our experiments, we employ five-fold stratified cross-validation to maintain class balance across folds.

4.2 Implementation Details

The training was done in mini-batches, with each mini-batch size set to 16. For each image in our training data, the image’s content is scaled to 1024×1024 pixels. As data augmentation, we used vertical and horizontal flips, as well as elastic deformation. We set a total of 1000 training iterations for each fold. We utilized Cross-Entropy as our loss [5]. An initial learning rate was set to 0.001 and we employed a decaying factor of 10 after 500 iterations. We used the SGD optimizer [20] with momentum set to 0.9. Dropout value was set to 0.25. We implemented our model in PyTorch [18].

Table 2. Quantitative performance analysis to detect abnormality using CBIS-DDSM dataset.

Model	AUC	Accuracy	Specificity	Precision	Recall	F1-Score
[15]	0.680	0.661	0.670	0.638	0.651	0.644
[24]	0.811	0.723	0.750	0.686	0.698	0.692
Ours - no descriptors	0.711	0.664	0.650	0.676	0.619	0.634
Ours	0.872	0.760	0.773	0.760	0.743	0.751

5 Results

We compare our multi-modal descriptor-based model (“Deep BI-RADS”) against several baselines: a descriptor-excluded variant of our own model, a multi-view Transformer baseline [24], and an advanced recent single-view Transformer approach with four branches [15]. The descriptor-excluded variant includes the

Table 3. The effect of different configurations for the inputs (Query, Keys, Value) to the view attention sub-layer in the multi-attention layer. The inputs can come from either the current view (C) or the opposite view (O) of the mammogram. Configuration 2, in which the Query and Keys inputs are from the opposite view, achieved the best overall performance.

Configuration	Q	K	V	AUC	Accuracy	Specificity	Precision	Recall	F1-Score
0	C	O	O	0.835	0.738	0.731	0.643	0.750	0.692
1	O	C	O	0.850	0.760	0.790	0.762	0.727	0.744
2	O	O	C	0.878	0.796	0.773	0.816	0.786	0.780
3	C	C	O	0.852	0.760	0.768	0.714	0.750	0.732
4	C	O	C	0.843	0.764	0.792	0.762	0.733	0.747
5	O	C	C	0.848	0.771	0.803	0.778	0.737	0.757

multi-attention layers, allowing us to evaluate the specific contribution of the BI-RADS descriptors. Comparing with a multi-view architecture helps assess the contribution of both the attention layers and the BI-RADS descriptors. The multi-view baseline employs a Transformer architecture to analyze pairs of unregistered mammograms from different views and achieves state-of-the-art results on the CBIS-DDSM. Comparison with a single-view Transformer evaluates the contribution of the multi-view architecture.

To ensure a fair comparison, we trained all models from scratch following their respective provided training protocols. We evaluate the models for classifying mass lesions, following common practice. The results are obtained using five-fold stratified cross-validation to maintain class balance across folds.

Table 2 presents the quantitative performance analysis. Our multi-modal approach achieves a higher AUC of 0.872 compared to 0.711 without incorporating the BI-RADS descriptors, demonstrating the benefits of integrating textual information. We also attain an AUC of 0.872 versus 0.811 for the baseline multi-view model, showcasing the advantages of our multi-attention fusion approach over prior multi-view only techniques.

Beyond AUC, utilizing BI-RADS descriptors enables consistent gains across accuracy, specificity, precision, recall, and F1-score on both tasks. Our approach increases recall from 0.619 to 0.743 compared to the baseline without BI-RADS. This demonstrates improved sensitivity in detecting true positive cases by incorporating textual descriptor classes.

Notably, the high F1 scores demonstrate that our model balances improved sensitivity with precision, rather than sacrificing one metric for the other. This indicates that our multi-modal methodology incorporates the radiologist context to enhance interpretation without introducing additional false positives.

Figure 2 presents the ROC curve for a single fold, summarizing the trade-off between the true positive rate and false positive rate for our model using different probability thresholds. Overall, our multi-modal method shows promise

for generalized breast abnormality detection by effectively combining visual and textual information.

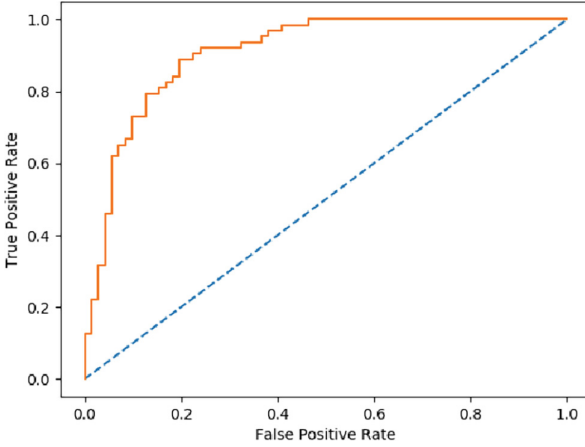


Fig. 2. ROC curve for our approach

5.1 Input to Multi-attention Layer

Table 3 represents different configurations (Fig. 3) for wiring the query (Q), and keys (K), and values (V). In both branches, the input to the multi-attention layer can be either from the current view (C) or from the opposite view (O). There are six possible configurations, as we always wire at least one input from the other view. Based on the results in Table 3, we conclude that wiring the query (Q) and keys (K) inputs to the attention layer from the opposite view (configuration 2) leads to the best performance, with the highest metrics.

Some observations:

- Wiring Q and K from the opposite view consistently outperforms wiring them from the current view (e.g., compare configurations 2 vs 3). This might suggest that the attention mechanism benefits from fusing information between the two views via the value input specifically.
- Based on how the inputs are interconnected between the two views, there is a noticeable difference in performance. This emphasizes the importance of effectively leveraging the two views.
- Wiring Q and K from the same view (configurations 2,3) performs better than wiring them from different views.
- Specificity is highest when wiring Q from the opposite view and K and V from the current view (configuration 5). However, other metrics like recall are lower in this configuration.

5.2 Number of Multi-attention Layers

The number of multi-attention layers primarily influences our model size. We trained configurations with 3, 5, 6, and 7 layers.

Table 4 presents the accuracy for each model across five folds. The 3-layer model underperformed, while the 5-layer model achieved the second-best results overall. The 6-layer configuration yielded the highest average accuracy, outperforming 5 layers. However, further increasing layers to 7 degraded performance, likely due to overfitting given the limited dataset size. In our implementation, we deploy 6 layers which achieved the optimal trade-off between model capacity and overfitting on this dataset.

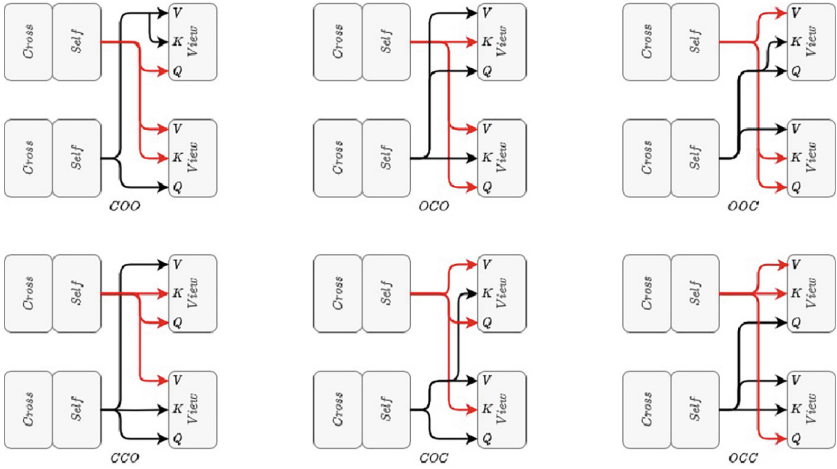


Fig. 3. The possible configurations of the inputs for the view attention sublayer in our multi-attention layer. Q, K, and V represent the Query, Keys, and Value respectively.

Table 4. Accuracy for different numbers of multi-attention layers. Results are across five folds.

Configuration	Average
3 layers	0.69 ± 0.015
5 layers	0.73 ± 0.019
6 layers	0.76 ± 0.010
7 layers	0.67 ± 0.018

5.3 Augmentations

We tested various data augmentation strategies to improve model generalization of our model. Table 5 presents test set performance for different augmentation configurations. The baseline with no augmentations underperformed all augmented models, indicating augmentations are beneficial. Adding random horizontal/vertical flips or elastic deformations to the baseline improved average accuracy. Resizing the images to 1024×1024 pixels achieved the best overall results. Interpolation sizes of 2048×2048 and 384×384 underperformed. Gaussian noise augmentation degraded performance, likely due to occluding meaningful mammographic details. The optimal configuration utilized interpolation upsampling to 1024×1024 pixels, which seems to balance overfitting and underfitting effects based on model capacity.

Table 5. Accuracy for different augmentation strategies.

Configuration	Average
Baseline w/o aug.	0.656 ± 0.025
Baseline + 384	0.702 ± 0.014
Baseline + 1024	0.72 ± 0.008
Baseline + 2048	0.646 ± 0.019
Baseline + h/vflip	0.666 ± 0.02
Baseline + elastic	0.688 ± 0.017
Baseline + Gaussian	0.612 ± 0.018

6 Conclusion

In this study we ask whether incorporating BI-RADS descriptors can improve deep learning for cancer detection. Our results provide a clear affirmative answer to this question. We presented a multi-modal approach that combines visual mammogram data with textual BI-RADS descriptors, utilizing a dual-branch architecture with iterative attention layers. Experiments on the CBIS-DDSM dataset demonstrated significant improvements over image-only models. These findings suggests that the fusion of features based on human expertise and automatically extracted features can lead to superior outcomes in cancer detection.

References

1. American College of Radiology: ACR BI-RADS® Atlas - Mammography. American College of Radiology, Reston, VA (2013)
2. American College of Radiology: ACR TI-RADS® Atlas. American College of Radiology, Reston, VA (2073)

3. Ba, L.J., Kiros, J.R., Hinton, G.E.: Layer normalization. CoRR abs/1607.06450 (2016). <http://arxiv.org/abs/1607.06450>
4. Chen, Y., et al.: Multi-view local co-occurrence and global consistency learning improve mammogram classification generalisation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III, pp. 3–13. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_1
5. De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. *Ann. Oper. Res.* **134**(1), 19–67 (2005)
6. Falconi, L.G., Maria Perez, W.G.A., Conci, A.: Transfer learning and fine tuning in breast mammogram abnormalities classification on CBIS-DDSM database **5**(2), 154–165 (2020)
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. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks (2016). <http://arxiv.org/abs/1603.05027>, ECCV 2016 camera-ready
9. Heath, M., et al.: Current status of the digital database for screening mammography. In: Digital Mammography, pp. 457–460. Springer, Dordrecht (1998). https://doi.org/10.1007/978-94-011-5318-8_75
10. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: general perception with iterative attention. In: International Conference on Machine Learning, pp. 4651–4664. PMLR (2021)
11. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big Transfer (BiT): General Visual Representation Learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 491–507. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_29
12. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **4**(1), 1–9 (2017)
13. Liu, H., et al.: A deep learning model integrating mammography and clinical factors facilitates the malignancy prediction of BI-RADS 4 microcalcifications in breast cancer screening. *Eur. Radiol.* **31**, 5902–5912 (2021)
14. Liu, Y., Zhang, F., Zhang, Q., Wang, S., Wang, Y., Yu, Y.: Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3811–3821 (2020)
15. Mo, Y., et al.: HoVer-Trans: anatomy-aware hover-transformer for ROI-free breast cancer diagnosis in ultrasound images. *IEEE Trans. Med. Imaging* (2023)
16. Nguyen, H.T.X., Tran, S.B., Nguyen, D.B., Pham, H.H., Nguyen, H.Q.: A novel multi-view deep learning approach for BI-RADS and density assessment of mammograms. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2144–2148 (2022). <https://doi.org/10.1109/EMBC48229.2022.9871564>
17. Nguyen, H.T., Tran, S.B., Nguyen, D.B., Pham, H.H., Nguyen, H.Q.: A novel multi-view deep learning approach for BI-RADS and density assessment of mammograms. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2144–2148. IEEE (2022)

18. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 32, pp. 8024–8035. Curran Associates, Inc. (2019)
19. Qiao, S., Wang, H., Liu, C., Shen, W., Yuille, A.L.: Weight standardization. *CoRR* abs/1903.10520 (2019). <http://dblp.uni-trier.de/db/journals/corr/corr1903.html#abs-1903-10520>
20. Robbins, H., Monro, S.: A stochastic approximation method. *Ann. Math. Stat.*, 400–407 (1951)
21. Shen, Y., Wu, N., Phang, J., Park, J.C., Liu, K., Tyagi, S., et al.: An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med. Image Anal.* **68**, 101908 (2021)
22. Tancik, M., et al.: Fourier features let networks learn high frequency functions in low dimensional domains. *Adv. Neural. Inf. Process. Syst.* **33**, 7537–7547 (2020)
23. Tsai, K.J., et al.: A high-performance deep neural network model for BI-RADS classification of screening mammography. *Sensors* **22**(3), 1160 (2022)
24. van Tulder, G., Tong, Y., Marchiori, E.: Multi-view analysis of unregistered medical images using cross-view transformers. In: de Bruijne, M., et al. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III*, pp. 104–113. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_10
25. van Tulder, G., Tong, Y., Marchiori, E.: Multi-view analysis of unregistered medical images using cross-view transformers. In: de Bruijne, M., et al. (eds.) *MICCAI 2021. LNCS*, vol. 12903, pp. 104–113. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_10
26. Vaswani, A., et al.: Attention is all you need. *Adv. Neural. Inf. Process. Syst.* **30**, 5998–6008 (2017)
27. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology (2018)
28. Wu, Y., He, K.: Group normalization. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII. Lecture Notes in Computer Science*, vol. 11217, pp. 3–19. Springer (2018). https://doi.org/10.1007/978-3-030-01261-8_1
29. Yan, Y., Conze, P.H., Lamard, M., Quellec, G., Cochener, B., Coatrieux, G.: Towards improved breast mass detection using dual-view mammogram matching. *Med. Image Anal.* **71**, 102083 (2021)
30. Zhang, B., Vakanski, A., Xian, M.: BI-RADS-Net: an explainable multitask learning approach for cancer diagnosis in breast ultrasound images. In: *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE (2021)



Histopathological Diagnosis of Meningioma and Solitary Fibrous Tumors Based on a Multi-scale Fusion Approach Utilizing Vision Transformer and Texture Analysis

Mohamed T. Azam¹, Hossam Magdy Balaha¹, Dibson D. Gondim²,
Akshitkumar Mistry³, Mohammed Ghazal⁴, and Ayman El-Baz¹(✉)

¹ Bioengineering Department, University of Louisville, Louisville, KY, USA
aselba01@louisville.edu

² Department of Pathology, University of Louisville School of Medicine, Louisville,
KY, USA

³ Department of Neurosurgery, University of Louisville School of Medicine,
Louisville, KY, USA

⁴ Electrical, Computer, and Biomedical Engineering Department, Abu Dhabi
University, Abu Dhabi, UAE

Abstract. Integration of Vision Transformer models with texture analysis presents a novel dual-stage approach for histopathology diagnosis. The proposed approach is mainly focused on discriminating between Meningioma (MEN) and Solitary Fibrous Tumor (SFT), two tumors known for their similar morphological characteristics. This approach leverages the inherent power of ViT models to capture global and local features from whole-slide images (WSIs), complementing this capability by integrating texture analysis techniques aimed at improving classification accuracy. Initially, ViT models are applied across three levels of WSI magnification, using rotationally and multi-scaled tiles to handle diverse scales and orientations inherent in histopathological imagery. ViT's attention mechanisms capture intricate details and spatial correlations within WSIs, offering a comprehensive view of histological structures. Concurrently, texture analysis methods including 3D-CLBP, 3D-GLCM, and 3D-GLRLM are used to extract the inherent patterns of the WSIs, such as homogeneity/inhomogeneity, morphology, and connectivity alongside the three RGB channels to capture the influence of color features. The scores obtained at the output of both stages are then fused and passed to a deep neural network, enabling a more reliable diagnosis. The experimental results show an accuracy of 93.42%, sensitivity of 92.15%, specificity of 94.73%, precision of 94.74%, balanced accuracy of 93.44%, and F1 score of 93.42%. These results elucidate the potential of the proposed approach in enhancing histopathology diagnostics.

Keywords: Histopathology · Whole Slide Images · Vision Transformers · Texture Analysis · Meningioma · Solitary fibrous tumors

1 Introduction

Differentiating meningioma from solitary fibrous tumors presents a frequent diagnosis challenge in neuropathology [28]. This distinction requires careful consideration and analysis due to the overlapping characteristics and complexities inherent in both conditions [5, 23]. Solitary fibrous tumor (SFT) is a rare, meningeal mesenchymal spindle-cell neoplasm that typically originates in the pleural cavity [13]. Notably, SFT tumors demonstrated aggressive biological behavior particularly when occurring in the central nervous system (CNS), with a reported recurrence rate close to 50% within a 5-year after surgical resection and a rate of up to 30% extraneural metastasis after initial surgery [26]. Meningioma (MEN), another type of meningeal neoplasm, originates from the meningeal layers surrounding either the brain or the spinal cord [4]. MEN tumors account for 37.6% of CNS primary tumors and around 50% of all benign brain tumors [24]. They are significantly more prevalent compared to SFT, constituting approximately 20% of all intracranial tumors [10]. Conversely to SFT, most MEN tumors are typically benign, leading to favorable long-term outcomes and lower rates of recurrence and metastasis [5]. The common clinical and imaging characteristics of SFT and MEN tumors, particularly in critical areas such as the CNS, highlight the challenge of precise diagnosis in tailoring suitable treatment plans and ensuring correct prognostic assessments [1, 21]. While Immunohistochemistry (IHC) serves as a valuable tool in resolving this differential diagnosis by identifying specific biomarkers [30], its effectiveness may be constrained in resource-limited settings and time-sensitive scenarios such as frozen section analyses and urgent surgery situations. Moreover, the need for specialized equipment and skilled personnel contributes to its associated high costs [3, 15]. Therefore, it is imperative to explore other alternative approaches to address this diagnostic challenge effectively.

Different artificial intelligence (AI) based approaches hold promise in distinguishing between various types of MEN and SFT tumors [11, 29]. Kong et al. [16] investigated the efficacy of machine learning (ML) models trained on MRI radiomics features to classify between intracranial solitary fibrous tumors (ISFTs) and angiomatous meningiomas (AMs) using a dataset of 268 patients. The study reported area under the curve (AUC) values of 0.917, 0.923, and 0.950 for the ML models based on radiomics, clinical, and fusion features, respectively. Furthermore, they achieved an AUC value of 0.786 based on the radiomics signature for histological stratification of ISFT. Le et al. [19] demonstrated the significance of texture features over clinical features in differentiating between malignant haemangiopericytoma (HPC) and angiomatous meningioma (AM). Utilizing a dataset of 67 cases, the support vector machine (SVM) classifier based on texture features extracted from enhanced T1WI achieved the best performance with an AUC of 0.90, surpassing SVM classifiers based on T2-FLAIR (AUC = 0.77) and DWI (AUC = 0.73). Notably, all texture-based SVM classifiers outperformed the clinical feature-based model, which achieved an AUC of 0.66. Dong et al. [8] utilized a 3D-MRI texture feature model to differentiate malignant intracranial SFT/HPC from AM, using a dataset of 97 patients with

SFT/HPC and 95 with AM. Their study incorporated various MRI modalities, including T1WI, T2WI, and contrast-enhanced T1WI. They reported the following AUC values: T1WI (AUC = 0.885), T2WI (AUC = 0.918), contrasted T1WI (AUC = 0.815), and combined sequence (AUC = 0.959). In the test set, these models achieved accuracies of 71.2%, 81.4%, 69.5%, and 83.1%, respectively.

Recently, transformers and other deep learning models have gained significant interest due to their outstanding performance across different Medical imaging-related tasks [18, 27]. Li et al. [20] developed an end-to-end DL model called (ViT-WSI), employing Vision Transformer (ViT) architecture on Whole Slide Images (WSI) for brain tumor analysis. This model accurately classifies tumor types and subtypes via weakly supervised learning. Through gradient-based analysis, the ViT-WSI model identifies three crucial histopathological features - IDH1 mutation, p53 mutation, and MGMT methylation - achieving patient-level AUC scores of 0.960, 0.874, and 0.845, respectively. Chen et al. [6] developed an MRI-based deep learning model to discriminate between Intracranial hemangiopericytoma/solitary fibrous tumor (SFT/HPC) and meningioma. The study utilized a pre-trained ResNet-50 Model on T1-contrast images using a dataset of 236 patients. They reported a promising accuracy result of .889 and an AUC of .91 in the validation set. Hossain et al. [12] explored various DL models, including VGG16, InceptionV3, VGG19, ResNet50, InceptionResNetV2, and Xception, to enhance the performance of multi-class brain tumor classification. They reported peak accuracies ranging from 93.58% to 94.5% for these models. Additionally, they proposed a transfer learning-based multiclass classification model called IVX16, which combines the strengths of the three best-performing models: VGG16, InceptionV3, and Xception. The IVX16 model attained a peak accuracy of 96.94%, outperforming individual models and demonstrating the effectiveness of their ensemble strategy.

To date, there have been exceedingly few AI-based studies published on imaging characteristics that can effectively differentiate between MEN and SFT tumors. Additionally, most studies predominantly focus on texture features derived from MRI modalities, overlooking the crucial role of histopathological WSIs as the gold standard in diagnosis. To the best of our knowledge, we are the first group to diagnose MEN and SFT tumors based on histopathology. This work extends our prior research on the application of ViT models to histopathology diagnostics [2]. Our previous study demonstrated that ViTs are capable of effectively analyzing histopathological images with good accuracy. Building on these results, we combine ViT with texture analysis techniques to enhance diagnostic performance and interpretability. This study presents a novel concurrent, two-stage approach designed to capture both global and local features within WSIs. In the first stage, the ViT model captures attention-based informative regions to differentiate MEN and SFT tumors. The second stage focuses on using ML models to perform diagnosis based on extracted texture features, including 3D Gray-Level Co-occurrence Matrix (3D-GLCM), 3D Gray-Level Run Length Matrix (3D-GLRLM), and 3D Circular Local Binary Pattern (3D-CLBP). Fusing the decision probabilities of both stages achieved a promising accuracy in

diagnosing MEN and SFT tumors. Utilizing this comprehensive methodology, which includes both deep learning-based and texture-based feature classification from multiple WSI magnification levels, can significantly enhance the reliability of tumor classification in histopathology.

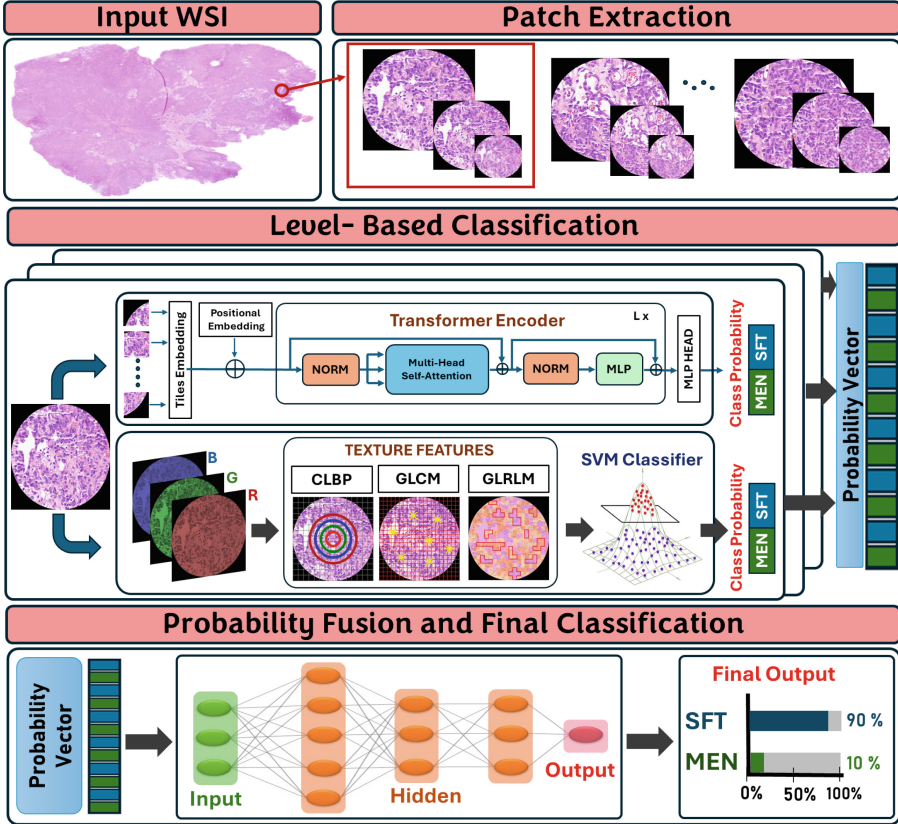


Fig. 1. Illustration of the proposed approach for classifying MEN and SFT tumors. The workflow depicts the process of patch extraction by tiling WSIs into circular patches, followed by a two-stage pipeline for ViT model and texture-based classification. Finally, a deep neural network is used for the final classification decision.

2 Methods

The proposed approach, as illustrated in Fig. 1, begins by tiling WSIs into circular patches extracted from three magnification levels within the WSI pyramid. Each patch undergoes a two-stage pipeline: The first stage employs the deep-learning ViT model, renowned for its adeptness in feature extraction through self-attention mechanisms, to classify different MEN and SFT extracted patches. Concurrently, the second stage focuses on extracting diverse texture features,

specifically 3D-CLBP, 3D-GLCM, and 3D-GLRLM, designed to capture variant patterns alongside the influence of color information within WSIs. These features are then amalgamated and fed into an ML classifier, i.e., SVM, to obtain the output scores/probabilities for the MEN and SFT classes. To accommodate the scale variations within WSIs, this two-stage process is applied to all patches across the three levels. Finally, all resulting output scores are fused and fed to a deep neural network for the final classification decision.

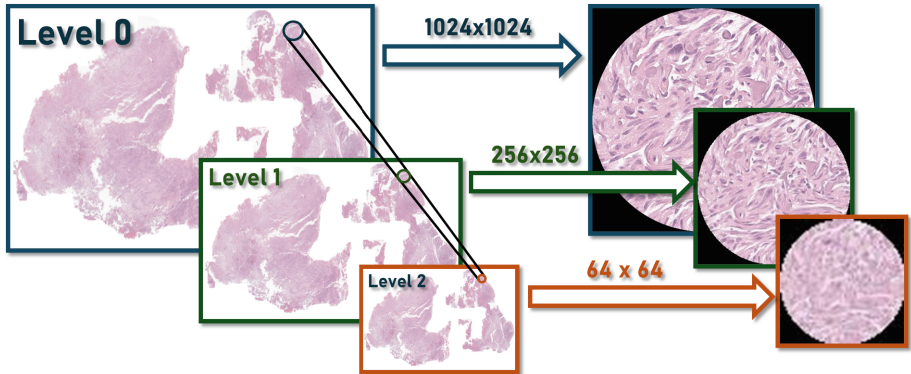


Fig. 2. Visualization of a WSI pyramid at magnification levels: Level 0 (40x), Level 1 (10x), and Level 2 (2.5x), along with corresponding centrally-aligned circular patches, depict the same physical area with sizes 1024×1024 , 256×256 , and 64×64 respectively.

2.1 WSI Patches Extraction

The proposed approach employs a multiscale strategy that utilizes patches extracted from multiple magnification levels, specifically (40x, 10x, and 2.5x), within WSIs. This results in generating a triple set of patches representing the same physical area; an example is shown in Fig. 2. To address the arbitrary orientations occurring during WSI capture, patches are extracted in a circular manner, ensuring rotation invariance. To enhance the quality of the dataset, patches with a background content exceeding 50% of patch area are systematically identified and excluded. This strategy yields a set of multiscale and rotationally invariant patches, facilitating an in-depth multiresolution analysis of diverse degrees of detail inherent in tissue samples within WSIs. Algorithm 1 outlines the main steps of the patch extraction process.

2.2 ViT Models Utilization

Employing ViT models in WSI analysis arises from its capability to filter and emphasize critical WSI parts/regions through self-attention mechanisms while also capturing internal correlations among features [9]. By stacking multiple

Algorithm 1. The main steps of Multi-level Patch Extraction Procedure

-
- 1: Utilize the WSI pyramid characteristic to extract levels (0, 1, and 2) corresponding to magnifications of 40x, 10x, and 2.5x, respectively.
 - 2: Extract patches of size (1024×1024) from the WSI at level 0.
 - 3: Discard patches where the background covers more than 50% of the area.
 - 4: **for each** acceptable patch **in** level 0 **do**:
 - 5: Calculate the centroid (*center*) of the patch.
 - 6: Extract a (256×256) corresponding patch from WSI level 1 centered at *center*.
 - 7: Extract a (64×64) corresponding patch from WSI level 2 centered at *center*.
 - 8: **end for**
 - 9: Store the extracted patches from levels 0, 1, and 2 for further analysis.
-

transformer layers, the ViT efficiently grasps the global context, enhancing comprehensive WSI analysis by considering both local and global features [17]. To initiate the ViT procedure, the WSI, represented as $I \in \mathbb{R}^{C \times H \times W}$, is partitioned into a collection of non-overlapping tiles T_1, T_2, \dots, T_m , where each tile $T_i \in \mathbb{R}^{C \times M \times M}$, with C representing the number of image channels (e.g., $C = 3$ for RGB channels), $H \times W$ indicating the WSI dimensions, and $M \times M$ represents the tile dimensions. Tiles containing backgrounds or noisy areas were excluded, resulting in a set of k tiles $\{T_1, T_2, \dots, T_k\}$ that exclusively encompass foreground tissues. By utilizing a linear projection layer and a position encoding layer, the histology features and positions of chosen patches are vectorized into a set of tokens $Z_0 = \{V_1, V_2, \dots, V_n\}$, where $V_i \in \mathbb{R}^E$ and E denote the vector size. The tokens are organized and fed into the transformer encoder, which consists of L stacked encoder blocks. Each transformer encoder block consists of two main components: multi-head self-attention (MSA) and a fully connected feed-forward multi-layer perceptron (MLP), as described by Eqs. 1 and 2, respectively. These components are augmented with residual skip connections and are preceded by a Layer Normalization (LN). Lastly, the first token of z_L , a learnable class token, is normalized and sent to the external head classifier layer for the class label i.e. MEN or SFT, prediction.

$$Z_l^* = \text{MSA}(\text{LN}(Z_{l-1})) + Z_{l-1}, \quad l \in [1, \dots, L], \quad (1)$$

$$Z_l = \text{MLP}(\text{LN}(Z_l^*)) + Z_l^*, \quad l \in [1, \dots, L], \quad (2)$$

The MSA block is pivotal in the transformer encoder. It comprises several heads that individually compute query-key-value scaled dot-product attention to learn attention weights. The output of the self-attention layer is determined by applying a softmax function to the scaled dot-product attention matrix as depicted in the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (3)$$

where, Q , K , and V are the query, keys, and values matrices, respectively, derived from the input embedded sequence by multiplying with learned matrices W_Q ,

W_K , and W_V . These matrices are utilized in the scaled dot-product attention mechanism to compute attention weights, and d denotes the dimensionality of the key vectors [32].

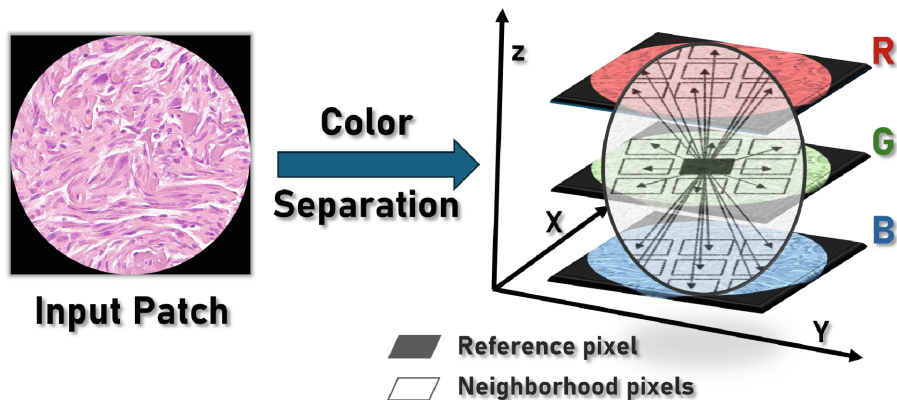


Fig. 3. Illustration showcasing Voxel-based interaction across the 26-neighborhood within the RGB channels for 3D texture features extraction.

2.3 Texture Features Extraction

Texture features serve as robust descriptors that capture various aspects of spatial arrangement and intensity patterns within volumetric data [31]. Considering WSI Multi-Level patches extracted in RGB images as 3D entities, we gain a unique perspective on different texture aspects such as homogeneity, morphology, and structural complexities among varying color intensities. This study employed three different texture analysis techniques: 3D-CLBP, 3D-GLCM, and 3D-GLRLM, for the comprehensive capturing of various texture aspects within WSIs. These techniques are essentially built upon considering the three color channels, i.e., RGB, of WSI patches as a 3D object, instead of converting the colored WSI to grayscale. Thus, enabling the incorporation of color influence along with texture variations aids in better discrimination between MEN and SFT tumors.

3D Circular Local Binary Pattern. The CLBP algorithm is an extension of Local Binary Pattern (LBP) that captures texture information in a circular region around each pixel enabling rotation invariance and adaptability in identifying texture patterns [14]. It works by comparing the intensity of neighboring pixels with the center pixel and encoding the result as a binary code. The CLBP computation begins by defining a circular neighborhood around the pixel of interest, extending to a specified radius R_{\max} . Within this circular region, the intensity of each neighboring pixel (x', y') is compared with the intensity of

the center pixel (x, y) . Based on this comparison, the CLBP value is generated according to the following equation:

$$CLBP(x, y) = \sum_{r=1}^{R_{\max}} \left(\sum_{(x', y') \in N_r(x, y)} \begin{cases} 2^{(r-1)} & \text{if } I_c(x', y') \geq I_c(x, y), \\ 0 & \text{otherwise} \end{cases} \right) \quad (4)$$

where $N_r(x, y)$ denotes the circular neighborhood of radius r centered at (x, y) , and $I_c(x', y')$ represents the intensity of the pixel at coordinates (x', y') . This equation is applied to the 3 channels (R, G, B) and the results are summed up across all radii within the neighborhood. After computing CLBP, we use percentiles derived from CDFs (ranging from the 10th to the 90th with a 10% increment) to create the feature vector for each WSI patch. This statistical method adeptly captures the overall distribution of pixel intensities, effectively encapsulating essential texture and structural features. In this study, utilizing the CLBP on multiscale WSI patches could effectively capture and encode the intricate texture detail and further improve the analysis and interpretation of different WSI patterns.

Table 1. A comprehensive list of the texture markers utilizing GLCM and GLRLM.

GLCM	GLRLM	
Contrast	Short Run Emphasis	Long Run Emphasis
Correlation	Grey Level Non-Uniformity	Run Length Non-Uniformity
Homogeneity	Run Percentage	Low Grey Level Run Emphasis
Angular Second Moment	High Grey Level Run Emphasis	Short Run Low Grey Level Emphasis
Energy	Short Run High Grey Level Emphasis	Long Run Low Grey Level Emphasis
Entropy	Long Run High Grey Level Emphasis	Run Entropy
Dissimilarity	Grey Level Variance	Run Length Variance

3D Gray Level Co-Occurrence Matrix. The GLCM is a statistical technique employed to represent the joint distribution of gray levels within an image, considering spatial connections [25]. In this study, different GLCM features were utilized, as listed in Table 1, to get a thorough examination of grayscale patterns like directional attributes, neighboring pixel values, and variances observed within diverse WSI patches. Building upon the traditional GLCM Method, we introduced 3D GLCM values by accounting for the three RGB channels of input patches. This expansion integrates color characteristics alongside grayscale ones, thereby enriching the texture representation within WSIs. To compute the GLCM for each WSI patch, a 256×256 matrix represents gray values. Each element denotes the frequency occurrence of a pair of combinations of gray values. GLCM values are derived by comparing the reference pixel to its 26 neighboring pixels: 8 pixels in the (G) channel, and 9 pixels in each of the upper (R) and lower (B) channels as shown in Fig. 3. This approach enables comprehensive spatial relationship analysis for each $3 \times 3 \times 3$ voxel grid, with a distance of 1 in the Z-plane and 1 in the XY plane. Angle analysis spans from 0° to 315° with a step of 45° to capture all directional dependencies.

3D Gray Level Run Length Matrix. The GLRM is a statistical matrix-based technique utilized to measure sequences of consecutive pixels with the same gray level value, known as gray level runs [7]. This technique is employed to describe different WSI regional heterogeneity information. Utilizing the 3D GLRLM for the RGB patches allows for considering the color variations alongside the spatial distribution of gray levels in the volumetric space. The GLRLM, a matrix of dimensions $L \times K$, is characterized by L gray levels and a maximum run length of K . It is constructed for a specific WSI patch by aggregating runs comprising pixels with gray level i and a run length of j . Each element (i, j) within the matrix represents the frequency of runs with gray level i and length j along a designated angle θ , where $\theta = (0^\circ, 45^\circ, 90^\circ, \text{ and } 135^\circ)$. This involves identifying consecutive horizontal (XY plane) and vertical (Z plane) sequences. Subsequently, a total of 16 GLRLM-based features were estimated, as listed in Table 1, to capture variant WSI structures.

2.4 Fusion and Final Classification

After employing the proposed two concurrent stages for level-based classification, the output probabilities from each stage across the three hierarchical WSI levels were aggregated into a probability vector. This vector contains level-based probability decisions that were then fed into a deep neural network (DNN) for the final diagnostic decision. The fusion process is designed to incorporate support from all WSI levels in the final diagnosis. Additionally, it leverages both the versatility of VIT as a deep learning model and SVM as a texture-based ML classifier to form complementary decisions. This aggregation ensures a complementary, reliable, and unbiased classification decision. The DNN architecture designed for SFT and MEN classification employs densely connected layers with ReLU activation and dropout regularization to prevent overfitting. It comprises a sequence of layers with 16, 32, 64, 32, and 16 neurons, concluding with a sigmoid activation unit. Hyperparameter tuning involves Adam, Adagrad, Nadam, and RMSprop optimizers, along with dropouts of 0.2, 0.3, and 0.4. Callbacks like ReduceLROnPlateau and EarlyStopping are used for training enhancement. Optimal settings include a learning rate of 0.001 with Adam optimizer, a dropout rate of 0.2, and epochs ranging from 30–80 through the 5-fold stratified cross-validation strategy.

3 Experimental Results

In this study, we utilized a multi-institutional dataset consisting of 92 cases (46 SFT and 46 MEN), all subjected to Hematoxylin and Eosin (H&E) staining. All patients provided their consent to participate. Expert pathologists confirmed the diagnoses and conducted a meticulous review of all pathology reports and histology slides. SFT diagnoses were validated using STAT-6 immunohistochemistry.

Initially, WSIs underwent preprocessing, involving the extraction of aligned patches representing the same physical area from three distinct levels within

the WSI hierarchy, facilitating comprehensive multiresolution analysis. Subsequently, a two-stage pipeline was employed to rigorously analyze patches at each level, integrating ML techniques and ViT models for precise diagnosis of SFT and MEN tumors. A diverse set of ML classifiers, including Support Vector Machines (SVM), Random Forest (RF), Adaboost, Logistic Regression (LogReg), and K-Nearest Neighbors (KNN), were employed. Moreover, for ViT, both the ‘vit-base-patch16-224’ and ‘vit-base-patch32-224’ models were used. It’s noteworthy that all models underwent fine-tuning to optimize their performance. To mitigate overfitting and ensure the proposed system robustness, the dataset is partitioned into five folds using stratified k-fold cross-validation. In this method, the dataset is divided into k equal-sized subsets, ensuring that each fold preserves the same distribution of classes as the original dataset, providing a more accurate assessment of the system’s performance [22]. For evaluating the system’s effectiveness, a comprehensive set of evaluation metrics was employed. These metrics include accuracy (ACC), sensitivity (SEN), specificity (SPC), precision (PRC), balanced accuracy (BAC), and F1-score. All results are reported in terms of Mean \pm Standard Deviation.

Table 2. Performance Evaluation of Different ML Classifiers Based on Texture Features Across Three Hierarchical Levels of WSI for Diagnosis of SFT and MEN Tumors.

Level	Classifier	ACC (%)	SEN (%)	SPC (%)	PRC (%)	BAC (%)	F1 (%)
Level 0	SVM	89.46 \pm 0.10	90.25 \pm 0.16	88.70 \pm 0.16	88.59 \pm 0.14	89.48 \pm 0.09	89.41 \pm 0.09
	RF	85.88 \pm 0.12	87.46 \pm 0.17	84.34 \pm 0.16	84.45 \pm 0.14	85.90 \pm 0.12	85.93 \pm 0.12
	KNN	83.20 \pm 0.15	84.74 \pm 0.15	81.71 \pm 0.16	81.83 \pm 0.15	83.23 \pm 0.14	83.26 \pm 0.14
	LOGReg	80.10 \pm 0.41	80.15 \pm 0.60	80.04 \pm 0.33	79.61 \pm 0.35	80.10 \pm 0.41	79.88 \pm 0.45
	Adaboost	79.20 \pm 0.15	78.62 \pm 0.40	79.76 \pm 0.25	79.06 \pm 0.16	79.19 \pm 0.15	78.84 \pm 0.19
Level 1	SVM	86.86 \pm 0.04	86.40 \pm 0.08	87.30 \pm 0.06	86.87 \pm 0.04	86.86 \pm 0.04	86.63 \pm 0.04
	RF	83.54 \pm 0.10	84.56 \pm 0.20	82.55 \pm 0.18	82.49 \pm 0.14	83.56 \pm 0.10	83.51 \pm 0.11
	KNN	81.48 \pm 0.09	82.76 \pm 0.17	80.24 \pm 0.13	80.29 \pm 0.10	81.50 \pm 0.09	81.51 \pm 0.10
	LOGReg	79.38 \pm 0.30	78.31 \pm 0.27	80.42 \pm 0.42	79.54 \pm 0.37	79.37 \pm 0.29	78.92 \pm 0.28
	Adaboost	78.70 \pm 0.12	77.64 \pm 0.42	79.74 \pm 0.41	78.84 \pm 0.26	78.69 \pm 0.12	78.24 \pm 0.15
Level 2	SVM	80.98 \pm 0.06	80.14 \pm 0.10	81.80 \pm 0.13	81.06 \pm 0.10	80.97 \pm 0.05	80.60 \pm 0.05
	RF	80.29 \pm 0.14	81.45 \pm 0.18	79.16 \pm 0.23	79.17 \pm 0.18	80.31 \pm 0.14	80.29 \pm 0.13
	KNN	77.90 \pm 0.14	79.13 \pm 0.16	76.71 \pm 0.19	76.76 \pm 0.16	77.92 \pm 0.14	77.93 \pm 0.14
	LOGReg	76.12 \pm 0.34	74.72 \pm 0.42	77.49 \pm 0.39	76.34 \pm 0.36	76.11 \pm 0.34	75.52 \pm 0.35
	Adaboost	76.62 \pm 0.06	73.48 \pm 0.40	79.68 \pm 0.30	77.86 \pm 0.16	76.58 \pm 0.06	75.61 \pm 0.14

Table 2 presents a comprehensive assessment of various ML classifiers across different hierarchical levels of WSIs-Level 0, Level 1, and Level 2-utilizing extracted texture features for the diagnosis of SFT and MEN tumors. Notably, the results highlight the superior performance of utilizing Level 0 of the WSI

hierarchy compared to Levels 1 and 2. This indicates that Level 0, offering the most granular and precise observations of cell structures in WSIs, captures different SFT and MEN texture variation patterns and intricate tissue characteristics essential for accurate tumor classification. The SVM classifier achieved an impressive ACC of 89.46%, closely followed by RF at 85.88% and KNN at 83.20%. Similarly, at Level 1 and Level 2, SVM maintained its superiority with ACC scores of 86.86% and 80.98%, respectively. The optimal hyperparameters for SVM classifier consist of a polynomial kernel of degree 2, auto kernel scaling, and a box constraint of 1. Notably, while SVM consistently outperforms other classifiers, there is a gradual decline in performance as the hierarchical level increases. This reflects the challenge of using higher WSI levels; as the view zooms out, image quality may degrade, hindering the identification of fine structures and texture variations within tissue samples. Nonetheless, SVM has demonstrated consistently superior results, showcasing its ability to recognize diverse texture patterns across all levels, establishing it as a reliable option for the precise diagnosis of MEN and SFT tumors.

To evaluate the performance of ViT models in diagnosing SFT and MEN tumors, Table 3 presents a comprehensive analysis of ViT-B16 and ViT-B32 across the three WSI hierarchical levels. At Level 0, the lowest and more detailed hierarchical level, ViT-B16 demonstrates superior overall performance with an ACC of 91.52%, outperforming ViT-B32, which achieves an ACC of 88.2%. However, as the hierarchical level increases, the performance gap between the two models narrows, with both models exhibiting a decline in performance metrics. At the intermediate level (Level 1), ViT-B16 maintains a slight advantage over ViT-B32, although both models show comparable performance. Notably, at Level 2, the highest hierarchical level, the accuracy of both models decreases due to the lower level of detail available, resulting in ViT-B16 and ViT-B32 achieving ACCs of 82.59% and 82.5%, respectively. The optimal hyperparameters for the ViT-B16 model consist of a batch size of 32 and a learning rate of 0.001, utilizing the Adam optimizer. These results demonstrate the ViT model’s efficacy in classifying MEN and SFT tumors across the varying levels of the WSI hierarchy.

Table 3. Performance Evaluation of ViT-B16 and ViT-B32 Models for SFT and MEN Tumor Diagnosis Across Three WSI Hierarchical Levels.

Level	ViT Model	ACC (%)	SEN (%)	SPC (%)	PRC (%)	BAC (%)	F1 (%)
Level 0	ViT-B/16	91.52 ± 0.49	91.21 ± 1.51	91.84 ± 1.91	92.04 ± 1.66	91.52 ± 0.5	91.61 ± 0.44
	ViT-B/32	88.2 ± 0.49	88.53 ± 1.37	87.87 ± 0.91	88.25 ± 0.68	88.2 ± 0.48	88.39 ± 0.56
Level 1	ViT-B/16	91.34 ± .62	90.33 ± 1.58	92.38 ± 1.1	92.43 ± 0.96	91.35 ± 0.62	91.36 ± 0.67
	ViT-B/32	87.72 ± 0.79	88.09 ± 1.56	87.34 ± 0.41	87.74 ± 0.38	87.72 ± 0.78	87.91 ± 0.87
Level 2	ViT-B/16	82.59 ± 1.18	85.18 ± 1.4	79.91 ± 1.69	81.36 ± 1.34	82.55 ± 1.18	83.22 ± 1.13
	ViT-B/32	82.5 ± 0.91	83.8 ± 1.58	81.16 ± 2.11	82.09 ± 1.58	82.48 ± 0.91	82.92 ± 0.85

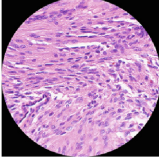
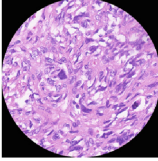
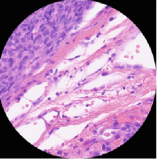
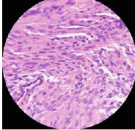
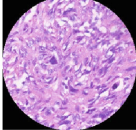
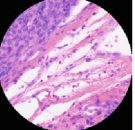
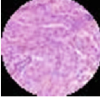


	Actual Label MEN	Actual Label SFT	Actual Label SFT
Level 0	 SVM → MEN [Prob = 0.79] ViT → MEN [Prob = 0.86]	 SVM → SFT [Prob = 0.95] ViT → SFT [Prob = 0.81]	 SVM → MEN [Prob = 0.95] ViT → SFT [Prob = 0.74]
Level 1	 SVM → SFT [Prob = 0.71] ViT → MEN [Prob = 0.82]	 SVM → SFT [Prob = 0.92] ViT → SFT [Prob = 0.52]	 SVM → MEN [Prob = 0.95] ViT → MEN [Prob = 0.62]
Level 2	 SVM → MEN [Prob = 0.62] ViT → MEN [Prob = 0.57]	 SVM → MEN [Prob = 0.69] ViT → MEN [Prob = 0.73]	 SVM → SFT [Prob = 0.84] ViT → SFT [Prob = 0.76]
	Fusion MEN [Prob =0.91]	Fusion SFT [Prob =0.85]	Fusion MEN [Prob =0.68]

Fig. 4. An illustration of the proposed model predictions over three instances, depicting SVM scores and ViT-B/16 scores at multiple WSI levels, alongside the final classification decision.

In addition to assessing the individual performance of each stage across the three WSI levels, we explored the potential benefits of combining the output scores of the best classifier from each pipeline, i.e., SVM and the ViT-B16 model, across the three magnification levels of the WSI hierarchy through a neural network fusion approach. This fusion strategy aimed to leverage the power of complementary information from each stage, facilitating the capture of both local and global features from various levels of granularity within WSIs. The results indicate that this fusion method achieved remarkable metrics with an ACC of $93.42\% \pm 0.42\%$, demonstrating a substantial improvement compared to individual classifiers. Moreover, the fusion approach yielded promising a SEN rate of $92.15\% \pm 0.8\%$ and an SPC rate of $94.73\% \pm 0.62\%$, indicating its capability to effectively identify both positive and negative instances of SFT and MEN tumors. Furthermore, the method exhibited a high PRC of $94.74\% \pm 0.57\%$, BAC of $93.44\% \pm 0.41\%$, and F1-score of $93.42\% \pm 0.43\%$, underscoring its robustness in maintaining a balance between true positives and true negatives while minimizing false classifications.

These results emphasize the power of combining complementary information to capture different texture variations within WSIs. It also, highlights the efficacy of ViT attention mechanisms on directing attention to significant WSI regions over multiscale, thus enabling the precise and dependable classification of SFT and MEN tumors in digital pathology. Figure 4 illustrates three examples of model predictions, showcasing the classification scores of SVM and ViT-B/16 models for each instance across different WSI scales. These scores contribute to the final classification decision.

4 Conclusion and Future Work

This paper proposed a novel two-stage multi-scale approach to diagnose Meningioma and Solitary Fibrous Tumors based on histopathological whole slide images (WSIs). The proposed approach incorporated the strengths of vision transformers and texture-based analysis techniques in capturing the variational patterns and spatial correlations between different regions within WSIs. By utilizing this approach over different WSI magnification levels, we enabled the capture of different local and global features for a more comprehensive analysis. Additionally, we utilized circular-aligned patches at various scales to ensure robust analysis of tissue samples to accommodate various scales and orientations of WSIs. Fusing all the decision scores through a Deep Neural Network achieved promising results with 93.42% accuracy, 92.15% sensitivity, 94.73% specificity, 94.74% precision, 93.44% balanced accuracy, and 93.42% F1 score. These results underscored the potential of the proposed approach in improving diagnostic accuracy and ultimately patient outcomes in the field of pathology. Future studies could investigate the sub-typing of these tumor types, expand the dataset to include other types, thus enhancing the model's generalizability, and explore the integration of other AI techniques that provide interpretability and significance of different WSI features.

References

1. Apra, C., et al.: Molecular description of meningeal solitary fibrous tumors/hemangiopericytomas compared to meningiomas: two completely separate entities. *J. Neurooncol.* **154**(3), 327–334 (2021)
2. Azam, M.T., et al.: A novel Vit-based multi-scaled and rotation-invariance approach for precise differentiation between meningioma and solitary fibrous tumor. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–4 (2024). <https://doi.org/10.1109/ISBI56570.2024.10635228>
3. Bharath, S., Sharma, D., Yadav, S.K., Shekhar, S., Jha, C.K.: A systematic review and meta-analysis of touch imprint cytology and frozen section biopsy and their comparison for evaluation of sentinel lymph node in breast cancer. *World J. Surg.* **47**(2), 478–488 (2023)
4. Buerki, R.A., Horbinski, C.M., Kruser, T., Horowitz, P.M., James, C.D., Lukas, R.V.: An overview of meningiomas. *Future Oncol.* **14**(21), 2161–2177 (2018)

5. Chen, T., et al.: Differentiating intracranial solitary fibrous tumor/hemangiopericytoma from meningioma using diffusion-weighted imaging and susceptibility-weighted imaging. *Neuroradiology* **62**, 175–184 (2020)
6. Chen, Z., Ye, N., Jiang, N., Yang, Q., Wanggou, S., Li, X.: Deep learning model for intracranial hemangiopericytoma and meningioma classification. *Front. Oncol.* **12**, 839567 (2022)
7. Dash, S., Senapati, M.R.: Gray level run length matrix based on various illumination normalization techniques for texture classification. *Evol. Intel.* **14**(2), 217–226 (2021)
8. Dong, J., et al.: Differential diagnosis of solitary fibrous tumor/hemangiopericytoma and angiomatous meningioma using three-dimensional magnetic resonance imaging texture feature model. *BioMed Res. Int.* **2020** (2020)
9. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
10. El-Abtah, M.E., Murayi, R., Lee, J., Recinos, P.F., Kshetry, V.R.: Radiological differentiation between intracranial meningioma and solitary fibrous tumor/hemangiopericytoma: a systematic literature review. *World Neurosurgery* **170**, 68–83 (2023)
11. Galldiks, N., et al.: Use of advanced neuroimaging and artificial intelligence in meningiomas. *Brain Pathol.* **32**(2), e13015 (2022)
12. Hossain, S., Chakrabarty, A., Gadekallu, T.R., Alazab, M., Piran, M.J.: Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification. *IEEE J. Biomedical Health Inform.* (2023)
13. Kataria, S.P., Bhutani, N., Kumar, S., Singh, G., Sen, R., Singh, I.: Solitary fibrous tumor of central nervous system masquerading as Meningioma: report of a rare case. *Int. J. Surg. Case Rep.* **54**, 10–14 (2019)
14. Ke-Chen, S., Yun-Hui, Y., Wen-Hui, C., Zhang, X.: Research and perspective on local binary pattern. *Acta Autom. Sinica* **39**(6), 730–744 (2013)
15. Kim, S.W., Roh, J., Park, C.S.: Immunohistochemistry for pathologists: protocols, pitfalls, and tips. *J. Pathol. Transl. Med.* **50**(6), 411 (2016)
16. Kong, X., Luo, Y., Li, Y., Zhan, D., Mao, Y., Ma, J.: Preoperative prediction and histological stratification of intracranial solitary fibrous tumours by machine-learning models. *Clin. Radiol.* **78**(3), e204–e213 (2023)
17. Li, J., Yan, Y., Liao, S., Yang, X., Shao, L.: Local-to-global self-attention in vision transformers. arXiv preprint [arXiv:2107.04735](https://arxiv.org/abs/2107.04735) (2021)
18. Li, X., et al.: Deep learning attention mechanism in medical image analysis: basics and beyonds. *Int. J. Network Dyn. Intell.*, 93–116 (2023)
19. Li, X., et al.: Presurgical differentiation between malignant haemangiopericytoma and angiomatous meningioma by a radiomics approach based on texture analysis. *J. Neuroradiol.* **46**(5), 281–287 (2019)
20. Li, Z., et al.: Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors. *IScience* **26**(1) (2023)
21. Liu, X., et al.: Differentiation of intracranial solitary fibrous tumor/hemangiopericytoma from atypical meningioma using apparent diffusion coefficient histogram analysis. *Neurosurg. Rev.* **45**(3), 2449–2456 (2022)
22. Mahesh, T., Geman, O., Margala, M., Guduri, M., et al.: The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. *Healthc. Anal.* **4**, 100247 (2023)

23. Ohba, S., et al.: Clinical and radiographic features for differentiating solitary fibrous tumor/hemangiopericytoma from meningioma. *World Neurosurgery* **130**, e383–e392 (2019)
24. Ostrom, Q.T., Cioffi, G., Waite, K., Kruchko, C., Barnholtz-Sloan, J.S.: CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018. *Neuro Oncol.* **23**(Supplement_3), iii1–iii105 (2021)
25. Park, Y., Guldmann, J.M.: Measuring continuous landscape patterns with Gray-Level co-occurrence matrix (GLCM) indices: an alternative to patch metrics? *Ecol. Ind.* **109**, 105802 (2020)
26. Schiariti, M., Goetz, P., El-Maghraby, H., Tailor, J., Kitchen, N.: Hemangiopericytoma: long-term outcome revisited. *J. Neurosurg.* **114**(3), 747–755 (2011)
27. Shamshad, F., et al.: Transformers in medical imaging: a survey. *Med. Image Anal.*, 102802 (2023)
28. Tao, X., Yan, X., Zhang, Y., Qin, S.: Intracranial solitary fibrous tumor mimicking meningioma. *J. Craniofacial Surgery* **34**(7), e688–e690 (2023)
29. Ugga, L., Spadarella, G., Pinto, L., Cuocolo, R., Brunetti, A.: Meningioma radiomics: at the nexus of imaging, pathology and biomolecular characterization. *Cancers* **14**(11), 2605 (2022)
30. Velázquez Vega, J.E., Ballester, L.Y., Schniederjan, M.J.: Tumors of the central nervous system. In: *Practical Oncologic Molecular Pathology: Frequently Asked Questions*, pp. 121–145 (2021). <https://doi.org/10.1007/978-3-030-73227-1>
31. Viswanathan, V.S., Toro, P., Corredor, G., Mukhopadhyay, S., Madabhushi, A.: The state of the art for artificial intelligence in lung digital pathology. *J. Pathol.* **257**(4), 413–429 (2022)
32. Wu, Y., Qi, S., Sun, Y., Xia, S., Yao, Y., Qian, W.: A vision transformer for emphysema classification using CT images. *Phys. Med. Biol.* **66**(24), 245016 (2021)



HFENet: High-Frequency Enhanced Network for Shape-Aware Segmentation of Left Ventricle in Pediatric Echocardiograms

Tianxiang Chen¹, Ziyang Wang², and Zi Ye³(✉)

¹ University of Science and Technology of China, Hefei, China
txchen@mail.ustc.edu.cn

² Department of Computer Science, University of Oxford, Oxford, UK
ziyang.wang@cs.ox.ac.uk

³ Institute of Intelligent Software, Guangzhou, China
yezi1022@gmail.com

Abstract. Automated ventricular function analysis can make health-care more consistent and available, especially where resources are scarce. However, current segmentation methods trained on adult heart ultrasounds cannot finely delineate the irregular shape of the left ventricle due to the ignorance of boundary feature exploration. To address this challenge, we introduce HFENet for shape-aware left ventricle segmentation. We propose a High-Frequency Enhancement Block (HFEB) that focuses on enhancing the high-frequency component, which is also the boundary area of left ventricles in pediatric echocardiograms. This way, the target boundary details can be explored during feature extraction. We propose space-frequency consistency loss to refine the shape of predicted masks further. Specifically, our new loss function incorporates spatial and frequency domain loss components to jointly refine predicted mask shapes in cases where current spatial-domain segmentation losses cannot be optimized further. Experiments carried out on two public datasets prove the superiority of the proposed HFENet in predicting the fineness of target shapes.

Keywords: Left ventricle · Semantic segmentation · Frequency domain · Lightweight

1 Introduction

With its rapid image acquisition, relative affordability, and non-reliance on ionizing radiation, echocardiography has become the most prevalent method for assessing children's congenital and acquired heart conditions [1]. Accurate segmentation of the left ventricle (LV) plays an essential role in this process, as it can enable the calculation of critical clinical metrics such as left ventricular mass, ejection fraction, and end-diastolic and end-systolic volumes [2].

However, manual delineation of the left ventricle in pediatric echocardiography faces notable challenges. Significant inter-observer variability and inter-modality

discordance can lead to inconsistent measurements, undermining diagnostic accuracy [3, 4]. Furthermore, the process is labor-intensive, adding to its inefficiency. These factors highlight the need for more standardized, automated approaches in clinical practice.

So far, deep learning has demonstrated notable success in enhancing the reliability and accuracy of left ventricular (LV) function assessment through echocardiography in adults, as evidenced by various studies [5]. Yet, it is more challenging in children due to varied anatomical abnormalities, heart rate, size, and cooperation ability. These factors contribute to a broad scope of spatial and temporal resolutions, impacting the overall quality of echocardiographic imaging [6]. Consequently, there is a degree of uncertainty regarding the generalization of machine learning models, which are primarily trained on adult datasets, to pediatric echocardiography, given these increased variabilities.

In this research, we utilize two distinct datasets from 4467 echocardiograms obtained from a total of 1958 pediatric patients at Lucile Packard Children’s Hospital Stanford between 2014 and 2021. This recent study by Reddy et al. has introduced these significant datasets with a gender distribution of 43% female [7], where the first apical four-chamber (A4C) dataset and the second parasternal short-axis (PSAX) dataset were extracted from the echocardiograms, resulting in 6449 two-dimensional images and 9001 images, respectively. These patients range in age from newborn to 18 years, offering a broad spectrum of pediatric cardiac profiles. By utilizing these datasets, this research is set to make a significant leap in pediatric cardiac care, aiming to strengthen the accuracy and efficiency of automating LV segmentation and, consequently, improving the overall quality of diagnosis and treatment in pediatric cardiology through innovative deep learning techniques.

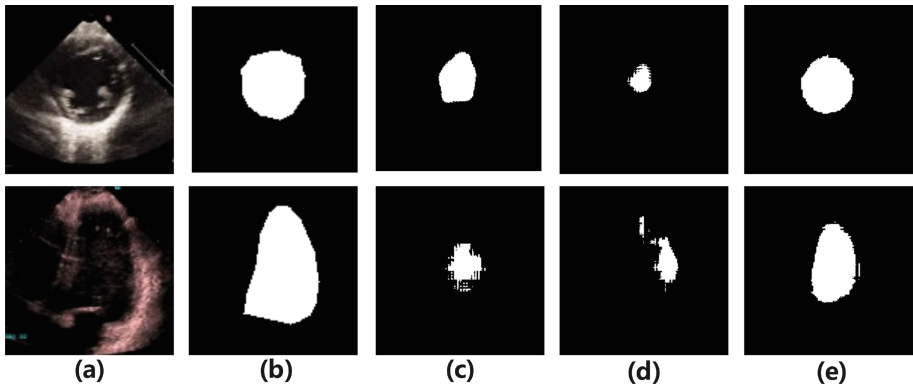


Fig. 1. Visualization of the left ventricle segmentation challenge (a) Input image. (b) Ground Truth. (c) Swin U-Net. (d) SpectFormer. (e) Ours (HFENet-3).

Benefiting from the Vision Transformer’s (ViT) [8] success applied in medical image tasks, several transformer architectures have been explored in the literature [9–12]. One of the most significant approaches has been Swin-Unet [13], a pure transformer-based U-shaped Encoder-Decoder network. In addition, it takes the Swin Transformer block as the fundamental unit for feature representation as well as long-range semantic information interactive learning [14].

However, when applying the present segmentation methods to segment the left ventricle in pediatric echocardiograms, the challenge exists since the irregular shape of left ventricles still cannot be well segmented since these methods pay insufficient attention to high-frequency boundary details, which can be seen in Fig. 1. It has been noticed that the critical importance of frequency domain analysis in computer vision is well-documented in the literature [15, 16], highlighting that low frequencies in images correspond to global structures and color and high frequencies reveal detailed attributes [17]. To highlight high-frequency boundary details, we get inspired by the works on frequency learning and then propose a High-Frequency Enhanced Network (HFENet). Our HFENet consists of an innovative module called the High-Frequency Enhancement Block (HFEB) that emphasizes high-frequency (HF) information, which is crucial for extracting object boundaries in segmentation tasks. In the block, we also employ cross-attention to mix in the low-frequency details, effectively achieving a balanced integration of detailed and broad features. Also, this study proposes a novel loss function called space-frequency consistency loss to further refine the left ventricle shapes with the aid of the frequency domain. Moreover, our contributions can be outlined as follows:

- We devise High-Frequency Enhancement Blocks (HFEB) embedded into our network encoder to highlight the high-frequency components of the extracted feature maps, which are also the boundaries of segmented objects. This way, a more precise shape-aware segmentation can be promoted during feature extraction.
- We propose space-frequency consistency loss, a novel loss function that integrates spatial domain loss with frequency domain loss to boost the shape-aware segmentation effect further, even though two predicted masks have very similar spatial structures
- Experiments demonstrate the advanced performance of our HFENet on two public datasets compared with other recently well-performed segmentation methods.

2 The Proposed Approach

The overall architecture of the HFENet proposed in this study, as illustrated in Fig. 2, consists of an encoder, bottleneck, decoder, and skip connections. The skip connections are adopted for fusing the multi-scale features from the encoder with the up-sampled features, just like the U-Net. We will elaborate on our model in more detail in the following subsections.

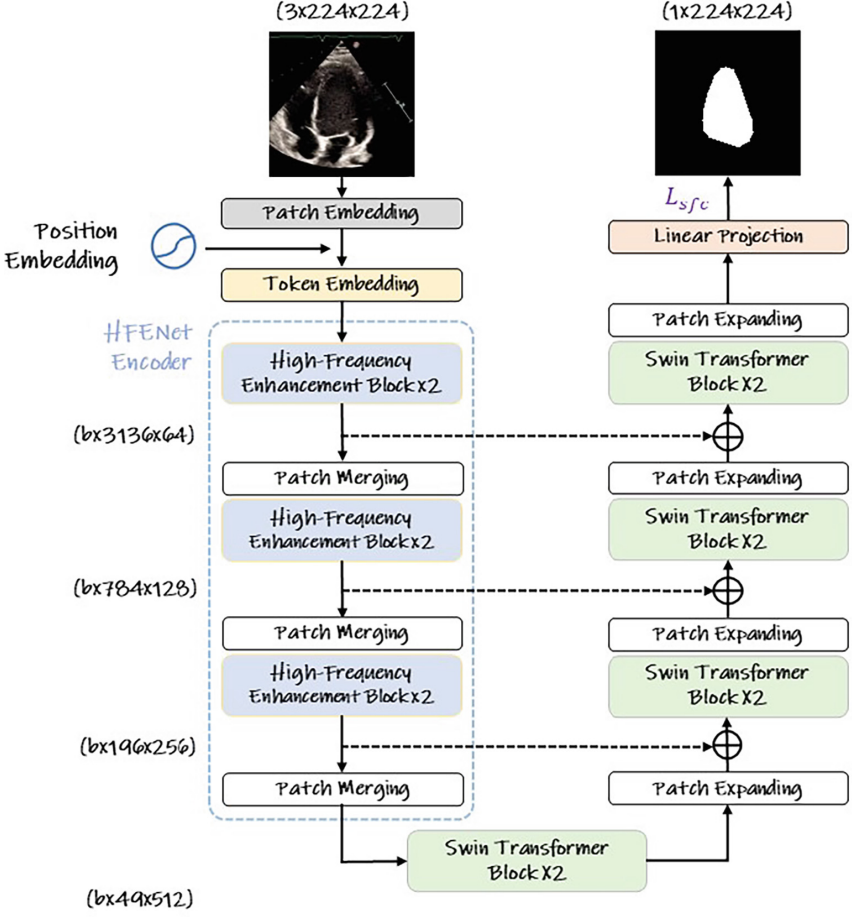


Fig. 2. The overall structure of the proposed model HFENet.

2.1 High-Frequency Enhancement Block

The High-Frequency Enhancement Block is structured as a combination of a High-Frequency Enhancement Module (HFEM) and a Swin Transformer Block (STB). Here, We introduce this novel HFEM, which aims to enhance high-frequency information processing for semantic segmentation tasks. Figure 3 presents the architecture of our process.

The HFEM comprises a Dual-Tree Complex Wavelet Transform (DTCWT) at the bottom that processes Low-Frequency (LF) and High-Frequency (HF) components, which are then subsequently processed through the Tensor Blending Method (TBM) and Einstein Blending Method (EBM) respectively, to derive Low-Frequency Representation (LFR) and High-Frequency Representation (HFR), modified by learnable weight matrices W_ϕ and W_ψ [18].

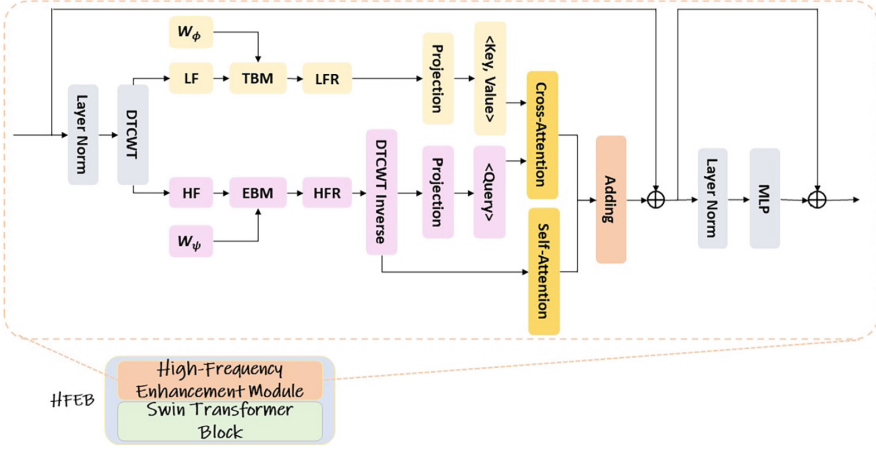


Fig. 3. The structure of the High-Frequency Enhancement Block.

One must emphasize that the Dual-Tree Complex Wavelet Transform Inverse (DTCWT Inverse) [18] is applied in this module solely to the High-Frequency Representation (HFR) by setting the low-frequency component to zero, effectively filtering out the low-frequency information and enhancing high-frequency details.

At the top, an ‘Adding’ process combines outputs from the Cross-Attention and Self-Attention modules, which are crucial in emphasizing fine-grained details and long-range dependencies within the data. The Self-Attention mechanism processes the high-frequency component in isolation. By applying this, the model can concentrate on specific areas of the image with significant textural details, efficiently amplifying the effect of this information. Accordingly, the formulation of Self-Attention can be expressed as follows:

$$\text{Self-Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q , K , and V indicate the query, key, and value matrices derived from the HFR after the DTCWT Inverse procedure.

With the HFEM, the Cross-Attention mechanism integrates signals across disparate frequency spectra. It takes a set of queries generated from the high-frequency component, later applying these queries modulates with the keys and values extracted from the low-frequency domain, facilitating the integration of global and fine-grained information. The simplified formulation of Cross-Attention is:

$$\text{Cross-Attention}(Q_{HF}, K_{LF}, V_{LF}) = \text{softmax}\left(\frac{Q_{HF}K_{LF}^T}{\sqrt{d_k}}\right)V_{LF} \quad (2)$$

where Q_{HF} is the query matrix derived from the HFR, K_{LF} and V_{LF} are the key and value matrices which can be originated from the LFR, respectively.

2.2 Swin Transformer Block

According to shifted window multi-head self-attention (SW-MSA), the Swin Transformer Block (STB) comprises two successive transformer blocks, where each multi-head attention module is substituted with a window-based attention module. It segments the image into discrete, non-overlapping patches, facilitating the application of self-attention mechanisms to understand inter-patch dependencies, thereby reducing the computational intensity.

2.3 Encoding Path

The encoder layer is designed to capture high-frequency details and focus on relevant spatial regions. Stacked HFEB blocks are systematically applied to embed the input image into a latent space and perform representation learning through successive stages. Like Swin-Unet, the patch merging layer can down-sample the spatial representation while upsampling the channel representation [13].

2.4 Decoding Path

The STB is pivotal in reconstructing spatial details from encoded features in the decoder architecture. The patch merging layer was replaced with the patching expanding layer to gradually increase the spatial dimensions while lowering the feature dimensions.

2.5 Space-Frequency Consistency Loss

The common loss functions for segmentation only calculate the loss in the space domain while ignoring space-frequency domain consistency. Specifically, similarity in space structure does not necessarily represent similarity in frequency. For example, as shown in Fig. 4, the two prediction masks (the boundaries are highlighted in red and green lines) are very similar in spatial shapes with GT in (a)

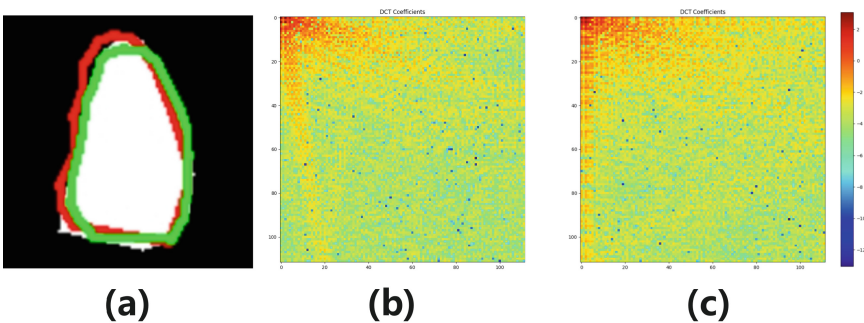


Fig. 4. The visual results of (a) Groundtruth + Predicted Mask A + Predicted Mask B all on the spatial domain (b) Predicted Mask A after DCT (c) Predicted Mask B after DCT.

differ a lot once transformed to the frequency domain by discrete cosine transform (DCT). Thereby we propose a space-frequency consistency loss, which contains a spatial domain loss \mathcal{L}_{space} and a frequency domain loss \mathcal{L}_{freq} . We adopt the traditional BCE loss as our \mathcal{L}_{space} . For \mathcal{L}_{freq} , we utilize DCT to project the predicted mask and GT mask into cosine components for different 2-dimensional frequencies and use the L2 norm to constrain the loss. We define \mathcal{L}_{freq} as

$$\mathcal{L}_{freq} = \sqrt{\|DCT(y) - DCT(\hat{y})\|^2} \quad (3)$$

where y and \hat{y} denote predicted mask and GT mask, respectively. By combining the two domain losses together, we can achieve better shape-aware segmentation quality than using spatial domain loss alone since frequency clues are explored. The final expression of our proposed space-frequency consistency loss is

$$\mathcal{L}_{sfc} = \mathcal{L}_{space} + \alpha \mathcal{L}_{freq} \quad (4)$$

where a hyper-parameter α is placed before \mathcal{L}_{freq} and is optimally set to 1 according to our ablation study.

3 Experiment

This section outlines the dataset deployed in our experimental investigation, describes the training and testing settings, and finally presents the experimental results, demonstrating the performance of our method and benchmarking it against contemporary methodologies.

3.1 Dataset

All experiments in the current section are carried out on the dataset comprising 4467 echocardiograms from 1958 pediatric patients (43% female, aged 0–18 years). The data gathered from Lucile Packard Children’s Hospital Stanford between 2014 and 2021 was divided into 80% for training, 10% for testing, and 10% for validation. It features 7643 grayscale two-dimensional video clips from A4C and PSAX views and 17600 labeled images derived from these echocardiograms.

3.2 The Experimental Set

The computational setup includes a single Tesla V100-32GB GPU, a 12-core CPU, and 61G of RAM. The system runs on an Ubuntu 18 environment with CUDA 11.0 and Pytorch 1.13 software.

The network’s training was conducted over 150 epochs, starting with an initial learning rate 1e-4. With the purpose of balancing computational efficiency and the model’s accuracy, batch sizes of 24 for training were chosen. The model’s performance was evaluated every five epochs, and a patience parameter of 10 was set for early stopping to avoid overfitting. Regarding the network’s structure, the depth was set up with layers in the configuration of [2, 2, 2, 2]. The multi-head attention mechanism was integral to this design, for which the number of heads was established as [3, 6, 12, 24].

3.3 Quantitative Comparison with Other Methods

Table 1. Performance comparison with other methods on PSAX and A4C.

Methods	Dataset PSAX		Dataset A4C	
	DSC	mIoU	DSC	mIoU
UNet_ fcn [19]	0.8624	0.7581	0.8312	0.7112
UNet_ deeplabv3 [19]	0.8682	0.7671	0.8336	0.7146
UNet_ pspnet [19]	0.8675	0.7660	0.8401	0.7243
Swin-Unet [13]	0.9113	0.9099	0.8920	0.8897
Spectformer [20]	0.9160	0.9144	0.9009	0.8979
PVT [21]	0.8910	0.8909	0.8820	0.8802
UniformerV2 [22]	0.9073	0.9059	0.8917	0.8894
Ours	0.9246	0.9215	0.9129	0.9096

Our study initially conducted comparative experiments using U-Net architecture with three different backbones - FCN, DeepLabV3, and PSPNet, alongside other SOTA segmentation models, including Swin-Unet [13], Spectformer [20], and UniformerV2 [22]. The findings presented in Table 1 indicate that our model significantly surpasses other models in DSC and mIoU, which represents superior shape-aware segmentation quality. In addition, the values in bold format are the highest numbers for the corresponding metrics.

3.4 Qualitative Comparison with Other Methods

We also provide a qualitative visualization of the left ventricle segmentation results in Fig. 5. Other methods more or less suffer flawed segmentation of target shapes due to a lack of attention paid to boundary information in pediatric echocardiograms. Our proposed method has the closest shape similarity with ground truths because it has advantages in providing more accurate and refined delineations of the shape of LV in pediatric echocardiograms.

3.5 Ablation Studies

Effect of High-Frequency Enhancement Block. We first conducted experiments to study the effect of varying the number of HFEB within the HFENet encoder architecture. As shown in Table 2, We incrementally integrated 1, 2, and 3 HFEBs into the encoder path’s first, second, and third stages, respectively, resulting in HFENet-1, HFENet-2, and HFENet-3. In cases where HFEBs were not used, pure STBs were employed as substitutes in these stages. Our findings indicate that adding HFEBs correlates with improvements across both datasets in both evaluation metrics (DSC and mIoU), demonstrating the positive effect of our proposed HFEB in boosting shape-aware segmentation during feature extraction.

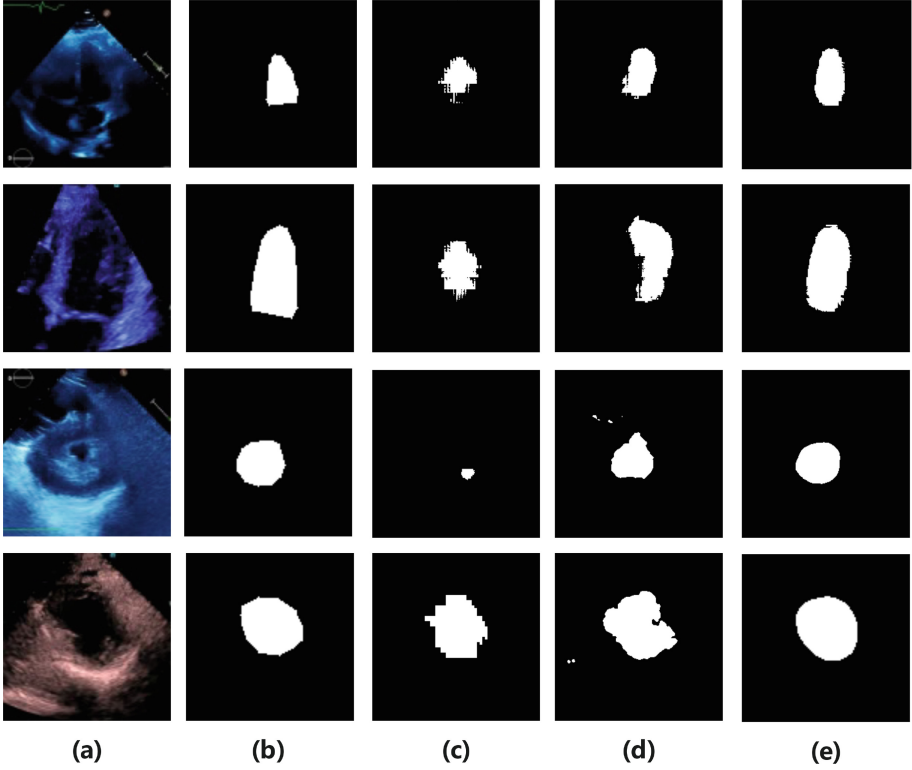


Fig. 5. Visual comparisons of different methods on the A4C dataset. (a) Input image. (b) Ground Truth. (c) Swin U-Net. (d) SpectFormer. (e) Ours (HFENet-3).

Table 2. Ablation study on performance metrics across PSAX and A4C datasets.

Methods	Dataset PSAX		Dataset A4C	
	DSC	mIoU	DSC	mIoU
HFENet-1	0.9200	0.9181	0.9044	0.9011
HFENet-2	0.9214	0.9193	0.9071	0.9038
HFENet-3	0.9220	0.9199	0.9129	0.9096
HFENet(Resize)-1	0.9192	0.9173	0.9047	0.9016
HFENet(Resize)-2	0.9209	0.9190	0.9059	0.9029
HFENet(Resize)-3	0.9246	0.9215	0.9098	0.9055

Table 3. Impact of varying α coefficients in hybrid loss functions.

α	0	0.1	1
DSC of A4C	0.9043	0.9074	0.9129

Table 4. Comparison of the ablation study and our method in terms of parameters, inference speed, and GFLOPs.

Methods	Params (in M)	IS in ms, PSAX	IS in ms, A4C	GFLOPs
UNet_ fcn [19]	29.060	11.024	11.116	12.660
UNet_ deeplabv3 [19]	29.060	11.750	12.228	12.710
UNet_ pspnet [19]	29.050	11.392	11.629	12.360
Swin-UNet [13]	28.146	10.980	11.623	10.075
Spectformer [20]	30.254	9.394	9.229	10.480
UniformerV2 [22]	30.040	10.939	10.882	9.801
HFENet-1	24.470	4.902	4.052	7.093
HFENet-2	25.746	6.540	6.035	8.089
HFENet-3	27.804	8.400	7.635	9.079
HFENet(Resize)-1	24.470	4.384	4.02	7.093
HFENet(Resize)-2	25.746	6.069	5.389	8.089
HFENet(Resize)-3	27.804	8.015	7.267	9.079

Note: IS stands for Inference Speed.

Secondly, we intended to replace the DTCWT Inverse operation in the HFEB with the Resize operation to determine whether the inverse operation is necessary. The motivation for this ablation is that the inversion can no longer recover the original input feature because we enhance the high-frequency part of the input feature, so we wonder whether inversion is still needed compared with the resize operation. We still follow the former HFEB number settings but replace inversion with a resize operation and denote this experiment setting as HFENet(Resize)-1, HFENet(Resize)-2, and HFENet(Resize)-3, respectively. It can also be observed that HFENet (resize) shows improved performance metrics as the HFEB number increases. Moreover, as Table 4 suggested, HFENet(Resize) offers a quicker inference speed than the standard HFENet, attributed to the resize operations being faster than the DTCWT Inverse operations.

Effect of Space-Frequency Consistency Loss. Table 3 shows the effect of our space-frequency consistency loss in A4C. $\alpha=0$ means no frequency domain loss is added to the final loss, and this condition performs the worst in Table 3. The reason is that only spatial loss is insufficient to determine which predicted mask is better once the two predicted masks yield very close IoU or MSE values. With the value of α growing from 0 to 0.1 and then to 1, the DSC value also rises in the A4C dataset, meaning that the frequency domain loss part indeed contributes to shape-aware segmentation since it can further align the predicted and GT features in the frequency domain to determine further which prediction is better in shape. This phenomenon complies with the motivation of our new loss design.

3.6 Model Complexity Analysis

Table 4 evaluates the scale and efficiency of the models. As the number of HFEBs increases within the HFENet series, there is an upward trend in both parameters and GFLOPs, suggesting a rise in model complexity and computational load. However, our proposed method demonstrates significantly lower GFLOPs than U-Net architectures with FCN, DeepLabV3, PSPNet backbones, and other SOTA ViT-based methods. Despite this, our segmentation performance greatly surpasses these methods. Furthermore, our model outperforms all comparison models in terms of parameter complexity and offers faster inference speed.

4 Conclusion

This work proposes an innovative deep network architecture, HFENet, for left ventricle segmentation in pediatric echocardiograms, explicitly focusing on high-frequency enhancement mechanisms. By integrating HFEB into the encoding phase, HFENet captures fine details more effectively and maintains a balance with the low-frequency domain through cross-attention. Combining MSE in the DCT domain, our customized loss function further refines the segmentation accuracy. Notably, HFENet features a lightweight architecture with small computational demands. Extensive experiments on both LV datasets demonstrate its robust performance in diverse imaging contexts.

References

1. Zhang, J., et al.: Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation* **138**(16), 1623–1635 (2018)
2. Ardila, D., et al.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**(6), 954–961 (2019)
3. Lang, R.M., et al.: Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American society of echocardiography and the European association of cardiovascular imaging. *Eur. Heart J.-Cardiovasc. Imaging* **16**(3), 233–271 (2015)
4. Huang, H., et al.: Accuracy of left ventricular ejection fraction by contemporary multiple gated acquisition scanning in patients with cancer: comparison with cardiovascular magnetic resonance. *J. Cardiovasc. Magn. Reson.* **19**, 1–9 (2017)
5. Madani, A., Arnaout, R., Mofrad, M., Arnaout, R.: Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit. Med.* **1**(1), 6 (2018)
6. Power, A., et al.: Echocardiographic image quality deteriorates with age in children and young adults with Duchenne muscular dystrophy. *Front. Cardiovasc. Med.* **4**, 82 (2017)
7. Reddy, C.D., Lopez, L., Ouyang, D., Zou, J.Y., He, B.: Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. *J. Am. Soc. Echocardiogr.* **36**(5), 482–489 (2023)
8. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations* (2020)

9. Deng, K., et al.: TransBridge: a lightweight transformer for left ventricle segmentation in echocardiography. In: Noble, J.A., Aylward, S., Grimwood, A., Min, Z., Lee, S.-L., Hu, Y. (eds.) ASMUS 2021. LNCS, vol. 12967, pp. 63–72. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87583-1_7
10. Zeng, Y., et al.: MAEF-Net: multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography. *Ultrasonics* **127**, 106855 (2023)
11. Azarmehr, N., Ye, X., Sacchi, S., Howard, J.P., Francis, D.P., Zolgharni, M.: Segmentation of left ventricle in 2D echocardiography using deep learning. In: Medical Image Understanding and Analysis: 23rd Conference, MIUA 2019, Liverpool, UK, July 24–26, 2019, Proceedings 23, pp. 497–504. Springer (2020). https://doi.org/10.1007/978-3-031-31407-0_7
12. Shoaib, M.A., et al.: An overview of deep learning methods for left ventricle segmentation. *Comput. Intell. Neurosci.* **2023**(1), 4208231 (2023)
13. Cao, H., et al.: Swin-Unet: Unet-Like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III, pp. 205–218. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_9
14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
15. Pan, Z., Cai, J., Zhuang, B.: Fast vision transformers with HiLo attention. *Adv. Neural. Inf. Process. Syst.* **35**, 14541–14554 (2022)
16. Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., Ding, S.: Detecting camouflaged object in frequency domain. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4504–4513. New Orleans, LA, USA (2022)
17. Chu, X., et al.: Twins: revisiting the design of spatial attention in vision transformers. *Adv. Neural. Inf. Process. Syst.* **34**, 9355–9366 (2021)
18. Patro, B.N., Agneeswaran, V.S.: Scattering vision transformer: spectral mixing matters. In: Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 54152–54166 (2023)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III, pp. 234–241. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Patro, B.N., Namboodiri, V.P., Agneeswaran, V.S.: SpectFormer: frequency and attention is what you need in a vision transformer (2023). <https://doi.org/10.48550/arXiv.2304.06446>
21. 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)
22. Li, K., et al.: UniFormerV2: unlocking the potential of image ViTs for video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1632–1643 (2023)



Chaos Theory Based Gravitational Search Algorithm For Medical Image Segmentation

Sajad Ahmad Rather^{1(✉)}, Partha Pratim Roy¹, and Sujit Das²

¹ Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee 247667, India

{sajad.pd,partha}@cs.iitr.ac.in

² Department of Computer Science and Engineering, National Institute of Technology, Telangana 506004, India

sujit.das@nitw.ac.in

Abstract. Multilevel thresholding plays a crucial role in image processing, with extensive applications in object detection, machine vision, medical imaging, and traffic control systems. It entails the partitioning of an image into distinct regions based on optimal pixel values. However, as the number of threshold levels increases, so does the computational cost for segmentation. To address this challenge, a novel method is proposed namely Chaos theory based Gravitational Search Algorithm (CGSA) for multilevel thresholding. CGSA combines the standard Gravitational Search Algorithm (GSA) for exploration with chaotic maps for exploitation of the complex pixel problem space. In this study, Kapur's entropy method is utilized to segment sample images into various partitions based on optimal pixel values. The effectiveness of CGSA in real-world scenarios is evaluated using COVID-19 chest CT scan imaging datasets from Kaggle database. The quality, symmetry, and consistency of the segmented output are assessed using metrics like Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Feature Similarity Index Measure (FSIM). Qualitative analysis includes convergence curves, segmented graphs, colormap images, and box plots. Statistical validation is conducted using the signed Wilcoxon rank sum test. Additionally, a comparison is made between CGSA's performance and that of eight state-of-the-art heuristic algorithms. The findings demonstrate the superior performance of CGSA, evidenced by its reduced computational time and enhanced image quality metrics values. Specifically, CGSA achieved SSIM of 0.81, FSIM of 0.82, and PSNR of 24.27, surpassing the performance of other competitive algorithms.

Keywords: Medical Imaging · Image Segmentation · Gravitational Search Algorithm · Chaos Theory · Kapur's Entropy · COVID-19

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78104-9_5.

1 Introduction

Image segmentation involves dividing an image into smaller regions of interest based on pixel intensity variations [1]. This process is especially crucial in medical imaging techniques like MRI and CT scans. Image segmentation techniques are utilized across a wide array of disciplines, including visual perception analysis, computational intelligence, and structure detection [2]. They provide a valuable tool for streamlining difficult procedures and increasing productivity. Image segmentation is a method used in computer vision that extracts features from pictures and uses them in further processing. MRI and CT images are frequently subjected to medical image segmentation to indicate the body region that requires attention. As a consequence, medical professionals may concentrate on treating the problematic location which may speed up the diagnosis procedure [3].

Image segmentation methods vary, but thresholding stands out as the most widely adopted due to its simplicity and effectiveness [4]. It primarily separates different portions of an image by setting a basic threshold value, making it easier to discern similar elements. Thresholding comes in two main forms: bi-level and multi-level. Bi-level thresholding, similar to binary thresholding, assigns a single intensity value to pixels below a specified threshold, whereas multilevel thresholding assigns multiple intensities hierarchically, resulting in a more intricate segmentation [5,6]. Noteworthy example of bi-level thresholding include Kapur's entropy method [7]. It is an advanced technique for image thresholding, aimed at differentiating between objects and background within an image [8]. The method involves evaluating a range of potential threshold values and computing the entropy for each. Entropy, in this context, is a quantitative measure of the information content or uncertainty within the image regions defined by the threshold. Specifically, Kapur's method seeks to identify the threshold that maximizes the total entropy of the segmented image, thereby achieving an optimal separation between the foreground and background regions. By maximizing entropy, this approach enhances the clarity and precision of the segmented image, thereby improving the effectiveness of subsequent tasks such as object detection and image analysis. In recent years, the emergence of the Coronavirus disease (COVID-19) in China in December 2019 has garnered significant attention [9]. Based on clinical trials and laboratory studies, Huang et al. [10] identified specific symptoms of COVID-19, including fever, cough, and shortness of breath. Declared a global health emergency by the World Health Organization (WHO), COVID-19 prompted worldwide preventive measures [11]. With over 600 million infections and nearly 7 million deaths in the past three years, COVID-19 has evolved into a global pandemic.¹

Several vaccines and medications have been developed to combat COVID-19. However, the list of COVID-19 variants is continually being updated with new mutations. The effectiveness of vaccination against these altered virus strains remains uncertain [12]. In the early stages of medical intervention and prevention, prompt and accurate detection of COVID-19 cases is crucial. Reverse Transcription Polymerase Chain Reaction (RT-PCR) is a commonly used technique

¹ <https://www.worldometers.info/coronavirus>.

for identifying COVID-19 infections, but its reliability is compromised by a relatively low rate of positive results and potential delays in detecting the virus for timely treatment [13]. Conversely, medical imaging tools like X-rays and CT scans offer higher detection rates and can reveal the virus within hours of infection, making them more precise in the early diagnosis of COVID-19. Automating the identification of COVID-19 using chest X-rays and CT scans could significantly mitigate the pandemic's impact on society. However, segmenting COVID-19 lesions on CT scans remains challenging due to their irregular shapes, diverse sizes, and indistinct boundaries between healthy and infected tissues [14]. Overcoming this challenge requires employing a range of image processing techniques to differentiate these properties based on similarities and differences. Therefore, in this study, we have implemented the Chaos theory-based Gravitational Search Algorithm (CGSA) for multi-level thresholding. The CGSA aims to address the limitations of Kapur's entropy scheme and provide accurate segmentation outcomes more quickly and with lower computational costs. The paper presents several key contributions:

1. **Introduction of Chaotic Gravitational Search Algorithm (CGSA) based segmentation methodology:** CGSA redefines image processing with its precise and efficient segmentation procedure. Its integration of chaotic maps enhances feature delineation and navigates complex search spaces, marking a significant advancement in heuristic segmentation techniques.
2. **Harnessing Chaos in segmentation:** At its core, CGSA strategically incorporates Chaos theory principles, mitigating local optima entrapment and optimizing segmentation accuracy. This integration establishes new standards for excellence in image processing methodologies.
3. **CGSA's application in Medical Imaging:** Beyond segmentation, CGSA finds applications in medical imaging, such as evaluating COVID-19 severity through chest CT scans. Rigorous comparative analyses underscore CGSA's superiority, positioning it as the preferred method for precise segmentation.

The organization of the paper proceeds as follows: In Sect. 2, a literature survey is conducted on the application of heuristic approaches in medical imaging. Section 3 outlines the methodology, including an explanation of GSA, chaotic maps, and the proposed CGSA-based image segmentation scheme. Following that, Sect. 4 presents the experimental results obtained from CT scan images. Finally, Sect. 5 concludes the study and discusses potential future research directions.

2 Literature Review

Image segmentation is a fundamental and challenging task in computer vision and image processing [3]. It plays a crucial role in various applications, such as object recognition, scene understanding, medical image analysis, and robotics [5]. The primary objective of image segmentation is to partition an input image into meaningful regions, each representing a distinct object or region of interest. Over the years, numerous image segmentation techniques have been proposed in the literature to address the diversity and complexity of images encountered in

real-world scenarios. One prominent and widely studied approach for image segmentation is multilevel thresholding [6]. Thresholding-based techniques involve dividing an image into regions based on pixel intensity levels. Multilevel thresholding extends this idea to partition images into multiple segments by selecting optimal intensity thresholds [7]. The effectiveness of multi-level thresholding lies in its simplicity and computational efficiency, making it a preferred choice in many image segmentation applications. The significance of multi-level thresholding in image segmentation has prompted numerous researchers to explore its application and propose various advancements in the field [8].

Image segmentation using multilevel thresholding has gained significant attention in medical imaging, particularly in the analysis of CT scan images for various diseases, including COVID-19. Several studies have explored the application of multilevel thresholding techniques to effectively segment lung regions and identify lung pathologies associated with COVID-19 infections [4]. In fact, the researchers like Su et al. [15] introduced CCMVO, an enhanced multiverse optimizer, for efficient processing of COVID-19 chest radiography. Incorporating horizontal and vertical search processes, the method outperformed other algorithms in image segmentation quality with reduced risk of stagnation. Validation through benchmark functions demonstrated the effectiveness in COVID-19 diagnosis using FSIM, PSNR, and SSIM metrics, offering a valuable tool for medical organizations. Liu et al. [16] proposed CLACO, a method combining ant colony optimization with Cauchy and greedy Levy mutations, enhancing COVID-19 X-ray image segmentation. This method utilized 2D Kapur's entropy as the fitness function for improved segmentation. Experimental results showed CLACO's superiority over variants and peer methods in benchmark functions. Similarly, Kumar Sahoo et al. [17] introduced Es-MFO, an enhanced Moth Flame Optimization algorithm, for accurate COVID-19 CT image classification and segmentation. It outperformed other techniques in benchmark tests, showing effectiveness in medical applications. Zhao et al. [18] introduced VMCSA, enhancing the Crow Search Algorithm with Variable Neighborhood Descent and Information Exchange Mutation methods. It outperforms alternatives in optimization and excels in COVID-19 X-ray image segmentation, showing robustness and superior outcomes. Likewise, the teams led by Houssein et al. [19] and Qi et al. [20] delved into the segmentation of COVID-19 X-ray images. These researchers proposed refined multilevel segmentation models employing swarm intelligence algorithms to enhance the accuracy of segmenting COVID-19 X-ray images. The literature review clearly reveals that Kapur's entropy scheme and Otsu's method are well-regarded as leading techniques for multi-level thresholding. These methods generally provide more flexibility and effectiveness compared to alternatives. However, while they perform well in simpler cases or with fewer thresholds, their efficiency declines as the number of thresholds increases because of greater computational requirements. This highlights a key issue with traditional thresholding methods: the computational burden. In contrast, heuristic algorithms present a viable solution due to their mathematical simplicity and faster convergence, which help lower computational costs and improve decision accuracy.

3 Methodology

The Chaotic Gravitational Search Algorithm (CGSA) is a robust amalgamation of the Gravitational Search Algorithm (GSA) with ten distinct chaotic maps, rendering it proficient in both exploratory and exploitative tasks. This section elucidates the mathematical framework that underpins the operational mechanics of CGSA.

3.1 Gravitational Search Algorithm (GSA)

Optimization algorithms often borrow concepts from nature or scientific principles to improve their efficiency. One such method, the Gravitational Search Algorithm (GSA), draws inspiration from physics. In GSA, the process begins by assigning masses to searcher agents, resembling how objects interact gravitationally in the physical world. This initialization stage is grounded in Newton's law of universal gravitation, which explains how the gravitational force between two masses is influenced by their mass and the distance between them. By simulating these gravitational interactions, GSA aims to guide the optimization process towards finding optimal solutions efficiently. Mathematically, Equation (1) calculates the gravitational force between masses 'x' and 'y' at time 't'.

$$\mathbf{f}_{xy^d(t)} = G(t) \frac{(m_{px}(t) \cdot m_{ay}(t))(\mathbf{x}_x^d(t) + \mathbf{x}_y^d(t))}{R_{xy}(t) + \varepsilon} \quad (1)$$

The dynamic interaction between the passive and active attractive masses within the scope of Equation (1) is evident through the variables $m_{ay}(t)$ and $m_{px}(t)$. Specifically, $m_{ay}(t)$ represents the gravitational pull exerted on a single point mass, while $m_{px}(t)$ illustrates the attractive force originating from a point mass situated within the gravitational field. Moreover, this dynamic relationship extends to the Euclidean metric R_{xy} , which depicts the spatial distance between these masses, and the infinitesimal constant ε , introduced to enhance numerical stability.

Achieving a delicate balance during the optimization process is crucial in the context of the GSA. The gravitational constant ' G ' plays a fundamental role in establishing this equilibrium, serving as the key factor in determining feasible positions within the vast solution space as outlined in Equation (2).

$$G(t) = G(t_0)e^{-\alpha \frac{CI}{MI}} \quad (2)$$

In the equation, $G(t_0)$ and $G(t)$ represent the starting and concluding values of the gravitational constant, correspondingly, with α serving as a minor coefficient. CI and MI denote the ongoing iteration and the maximum iterations, respectively. Determining mass involves accounting for active, passive, and inertia masses, wherein a greater mass signifies a heightened gravitational pull. The gravitational mass $M_x(t)$ is ascertained by Equation (3) when the active m_{ax} , passive m_{px} , and inertia m_{ix} masses equate.

$$m_{ax} = m_{px} = m_{ix} = M_x \quad (3)$$

$$m_x(t) = \frac{fit_x(t) - worst(t)}{best(t) - worst(t)} \quad (4)$$

$$M_x(t) = \frac{m_x(t)}{\sum_{y=1}^m m_x(t)} \quad (5)$$

The fitness measure, $m_x(t)$, as elaborated in Equation (4), appraises the performance of masses within a particular context. Factors such as $best(t)$ and $worst(t)$ within $fit_x(t)$ specify whether the objective leans towards minimizing or maximizing. In the context of Equation (5), the variable ‘m’ signifies the quantity of masses intricately distributed across a spatial domain with multiple dimensions, actively engaging in complex gravitational interactions. Additionally, Equation (6) is utilized by classical mechanics to determine the cumulative gravitational force.

$$\mathbf{f}_x^d(t) = \sum_{y=1, y \neq x}^m \gamma_y \mathbf{f}_{xy}^d(t) \quad (6)$$

In Equation (6), γ_y is a stochastic variable. As per Equation (1), it’s clear that more massive masses yield a heightened gravitational influence. Furthermore, the exploration spans the complete search domain, unveiling viable localities. This comprehensive exploration is crucial for upholding solution quality, facilitated by employing the *kbest* strategy as delineated in Equation (7). In the context of a physical system undergoing acceleration, a consequential force is invariably produced. Within the solution space, masses consistently apply force to each other, resulting in the manifestation of acceleration $\mathbf{a}_x^d(t)$. This acceleration steers solutions toward feasible regions. In Equation (8), $\mathbf{f}_x^d(t)$ signifies the mutual force exerted by masses.

$$\mathbf{f}_x^d(t) = \sum_{y=kbest, y \neq x}^m \gamma_y \quad (7)$$

$$\mathbf{a}_x^d(t) = \frac{\mathbf{f}_x^d(t)}{m_{ix}(t)} \quad (8)$$

In the GSA framework, the inertial mass is symbolized as $m_{ix}(t)$. Each individual point mass is characterized by both a positional attribute and a velocity. Conversely, towards the termination of the iterative procedure, a singular mass exhibiting a pronounced gravitational field prevails. Consequently, the determination of velocity $\mathbf{v}_x^d(t)$ and position $\mathbf{x}_x^d(t)$ is paramount for the identification of an optimal solution, as delineated in Equation (9) and Equation (10).

$$\mathbf{v}_x^d(t+1) = \gamma_y \mathbf{v}_x^d(t) + \mathbf{a}_x^d(t) \quad (9)$$

$$\mathbf{x}_x^d(t+1) = \mathbf{x}_x^d(t) + \mathbf{v}_x^d(t+1) \quad (10)$$

3.2 Chaos Theory

In the context of various metaheuristic algorithms, the use of extensive and diverse random number sequences is of paramount importance. When randomly generated numbers cluster within a specific range or repeatedly produce identical values, there is an increased risk of the algorithm becoming trapped in local optima. To mitigate this risk, it is crucial that the generated numbers exhibit diversity and have a spread spectrum. This diversity ensures a broader exploration of the solution space, helping the algorithm avoid getting stuck in sub-optimal solutions and improving its ability to discover more globally optimal solutions [21].

Table 1. Mathematical formulas describing chaotic mappings

Chaotic function	Chaotic map	Limits
Chebyshev	$x_{i+1} = \cos(i \cos^{-1}(x_i))$	(1,-1)
Circle	$x_{i+1} = \text{mod} (x_i + b - (\frac{a}{2\pi}) \sin(2\pi x_i), 1),$ a=0.5, b=0.2	(0,1)
Gauss	$x_{i+1} = \begin{cases} 1 & \text{if } x_i = 0 \\ \frac{1}{\text{mod}(x_i, 1)} & \text{otherwise} \end{cases}$	(0,1)
Iterative	$x_{i+1} = \sin\left(\frac{a\pi}{x_i}\right), a = 0.7$	(-1,1)
Logistic	$x_{i+1} = ax_i(1 - x_i), a = 0$	(0,1)
Piecewise	$x_{i+1} = \begin{cases} \frac{x_i}{P} & \text{if } 0 \leq x_i \leq P \\ \frac{x_i - P}{0.5 - P} & \text{if } P \leq x_i \leq 0.5, \quad P = 0.4 \\ \frac{1 - P - x_i}{0.5 - P} & \text{if } 0.5 \leq x_i < 1 - P \\ \frac{1 - x_i}{P} & \text{if } 1 - P \leq x_i < 1 \end{cases}$	(0,1)
Sine	$x_{i+1} = \frac{a}{4} \sin(\pi x_i), a = 4$	(0,1)
Singer	$x_{i+1} = \mu(7.86x_i - 23.31x_i^2 + 28.75x_i^3 - 13.302875x_i^4),$ $\mu = 2.3$	(0,1)
Sinusoidal	$x_{i+1} = ax_i^2 \sin(\pi x_i), a = 2.3$	(0,1)
Tent	$x_{i+1} = \begin{cases} \frac{x_i}{0.7} & \text{if } x_i < 0.7 \\ \frac{10}{3}(1 - x_i) & \text{if } x_i \geq 0.7 \end{cases}$	(0,1)

The foundational tenets of methodologies underpinned by chaos are contingent upon the utilization of functions representing discrete-time systems characterized by their inherently chaotic behavior. The theoretical framework posits that the numerical outputs emanating from chaotic maps intrinsically lack predictability, exhibiting prominent spread-spectrum attributes and non-periodic behavior. In their integration within metaheuristic algorithms, chaotic maps supplant traditional random variables, deviating from the conventional practices inherent in standard approaches [22, 23]. The construction of a chaotic sequence involves the aggregation of chaotic variables employed in a given iteration. The deliberate integration of chaotic sequences is aimed at endowing the algorithm with the capacity to transcend local minima during the pursuit of the global minimum, thereby underscoring its inherent adaptability. This intentional

amalgamation empowers heuristic algorithms to diverge from impractical regions within the vast expanse of the search space [24]. Consequently, the initiation of metaheuristic algorithms with chaotic maps is envisaged to expedite and refine the process of ascertaining the initial random number sequence, culminating in the generation of more efficacious solutions for optimization problems [25]. The chaotic maps subjected to scrutiny in this study are detailed in Table 1, with their corresponding random patterns illustrated in Fig. 1.

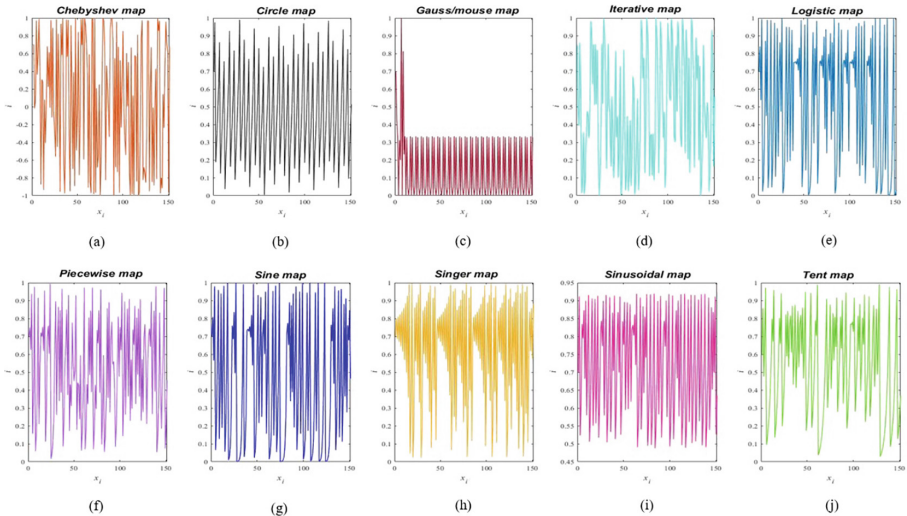


Fig. 1. The plots of the ten chaotic maps shown in (a)-(j) clearly illustrate the inherent random variations in the behavior of chaotic systems. Each plot captures the unpredictable fluctuations over time, highlighting the complexity and sensitivity to initial conditions that characterize chaotic dynamics. These visual representations underscore the diverse and intricate patterns that emerge from chaotic maps, providing a clear depiction of their erratic and non-linear nature.

3.3 Image Segmentation Using Chaotic Gravitational Search Algorithm (CGSA)

In this section, a novel strategy for image segmentation is proposed, integrating the Gravitational Search Algorithm (GSA) with chaos theory. The inherent challenges of the GSA technique, such as its slow convergence and susceptibility to getting stuck in local minima, necessitate innovative solutions. To address these issues, ten different chaotic maps are employed to enhance the performance of the standard GSA framework. A noteworthy characteristic of chaotic maps is their sensitivity to initial parameters. Even small adjustments to these parameters can lead to significant changes in the output. This observation highlights the potential of chaotic maps to introduce variability and dynamism into optimization algorithms, thereby augmenting their ability to explore solution spaces effectively and mitigate convergence challenges. Furthermore, chaotic normalization

ensures an appropriate balance of exploration and exploitation by analytically computing Equation (11).

$$C_i^{\text{norm}}(t) = \frac{(C_i(t) - a) \cdot (d - c)}{(b - a)} + c \tag{11}$$

In Equation (11), (a, b) is the chaos map range, i is an index of chaos, and (c, d) are chaos normalized intervals where c equals zero and d is measured by Equation (12).

$$d = \frac{MI - CI}{MI} \times (Max - Min) \tag{12}$$

Such that Max and Min are 20 and $1E-10$, respectively, indicating adaptive intervals. The gravitational constant (G) plays a central role in the standard GSA, determining the magnitude of the gravitational field, as elucidated in Equation (2). Achieving a delicate balance between intensification and diversification phases hinges on maintaining an appropriate equilibrium between G and other parameters. During the initialization phase, a rapid decrease in G facilitates exploration of the search space. Conversely, in the later iterations, G remains constant, favoring the selection of solutions in proximity to the global optimum. Due to its pivotal role in simplifying both exploration and exploitation phases, G is regarded as the primary controlling parameter in our approach.

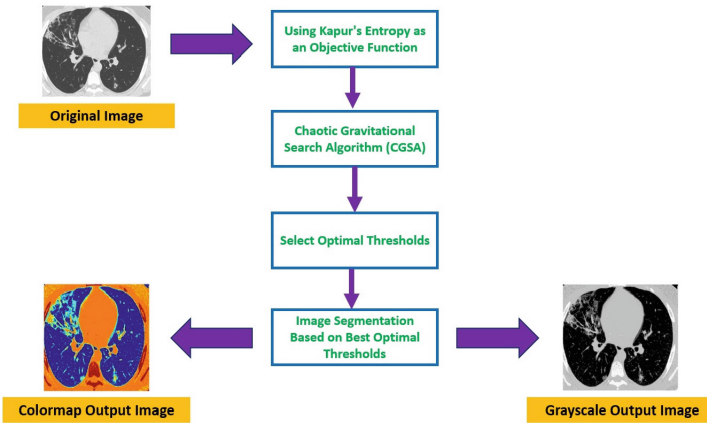


Fig. 2. The figure clearly delineates the segmentation procedure, beginning with the acquisition of the test image. Kapur’s entropy is employed as the objective function to partition the image into regions containing the maximum information content. Subsequently, the parameters of the CGSA algorithm are iteratively adjusted to expedite the optimization process. This results in a final segmented image characterized by high-intensity pixels.

The CGSA integrates the GSA's gravitational constant with the random sequences of the chaotic maps. Thus, Equation (2) is integrated with Equation (11) to obtain chaotic gravitational constant ($G^C(t)$).

$$G^C(t) = C_i^{\text{norm}}(t) + G(t_0)e^{(-\alpha \frac{CI}{MT})} \quad (13)$$

Equation (13) clearly demonstrates that $G^C(t)$ possesses all the essential characteristics required to enhance both the intensification and diversification aspects of the standard GSA. The main advantage of CGSA is taking candidate solutions away from the local minima areas. Besides, it maintains the appropriate balance between exploration and exploitation stages as well as increases the convergence rate. The flowchart of CGSA based multilevel thresholding is shown in Fig. 2.

4 Experimental Results and Discussion

The empirical investigation scrutinized ten distinct variants of the Gravitational Search Algorithm (GSA) augmented with chaos, denoted CGSA1 to CGSA10, in the context of image segmentation, particularly focusing on COVID-19 chest CT scans. These scans, sourced from the Kaggle database² exhibited pixel values ranging from 0 to 255, as depicted in Fig. 3. This dataset includes 1252 CT scans from patients with COVID-19 and 1230 CT scans from patients without the infection, totaling 2482 scans. A comparative evaluation was conducted against a repertoire of heuristic algorithms, including PSO [26], GSA [27], PSOGSA [28], SCA [29], SSA [30], DE [31], BBO [32], and CPSOGSA [33–35], with initial parameters drawn from their respective seminal works. Simulation outcomes were based on 20 searcher agents and 300 iterations per algorithm run, with termination criteria set at 10% identical outcomes. Computational experiments were conducted on a 3.40 GHz i7 Intel processor, utilizing MATLAB R2020a.

The quantitative assessment of the segmented image entails the systematic application of distinct mathematical metrics, namely the PSNR, SSIM, and FSIM. These metrics serve as meticulous instruments for the exhaustive examination and scrutiny of the inherent quality of the segmented output. At its core, PSNR plays a pivotal role in discerning the legitimacy of the segmented image by intricately analyzing the nuanced variations in threshold values across successive iterations within the optimization process. The precise mathematical formulation governing PSNR is rigorously encapsulated in Equation (14).

$$\text{PSNR} = 10 \log_{10} \left(\frac{255^2}{\text{MSE}} \right) \quad (14)$$

$$\text{MSE} = \frac{1}{RC} \sum_{i=1}^R \sum_{j=1}^C (I(i, j) - O(i, j))^2 \quad (15)$$

² <https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset>.

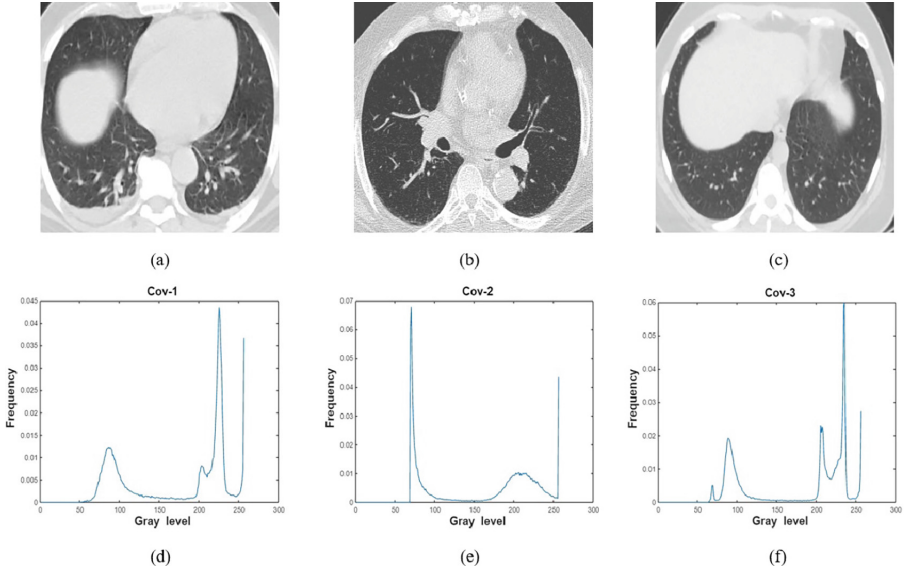


Fig. 3. The grayscale backgrounds in (a)-(c) provide a detailed view of the overall image characteristics, showcasing the texture and structure. Meanwhile, the histograms in (d)-(f) depict pixel intensity distributions, highlighting the contrast and brightness variations across the images.

Equation (15) encapsulates the intricate dimensions denoted by R and C , representing rows and columns within the matrix, respectively. Meanwhile, the symbols I and O delineate the conventional input image and the resulting segmented output image. On a parallel note, SSIM emerges as a nuanced metric, delving into the subtleties of both uniformity and resemblance between the segmented image and a predefined input reference image. A discernibly elevated SSIM value signifies the preeminence of pixels of superior quality within the segmented image. The mathematical equation for SSIM is shown in Equation (16).

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{16}$$

The mean luminance values of the original and segmented images are denoted by μ_x and μ_y , respectively. The standard deviations of the original and segmented images are represented by σ_x^2 and σ_y^2 . Covariance between the original and segmented images is denoted by σ_{xy} , and $\langle c_1, c_2 \rangle$ signify minor constants in the formula. Similarly, FSIM serves as a sophisticated image assessment metric, meticulously evaluating the quality of a segmented image by scrutinizing pixels within the local neighborhood. This comprehensive process involves the analysis of pivotal elements for discerning optimal pixel values and gauging the precision of the

resulting segmented image. The intricate mathematical expression for FSIM is delineated in Equation (17).

$$\text{FSIM}(x) = \frac{\sum_{x \in \Omega} S_L(x) \text{PC}_m(x)}{\sum_{x \in \Omega} \text{PC}_m(x)} \quad (17)$$

Ω signifies the pixel search area and $S_L(x)$ denotes image similarity while PC shows the phase congruency of the image. As the task of image segmentation involves maximizing a certain objective, a heightened FSIM value indicates the adeptness in pinpointing optimal pixels within the solution space, showcasing effective segmentation capabilities.

Furthermore, it's important to carefully analyze the results using statistics. To understand how well the Heuristic Algorithms (HAs) performed, we use a method called the Signed Wilcoxon Rank-sum Test [36]. This test helps us compare the performance of the algorithms. The null hypothesis (H_0) suggests that CGSA doesn't provide the best pixel values for the image, while the alternative hypothesis (H_1) suggests that CGSA does give good pixel values. The P-values are calculated for all the algorithms to see if they support H_1 . If the P-value is less than 0.05, it means H_1 is likely true. If the P-value is close to 1, it means the peer algorithm performs similarly to the best one.

4.1 Experimental Analysis of Medical Images

The CGSA is a powerful heuristic method designed to handle complex problems by navigating through nonlinear problem spaces and finding the best solutions. Its unique hybrid design combines robust exploration and exploitation mechanisms, making it well-equipped to tackle challenging and intricate problem landscapes. To validate its efficacy, CGSA has been applied to chest CT scan database images, serving as a challenge for its ability to determine optimal image thresholds and furnish precise values for key image quality assessment metrics such as PSNR, SSIM, and FSIM. Furthermore, it is of keen interest to scrutinize how CGSA copes with escalating computational demands as image thresholds increase. The comprehensive experimental analysis, focusing on the chest CT scan images, has been documented in Table 2. Noteworthy findings reveal that CGSA versions, alongside DE and PSOGSA, have demonstrated superior performance in identifying optimal thresholds. Additionally, CGSA exhibits commendable results across PSNR, SSIM, and FSIM metrics. In contrast, algorithms such as PSO and SSA exhibit elevated Standard Deviation (SD) and Mean Square Error (MSE) values, indicative of potential outliers within the segmented output. Furthermore, computational intensive methods like SCA, SSA, DE, and BBO necessitate considerable time investments to ascertain optimal pixel configurations within the problem space. Importantly, the results underscore CGSA's effi-

Table 2. Simulation outcomes for CT scan image processing employing classical, hybrid, and contemporary Heuristic Algorithms (HAs)

Method	k	SD	MSE	PSNR	SSIM	FSIM	BV	Run Time(s)	P-Value
GSA	2	0.11	8507.33	8.83	0.52	0.60	15.75	7.5807	0.0020
	3	0.05	4161.37	11.93	0.48	0.66	17.16	11.2454	0.0020
	5	0.23	2914.64	13.48	0.55	0.73	22.52	18.4125	0.0020
	7	1.03	3985.53	12.12	0.49	0.67	27.13	19.6508	0.0020
	10	0.31	2782.01	13.68	0.64	0.69	33.24	28.6723	0.0020
PSO	2	0.29	29370.51	3.45	0.10	0.41	10.45	6.0663	0.0020
	3	0.56	26891.35	3.83	0.16	0.41	14.32	8.3811	0.0020
	5	1.14	11239.81	7.62	0.51	0.61	19.66	16.1352	0.0020
	7	1.12	13700.89	6.76	0.43	0.58	28.19	17.6534	0.0020
	10	0.97	15330.86	6.27	0.39	0.56	32.30	24.5230	0.0020
PSO GSA	2	0.88	21516.36	4.80	0.27	0.50	10.19	5.8473	0.0020
	3	0.89	21781.24	4.75	0.27	0.48	11.54	8.5503	0.0020
	5	0.43	13714.78	6.75	0.42	0.57	21.87	15.2910	0.0020
	7	0.87	14296.90	6.57	0.41	0.57	21.28	16.2218	0.0020
	10	0.49	2225.70	14.65	0.77	0.76	36.03	23.1176	0.0020
CPSO GSA	2	0.35	27806.03	3.68	0.13	0.41	09.71	6.2928	0.0020
	3	0.61	2125.45	4.85	0.28	0.44	14.95	7.9160	0.0020
	5	0.74	24827.08	4.18	0.20	0.41	22.15	13.3482	0.0020
	7	0.29	7117.84	9.60	0.65	0.68	24.76	16.3624	0.0020
	10	0.82	6980.08	9.69	0.66	0.68	34.40	25.1694	0.0020
BBO	2	0.72	11705.57	7.44	0.50	0.60	15.83	8.1842	0.0020
	3	1.20	2236.27	14.63	0.58	0.68	19.59	10.9221	0.0098
	5	1.07	1725.86	15.76	0.69	0.70	22.98	17.6534	0.0020
	7	0.53	436.06	21.73	0.75	0.74	30.21	20.9856	0.0020
	10	1.62	1780.38	15.62	0.67	0.70	37.31	29.4807	0.0020
DE	2	1.63	3902.63	12.21	0.52	0.56	16.62	8.3519	0.0059
	3	2.63	8708.85	8.73	0.43	0.63	18.79	11.2949	0.0020
	5	1.43	3831.97	12.29	0.40	0.61	26.34	17.1391	0.0098
	7	1.47	575.02	20.53	0.79	0.77	32.57	21.6220	0.0020
	10	1.92	482.32	21.29	0.79	0.77	39.34	33.9167	0.0020
SCA	2	1.61	4489.83	11.60	0.50	0.69	16.36	8.2530	0.0137
	3	1.50	1708.80	15.80	0.65	0.66	19.10	11.2949	0.0050
	5	2.04	1138.83	17.56	0.70	0.69	25.82	17.4521	0.0020
	7	2.61	0634.17	20.10	0.80	0.78	33.70	21.1952	0.0020
	10	3.16	0305.68	23.27	0.83	0.83	39.97	30.2570	0.0020
SSA	2	5.25	33672.14	2.85	0.00	0.41	13.36	6.9575	0.0020
	3	3.89	31154.95	3.19	0.01	0.50	17.54	11.2088	0.0020
	5	3.67	4968.06	11.16	0.45	0.62	26.34	22.5873	0.0020
	7	3.87	4968.06	11.16	0.45	0.62	26.34	22.5873	0.0020
	10	12.58	31154.95	3.19	0.01	0.50	28.23	30.2576	0.0020
CGSA1	2	1.91	3412.81	12.79	0.53	0.71	16.90	2E-06	0.3594
	3	0.61	2487.68	14.17	0.56	0.64	21.23	2E-06	00001
	5	1.32	710.19	19.61	0.79	0.79	28.65	2E-06	00001
	7	2.70	484.58	21.27	0.81	0.80	37.67	2E-06	0.4316
	10	2.41	839.10	18.89	0.76	0.78	46.90	2E-06	0.4316

(continued)

Table 2. (continued)

Method	k	SD	MSE	PSNR	SSIM	FSIM	BV	Run Time(s)	P-Value
CGSA2	2	0.39	5934.26	10.39	0.50	0.68	16.13	1E-06	0.0371
	3	0.63	2902.26	13.50	0.56	0.73	21.24	2E-06	0.1602
	5	0.97	968.65	18.26	0.76	0.77	28.68	2E-06	0.3750
	7	2.30	938.89	18.40	0.77	0.78	36.53	2E-06	0.0488
	10	2.79	480.32	21.31	0.78	0.81	49.61	1E-06	0.1934
CGSA3	2	0.26	3070.81	13.25	0.54	0.70	16.13	2E-06	0.0195
	3	0.35	2585.22	14.00	0.55	0.69	19.89	2E-06	0.0840
	5	1.39	2529.68	14.10	0.58	0.73	27.17	2E-06	0.0039
	7	2.59	2441.15	14.30	0.59	0.73	33.56	2E-06	0.0098
	10	3.70	1656.01	15.96	0.68	0.73	43.79	2E-06	0.0039
CGSA4	2	1.67	5756.22	10.52	0.50	0.68	16.90	1E-06	0.0645
	3	0.83	1753.62	15.69	0.65	0.67	21.23	2E-06	0.0195
	5	1.63	634.83	20.10	0.80	0.79	28.63	2E-06	0.6250
	7	2.18	818.63	18.99	0.77	0.78	37.94	2E-06	00001
	10	2.52	603.13	20.32	0.77	0.77	49.26	3E-06	00001
CGSA5	2	0.41	2508.22	12.67	0.52	0.71	16.92	2E-06	0.3223
	3	0.26	5962.38	10.37	0.43	0.52	21.22	1E-06	0.1309
	5	1.58	1314.90	16.94	0.72	0.75	28.62	1E-06	0.3750
	7	2.44	518.73	20.98	0.81	0.81	35.33	2E-06	0.1934
	10	1.88	795.33	19.12	0.78	0.78	47.12	2E-06	0.1602
CGSA6	2	0.29	6658.29	9.87	0.41	0.52	16.92	2E-06	0.2324
	3	0.39	1423.29	16.59	0.70	0.73	21.24	2E-06	0.4316
	5	0.01	1348.23	16.83	0.71	0.73	28.65	2E-06	0.3223
	7	2.01	1035.25	17.98	0.74	0.77	37.25	2E-06	00001
	10	2.28	242.99	24.27	0.82	0.81	47.77	2E-06	0.2324
CGSA7	2	0.39	3424.54	12.78	0.52	0.71	16.89	2E-06	0.3224
	3	0.88	2988.46	13.37	0.51	0.63	21.21	2E-06	0.0645
	5	0.98	1205.61	17.31	0.71	0.76	28.64	1E-06	00001
	7	2.25	607.20	20.29	0.78	0.79	35.37	2E-06	0.0195
	10	3.22	469.83	21.41	0.77	0.79	49.02	3E-06	0.8457
CGSA8	2	0.64	3520.08	12.66	0.52	0.71	16.90	2E-06	00001
	3	0.32	2494.82	14.16	0.56	0.70	19.81	1E-06	0.1309
	5	1.43	682.08	19.79	0.78	0.79	28.61	2E-06	0.7695
	7	2.15	1186.98	17.38	0.72	0.73	35.36	2E-06	0.4922
	10	2.48	489.08	21.23	0.78	0.76	49.28	1E-06	0.9219
CGSA9	2	0.88	3450.24	12.75	0.52	0.71	16.93	1E-06	0.6953
	3	0.66	1663.81	15.91	0.67	0.72	21.25	3E-06	0.4316
	5	0.83	774.26	19.24	0.78	0.74	26.98	2E-06	0.0273
	7	2.28	793.21	19.13	0.76	0.76	36.81	2E-06	0.7695
	10	2.91	295.36	23.42	0.78	0.82	47.18	2E-06	0.9219
CGSA10	2	0.29	5130.79	11.02	0.43	0.51	16.90	8.8356	0.0840
	3	0.44	2679.72	13.84	0.56	0.70	21.23	9.8946	0.1602
	5	0.53	4538.25	11.56	0.47	0.52	28.64	17.0167	0.1602
	7	1.47	2255.33	14.59	0.59	0.73	38.01	19.5191	0.8457
	10	3.46	1063.59	17.86	0.76	0.77	49.16	27.8105	0.6250

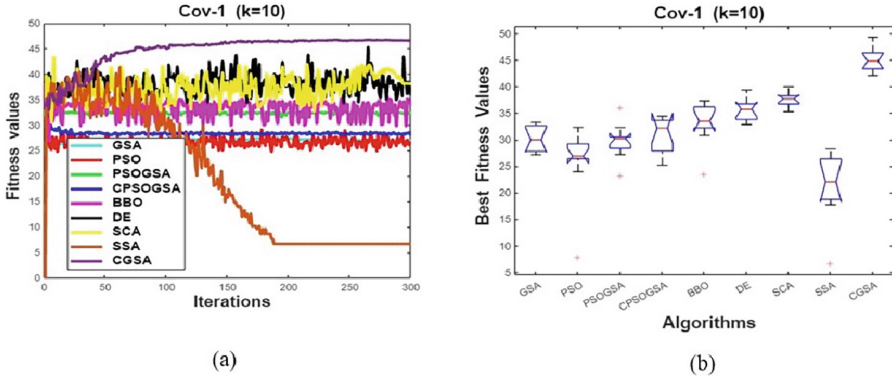


Fig. 4. The convergence curves in (a) show that CGSA achieves the best performance with higher values for Kapur’s objective function. Furthermore, the boxplots in (b) reveal that the CGSA has highest values for the lower and upper quartiles indicating a greater amount of information content.

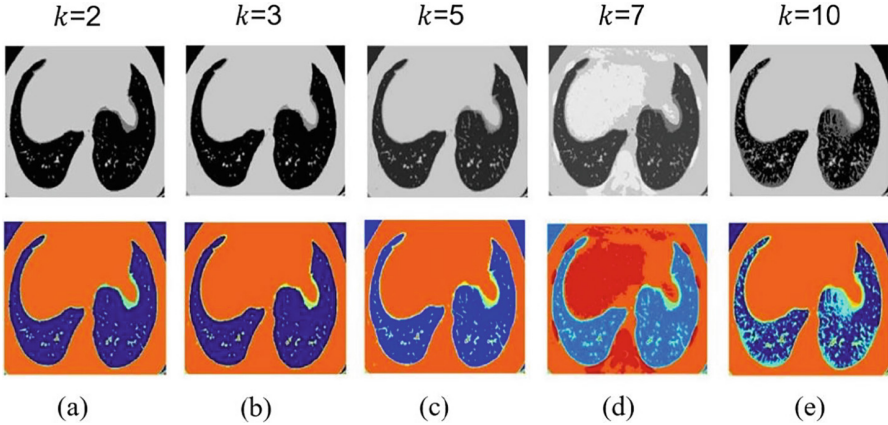


Fig. 5. The images (a)-(e) illustrates the segmented output acquired through the CGSA across a spectrum of $k = 2,3,5,7,10$ values. Colormap representations highlight the distribution of pixel intensities, while the histogram curves portray the frequency distribution of pixel values within the segmented images.

ciency, evidenced by its reduced CPU time requirements for image segmentation compared to alternative methodologies.

In a similar manner, we assess the local exploitation capabilities of the competitive methods using convergence curves, depicted in Fig. 4. It is evident that CGSA, DE, and SCA exhibit swift convergence rates, indicating their adeptness in navigating complex search spaces and identifying optimal pixel configurations. Notably, CGSA stands out with its faster convergence rate compared to its counterparts, highlighting its effectiveness in achieving feasible outcomes.

Furthermore, the box plot analysis underscores CGSA's superior performance, as evidenced by its favorable values for the Kapur's objective function. Additionally, CGSA displays fewer outliers relative to its peers, affirming its proficiency in image segmentation.

Furthermore, in Fig. 5, we can clearly observe the impact of the Coronavirus on the lungs through the segmented images generated by CGSA. These images provide a visual representation of the severity of lung involvement. Notably, the colormap visualization highlights areas of consolidation in the lungs, with the right lung appearing to be more heavily affected, indicating significant infection. In the generated segmented images, different colors represent various lung conditions: yellowish areas depict bronchi filled with fluid or pus, light blue indicates fibrosis or the thickening of lung tissue, and blue represents healthy lung tissue. This visual depiction offers valuable insights into the pathological changes caused by the Coronavirus in the lungs.

5 Conclusion and Future Directions

This investigation utilizes the Chaos theory-based Gravitational Search Algorithm (CGSA) to perform the multilevel thresholding of grayscale images. Through the integration of Kapur's entropy scheme with CGSA, the algorithm can effectively pinpoint optimal pixels within the search space. The performance evaluation of this approach is conducted using chest CT scan images, employing metrics like SSIM, FSIM, PSNR, MSE, and CPU time. Furthermore, the study undertakes a comparative analysis between the outcomes of CGSA and those of other competitive algorithms.

The findings emphasize how CGSA outperforms other methods, showing its exceptional ability to find the optimal pixel values across various intensity levels without needing high computing power. Additionally, achieving the best PSNR, SSIM, and FSIM values indicates that the images it produces are well-balanced, high-quality, and consistent. Notably, CGSA is effective at identifying both large and small irregular areas in CT scan images, which could improve the diagnosis of COVID-19 patients. This highlights CGSA's potential usefulness in practical applications involving image processing. Furthermore, Differential Evolution (DE) and Sine Cosine Algorithm (SCA) also yield favorable outcomes for image thresholds. Future research avenues may involve the application of CGSA to segment color images instead of grayscale benchmarks and the exploration of alternative fitness functions, such as Otsu's variance scheme and Renyi entropy methods, for CGSA-based image segmentation.

References

1. Luo, S., et al.: Meta-seg: a survey of meta-learning for image segmentation. *Pattern Recogn.* **126**, 108586 (2022)
2. Golzari Oskouei, A., et al.: CGFFCM: cluster-weight and group-local feature-weight learning in fuzzy C-means clustering algorithm for color image segmentation. *Appl. Soft Comput.* **113**, 108005 (2021)

3. Fournel, J., et al.: Medical image segmentation automatic quality control: a multi-dimensional approach. *Med. Image Anal.* **74**, 102213 (2021)
4. Chakraborty, S., et al.: Biomedical image segmentation using fuzzy multilevel soft thresholding system coupled modified cuckoo search. *Biomed. Signal Process. Control* **72**, 103324 (2022)
5. Houssein, E.H., et al.: An improved opposition-based marine predators algorithm for global optimization and multilevel thresholding image segmentation. *Knowl.-Based Syst.* **229**, 107348 (2021)
6. Zhao, D., et al.: Ant colony optimization with horizontal and vertical crossover search: fundamental visions for multi-threshold image segmentation. *Expert Syst. Appl.* **167**, 114122 (2021)
7. Cao, X., Li, et al.: A robust parameter-free thresholding method for image segmentation. *IEEE Access* **7**, 3448–3458 (2019)
8. Kotte, S., et al.: An efficient approach for optimal multilevel thresholding selection for gray scale images based on improved differential search algorithm. *Ain Shams Eng. J.* **9**(4), 1043–1067 (2018)
9. Khalilpourazari, S., et al.: Modeling and optimization of multi-item multi-constrained EOQ model for growing items. *Knowl.-Based Syst.* **164**, 150–162 (2019)
10. Huang, C., et al.: Clinical features of patients infected with 2019 novel coronavirus in Wuhan. *China Lancet* **395**(10223), 497–506 (2020)
11. World Health Organization: laboratory testing for coronavirus disease 2019 (COVID-19) in suspected human cases, pp. 1-7 (2020)
12. Toyoshima, Y., et al.: SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *J. Hum. Genet.* **65**(12), 1075–1082 (2020)
13. Munusamy, H., et al.: FractalCovNet architecture for COVID-19 Chest X-ray image classification and CT-scan image segmentation. *Biocybern. Biomed. Eng.* **41**(3), 1025–1038 (2021)
14. Singh, P., et al.: A quantum-clustering optimization method for COVID-19 CT scan image segmentation. *Expert Syst. Appl.* **185**, 115637 (2021)
15. Su, H., et al.: Multilevel threshold image segmentation for COVID-19 chest radiography: a framework using horizontal and vertical multiverse optimization. *Comput. Biol. Med.* **146**, 105618 (2022)
16. Liu, L., et al.: Ant colony optimization with Cauchy and greedy Levy mutations for multilevel COVID-19 X-ray image segmentation. *Comput. Biol. Med.* **136**, 104609 (2021)
17. Sahoo, S.K., et al.: Self-adaptive moth flame optimizer combined with crossover operator and Fibonacci search strategy for COVID-19 CT image segmentation. *Expert Syst. Appl.* **227**, 120367 (2023)
18. Zhao, S., et al.: Boosted crow search algorithm for handling multi-threshold image problems with application to X-ray images of COVID-19. *Expert Syst. Appl.* **213**, 119095 (2023)
19. Houssein, E.H., et al.: An efficient multi-thresholding based COVID-19 CT images segmentation approach using an improved equilibrium optimizer. *Biomed. Signal Process. Control* **73**, 103401 (2022)
20. Qi, A., et al.: Directional mutation and crossover boosted ant colony optimization with application to COVID-19 X-ray image segmentation. *Comput. Biol. Med.* **148**, 105810 (2022)
21. Mirjalili, S., et al.: Chaotic gravitational constants for the gravitational search algorithm. *Appl. Soft Comput.* **53**, 407–419 (2017)

22. Alatas, B.: Chaotic bee colony algorithms for global numerical optimization. *Expert Syst. Appl.* **37**(8), 5682–5687 (2010)
23. Li, C., et al.: Parameters identification of chaotic system by chaotic gravitational search algorithm. *Chaos, Solitons Fract.* **45**(4), 539–547 (2012)
24. Mingjun, J., et al.: Application of chaos in simulated annealing. *Chaos, Solitons Fract.* **21**(4), 933–941 (2004)
25. Gandomi, A.H., et al.: Firefly algorithm with chaos. *Commun. Nonlinear Sci. Numer. Simul.* **18**(1), 89–98 (2013)
26. Kennedy, J., et al.: Particle swarm optimization. In: *Proceedings of ICNN'95-International Conference on Neural Networks*, pp. 1942–1948. IEEE, Australia (1995)
27. Rashedi, E., et al.: GSA: a gravitational search algorithm. *Inf. Sci.* **179**(13), 2232–2248 (2009)
28. Mirjalili, S., et al.: A new hybrid PSOGSA algorithm for function optimization. In: *Proceedings of the 2010 International Conference on Computer and Information Application*, pp. 374–377. IEEE, China (2010)
29. Mirjalili, S.: SCA: a sine cosine algorithm for solving optimization problems. *Knowl.-Based Syst.* **96**, 120–133 (2016)
30. Mirjalili, S., et al.: Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. *Adv. Eng. Softw.* **114**, 163–191 (2017)
31. Storn, R., et al.: Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.* **11**, 341–359 (1997)
32. Simon, D.: Biogeography-based optimization. *IEEE Trans. Evol. Comput.* **12**(6), 702–713 (2008)
33. Rather, S. A., et al.: Hybridization of constriction coefficient based particle swarm optimization and gravitational search algorithm for function optimization. In: *Proceedings of the International Conference on Advances in Electronics, Electrical & Computational Intelligence (ICAEEC)*, pp. 1–10. Elsevier, India (2020)
34. Rather, S.A., et al.: Constriction coefficient based particle swarm optimization and gravitational search algorithm for multilevel image thresholding. *Expert. Syst.* **38**(7), e12717 (2021)
35. Rather, S.A., et al.: A hybrid constriction coefficient-based particle swarm optimization and gravitational search algorithm for training multi-layer perceptron. *Int. J. Intell. Comput. Cybern.* **13**(2), 129–165 (2020)
36. Wilcoxon, F.: Individual comparisons by ranking methods. *Biometrics Bull.* **1**(6), 80–83 (1945)



Integrated Grading Framework for Histopathological Breast Cancer: Multi-level Vision Transformers, Textural Features, and Fusion Probability Network

Hossam Magdy Balaha¹(✉), Khadiga M. Ali², Ali Mahmoud¹,
Mohammed Ghazal³, and Ayman El-Baz¹

¹ Department of Bioengineering, J.B. Speed School of Engineering,
University of Louisville, Louisville, KY, USA
hmbala01@louisville.edu

² Pathology Department, Faculty of Medicine, Mansoura University,
Mansoura, Egypt

³ Electrical, Computer, and Biomedical Engineering Department, Abu Dhabi
University, Abu Dhabi, UAE

Abstract. Breast cancer (BC) remains a significant global health concern, necessitating accurate and efficient diagnostic approaches. In this study, we propose a comprehensive framework that integrates feature extraction, selection, and classification using Support Vector Machines (SVM) along with hyperparameter optimization. Additionally, we employ multi-level Vision Transformers (ViTs) for patch classification, aiming to capture both local and global information from histopathological slides. Moreover, we introduce a Fusion Probability Network (FPN) to combine the outputs of SVM and ViTs. Through this multi-faceted approach, we aim to improve diagnostic performance and contribute to more effective BC diagnosis. Sensitivity analyses and ablation studies across various sample sizes confirm the framework's effectiveness. Results show high accuracy (up to 96.50%), precision (up to 93.33%), recall (up to 93%), specificity (up to 97.67%), F1 score (up to 92.99%), and balanced accuracy (up to 95.33%). Ablation studies highlight the significance of the feature extraction pipeline in enhancing the framework's effectiveness and robustness, as well as its adaptability to diverse patch morphologies. Overall, our study offers promising avenues for improving BC grading, with potential implications for enhancing clinical decision-making and patient outcomes.

Keywords: Breast Cancer (BC) · Fusion Probability Network (FPN) · Multi-level Vision Transformers · Texture Analysis

1 Introduction

Breast cancer (BC) is a pervasive global health issue affecting millions of women worldwide [2]. Traditional imaging approaches have limitations, including subjective interpretation, false positives, and difficulties in distinguishing between

benign and malignant lesions [6, 10, 20]. These limitations underscore the urgent need for innovative artificial intelligence (AI) approaches to enhance the efficiency of BC diagnosis and treatment planning [12, 29].

Recent advancements in AI have revolutionized BC diagnosis. Various approaches have emerged, aiming for objective and precise diagnosis using AI [21, 33]. For instance, Wang et al. [26] proposed DeepGrade, focusing on histological grading of breast tumors, specifically re-stratifying NHG 2 cases. Joseph et al. [4] utilized handcrafted feature techniques to extract texture, shape, and color features, employing a deep neural network for classification. Additionally, Wetstein et al. [28] employed a deep learning-based BC grading model to differentiate between low/intermediate and high-grade tumors, while also predicting nuclear, mitotic, and tubular grade characteristics. *Despite these advancements, key gaps remain, prompting the development of our framework. Prior studies often focused solely on histological grading or image feature extraction, lacking integration into a cohesive framework. Furthermore, there is a gap in feature selection and integration from histopathological slides, with few standardized methods available. Additionally, existing approaches may lack scalability and robustness across different magnification levels of histopathological slides.*

To address these gaps, a novel framework for enhancing BC diagnosis accuracy and efficiency is proposed. It involves patch extraction from histopathological slides at three magnification levels to capture global and localized tissue features. A feature extraction pipeline extracts statistical, shape, and texture features. Non-significant features are filtered out using one-way analysis of variance (ANOVA). Classification is conducted using a support vector machine (SVM), with hyperparameters tuned using the tree-structured Parzen estimator (TPE). Simultaneously, patches from all magnification levels are fed into vision transformers (ViT). Probabilities from both the SVM and ViT models are input to a deep neural fusion probability network (FPN). This network integrates probabilities from each model to categorize the cellularity condition.

2 Materials

The dataset used in this study is the Post-NAT-BRCA dataset [13], a comprehensive collection of high-resolution microscopic images and clinical data from breast resections in patients with residual invasive breast cancer following neoadjuvant therapy (NAT). This dataset includes annotations for tumor cellularity, which is a crucial parameter for calculating the Residual Cancer Burden Index (RCBi), a tool used to assess response to NAT. The images are annotated with various cell types, including malignant, healthy, and lymphocyte cells, aiding in the development of cell segmentation algorithms [19].

The dataset comprises 96 Whole Slide Images (WSIs) stained with Hematoxylin and Eosin (H&E), extracted from 54 patients. We systematically categorized annotated patches into four separate groups based on their cellularity. The categories include low grade, moderate grade, high grade, and normal cellularity. Patches of size 64×64 pixels were extracted from each WSI using the

corresponding annotation file, ensuring no overlap between patches. In total, more than 100,000 patches were extracted.

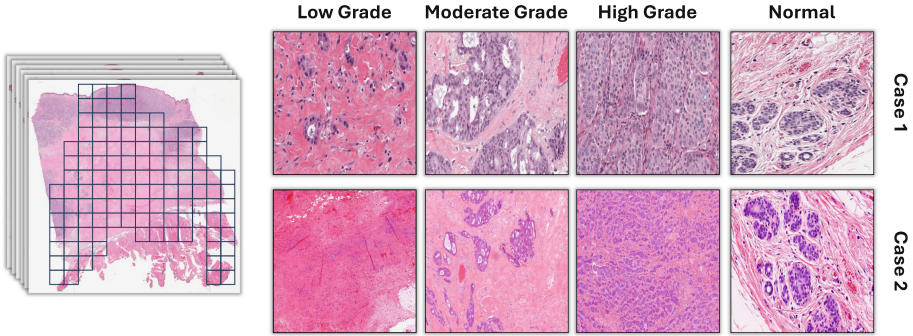


Fig. 1. Samples from the utilized dataset in the current study.

3 Methods

The proposed framework, visualized in Fig. 2, employs histopathological slides to diagnose BC. It begins by extracting patches from three magnification levels. Next, a feature extraction pipeline extracts statistical, shape, and texture features from the top magnification level. Subsequently, the p-value is computed using ANOVA to filter out non-significant features. Following this, feature selection is performed to identify the most promising features among the significant ones. Classification is carried out using SVM, and hyperparameters are tuned using the TPE. Concurrently, patches from the three magnification levels are input to ViT. The probabilities from the four models (one for numerical features and three for the magnification levels) are then input to the FPN. The FPN combines the probabilities from each model to make a final decision.

3.1 Numerical Features Extraction from Histopathological Patches

Aim and Hypothesis 1: This study aims to comprehensively capture the diverse characteristics of histopathological BC slides by integrating various features, including shape features (SF), texture descriptors such as Gray-Level Co-occurrence Matrix (GLCM) and Gray-Level Run Length Matrix (GLRLM), Statistical Feature Matrix (SFM), Local Binary Patterns (LBP), and global shape descriptors known as Hu Moments.

We hypothesize that this approach will significantly enhance the classification of BC from histopathological patches, leading to higher accuracy and reliability in distinguishing between different tissue types and pathological conditions associated with BC.

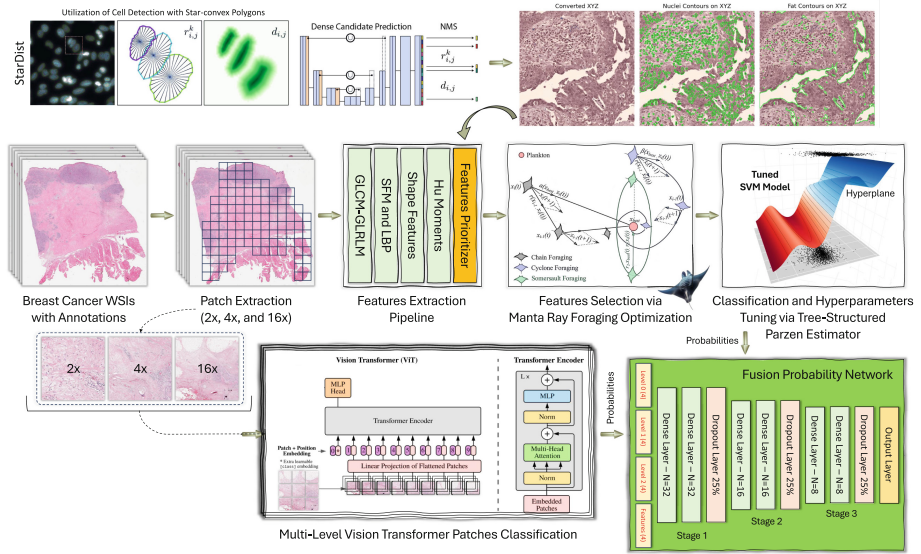


Fig. 2. The proposed comprehensive framework for breast cancer diagnosis from histopathological slides.

Theory and Implementation: The patches undergo conversion from RGB to XYZ color space, offering advantages over traditional RGB mode. XYZ mode ensures standardized color representation, aiding in detecting pathological features. It maintains consistency for human perception, ensuring faithful color reproduction and interpretation. Additionally, separating luminance from chromaticity allows precise adjustments of brightness and contrast, crucial for highlighting subtle tissue nuances while preserving color fidelity [15].

After patch extraction, nuclei, fat, and other tissue components are segmented as presented in Fig. 3. Nuclei segmentation utilizes StarDist [27] to accurately delineating nuclei boundaries. Fat segmentation combines blurring and thresholding techniques to isolate adipose tissue regions within histopathological samples. Lastly, other tissue components are segmented by identifying regions unrelated to nuclei or fats.

Statistical, shape, and texture features are extracted from different tissue types, including shape, GLCM, GLRLM, LBP, SFM, and Hu moments features. Shape features accurately characterize geometric properties, while GLCM captures subtle textural patterns by quantifying spatial relationships between pixel intensities, aiding in detecting irregular cell arrangements [8]. GLRLM quantifies the length and frequency of homogeneous pixel runs, providing insight into tissue texture [17]. SFM integrates shape and texture information, enhancing discrimination between tissue types [16]. LBP captures local pixel intensity patterns, aiding in detecting cancerous regions [1]. Hu Moments capture global shape features, aiding in differentiation between tissue types and pathological conditions [23].

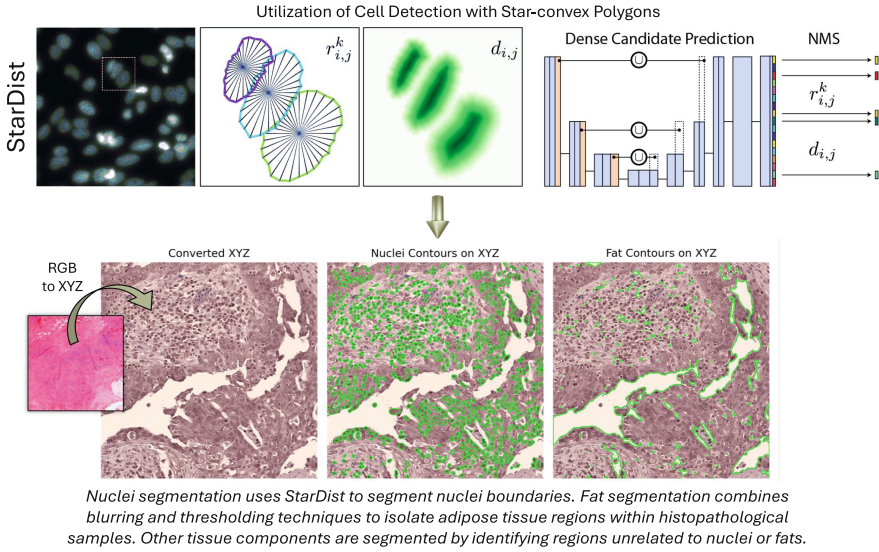


Fig. 3. Visualization of the steps to extract the nuclei, fat, and other tissue components.

Significance and Elimination: After extracting features from the histopathology patches, a set of 70 features was obtained. To enhance the robustness and scientific rigor of our analysis, we conducted an ANOVA test to assess the significance of each feature. Features with a significance $p < 0.05$ were retained, as they demonstrate a higher degree of association with the underlying biological phenomena we are investigating. The remaining number of features is 57, reflecting the subset of features that exhibited statistically significant differences.

3.2 Numerical Features Selection via Manta Ray Foraging Optimization (MRFO)

Aim and Hypothesis 2: Employing Manta Ray Foraging Optimization (MRFO) for feature selection post-extraction [32], we aimed to identify the most informative features crucial for BC classification from histopathological patches. We hypothesized that MRFO could efficiently select a subset of features, reducing computational complexity, enhancing interpretability, and improving generalization performance on unseen data.

Theory and Implementation: The foraging behavior of manta rays is simulated to explore the feature space and select promising feature subsets as summarized in Algorithm 11. Let \mathbb{X} represent the set of features extracted from histopathological patches, where each feature is denoted by $x_i \in \mathbb{X}$. A population of manta ray agents (\mathbb{M}) (sized $N = |\mathbb{M}|$) is initialized where each agent

($m_j \in \mathbb{M}$) representing a potential feature subset. Each manta ray agent (m_j) is characterized by a binary vector, b_{m_j} , of length $|\mathbb{X}|$, where $b_{m_j}[i] = 1$ if feature (x_i) is selected and $b_{m_j}[i] = 0$ otherwise. The quality of each feature subset is evaluated using the fitness function (\mathcal{F}) that measures the classification performance using a classification algorithm (SVM in our study) to the subset of selected features represented by m_j . Let $\mathcal{F}(b_{m_j})$ denote the fitness value of agent indexed at j , the goal is to select a subset of features (\mathbb{S}) that maximizes the classification performance: $\mathbb{S} = \mathbb{M}[\operatorname{argmax}_j (\mathcal{F}(b_{m_j}))]$ [5].

Algorithm 1: Manta Ray Foraging for Feature Selection

Input: \mathbb{X} : Set of extracted features
Output: \mathbb{S} : Selected subset of features

- 1 Initialize population \mathbb{M}
- 2 **for** each agent $m_j \in \mathbb{M}$ **do**
- 3 Initialize binary vector b_{m_j}
- 4 **end**
- 5 **for** each iteration **do**
- 6 **for** each agent $m_j \in \mathbb{M}$ **do**
- 7 Evaluate fitness $\mathcal{F}(b_{m_j})$ using SVM classification performance
- 8 **end**
- 9 Update agent positions to simulate foraging behavior
- 10 **end**
- 11 $\mathbb{S} \leftarrow \mathbb{M}[\operatorname{argmax}_j (\mathcal{F}(b_{m_j}))]$

3.3 Numerical Features Classification and Hyperparameters Optimization via Tree-Structured Parzen Estimator

Aim and Hypothesis 3: Our goal is to optimize the BC classification model by tuning hyperparameters using the TPE algorithm [14, 18], and assessing the effectiveness of SVM in classifying BC from histopathological patches post-tuning. *We hypothesize that systematic exploration of the hyperparameter space will identify optimal configurations for SVM, enhancing classification accuracy. We anticipate that SVM’s capability to find optimal hyperplanes in high-dimensional feature spaces will enable effective discrimination between tissue types associated with BC, improving diagnostic performance.*

Theory and Implementation: The TPE algorithm iteratively explores the hyperparameter space to identify promising hyperparameters (\mathbb{H}) as summarized in Algorithm 12. It starts with initializing a population \mathbb{P} of hyperparameter configurations sampled from the search space. Each configuration ($p_i \in \mathbb{P}$) is evaluated using the model (\mathcal{F}) on the testing dataset, computing a performance metric $Q(p_i)$, such as accuracy.

Probabilistic models are constructed to capture the relationship between hyperparameters and model performance, with PDFs $p(x|y = 1)$ for good configurations and $p(x|y = 0)$ for bad ones. Configurations maximizing the probability ratio $\frac{p(x|y=0)}{p(x|y=1)}$ are selected. The models are updated based on evaluated configu-

rations, refining the PDFs to guide the search towards promising regions. This process repeats for a predefined number of iterations (\bar{T}).

Let \mathbb{W} represent the feature matrix containing extracted features, and y the corresponding class labels. \mathbb{W} is scaled to ensure all features have similar scales, preventing dominance during optimization. The SVM model is trained using preprocessed feature matrix \mathbb{W} and class labels y , aiming to find the optimal hyperplane that maximally separates data points of different classes while maximizing the margin.

Mathematically, this is expressed as: $\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \times \sum_{i=1}^N \eta_i$ subject to $y_i \times (\mathbf{w}^T \times \mathbf{x}_i + b) \geq 1 - \eta_i$ and $\eta_i \geq 0$, where \mathbf{w} represents the weight vector, b the bias term, C the regularization parameter, η_i slack variables, and N the number of samples.

Algorithm 2: TPE Algorithm for Hyperparameter Optimization

Input: \mathbb{H} : Hyperparameter space, \bar{T} : Number of iterations

Output: Optimal hyperparameters

- 1 Initialize population \mathbb{P} of hyperparameter configurations
 - 2 **for** each configuration $p_i \in \mathbb{P}$ **do**
 - 3 | Evaluate performance $Q(p_i)$ using model \mathcal{F} on testing dataset
 - 4 **end**
 - 5 Construct PDFs $p(x|y = 1)$ for good configurations and $p(x|y = 0)$ for bad configurations
 - 6 **for** each iteration $t = 1$ to \bar{T} **do**
 - 7 | Select configurations maximizing $\frac{p(x|y=0)}{p(x|y=1)}$
 - 8 | Evaluate selected configurations p_i and update performance $Q(p_i)$ using $Q(p_i) = \mathcal{F}(p_i | \text{testing dataset})$
 - 9 | Update PDFs $p(x|y = 1)$ and $p(x|y = 0)$ based on evaluated configurations using:
 - 10 |
$$p(x|y = 1) = \frac{\sum_{p_i \in \mathbb{P}, Q(p_i) > Q_{\text{threshold}}} \delta(x - p_i)}{\sum_{p_i \in \mathbb{P}} \delta(x - p_i)}$$
 - 11 |
$$p(x|y = 0) = \frac{\sum_{p_i \in \mathbb{P}, Q(p_i) \leq Q_{\text{threshold}}} \delta(x - p_i)}{\sum_{p_i \in \mathbb{P}} \delta(x - p_i)}$$
 - 12 **end**
-

3.4 Multi-level Vision Transformers Classification: Patches

Aim and Hypothesis 4: This study explores the utilization of multi-level Vision Transformers (ViTs) for image patch classification [9, 31]. The aim is to discriminate between tissue structures effectively using ViTs. *Our hypothesis suggests that multi-level ViTs can capture both local and global information from image patches, facilitating accurate BC classification from histopathological slides. We anticipate that the self-attention mechanism in ViTs will enable discernment of subtle malignancy indicators, enhancing diagnostic performance.*

Theory and Implementation: The need for multi-level ViTs in image patch classification stems from the complexity and heterogeneity of histopathological images. Traditional CNNs may struggle to capture long-range dependencies and contextual information crucial for accurate classification. ViTs address this need

by leveraging self-attention mechanisms to capture global context and relationships between different parts of the image patch.

Histopathological slides are partitioned into smaller image patches, denoted by I . Each patch I_i is processed by the multi-level ViT to extract hierarchical features, represented by F_i . These features are organized hierarchically across multiple levels, capturing both local and global information. Let $F_i(l)$ denote the features at level l for patch I_i , where l ranges from 1 to L , the total number of levels. The self-attention mechanism in ViT computes Query (Q), Key (K), and Value (V) vectors, where $Q_i = x_i \times W_Q$, $K_i = x_i \times W_K$, and $V_i = x_i \times W_V$ respectively.

These vectors are used to compute attention scores: $A = \text{SoftMax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right)$, which weight the values ($\text{Attention}(Q, K, V) \times V = A_i \times V$), where d_k denotes the dimension of the key vectors, and N is the number of patches. The hierarchical features $F_i(l)$ are input into a classification head to predict the probability of malignancy for each image patch. Let p_i denote the predicted probability of malignancy for patch I_i . These probabilities are compared with ground truth labels y_i to compute the loss function L . The objective is to minimize L to improve the model's classification performance.

3.5 Fusion Probability Network (FPN)

Aim and Hypothesis 5: This study aims to enhance overall classification accuracy by employing the Fusion Probability Network (FPN) to combine the probabilities from multi-level ViT and SVM models for BC classification [11]. The FPN integrates predictions from individual models, leveraging their complementary strengths to enhance overall performance. *We hypothesize that integrating these models will improve BC diagnosis classification performance. We anticipate that FPN's ability to effectively combine diverse model strengths will lead to more robust and reliable predictions.*

The need for FPN arises from effectively leveraging multiple models for BC classification. Unlike majority voting, a simple ensemble method, FPN considers the confidence or uncertainty associated with each model's prediction. It combines probabilities from different models in a weighted or attention-based manner, ensuring a more sophisticated approach that fully exploits the strengths of individual models, enhancing classification accuracy by considering both predictions and their reliability.

In using FPN to combine the probabilities of the multi-level ViT and SVM models, several steps are involved (see Algorithm 13). First, both the multi-level ViT and SVM model generate probabilistic outputs for each image patch in the dataset. These probabilities, denoted as P_{ViT} and P_{SVM} respectively, represent the likelihood of each image patch belonging to the positive class (i.e., BC). Next, the probabilities generated by the ViT and SVM models are concatenated to create a combined feature vector F_{FPN} . This combined feature vector aggregates the probabilistic outputs of both models and serves as input to the Fusion Probability Network.

Mathematically, this concatenation can be represented as $F_{\text{FPN}} = [P_{\text{ViT}}, P_{\text{SVM}}]$. The FPN is a neural network comprising multiple layers of neurons. It takes the combined feature vector F_{FPN} as input and learns to fuse the probabilistic outputs of the ViT and SVM model to generate a final probability P_{FPN} for each image patch. The network’s architecture and parameters are optimized during training using a labeled dataset, where the ground truth labels are used to compute the loss function \mathcal{L} .

The discrepancy between the fused probability estimates $P_{\text{FPN}}(I_i)$ and ground truth labels y_i can be expressed as: $\min \sum_{i=1}^N (\mathcal{L}(P_{\text{FPN}}(I_i), y_i))$. The decision from the FPN can be expressed as: $\underset{c}{\operatorname{argmax}} (P_{\text{FPN},c}(I_i))$ where I_i represents the i -th image patch, $P_{\text{FPN},c}(I_i)$ denotes the probability estimate generated by the FPN for class c for the i -th image patch, and argmax_c denotes the class c that maximizes the probability estimate, indicating the predicted class for the image patch.

Algorithm 3: FPN for Combining Probabilities from ViT and SVM Models

Input: Image patches, ground truth labels y

Output: Predicted class for each image patch

- 1 Generate probabilistic outputs P_{ViT} from ViT model
 - 2 Generate probabilistic outputs P_{SVM} from SVM model
 - 3 Concatenate probabilities to form combined feature vector $F_{\text{FPN}} = [P_{\text{ViT}}, P_{\text{SVM}}]$
 - 4 Initialize Fusion Probability Network (FPN)
 - 5 **for each iteration during training do**
 - 6 Input F_{FPN} to FPN
 - 7 Compute fused probability P_{FPN} for each image patch
 - 8 Compute loss $\mathcal{L}(P_{\text{FPN}}(I_i), y_i)$
 - 9 Update FPN parameters to minimize loss
 - 10 **end**
 - 11 **for each image patch I_i do**
 - 12 Predict class $\underset{c}{\operatorname{argmax}} (P_{\text{FPN},c}(I_i))$
 - 13 **end**
-

4 Experiments

The study’s software setup uses Python via Anaconda on a Windows 11 operating system. Hardware includes an 8 GB NVIDIA GPU, 256 GB of RAM, and an Intel Core i7 processor. The experiments were run for 1,000 stochastic trials. The mean metrics and 95% confidence interval are reported. As mentioned, patches of size 64×64 pixels were extracted from each WSI using the corresponding annotation file, ensuring no overlap between patches. In total, more than 100,000 patches were extracted.

The suggested approach is applied on different sample sizes (i.e., number of patches) to study the performance sensitivity as presented in Table 1. From it, accuracy ranged from 95.36% to 96.50% across sample sizes of 10K, 25K, 50K,

and 100K. Precision values ranged from 91.07% to 93.33%, with the highest precision achieved at a sample size of 10K.

Similarly, recall rates varied from 90.68% to 93.00%, with the highest recall also recorded at a sample size of 10K. Specificity remained consistently high, ranging from 96.90% to 97.67%, indicating the model’s ability to accurately identify negative cases. The F1 score, a harmonic mean of precision and recall, ranged from 90.73% to 92.99%. Balanced accuracy (BAC) values were observed between 93.79% and 95.33%, with the highest BAC achieved at a sample size of 10K.

Moreover, the results showed that the system can work with different sample sizes, highlighting the robustness of the proposed approach across varied dataset sizes and validating the hypotheses mentioned in the current study.

Table 1. Performance sensitivity analysis resulted from the suggested framework on different sample sizes. The experiments are run for 1,000 stochastic trials. The mean metrics and 95% confidence interval are reported.

Sample Size	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 (%)	BAC (%)
10K	96.50 ± 0.23	93.33 ± 0.16	93.00 ± 0.12	97.67 ± 0.35	92.92 ± 0.28	95.33 ± 0.32
25K	95.95 ± 0.27	92.18 ± 0.22	91.90 ± 0.21	97.30 ± 0.18	91.88 ± 0.17	94.60 ± 0.28
50K	96.50 ± 0.31	93.22 ± 0.12	93.00 ± 0.15	97.67 ± 0.28	92.99 ± 0.31	95.33 ± 0.10
100K	95.36 ± 0.12	91.07 ± 0.29	90.68 ± 0.11	96.90 ± 0.18	90.73 ± 0.22	93.79 ± 0.37

4.1 Ablation Studies

Two ablation studies were conducted to validate the proposed approach and assess its generalizability. These studies included (1) removing numerical features, relying solely on patches and Vision Transformers (ViTs); and (2) altering patch spaces through morphological transformations such as rotation, blurring, and noise introduction.

First Ablation Study: The target of the first ablation study is to assess the impact of excluding the feature pipeline on the performance metrics of the framework. The question is to evaluate how the absence of the feature pipeline affects key metrics compared to the results of the whole system.

Results in Table 2 (*the first group*) show sensitivity analysis without the feature pipeline on the testing subset. *Result:* Comparing with Table 1, notable differences in performance metrics are observed when the feature pipeline is not utilized. Accuracy decreases by approximately 3–4%, precision by 5–7%, recall by 4–6%, F1 score by 3–6%, and BAC by 4–5%.

Insights: These performance decreases underscore the significant importance of the feature pipeline in enhancing overall effectiveness and robustness of the framework.

Second Ablation Study: The target of the second ablation study is to evaluate how well the system performs under different patch morphologies; and understand how these variations affect the system’s ability to diagnose and classify patches. Additionally, it investigates the influence of the feature pipeline on system performance, comparing results with and without it.

Performance results, shown in Table 2, include outcomes after applying random rotations within $[-90^\circ, 90^\circ]$ (*the second two groups*) and random blurring within $[(1, 1) \rightarrow (5, 5)]$ (*the last two groups*). *Result:* The system effectively diagnoses and classifies patches regardless of rotation or blurring, indicating rotation and/or blurring invariance.

Insights: These findings underscore the system’s robustness and adaptability to variations in patch orientations and morphology, enhancing its suitability for diverse imaging scenarios.

Table 2. Performance sensitivity analysis resulted from the suggested framework on different sample sizes in the ablation studies. The experiments are run for 1,000 stochastic trials. The mean metrics and 95% confidence interval are reported.

Features Pipeline	Sample Size	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 (%)	BAC (%)
Not Included	10K	89.50 ± 0.36	81.39 ± 0.27	79.00 ± 0.18	93.00 ± 0.47	79.18 ± 0.39	86.00 ± 0.24
	25K	92.70 ± 0.42	86.37 ± 0.35	85.40 ± 0.21	95.13 ± 0.38	85.46 ± 0.49	90.27 ± 0.33
	50K	93.35 ± 0.28	87.32 ± 0.49	86.70 ± 0.32	95.57 ± 0.14	86.68 ± 0.27	91.13 ± 0.41
	100K	93.16 ± 0.17	86.92 ± 0.21	86.33 ± 0.45	95.44 ± 0.29	86.41 ± 0.38	90.88 ± 0.19
Not Included	10K + R	86.88 ± 0.29	77.18 ± 0.42	73.75 ± 0.36	91.25 ± 0.18	73.19 ± 0.24	82.50 ± 0.31
	25K + R	91.10 ± 0.48	84.16 ± 0.39	82.20 ± 0.27	94.07 ± 0.45	81.61 ± 0.37	88.13 ± 0.22
	50K + R	92.25 ± 0.33	85.66 ± 0.27	84.50 ± 0.41	94.83 ± 0.29	84.05 ± 0.38	89.67 ± 0.47
	100K + R	91.21 ± 0.22	83.65 ± 0.38	82.43 ± 0.25	94.14 ± 0.47	82.41 ± 0.29	88.28 ± 0.34
Included	10K + R	87.63 ± 0.27	77.16 ± 0.35	75.25 ± 0.43	91.75 ± 0.39	74.99 ± 0.26	83.50 ± 0.47
	25K + R	91.95 ± 0.31	85.41 ± 0.24	83.90 ± 0.28	94.63 ± 0.22	83.74 ± 0.39	89.27 ± 0.32
	50K + R	93.45 ± 0.42	87.47 ± 0.37	86.90 ± 0.45	95.63 ± 0.28	86.70 ± 0.31	91.27 ± 0.48
	100K + R	92.34 ± 0.39	85.63 ± 0.26	84.68 ± 0.33	94.89 ± 0.42	84.67 ± 0.37	89.78 ± 0.29
Not Included	10K + B	90.00 ± 0.28	83.85 ± 0.36	80.00 ± 0.21	93.33 ± 0.45	79.67 ± 0.32	86.67 ± 0.19
	25K + B	92.40 ± 0.42	86.18 ± 0.27	84.80 ± 0.38	94.93 ± 0.29	84.65 ± 0.37	89.87 ± 0.24
	50K + B	92.95 ± 0.31	86.98 ± 0.39	85.90 ± 0.24	95.30 ± 0.18	85.75 ± 0.43	90.60 ± 0.29
	100K + B	93.41 ± 0.17	87.51 ± 0.24	86.83 ± 0.32	95.61 ± 0.41	86.88 ± 0.19	91.22 ± 0.37
Included	10K + B	95.50 ± 0.36	91.33 ± 0.23	91.00 ± 0.29	97.00 ± 0.18	90.99 ± 0.41	94.00 ± 0.28
	25K + B	93.60 ± 0.27	87.97 ± 0.35	87.20 ± 0.32	95.73 ± 0.19	87.24 ± 0.38	91.47 ± 0.25
	50K + B	94.85 ± 0.45	90.42 ± 0.28	89.70 ± 0.37	96.57 ± 0.24	89.74 ± 0.33	93.13 ± 0.29
	100K + B	94.21 ± 0.39	89.28 ± 0.34	88.40 ± 0.26	96.11 ± 0.43	88.47 ± 0.31	92.25 ± 0.27

R: Rotation, B: Blurring, BAC: Balanced Accuracy, Included: Features pipeline is included, and Not Included: Features pipeline is omitted.

4.2 Comparison with Related Studies

The proposed approach is tested using an external dataset called the Breast Cancer Histopathological Database (BreakHis) [24] to examine its validation and applicability to different benchmarks. Table 3 presents a comparative analysis of the findings from this research with those of other studies conducted on the BreakHis dataset. The outcomes reveal notable enhancements compared to earlier related studies conducted on BreakHis. Our study achieved high accuracy, precision, recall, specificity, and F1 score with small variances across 25 stochastic trials. Compared to previous research, our results demonstrate robust performance, indicating potential advancements in the field of breast cancer diagnosis using these metrics.

Table 3. Comparison of this study’s results with those of other related studies on the BreakHis dataset.

Study	Year	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 (%)
Seo et al. [22]	2022	86.80	90.90	93.20	-	89.90
Agarwal et al. [3]	2022	94.67	92.60	80.52	-	85.21
Balasubramanian et al. [7]	2024	98.43	-	-	-	-
Xiao et al. [30]	2024	92.00	-	-	-	-
Taheri et al. [25]	2024	95.10	-	-	-	-
Current Study	2024	98.88 ± 0.58	98.65 ± 0.77	96.54 ± 0.81	98.46 ± 0.69	96.24 ± 0.74

The experiment of the current study is run for 25 stochastic trials. The mean metrics and 95% confidence interval are reported.

4.3 Time Complexity Analysis

The feature extraction process involves several computational steps to analyze histopathological patches. Initially, patches are extracted from multiple magnification levels, requiring operations proportional to the number of pixels in each patch. Feature extraction algorithms like GLCM, GLRLM, LBP, SFM, and Hu Moments compute statistical and textural features, with time complexity generally dependent on patch size and feature extraction method. For instance, GLCM and GLRLM computations involve traversing image pixels and constructing matrices based on pixel relationships, typically resulting in time complexities of $O(N^2)$, where N is the number of pixels. LBP and SFM computations also operate at similar complexities due to pixel-level operations and matrix manipulations. Hu Moments, involving image moment calculations, can vary but often have complexities around $O(N^2)$ per patch.

MRFO for feature selection iteratively evaluates feature subsets’ quality using SVM classification performance. The computational complexity of MRFO primarily lies in evaluating each feature subset’s fitness, which involves training SVM models on subsets of features. Given M possible feature subsets and SVM training complexity $O(M \times N \times d)$, where N is the number of training instances and d is the dimensionality of the feature space, MRFO’s overall complexity is influenced by the number of iterations and population size, typically $O(T \times M \times N \times d)$.

TPE optimizes SVM hyperparameters by iteratively exploring and updating probability distributions over the hyperparameter space. The complexity involves evaluating configurations’ performances using SVM, adjusting PDFs, and selecting new configurations, typically $O(T \times N \times C)$, where T is the number of iterations, N is the number of configurations, and C is the SVM’s training complexity.

Multi-level ViTs process image patches by extracting hierarchical features through self-attention mechanisms. The time complexity is influenced by the number of patches P , levels L , and self-attention operations per patch, typically $O(P \times L \times N^2)$, where N is the patch size. ViTs’ training also involves backpropagation through multiple layers, affecting overall complexity.

The FPN integrates probabilities from ViT and SVM models to enhance classification accuracy. The complexity includes concatenating and processing probabilistic outputs from each model, typically $O(P \times C)$, where P is the number of image patches and C is the number of classes. Training FPN involves optimizing network parameters via backpropagation, contributing additional complexity depending on network architecture and iterations.

The overall complexity combines these components' computational efforts: $O(M \times N \times d + T \times N \times C + P \times L \times N^2 + P \times C)$ where M is number of feature subsets (influenced by the number of patches and feature extraction methods), N is number of pixels per patch (influenced by feature extraction and SVM), d is SVM training complexity, T is number of hyperparameter optimization iterations, C is number of classes, P is number of patches, and L is number of levels in ViTs.

4.4 Clinical Relevance in Enhancing the Diagnosis

BC diagnosis heavily relies on accurate interpretation of histopathological slides. The proposed framework integrates advanced computational methods to enhance diagnostic accuracy and reliability. By leveraging multi-level feature extraction from histopathological patches and employing state-of-the-art classification techniques, the framework aims to improve the discrimination between different tissue types and pathological conditions associated with BC. This comprehensive approach not only enhances the efficiency of diagnosis but also supports clinicians in making informed decisions based on robust quantitative analysis of tissue characteristics.

The utilization of diverse feature extraction techniques, including statistical, shape, and texture features, coupled with advanced machine learning algorithms such as SVM and ViT, underscores the potential to achieve more precise and consistent BC diagnosis. By integrating these computational methods, the framework seeks to mitigate interpretational variability and enhance the overall clinical workflow, contributing to improved patient outcomes and management strategies.

5 Conclusions and Future Directions

This study proposed a comprehensive framework, encompassing feature extraction, selection, classification, hyperparameter optimization, and model prediction fusion, aimed to enhance the performance and reliability for BC grading. Promising results were achieved, with the best-performing model reaching an accuracy of 96.50%, demonstrating the effectiveness of our approach across varying sample sizes. The integration of advanced feature extraction techniques facilitated a comprehensive characterization of histopathological patches, capturing crucial tissue characteristics indicative of BC. multi-level ViTs efficiently extracted hierarchical features, while the FPN effectively combined model predictions, improving overall classification performance. Ablation studies underscored the importance of the feature extraction pipeline's robustness and adaptability to diverse

patch morphologies. Despite variations in patch orientations and morphologies, consistent performance was observed, highlighting the system's reliability and diagnostic capabilities.

Future directions include expanding the dataset to improve model generalizability across diverse breast cancer subtypes and integrating multi-modal data, such as genomic and radiological information, to enhance classification accuracy. Developing a real-time application for clinical use, improving model explainability, and incorporating longitudinal data for prognostic purposes are key next steps. Collaboration with clinicians will refine the model for practical use, while addressing regulatory and ethical concerns, such as data privacy and bias, remains essential. Continuous model improvement through ongoing learning will ensure it stays accurate and relevant.

References

1. Abd Ghani, M.K., et al.: Decision-level fusion scheme for nasopharyngeal carcinoma identification using machine learning techniques. *Neural Comput. Appl.* **32**(3), 625–638 (2020). <https://doi.org/10.1007/s00521-018-3882-6>
2. Aboudessouki, A., et al.: Automated diagnosis of breast cancer using deep learning-based whole slide image analysis of molecular biomarkers. In: 2023 IEEE International Conference on Image Processing (ICIP), pp. 2965–2969 (2023). <https://doi.org/10.1109/ICIP49359.2023.10222479>
3. Agarwal, P., Yadav, A., Mathur, P.: Breast cancer prediction on BreakHis dataset using deep CNN and transfer learning model. In: Nanda, P., Verma, V.K., Srivastava, S., Gupta, R.K., Mazumdar, A.P. (eds.) *Data Engineering for Smart Systems: Proceedings of SSIC 2021*, pp. 77–88. Springer Singapore, Singapore (2022). https://doi.org/10.1007/978-981-16-2641-8_8
4. Ameh Joseph, A., Abdullahi, M., Junaidu, S.B., Hassan Ibrahim, H., Chiroma, H.: Improved multi-classification of breast cancer histopathological images using handcrafted features and deep neural network (dense layer). *Intell. Syst. Appl.* **14**, 200066 (2022). <https://doi.org/10.1016/j.iswa.2022.200066>
5. Baghdadi, N.A., Malki, A., Balaha, H.M., AbdulAzeem, Y., Badawy, M., Elhosseini, M.: Classification of breast cancer using a manta-ray foraging optimized transfer learning framework. *PeerJ Comput. Sci.* **8**, e1054 (2022)
6. Balaha, H.M., Antar, E.R., Saafan, M.M., El-Gendy, E.M.: A comprehensive framework towards segmenting and classifying breast cancer patients using deep learning and Aquila optimizer. *J. Ambient. Intell. Humaniz. Comput.* **14**(6), 7897–7917 (2023). <https://doi.org/10.1007/s12652-023-04600-1>
7. Balasubramanian, A.A., et al.: Ensemble deep learning-based image classification for breast cancer subtype and invasiveness diagnosis from whole slide image histopathology. *Cancers* **16**(12), 2222 (2024)
8. Corrias, G., Micheletti, G., Barberini, L., Suri, J.S., Saba, L.: Texture analysis imaging “what a clinical radiologist needs to know”. *Eur. J. Radiol.* **146**, 110055 (2022). <https://doi.org/10.1016/j.ejrad.2021.110055>
9. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale (2020)

10. Gamal, A., et al.: A novel machine learning approach for predicting neoadjuvant chemotherapy response in breast cancer: integration of multimodal radiomics with clinical and molecular subtype markers. *IEEE Access* **12**, 104983–105003 (2024). <https://doi.org/10.1109/access.2024.3432459>
11. Jiang, X., Hu, Z., Wang, S., Zhang, Y.: Deep learning for medical image-based cancer diagnosis. *Cancers* **15**(14), 3608 (2023). <https://doi.org/10.3390/cancers15143608>
12. Lima, Z.S., Ebadi, M.R., Amjad, G., Younesi, L.: Application of imaging technologies in breast cancer detection: a review article. *Open Access Macedonian J. Med. Sci.* **7**(5), 838–848 (2019). <https://doi.org/10.3889/oamjms.2019.171>
13. Martel, A., Nofech-Mozes, S., Salama, S., Akbar, S., Peikari, M.: Assessment of residual breast cancer cellularity after neoadjuvant chemotherapy using digital. *Pathology* (2019). <https://doi.org/10.7937/TCIA.2019.4YIBTJNO>
14. Michael, E., Ma, H., Li, H., Qi, S.: An optimized framework for breast cancer classification using machine learning. *BioMed Res. Int.* **2022**, 8482022 (2022). <https://doi.org/10.1155/2022/8482022>
15. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(7), 3523–3542 (2022). <https://doi.org/10.1109/TPAMI.2021.3059968>
16. Murtaza, G., et al.: Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artif. Intell. Rev.* **53**(3), 1655–1720 (2020). <https://doi.org/10.1007/s10462-019-09716-5>
17. Osapoetra, L.O., Chan, W., Tran, W., Kolios, M.C., Czarnota, G.J.: Comparison of methods for texture analysis of QUS parametric images in the characterization of breast lesions. *PLoS ONE* **15**(12), e0244965 (2020). <https://doi.org/10.1371/journal.pone.0244965>
18. Ozaki, Y., Tanigaki, Y., Watanabe, S., Nomura, M., Onishi, M.: Multiobjective tree-structured parzen estimator. *J. Artif. Intell. Res.* **73**, 1209–1250 (2022)
19. Peikari, M., Salama, S., Nofech-Mozes, S., Martel, A.L.: Automatic cellularity assessment from post-treated breast surgical specimens. *Cytometry. Part A: J. Int. Soc. Anal. Cytol.* **91**(11), 1078–1087 (2017). <https://doi.org/10.1002/cyto.a.23244>
20. Sabry, M., et al.: A vision transformer approach for breast cancer classification in histopathology. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–4. IEEE (2024)
21. Saleh, G.A., et al.: Impact of imaging biomarkers and AI on breast cancer management: a brief review. *Cancers* **15**(21), 5216 (2023)
22. Seo, H., Brand, L., Barco, L.S., Wang, H.: Scaling multi-instance support vector machine to breast cancer detection on the BreakHis dataset. *Bioinformatics* **38**(Supplement_1), i92–i100 (2022)
23. Sharma, S., Mehra, R.: Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight. *J. Digit. Imaging* **33**(3), 632–654 (2020). <https://doi.org/10.1007/s10278-019-00307-y>
24. Spanhol, F., Oliveira, L., Petitjean, C., Heutte, L.: Breast cancer histopathological database (breakhis) (2021)
25. Taheri, S., Golrizkhatami, Z., Basabrain, A.A., Hazzazi, M.S.: A comprehensive study on classification of breast cancer histopathological images: binary versus multi-category and magnification-specific versus magnification-independent. *IEEE Access* (2024)

26. Wang, Y., et al.: Improved breast cancer histological grading using deep learning. *Ann. Oncol.* **33**(1), 89–98 (2022). <https://doi.org/10.1016/j.annonc.2021.09.007>
27. Weigert, M., Schmidt, U.: Nuclei instance segmentation and classification in histopathology images with Stardist. In: 2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC), pp. 1–4 (2022). <https://doi.org/10.1109/ISBIC56247.2022.9854534>
28. Wetstein, S.C., et al.: Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. *Sci. Rep.* **12**(1), 15102 (2022). <https://doi.org/10.1038/41598-022-19112-9>, publisher: Nature Publishing Group
29. Wilkinson, L., Gathani, T.: Understanding breast cancer as a global health concern. *Br. J. Radiol.* **95**(1130), 20211033 (2022). <https://doi.org/10.1259/bjr.20211033>
30. Xiao, M., Li, Y., Yan, X., Gao, M., Wang, W.: Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example. In: Proceedings of the 2024 7th International Conference on Machine Vision and Applications, pp. 145–149 (2024)
31. Xu, H., et al.: Vision transformers for computational histopathology. *IEEE Rev. Biomedical Eng.* (2023)
32. Zhao, W., Zhang, Z., Wang, L.: Manta ray foraging optimization: an effective bio-inspired optimizer for engineering applications. *Eng. Appl. Artif. Intell.* **87**, 103300 (2020). <https://doi.org/10.1016/j.engappai.2019.103300>
33. Zou, Y., Zhang, J., Huang, S., Liu, B.: Breast cancer histopathological image classification using attention high-order deep network. *Int. J. Imaging Syst. Technol.* **32**(1), 266–279 (2022)



Dual-MambaNet: A Lightweight Dual-Branch Brain Image Segmentation Network Based on Local Attention and Mamba

Feifei Zhang^{1,2}, Fei Shi^{1,2(✉)}, Dayong Ren^{3(✉)}, Zhenhong Jia^{1,2},
and Jianyi Wang^{1,2}

¹ School of Computer Science and Technology, Xinjiang University,
Xinjiang 830046, Urumqi, China
sigofei@xju.edu.cn

² Key Laboratory of Signal Detection and Processing, Xinjiang University,
Xinjiang 830046, Urumqi, China

³ National Key Laboratory for Novel Software Technology, Nanjing University,
Nanjing 210023, China
rdyedu@gmail.com

Abstract. Brain tissue segmentation is critical for diagnosing and treating brain diseases. While Mamba-based models excel in the medical field, they face performance bottlenecks with high-resolution MRI images, often losing local feature information in complex texture structures. To address these challenges and enable deployment in resource-limited settings, we propose Dual-MambaNet, a lightweight segmentation model based on Mamba. In Dual-MambaNet, we introduce the Outlook attention module to capture local complex textures and structures in brain MRI images. Subsequently, we combined it with the Mamba block to construct a feature extractor (FE) encoder layer to couple local and global features. Additionally, we integrate dual decoder branches and a multi-level pixel contrastive loss function (MPCL) to better integrate local and global features. This method optimizes global feature representation by refining local complex textures and structural details, effectively capturing multi-level features in MRI images. Experimental results on public brain MRI datasets OASIS1 and MRBrainS13 demonstrate that Dual-MambaNet achieves high segmentation accuracy with minimal parameters and computational complexity, making it suitable for deployment in resource-limited medical environments.

Keywords: MRI · Brain tissue segmentation · Mamba · MPCL

This work was supported by National Natural Science Foundation of China (No. 62261053) and Tianshan Talent Training Project - Xinjiang Science and Technology Innovation Team Program (2023TSYCTD0012).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15328, pp. 92–107, 2025.
https://doi.org/10.1007/978-3-031-78104-9_7

1 Introduction

The rapid diagnosis of brain and nervous system disorders depends on healthcare professionals' expertise and professional skills. The analysis of brain magnetic resonance imaging (MRI) by healthcare professionals requires a significant amount of time and effort. In recent years, with the rapid advancement of computer technology, the use of computer-assisted techniques has improved the speed of segmentation and diagnosis of magnetic resonance imaging (MRI), enhancing the efficiency of medical diagnosis [10].

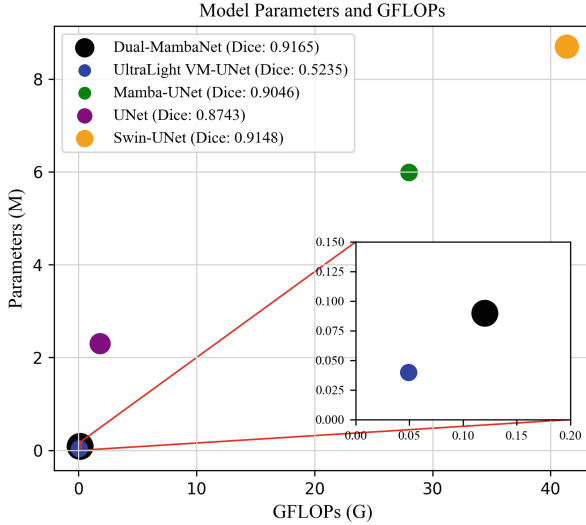


Fig. 1. Comparison of model parameters (Million) and GFLOPs (G) (The size of the circles represents the average Dice score on the OASIS1 dataset (↑)).

Convolutional neural networks(CNN), represented by UNet [14], are widely applied in medical image segmentation. CNN excel at capturing local features but may struggle to utilize global contextual information, which can lead to lower segmentation accuracy [2]. Inspired by self-attention mechanisms in natural language processing, Vision Transformer (ViT) was the first to apply multi-head attention mechanisms to visual tasks [3]. Due to its excellent capability in extracting global context, ViT is widely applied in medical image segmentation. However, its quadratic complexity can lead to high computational costs, especially in high-resolution medical image segmentation. In resource-constrained medical environments, these high computational costs pose challenges for model deployment. Therefore, there is an urgent need in medical image segmentation for lightweight algorithms that can achieve high accuracy.

Recently, advancements in state space models (SSM) [12] have provided new insights into lightweight medical image segmentation algorithms. The linear complexity of state space models and their excellent capability to model long-range

relationships have led to their widespread application in medical image segmentation, with Mamba as a representative example [4]. To facilitate model deployment in medical environments and improve segmentation accuracy, a series of lightweight Mamba models have been proposed to facilitate deployment in medical environments. LightM-UNet [8] significantly reduces model parameters while maintaining high segmentation accuracy to ensure feasibility for deployment in medical environments. UltraLight VM-UNet [18] introduces the Parallel Vision Mamba (PVM) module, resulting in a more lightweight model that ensures accuracy in skin lesion segmentation tasks.

Despite these studies alleviating the issues of complexity and computational cost to some extent, existing models still face performance bottlenecks when processing high-resolution MRI images, often losing local feature information in complex texture structures. To address these challenges and enable deployment in resource-limited environments, we propose Dual-MambaNet, a lightweight segmentation model based on Mamba. In Dual-MambaNet, we introduce the Outlook attention module to capture local complex textures and structures in brain MRI images. Subsequently, we combine it with the Mamba block to construct a feature extractor (FE) encoder layer, coupling local and global features. Additionally, we propose dual decoder branches and a multi-level pixel contrastive loss function (MPCL) to integrate local and global features better. This approach optimizes global feature representation by refining local complex textures and structural details, effectively capturing multi-level features in MRI images. Figure 1 compares the parameters and GFLOPs of Dual-MambaNet and other models (UNet, Swin-UNet, Mamba-UNet and UltraLight VM-UNet). As shown, Dual-MambaNet maintains high accuracy while having lower parameters and GFLOPs, facilitating its deployment in resource-limited medical environments.

In this paper, our contributions are as follows:

1. This paper designs a feature extractor (FE) as the encoder part, which extracts structural features through spatial transformation operations achieved by adaptive long-range and short-range computations. Specifically, Mamba is used for extracting global contextual information, while the local attention mechanism (Outlook attention) captures local features.
2. This paper employs a dual-branch decoder to strengthen the coupling of information at different levels and enhance the model’s ability to couple global and local features.
3. A multi-level pixel contrastive loss function (MPCL) is proposed to optimize the coupling of the model’s low-level and high-level features.
4. This paper proposes a lightweight model for brain MRI image segmentation. The model maintains high segmentation accuracy with a minimal increase in the number of parameters and GFLOPs.

2 Related Work

With the development of artificial intelligence, deep learning has been widely applied to medical image segmentation. Convolutional neural networks (CNN)

have been extensively used for image segmentation tasks [9]. UNet [14] has been widely applied in medical image segmentation due to its symmetric encoder-decoder architecture and skip connections [7]. The encoder and decoder of UNet can extract features at different levels, and the skip connections facilitate efficient transformation between these levels. However, this simple fusion method can only partially exploit these features, inevitably creating a semantic gap between features at different levels. To bridge this semantic gap, UNet++ [22] enhances the fusion of high-level and low-level information by adding convolutional layers within the skip connections. Building on this, UNet3+ [6] achieves more accurate segmentation by integrating multi-scale high-level and low-level features. However, methods based on CNNs can only extract local information and need more capture global contextual information.

To enhance the extraction of global contextual information, TransUNet [2] combines Transformer with UNet, achieving higher accuracy in medical image segmentation. Building on this, UTNet [5] employs multi-scale Transformers to fuse high-level and low-level features better. Given the suitability of ViT [3] for visual tasks and its robust feature extraction capabilities, many studies have integrated ViT and its variants with UNet, yielding improved results.

The linear complexity and long-range relationship modelling capability of state space models have recently led to the widespread application of Mamba models in medical image segmentation. Studies have shown that Mamba is effective in image segmentation [20]. VMamba [23] introduced a hierarchical visual backbone network and Cross-Scan Module (CSM) based on Mamba, making Mamba more suitable for 2D image tasks. Mamba-UNet [17], based on the Swin-UNet architecture, applies pure visual Mamba modules (VSS) to medical image segmentation, outperforming CNN- and Transformer-based models. LightM-UNet [8] was proposed to explore more lightweight models, significantly reducing the number of model parameters. UltraLight VM-UNet [18] verified the significant impact of channel count on model parameters and introduced the Parallel Vision Mamba module (PVM), achieving a more lightweight model while ensuring accuracy in skin lesion segmentation.

Although CNN-based methods can accomplish complex medical image segmentation tasks, they often fail to fully utilize global contextual information, resulting in poor feature extraction capability and, consequently, lower segmentation performance. On the other hand, transformer-based algorithms can effectively extract global contextual information, but they tend to be complex with high computational complexity. This paper proposes a lightweight medical image segmentation model based on Mamba (Dual-MambaNet) to address information loss issues in brain MRI image segmentation and the difficulty of deploying complex models in resource-constrained medical environments.

3 Method

3.1 Architecture Overview

Figure 2 shows the overall architecture of Dual-MambaNet, which consists of a 6-layer encoder, a 3-layer low-level feature decoder, and a 6-layer high-level feature

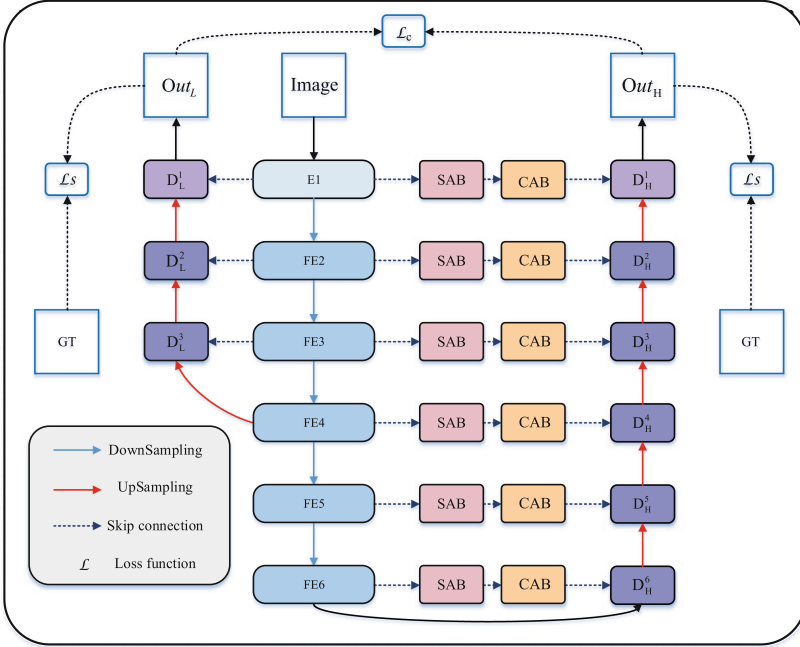


Fig. 2. The overall architecture of Dual-MambaNet.

decoder, forming an asymmetric U-shaped network. In the encoder part, except for the first encoder layer, which uses convolutional layers and local attention mechanisms, all other layers use the feature extractor (FE), as shown in Fig. 3(b). In the decoder part, except for the last decoder layer, all other layers use the parallel Mamba layer (PVM) [18]. The number of channels in each layer of the encoder and decoder structures are [8, 16, 24, 32, 48, 64].

Skip connections use the channel attention bridge (CAB) and spatial attention bridge (SAB) as proposed in [15]. The SAB module includes max pooling, average pooling, and dilated convolutions with shared weights. The CAB module includes global average pooling, concatenation operations, fully connected layers, and a Sigmoid activation function. In the skip connections of the low-level decoder, attention bridges are not used to avoid over-decoding due to the large gap between shallow features (which contain better detail information) and deep features (which contain more semantic information).

The model has two final outputs: the left decoder branch outputs low-level features, and the right decoder outputs high-level features. The low-level output information is also used to finely optimize the high-level output through a pixel-level contrastive loss function, improving the model’s segmentation accuracy.

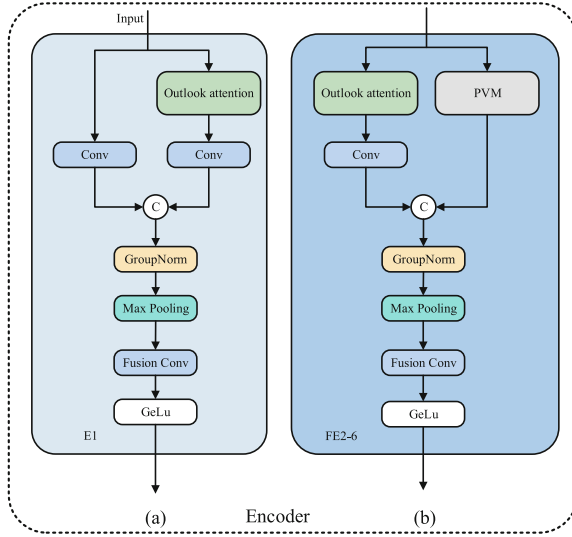


Fig. 3. Encoder Structure Diagram (a) E1: First Layer Encoder Structure. (b) FE2-FE6: Feature Encoder (FE) Constructed.

3.2 Parallel Vision Mamba Layer (PVM)

The Mamba module used in this paper is the Parallel Vision Mamba layer (PVM) proposed in [18], which is based on the Mamba module [4]. Its structure is shown in Fig. 4(b). In [18], the impact of the number of channels on the Mamba parameter count was extensively discussed, demonstrating that the number of channels has an exponential effect on the Mamba parameter count. Based on this conclusion, the PVM Layer was proposed. The structure of PVM is shown in Fig. 4(a). PVM mainly combines Mamba with residual connections and adjustment factors. The feature token X (with C channels) first passes through a LayerNorm layer and is then divided into four sub-features (each with $C/4$ channels) along the channel dimension. Each sub-feature is then fed into Mamba, and the outputs are subjected to residual and adjustment operations to optimize the ability to capture long-range spatial information. Finally, the four features are concatenated along the channel dimension to form X_{out} , which has the exact dimensions as the original input X . X_{out} , then undergoes LayerNorm and linear projection operations to transform it to the exact dimensions as the original image. This allows Mamba to enhance the capture of long-range spatial relationships without introducing additional parameters and computational complexity.

3.3 Outlook Attention

The Outlook attention in this paper is based on [21], and its specific structure is shown in Fig. 5. Specifically, for each spatial position (i, j) , Outlook Attention calculates its similarity with all neighbours within a local window of size

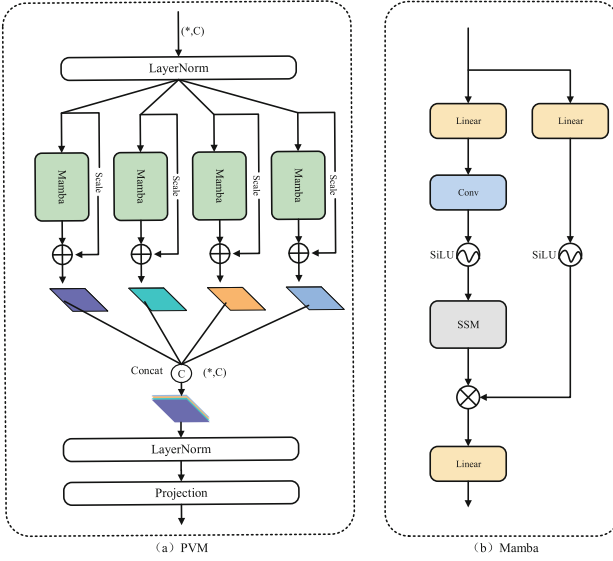


Fig. 4. (a) Parallel Vision Mamba (PVM) structure diagram. (b) Mamba Block.

$K \times K$ centred at (i, j) . Unlike self-attention, which requires a Query-Key matrix multiplication to compute attention, Outlook Attention simplifies this process through a reshaping operation.

Formally, given the input X , each C -dimensional token is first projected using two linear layers $W_A \in \mathbb{R}^{C \times K^4}$ and $W_V \in \mathbb{R}^{C \times C}$ into the outlook weights $A \in \mathbb{R}^{H \times W \times K^4}$ and value representations $V \in \mathbb{R}^{H \times W \times C}$, respectively. Let $V_{\Delta_{i,j}} \in \mathbb{R}^{C \times K^2}$ denote the values within the local window centred at (i, j) . This process is represented by Eq. 1.

$$V_{\Delta_{i,j}} = \left\{ V_{i+p-\lfloor \frac{K}{2} \rfloor, j+q-\lfloor \frac{K}{2} \rfloor} \right\}, \dots, 0 \leq p, q < K, \tag{1}$$

where, $\lfloor \frac{K}{2} \rfloor$ represents the floor function of $\frac{K}{2}$.

In Outlook attention, the outlook weights at position (i, j) can be directly used as the aggregated attention weights by reshaping them into $\hat{A}_{i,j} \in \mathbb{R}^{K^2 \times K^2}$, followed by the softmax activation function. Consequently, the value projection process can be written as Eq. 2:

$$Y_{\Delta_{i,j}} = \text{MatMul} \left(\text{Softmax}(\hat{A}_{i,j}), V_{\Delta_{i,j}} \right), \tag{2}$$

where $\hat{A}_{i,j}$ is the reshaped outlook weights, and $V_{\Delta_{i,j}}$ represents the values within the local window centred at (i, j) .

Outlook attention densely aggregates the projected value representations. By summing the differently weighted values from the same position across different local windows, the output result is obtained as shown in Eq. 3:

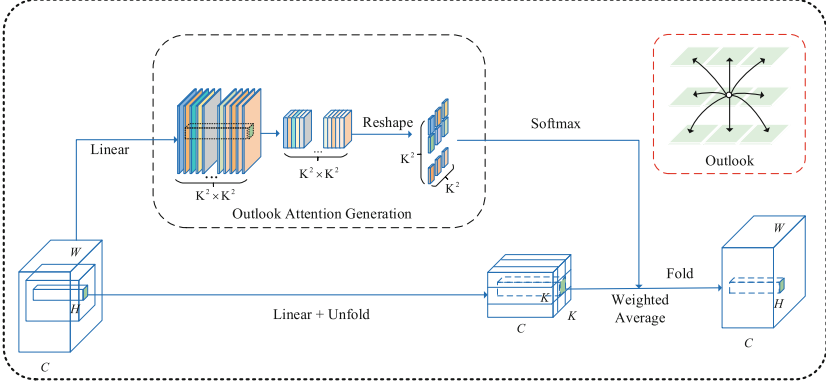


Fig. 5. Outlook attention structure diagram.

$$\tilde{Y}_{i,j} = \sum_{0 \leq m, n < K} Y_{\Delta_{i+m-\lfloor \frac{K}{2} \rfloor, j+n-\lfloor \frac{K}{2} \rfloor}}^{i,j}. \quad (3)$$

3.4 Loss Function

Pixel Level Contrastive Learning. Contrastive learning is widely used in self-supervised learning. Its main idea is discrimination between positive and negative samples, primarily achieved by using a metric function to encourage the network to bring positive samples closer while pushing negative samples apart. In medical image segmentation, contrastive learning addresses the critical issue of sparse annotated samples in datasets while enhancing the model’s generalization ability and augmenting the model’s capacity for feature extraction [19].

Dual-MambaNet uses two decoders: a low-level decoder and a high-level decoder. The output of the low-level decoder is considered the segmentation result for generating pseudo-labels, while the output of the high-level decoder is regarded as the segmentation result for the accurate labels. Inspired by [16], we propose using our novel improved multi-level pixel contrastive loss function(MPCL) between the outputs of the two decoders. This approach optimizes the output of the high-level decoder based on the output of the low-level decoder, thereby improving the final output of the model.

Considering that in brain images, each tissue has a relatively small size and many pixels belong to the background, these background pixels do not provide sufficient features for the network. Therefore, we propose using adaptive average pooling to filter out unimportant background pixels, enhancing the model’s feature extraction capability. Additionally, we apply L2 regularization on the channel dimension to sparsify the features, thereby improving the model’s generalization and robustness. Specifically, our proposed multi-level pixel contrastive loss function(MPCL) can be expressed as Eq. 4:

$$\mathcal{L}_{\text{MPCL}} = \frac{\sum \|(G(D_\theta(D_L \cup D_H)), G(D_\theta(D_H)))\|_2^2}{N}, \quad (4)$$

where D_θ is the decoder using AdaptiveAvgPool, G is the L2 regularization operation along the channel axis, and N is the number of input data. D_L and D_H represent the outputs of the low-level and high-level decoders, respectively, and \cup denotes the union operation. To effectively utilize the low-level decoder output to optimize the high-level decoder output, we consider the high-level decoder output as the low-level decoder output to maximize the distance between different level outputs, thereby improving the model's performance.

Total Loss. In our model, decoders and labels use standard cross-entropy loss (\mathcal{L}_{CE}) and Dice loss functions (\mathcal{L}_{Dice}). A multi-level pixel contrastive loss function (MPCL) is also used between the outputs of the two decoders.

The total loss is defined as Eq. 5:

$$\mathcal{L}_{\text{total}} = \lambda \left(\frac{\frac{\mathcal{L}_{Dice}^L + \mathcal{L}_{CE}^L}{2} + \frac{\mathcal{L}_{Dice}^H + \mathcal{L}_{CE}^H}{2}}{2} \right) + (1 - \lambda) \mathcal{L}_{\text{MPCL}}, \quad (5)$$

where \mathcal{L}_{Dice}^L and \mathcal{L}_{Dice}^H represent the Dice losses for the low-level and high-level feature outputs, respectively. Similarly, \mathcal{L}_{CE}^L and \mathcal{L}_{CE}^H represent the cross-entropy losses for the low-level and high-level feature outputs, respectively. $\mathcal{L}_{\text{MPCL}}$ represents the multi-level pixel contrastive loss function between the high-level and low-level feature outputs. The weighting factor λ , set empirically to 0.9, balances the contributions between the contrastive loss and the other loss functions. As shown in Fig. 6, we illustrate the process of the proposed dual-branch decoder and multi-level pixel contrastive loss function collaboratively optimizing the final output. In Fig. 6, D_H represents the high-level decoder branch, and D_L represents the low-level decoder branch.

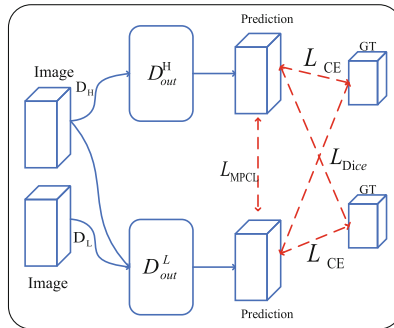


Fig. 6. A diagram of the dual-branch decoder framework based on multi-level pixel contrastive learning.

4 Experiments and Results

4.1 Datasets

OASIS1: The OASIS-1 dataset [11] used in this experiment is from the Open Access Series of Imaging Studies (OASIS). It comprises 421 subjects aged between 18 and 96 years. Each subject has a T1-weighted magnetic resonance imaging (MRI) scan. The dataset labels classify brain tissue into the cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM).

MICCAI 2013 MR BRAIN IMAGE SEGMENTATION: The MRBrainS13 challenge dataset consists of 20 subjects acquired using a 3.0T Philips Achieva MR scanner at the University Medical Center Utrecht, Netherlands [13]. The dataset includes multi-sequence MRI brain scans, such as T1, T1-IR and T2-FLAIR used for the challenge. The dataset labels classify brain tissue into the cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM).

4.2 Implementation Details

All experiments were conducted on a GeForce RTX 3090Ti GPU system with 24GB memory and Ubuntu 22.04, Python 3.8.19, PyTorch 2.2.0, and CUDA 11.8. The model is used for 2D medical image segmentation. We randomly split the two datasets into training, testing, and validation sets in an 8:1:1 ratio. All images were normalized and resized to 224×224 , and data augmentation techniques, including vertical flip, horizontal flip, and random rotation, were applied. The Dual-MambaNet model was trained for 40,000 iterations with a batch size 24. The AdamW optimizer was used with a learning rate of $1e-4$ and a weight decay set to $1e-4$. Network performance was evaluated on the validation set every 200 iterations, and model weights were saved only when the new best performance was achieved on the validation set.

4.3 Comparison Methods

To ensure a fair comparison, the baseline methods (UltraLight VM-UNet) [18], Mamba-UNet [17], UNet [14], and Swin-Unet [1] were also trained under the same hyperparameter configurations without loading pre-trained models. We directly compared Dual-MambaNet with the baseline method (UltraLight VM-UNet) and other methods based on CNN, Transformer and Mamba.

4.4 Evaluation Metrics

This study also employed three objective evaluation metrics for quantitative comparison of our proposed method: (1) Similarity Measurement: Dice coefficient (denoted by an upward arrow \uparrow), where values closer to 1 indicate better performance. (2) Difference Measurements: Hausdorff Distance (HD) 95% and Average Surface Distance (ASD) (both denoted by a downward arrow \downarrow), where lower values are better, indicating higher similarity between the predicted segmentation and the ground truth.

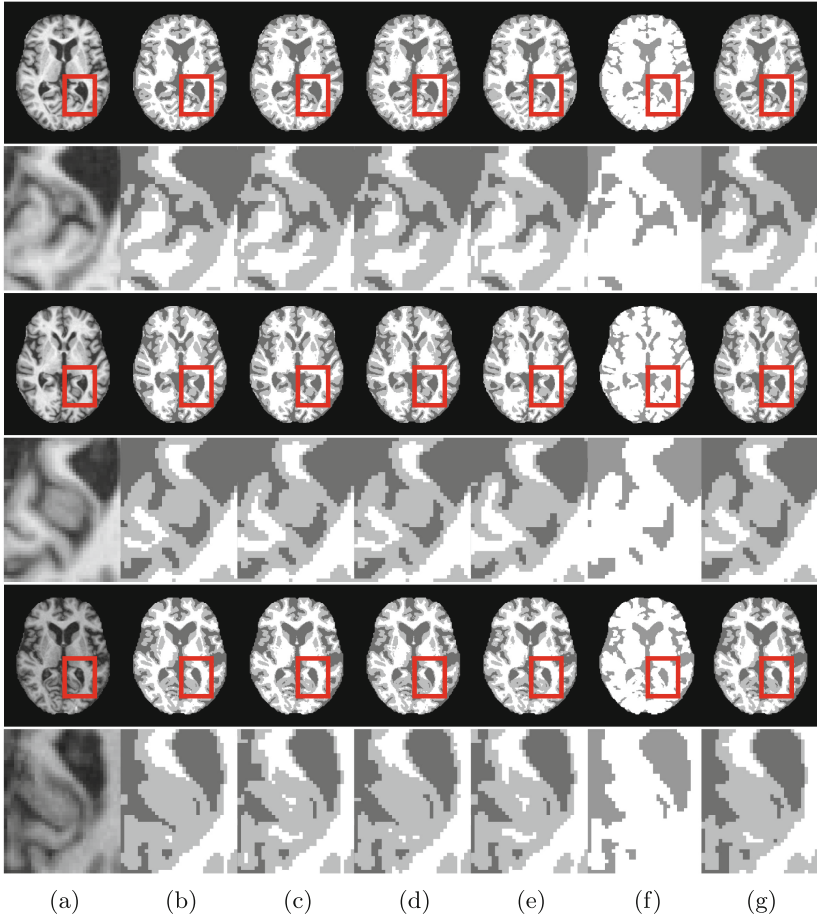


Fig. 7. Segmentation results comparison among different models on the OASIS1 dataset, with localized zoom-in comparison. (a) Image. (b) GT. (c) Mamba-UNet. (d) Swin-UNet. (e) UNet. (f) UltraLight VM-UNet. (g) Dual-MambaNet.

4.5 Qualitative Results

Figure 7 and Fig. 8 present three randomly selected original image samples from the OASIS1 and MRBrainS13 datasets. They compare the segmentation results of all baseline methods, including Dual-MambaNet, on the OASIS1 and MRBrainS13 datasets, along with zoomed-in views of local details.

As shown in the results of Fig. 7 and Fig. 8, as well as the enlarged views of local details, Dual-MambaNet can segment all categories completely compared to the Baseline (UltraLight VM-UNet). It can extract local features better while also capturing high-level semantic information. Compared to other classic models based on CNN, Transformer, and Mamba, Dual-MambaNet can also fully extract

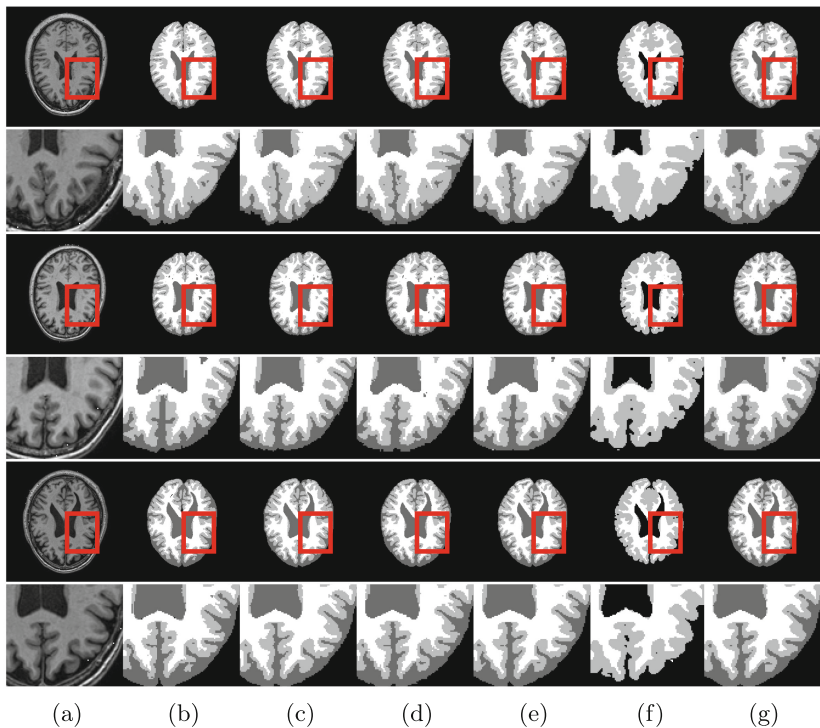


Fig. 8. Segmentation results comparison among different models on the MRBrainS13 dataset, with localized zoom-in comparison. (a) Image. (b) GT. (c) Mamba-UNet. (d) Swin-UNet. (e) UNet. (f) UltraLight VM-UNet. (g) Dual-MambaNet.

local features. As seen in the enlarged local view in Fig. 7, Dual-MambaNet can recognize more complex local features while maintaining the integrity of global information. As shown in the enlarged local view in Fig. 8, Dual-MambaNet can better recognize edge information and maintain the integrity of global features. In the above analysis, Dual-MambaNet can fully extract global features while better capturing complex textures and structural features such as edges.

4.6 Quantitative Results

Table 1 and Table 2 directly compare Dual-MambaNet with other segmentation networks on the OASIS1 and MRBrainS13 datasets, respectively, including similarity and difference metrics. The best-performing results are highlighted in bold, and '-' indicates that the model did not segment that category.

Quantitative results indicate that on large-scale datasets, Dual-MambaNet performs comparably to CNN-based and Transformer-based models on some metrics while surpassing classical models and the latest Vision Mamba models on others. Additionally, Dual-MambaNet has lower parameter counts and GFLOPs.

Table 1. Comparison of objective evaluation metrics of models on the OASIS1 dataset.

Model	Dice(\uparrow)			HD95(\downarrow)			ASD(\downarrow)			Para(M)	GFLOPs(G)
	CSF	GM	WM	CSF	GM	WM	CSF	GM	WM		
Mamba-UNet	0.9118	0.9114	0.8905	1.2102	1.7833	3.4910	0.3183	0.5163	1.0448	28.00	5.99
UNet	0.8719	0.8717	0.8793	1.4852	1.9535	4.0379	0.3059	0.6240	1.4087	1.81	2.3
Swin-UNet	0.9160	0.9187	0.9098	1.2913	1.2347	1.8963	0.2536	0.4269	0.4949	41.34	8.71
UltraLight VM-UNet	0.8528	0.7177	-	1.6120	6.5613	-	0.5413	0.4943	-	0.049	0.04
Dual-MambaNet	0.9161	0.9198	0.9136	1.1439	1.3626	2.5324	0.3222	0.3771	0.2542	0.10	0.08

Table 2. Comparison of objective evaluation metrics of models on the MRBrainS13 dataset.

Model	Dice(\uparrow)			HD95(\downarrow)			ASD(\downarrow)			Para(M)	GFLOPs(G)
	CSF	GM	WM	CSF	GM	WM	CSF	GM	WM		
Mamba-UNet	0.6655	0.6918	0.7150	2.3588	3.4587	4.2645	0.5810	1.0214	1.3920	28.00	5.99
UNet	0.6614	0.7003	0.7327	2.1166	2.5988	5.0152	0.4488	0.7471	1.7367	1.81	2.3
Swin-UNet	0.6683	0.7045	0.7417	1.7946	1.5695	4.5107	0.4843	0.4716	1.3532	41.34	8.71
UltraLight VM-UNet	-	0.6219	0.6895	-	4.8375	4.7374	-	1.4408	1.6855	0.049	0.04
Dual-MambaNet	0.6697	0.7077	0.7199	2.3193	2.5511	4.2528	0.5565	0.7432	1.5689	0.10	0.08

Compared to UltraLight VM-UNet, Dual-MambaNet significantly improves segmentation accuracy with a slight increase in complexity. Dual-MambaNet also demonstrates good generalization ability and robustness on small datasets, accurately predicting segmentation masks. Despite a slight increase in parameters and GFLOPs compared to the baseline model (UltraLight VM-UNet), Dual-MambaNet significantly enhances segmentation performance. Dual-MambaNet achieves higher segmentation accuracy than other methods while maintaining lower parameter counts and GFLOPs.

4.7 Ablation Study

Dual-MambaNet involves three key components: 1) Outlook Attention; 2) Dual Decoder Branches; 3) Multi-Level Pixel Contrastive Loss(MPCL). We compare the parts proposed in this study through ablation studies. To validate the effectiveness of the proposed model and its improvements, extensive ablation experiments were conducted on the MRBrainS13 dataset, using Dice and HD95, to evaluate the performance of each component quantitatively. The best-performing values are highlighted in bold, and ‘-’ indicates that the model did not segment that category. The results are shown in Table 3. In this table, ‘Atten’ represents the improved Outlook attention, ‘Double’ represents the dual decoder branches structure, and ‘MPCL’ represents the multi-level pixel contrastive loss function. Additionally, \checkmark indicates that the component is used and \times indicates that the component is not used.

Table 3. Comparison of Ablation Experiment Results.

Model			Dice(\uparrow)			HD95(\downarrow)		
Double	Atten	MPCL	CSF	GM	WM	CSF	GM	WM
\times	\times	\times	-	0.6219	0.6895	-	4.8375	4.7374
\checkmark	\times	\times	0.6348	0.6627	0.6672	2.9957	3.9822	7.0246
\checkmark	\times	\checkmark	0.6414	0.6629	0.6845	2.9074	3.3953	5.3508
\checkmark	\checkmark	\times	0.6370	0.6727	0.6787	3.0993	3.4120	6.8035
\times	\checkmark	\times	0.6321	0.6542	0.6654	3.0875	3.9764	6.9864
\checkmark	\checkmark	\checkmark	0.6697	0.7077	0.7199	2.3193	2.5511	4.2528

As shown in Table 3, although using local attention alone for feature extraction improves the model’s accuracy, the Dual-MambaNet with the dual-branch decoder captures the complex features of brain MRI images more effectively. This indicates that the dual-branch decoder enhances the model’s ability to couple multi-level information, thereby improving segmentation accuracy. Furthermore, using the pixel-level contrastive loss function for output optimization further improves segmentation accuracy, demonstrating that this loss function strengthens the coupling ability of the dual-branch decoder. While using any single component alone can improve segmentation performance, the model achieves the best performance when all three components are combined. These results show that Dual-MambaNet can segment brain MRI images with high accuracy.

5 Conclusion

This paper addresses the performance bottlenecks and loss of local feature information in brain tissue segmentation of high-resolution MRI images by proposing the Dual-MambaNet model. This model combines the Outlook attention module with Mamba to construct a feature extractor (FE) encoder layer, effectively connecting local and global features. Additionally, dual decoder branches and a multi-level pixel contrastive loss function (MPCL) are introduced to optimize feature representation. Experimental results on the OASIS1 and MRBrainS13 datasets demonstrate that Dual-MambaNet achieves high segmentation accuracy with lower parameters and GFLOPs, making it suitable for deployment in resource-limited medical environments. This research provides a promising solution for medical image segmentation under constrained computational resources.

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218. Springer (2022). https://doi.org/10.1007/978-3-031-25066-8_9

2. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint [arXiv:2102.04306](https://arxiv.org/abs/2102.04306) (2021)
3. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
4. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint [arXiv:2312.00752](https://arxiv.org/abs/2312.00752) (2023)
5. Hatamizadeh, A., et al.: Unetr: transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
6. Huang, H., et al.: Unet 3+: a full-scale connected unet for medical image segmentation. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055–1059. IEEE (2020)
7. Li, Z., Zhang, C., Zhang, Y., Wang, X., Ma, X., Zhang, H., Wu, S.: Can: context-assisted full attention network for brain tissue segmentation. *Med. Image Anal.* **85**, 102710 (2023)
8. Liao, W., Zhu, Y., Wang, X., Pan, C., Wang, Y., Ma, L.: Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. arXiv preprint [arXiv:2403.05246](https://arxiv.org/abs/2403.05246) (2024)
9. Liu, C., et al.: Auto-deeplab: hierarchical neural architecture search for semantic image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 82–92 (2019)
10. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1013–1023 (2021)
11. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* **19**(9), 1498–1507 (2007)
12. Mehta, H., Gupta, A., Cutkosky, A., Neyshabur, B.: Long range language modeling via gated state spaces. arXiv preprint [arXiv:2206.13947](https://arxiv.org/abs/2206.13947) (2022)
13. Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Biessels, G.J., Viergever, M.A.: MR Brain Segmentation Challenge 2013 Data (2024). <https://doi.org/10.34894/645ZIN>
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18, pp. 234–241. Springer (2015)
15. Ruan, J., Xiang, S., Xie, M., Liu, T., Fu, Y.: Malunet: A multi-attention and lightweight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1150–1156. IEEE (2022)
16. Wang, Z., Ma, C.: Semi-mamba-unet: Pixel-level contrastive cross-supervised visual mamba-based unet for semi-supervised medical image segmentation. arXiv preprint [arXiv:2402.07245](https://arxiv.org/abs/2402.07245) (2024)
17. Wang, Z., Zheng, J.Q., Zhang, Y., Cui, G., Li, L.: Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv preprint [arXiv:2402.05079](https://arxiv.org/abs/2402.05079) (2024)
18. Wu, R., Liu, Y., Liang, P., Chang, Q.: Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. arXiv preprint [arXiv:2403.20035](https://arxiv.org/abs/2403.20035) (2024)

19. You, C., Zhou, Y., Zhao, R., Staib, L., Duncan, J.S.: Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Trans. Med. Imaging* **41**(9), 2228–2237 (2022)
20. Yu, W., Wang, X.: Mambaout: Do we really need mamba for vision? arXiv preprint [arXiv:2405.07992](https://arxiv.org/abs/2405.07992) (2024)
21. Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S.: Volo: Vision outlooker for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(5), 6575–6586 (2022)
22. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: a nested u-net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pp. 3–11. Springer (2018)
23. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: efficient visual representation learning with bidirectional state space model. arXiv preprint [arXiv:2401.09417](https://arxiv.org/abs/2401.09417) (2024)



Location Matters: Harnessing Spatial Information to Enhance the Segmentation of the Inferior Alveolar Canal in CBCTs

Luca Lumetti¹, Vittorio Pipoli^{1,2}, Federico Bolelli^{1(✉)}, Elisa Ficarra¹,
and Costantino Grana¹

¹ University of Modena and Reggio Emilia, Modena, Italy

{luca.lumetti,federico.bolelli,elisa.ficarra,costantino.grana}@unimore.it

² University of Pisa, Pisa, Italy
vittorio.pipoli@phd.unipi.it

Abstract. The segmentation of the Inferior Alveolar Canal (IAC) plays a central role in maxillofacial surgery, drawing significant attention in the current research. Because of their outstanding results, deep learning methods are widely adopted in the segmentation of 3D medical volumes, including the IAC in Cone Beam Computed Tomography (CBCT) data. One of the main challenges when segmenting large volumes, including those obtained through CBCT scans, arises from the use of patch-based techniques, mandatory to fit memory constraints. Such training approaches compromise neural network performance due to a reduction in the global contextual information. Performance degradation is prominently evident when the target objects are small with respect to the background, as it happens with the inferior alveolar nerve that develops across the mandible, but involves only a few voxels of the entire scan. In order to target this issue and push state-of-the-art performance in the segmentation of the IAC, we propose an innovative approach that exploits spatial information of extracted patches and integrates it into a Transformer architecture. By incorporating prior knowledge about patch location, our model improves state of the art by ~ 2 points on the Dice score when integrated with the standard U-Net architecture. The source code of our proposal is publicly released.

Keywords: Inferior Alveolar Canal · 3D Segmentation · CBCT · Transformers · Patch-based Learning

1 Introduction

The presence of the Inferior Alveolar Nerve (IAN) represents a challenge for maxillofacial surgery. Such a nerve crosses the Inferior Alveolar bone Canal (IAC) and supplies sensation to the lower teeth, lips, and chin. For this reason, IAN position (Fig. 1) must be carefully identified before surgical intervention (e.g., implant placement and molar extraction) to prevent aches, pain, and temporary or permanent paralysis [33]. Usually, the preoperative treatment planning

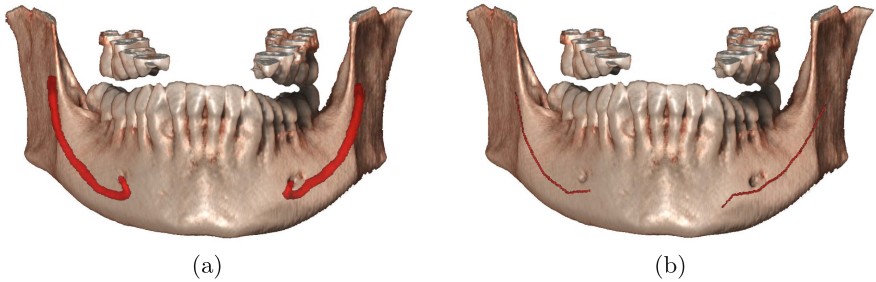


Fig. 1. CBCT with the IAC marked in red. (a) contains a 3D dense annotation, while (b) contains a 2D sparse annotation obtained from a panoramic view of the mandible and later re-projected to the 3D space (Color figure online).

is based on IAC segmentation performed on 3D data acquired with Cone Beam Computer Tomography (CBCT). Nevertheless, producing 3D annotations for 3D data is dramatically challenging and time-consuming. Hence, the standard practice consists of extracting 2D panoramic views where the surgeon can annotate the approximate position of the IAC drawing 2D curves. Despite this procedure being effective most of the time, having the 3D segmentation of the IAC would crucially improve the precision of the surgery planning, minimizing the likelihood of errors during surgery operations.

Recent advancements in deep learning have significantly impacted multiple domains, including medical imaging, particularly through methods based on Convolutional Neural Networks (CNNs) [11, 14, 15, 23–26]. Among them, of the most popular is U-Net [27], an encoder-decoder architecture with skip connections capable of extracting deep features while trying to retain as many fine-grained details as possible [10]. As well, many U-Net-based approaches for the automatic segmentation of the IAC [5, 16, 30] have been recently published, also thanks to the public availability of a 3D-annotated dataset [4].

Despite the great success of CNN in medical imaging, the rise of Transformer architectures [31] stands as a turning point. Representing the standard of Natural Language Processing since 2017 and deeply affecting the Computer Vision field since 2020 [8], Transformer-based architectures demonstrate dominance in several tasks due to their capability of modeling long-range interactions [6, 21, 22, 28]. This is in contrast with the *CNN locality bias*, which instead forces the modeling of local interactions that lie within the CNN sliding kernels [8]. For this reason, researchers are developing strategies to improve U-Net-based architectures [27] by integrating some Transformer layers to enhance long-range interactions, with encouraging results [9, 29]. In this work, we investigate innovative and effective ways to improve such an integration.

Regardless of the adopted method, processing 3D volumes leads to severe memory constraints, making the segmentation of a single 3D scan in one shot a prohibitive operation. Meanwhile, decreasing the resolution of such 3D images with downsampling techniques is counterproductive because fine-grained details

are needed to improve the segmentation quality. Hence, the only solution to solve both problems is splitting the 3D scans into multiple patches that will be processed separately, without losing detailed information. The literature refers to the aforementioned procedure as *patch-based learning*. Even if patch-based learning allows the training of deep neural networks with standard hardware resources, it must be mentioned that it forces the model to focus only on a fraction of the total information at a time, losing global context (e.g., the position of the examined patch with respect to the other patches of the 3D volume). In this research, we aim at mitigating this phenomenon with Transformers [31].

Paper Contribution. We present an innovative 3D segmentation model enhanced by a memory-augmented Transformer encoder that effectively harnesses absolute spatial coordinates, addressing the challenges of patch-based training.

Specifically, our proposal evolves from the standard 3D U-Net architecture by incorporating a memory-augmented Transformer in the bottleneck. By leveraging the inherent capacity of Transformers to model interactions between all pairs of elements within a given sequence, we aim to enhance the flow of information among the elements of the U-Net bottleneck, thereby increasing contextualization. Moreover, we harness such a flow of information to effectively inject contextual information related to the processed patches, i.e., their position within the entire volume, thus mitigating issues associated with patch-based learning. The “memory” is an additional refinement that supports the model in retaining crucial prior concepts that may be challenging to be directly extracted from image features, but are nonetheless valuable for interpretation. In summary, the key contributions of this paper are as follows:

- i) We propose a memory-augmented Transformer module that harnesses absolute spatial coordinates, mitigating issues related to patch-based learning;
- ii) We design an U-Net-based deep learning architecture integrating our proposed module and tailored for 3D IAC segmentation, outperforming state of the art on the selected segmentation task of ~ 2 Dice points;
- iii) The source code of our proposal is publicly released¹ to allow the replication of the experiments and foster future research advancements.

2 Related Works

While classical computer vision approaches have made significant contributions in the past [1, 2, 13, 19, 32], today, the most successful models for the segmentation of the IAC are based on machine learning and deep learning.

Notably, Jaskari *et al.* [12] presented one of the pioneering applications of deep learning for mandibular canal segmentation. Their approach involved training a convolutional network using a dataset of coarsely annotated 3D scans. This

¹ https://github.com/AImagelab-zip/alveolar_canal

deep learning approach demonstrated superior performance compared to previous methods relying on Statistical Shape Models. However, it encountered limitations due to the lack of finely annotated voxel-level data and the sub-optimal quality of segmentation masks generated automatically from coarse annotations.

Cipriano *et al.* [5] introduced a significant breakthrough by proposing the first publicly available dataset of 3D annotated CBCT scans of the human jaw, named *Maxillo*, alongside a deep learning model for the 3D segmentation of the IAC, *PosPadUNet3D*. This marks a substantial advancement in publicly accessible datasets for the segmentation of the inferior alveolar canal. The *Maxillo* dataset has been later extended with the 2023 MICCAI ToothFairy Challenge.²

Additionally, in [30], Usman *et al.* proposed a two-stage approach also based on the U-Net architecture. On the hypothesis that the predominant challenge in segmenting the inferior alveolar canal relates to the class imbalance between the mandibular canal and the background, they initially apply a CNN to isolate the regions of the input volume where the canal is likely to be located, reducing background interference. Then, leveraging U-Net architecture, the segmentation of the mandibular canal is performed exclusively within the extracted regions.

The latest approach tested on public data is contributed by Zhao *et al.* [34] and, similarly to [30], it works in a two-stage fashion. Firstly, the mandibular centerline is extracted via automatic segmentation of the mandible and localization of the mandibular and mental foramen. The sub-volumes containing the mandibular canal information are then obtained using a double reflection method based on the Frenet frame. Secondly, the extracted sub-volumes are fed into a U-Net-based 3D segmentation network, and the topology of the mandibular canal is constrained with the cIDice. To conclude the segmentation process, the prediction masks are inversely transformed back into the original CBCT images.

2.1 Patch-Based Learning

All of the aforementioned solutions employ a patch-based learning strategy. Indeed, when targeting complex, high-dimensional inputs or when the computation resources available are limited or should be kept so, patch-based learning is the only viable approach. The segmentation or classification in whole-slide images, as well as the segmentation of anatomical structures in 3D volumes, are noticeable medical imaging applications requiring such a kind of learning procedure. Indeed, feeding a neural network with gigapixel images or hundreds of millions of voxels coming from 3D volumes is not a feasible approach.

To meet memory constraints, the simple downsampling of the input data is counterproductive whenever the preservation of fine-grained details is crucial. A common approach consists of training neural networks using subsets extracted from the original data [3, 12]. Such an approach, known as patch-based learning, mitigates memory constraints but also leads to a loss of global information due to restricted (patch-limited) receptive fields. Moreover, ambiguity in segmenting objects situated at the intersections of multiple patches may arise, causing

² <https://toothfairy.grand-challenge.org>

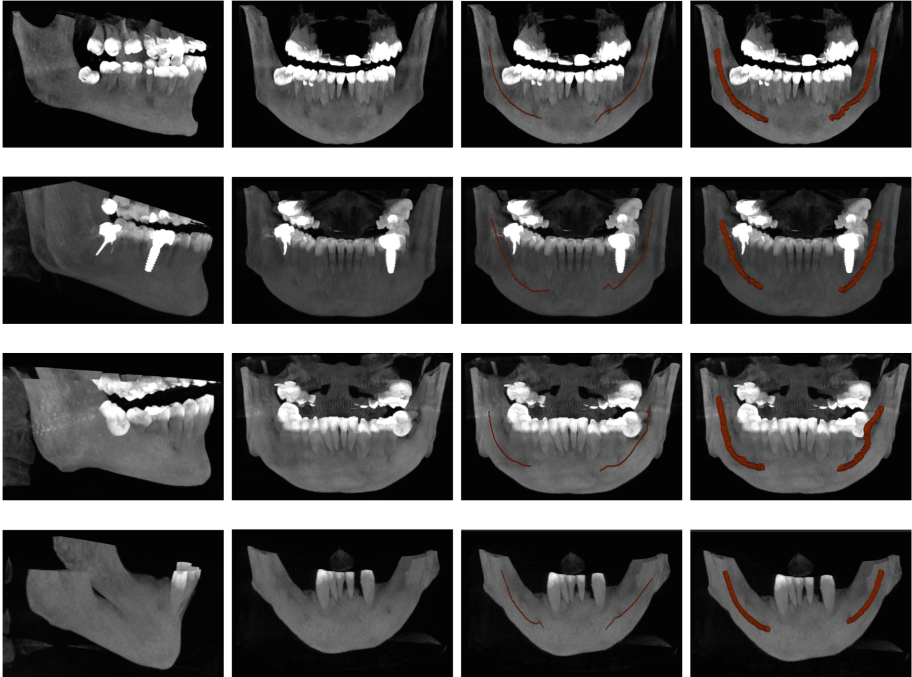


Fig. 2. Sample data from the ToothFairy dataset. Each line of the image contain a different patient, from left to right you can see left-side and frontal views of the CBCT volume, sparse and dense annotations of the inferior alveolar nerve.

potential artifacts around patch boundaries. When the object to be segmented is small in comparison to the entire volume, as it happens in the segmentation of the IAC, the aforementioned challenges become particularly prominent.

A first proposal to overcome the patch-based learning drawbacks in the segmentation of the IAN is introduced in [5] with the *PosPadUNet3D*. The authors suggested leveraging the positional information from the coordinates of extracted patches by simply projecting and concatenating these coordinates within the network bottleneck. Although this approach demonstrated some improvements in performance, the aforementioned major issues still persisted. Unlike *PosPadUNet3D*, our approach harnesses the information flow of Transformers, semantically conditioning the bottleneck embedding based on the spatial information instead of a simple feature concatenation.

3 Dataset

The maxillofacial dataset employed in our experiments is an improved version of the *Maxillo* dataset introduced by Cipriano *et al.* [4]. Such an improvement,

known as *ToothFairy* dataset, was part of the homonymous MICCAI 2023 challenge hosted on the *Grand Challenge* platform.³

All of the 3D CBCT volumes of the ToothFairy dataset were collected from the Affidea center in Modena, Italy, part of a leading pan-European healthcare group specializing in advanced diagnostics, outpatient services, laboratory analyses, physiotherapy and rehabilitation, and cancer diagnosis and treatment. The scans were acquired using the NewTom/NTVGiMK4 CBCT device, with acquisition parameters set at 3 mA, 110 kV, and 0.3 mm cubic voxels. The dataset is publicly available after user registration:⁴ such availability, along with the public release of the source code, ensures full reproducibility of our experiments and verification of our claims.

The annotation process was initially performed by diagnostic technicians responsible for the examinations, providing what we refer to as “sparse annotations” (Fig. 1b): the upper boundary of the canal is marked along the entire dental arch, offering a useful approximation of the nerve position. Such annotations are performed on 2D panoramic views of the jawbone and are routinely used in surgical practice to measure the height and depth of implant placement sites, thereby avoiding injuries to the inferior alveolar nerve.

Instead, the 3D annotations (in the following also referred to as “dense annotations”) of the ToothFairy dataset (Fig. 1a) have been created using an updated version of the IACAT tool [18], specifically version 2.0 developed in [17], by a team of medical experts with over five years of experience in the maxillofacial field.

All of the 443 volumes in the ToothFairy dataset are paired with the 2D sparse annotation. For a subset of 153 scans, the 3D dense annotation is also provided. For what concerns volume shapes, the average size in the dataset is $169 \times 342 \times 370$, while minimum and maximum volumes have respectively $148 \times 265 \times 312$ and $178 \times 423 \times 463$ dimensions. Sample images of the employed dataset are reported in Fig. 2.

4 Methods

This paper proposes a novel U-Net-based deep learning model for the segmentation of the IAC. Specifically, we devise a module that harnesses memory-augmented Transformer layers for modeling long-range interactions and integrating absolute positional information to mitigate issues related to patch-based learning. All the details concerning our proposed methodology can be found in Sect. 4.1.

In our work, a two-step training procedure is employed to exploit both volumes that are annotated in 3D and those that are annotated only in 2D, improving overall segmentation performances. An in-depth explanation of this training procedure can be found in Sect. 4.2.

Finally, the Hann-based post-processing employed in our pipeline is described in Sect. 4.3.

³ <https://toothfairy.grand-challenge.org/>

⁴ <https://ditto.ing.unimore.it/toothfairy/>

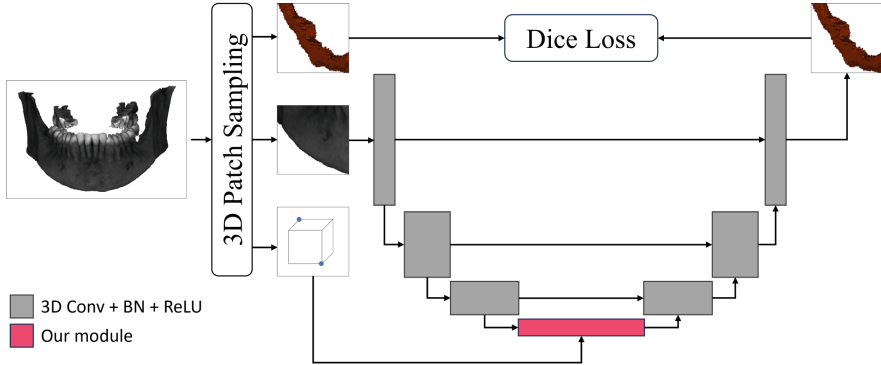


Fig. 3. Proposed Transformer module integrated in the standard 3D U-Net architecture. A detailed visualization of our module is reported in Fig. 4.

4.1 The Proposed Approach

We design a novel deep-learning model to address the limitations associated with patch-based learning through the utilization of Transformers capable of exploiting contextual information. More specifically, we propose a module based on Transformer encoder blocks, accompanied by learned embedding representations for positional encoding, and integrate it in the bottleneck of the well-known U-Net architecture (Fig. 3). By capitalizing on the inherent capacity of Transformers to model interactions between all pairs of elements within a given sequence, we aim to enhance the flow of information among the elements of the U-Net bottleneck. Moreover, we leverage this to effectively inject contextual information related to the processed patches.

In practice, we introduce a specialized token that captures the absolute position of the patch within the original volume, referred to as [ABS]. This is accomplished by projecting the 3D coordinates of two opposite corners of the patch into the bottleneck dimensional space, exploiting a learnable matrix of dimension $6 \times d_{model}$, where 6 are the numbers identifying the position of the patch within the entire volume and d_{model} is the number of channels in the U-Net bottleneck (Fig. 4). Subsequently, we concatenate this token with the remaining elements of the bottleneck, allowing its information to influence their representations through the Transformer encoder.

It is worth noticing that Transformers already employ positional encoding to describe the location of a token in a sequence. Such an encoding provides information about the position of (groups of) voxels within the current patch only. Instead, our [ABS] token encodes the position of a patch with respect to the entire volume. However, the inbuilt positional encoding of the Transformer architecture must not be applied to the [ABS] because all of the other tokens should be able to employ its information independently from their position. To achieve this goal, the Transformer inbuilt positional encoding is summed only to the tokens representing volume information. Again, this ensures that

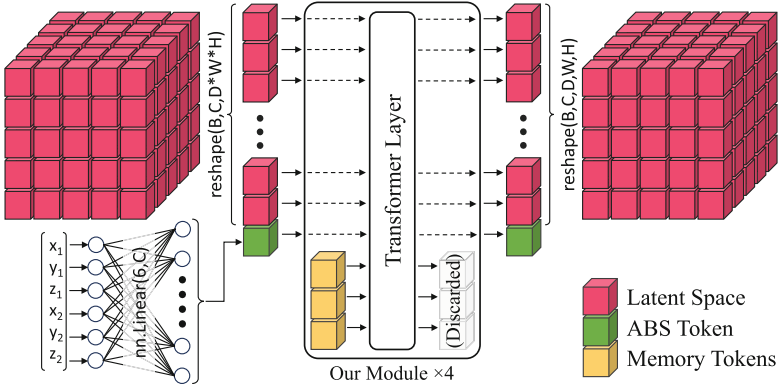


Fig. 4. The proposed module. B , C , D , W , and H represent respectively batch size, channels, depth, height, and width. The patch coordinates $[x_1, y_1, z_1, x_2, y_2, z_2]$ are projected using a linear layer to produce the [ABS] token. The activation map obtained in the bottleneck of U-Net before the first transposed convolution (pink blocks) is flattened across the spatial dimension and concatenated with the [ABS] token. The resulting tensor is fed to a cascade of four Transformer layers: for each layer a new set of memory token is concatenated to the input sequence and discarded from the output so that the sequence length does not vary. After the Transformer layers, the [ABS] token is removed and the remaining output is reshaped back to the original spatial dimensionality. (Color figure online)

the [ABS] token remains positionally untied from the rest of the sequence. This disentangled approach allows each element to pay attention to the special token’s information and vice versa, regardless of its position in the sequence.

Additionally, we enriched the proposed module with *memory*. The integration of Transformer memory has demonstrated considerable effectiveness in tasks such as image captioning [7]. This mechanism enables the Transformer to retain crucial prior concepts that may be challenging to be directly extracted from image features, but are nonetheless valuable for interpretation. Recognizing the applicability of this approach to the patch-based learning paradigm, wherein each patch is extracted from a wider context, we harness the power of Transformer memory to incorporate external information, thus enhancing the processing of individual patches. A graphical summary of this process is provided in Fig. 4.

4.2 Model Training

With the aim of also leveraging volumes with only 2D sparse annotations available, we adopt a two-step procedure composed of an initial step called “deep label expansion” or “generation phase” and a second one that consists of a standard segmentation training. In the deep label expansion, the network is trained using CBCT volumes paired with their corresponding sparse 2D labels to generate dense 3D annotations. Again, the rationale behind this operation is to obtain a model that can leverage the sparse 2D labels (available for all the volumes in

the Maxillo dataset) to create dense synthetic 3D annotations when they are not available.

The second step consists of merging the initial training set provided with “true” labels with the synthetically annotated CBCT volumes, generated by the deep label expansion. Thus, a total of 420 3D annotated volumes is obtained as a train set. Still, 8 and 15 volumes from the non-synthetic 3D dataset are available for the validation and test set respectively. Using the above-mentioned set of data, our segmentation model is trained to output 3D masks representing the inferior alveolar canal, and consequentially evaluated.

In other words, our pipeline leverage the proposed model twice, changing only the input data. A first instance is used to extend the amount of 3D IAC annotations by learning to “expand” the available 2D labels. The second instance is trained to predict a 3D segmentation of the IAC starting from a virgin scan. Test data of both instances are never seen during training.

4.3 Post-processing

Even if the proposed memory-augmented Transformer-based encoder with [ABS] token mitigates the lack of global information in patch-based learning and reduces the segmentation ambiguity on patch borders, we still need to deal with noise and artifacts generated at patch boundaries (Fig. 5). Taking inspiration from audio encoding [20], we introduced a post-processing algorithm based on the Hann windows function to tackle the presence of artifacts near patch edges. The Hann window function is defined as:

$$W_{\text{Hann}}(i) = \frac{1}{2} \left(1 - \cos \frac{2\pi i}{I} \right) \quad (1)$$

where i is an element in the considered interval I . This function is symmetric, peaking at 1 in the middle of the window and tapering to 0 at the edges. The sum of two Hann windows, each shifted by $\frac{I}{2}$ (50%), is equivalent to a rectangular window of width I and height 1:

$$W_{\text{Hann}}(i) + W_{\text{Hann}}\left(i + \frac{I}{2}\right) = 1 \quad (2)$$

Such a property is exploited in audio encoding to eliminate border artifacts by multiplying the Hann window with frames that overlap by 50%, before summing them together.

While this approach is defined in 1D for audio, we extended it to multiple dimensions, and applied it to the 3D segmented patches produced by our model. Thus, we are able to reduce the aforementioned noise on patch borders and improve the overall performance.

The effects of the proposed 3D extension of the Hann filtering to the segmented patches produced by our model are depicted in Fig. 5.

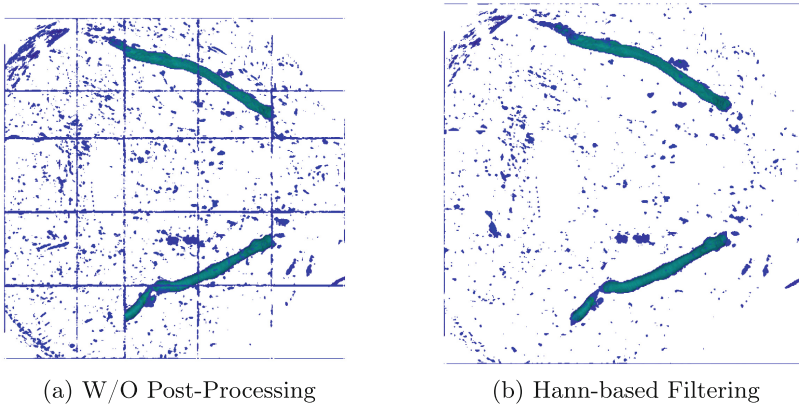


Fig. 5. Effects of Hann-based filtering on an axial plane extracted from a predicted volume. On the left (a), the prediction of our model without post-processing. On the right (b), the effect of the proposed Hann-based post-processing on the same model output. In both images, blue represents logits that have a value higher than 10^{-4} . The post-processing significantly reduces artifacts that appear close to patch borders. Even if most of these artifacts do not cause any issues, the ones that are close to the IAC badly influence the final segmentation. (Color figure online)

5 Experiments and Results

Section 5.1 defines the details of the adopted patch-based learning procedure, alongside with our experimental setting. We compare our proposal with state-of-the-art models in Sect. 5.2 and conduct an ablation study to highlight the contribution of the absolute token [ABS] and the memory of the Transformer in Sect. 5.3. Finally, Sect. 5.4 provides some visualizations of our model predictions, discussing its strengths and weaknesses.

5.1 Experimental Setting

Since we adopted a patch-based learning approach, we fed our model with patches of $120 \times 120 \times 120$ instead of the entire volume as a whole. During training we extracted patches with random uniform sampling, while during inference patches are extracted with an overlap of 50% in all the dimensions.

For what concerns the hardware resources, we trained our model in a distributed fashion, exploiting two NVIDIA Quadro RTX 5000 GPUs. The time needed for a complete training is approximately 16h with a batch size of 2.

5.2 Comparison with the State of The Art

In order to compare our proposal with the latest advances in the segmentation of the inferior alveolar nerve [30,34], Table 1 is provided. Both [30,34] leverage a two-stage approach that aims at filtering out background data before actually

Table 1. Comparison of our proposed model with the state of the art on IAC segmentation.

Dataset	Method	IoU	Dice
Maxillo	Usman <i>et al.</i> [30]	–	0.770
	Cripriano <i>et al.</i> [5]	0.650	0.790
	Zhao <i>et al.</i> [34]	–	0.810
	Ours	0.704	0.824
ToothFairy	Ours	0.710	0.831

performing the canal segmentation. In doing so, [30] makes use of a CNN-based approach that performs worse than both the positional encoding proposed in [5], and the non-deep two-stage approach based on the Frenet frame described in [34]. In Table 1 the proposed (complete) model is trained from scratch by means of both 3D “true” label and synthetically generated labels obtained from the deep label expansion phase. For a fair comparison, we performed the training twice, using only the Maxillo dataset (the dataset employed by competitors) and the complete ToothFairy dataset (our reference dataset). The test set of the two datasets matches, being one the extension of the other. The comparative evaluation provided confirms that our proposal outperforms the state-of-the-art competitors on the public dataset, by setting a new upper bound for IAC segmentation.

5.3 On the Effectiveness of the ABS Token and Memory

To showcase the contribution of each model component, we perform our evaluation by progressively including them in Table 2. We performed 10 experiments for each setup,⁵ but focused only on the deep label expansion phase of the training, thus limiting the number of experiments without losing generality in the conclusion raised (Sect. 4.2). It is worth noticing that any improvement in the deep label expansion step will benefit the whole segmentation pipeline. Moreover, since the model employed in the two phases is the same, the contributions of each proposed component can already be inferred during the generation phase.

At first glance, the comparison between the first two table lines might imply a lack of efficacy of the Transformer architecture. However, it is crucial to note that PosPadUNet3D incorporates absolute positional information from the original volume, which is not the case for TransPosPadUNet3D, which simply relies on a Transformer module introduced in the bottleneck of the U-Net architecture.

Introducing the [ABS] token to TransPosPadUNet3D (third line of Table 2) enhances its performance, already improving with respect to PosPadUNet3D and consistently demonstrating the effect of the proposed ABS token. Furthermore, the performance of TransPosPadUNet3D shows a progressive improvement, initially with the integration of memory tokens, and subsequently through the

⁵ Experiments on the same setup differ only in the initialization seed.

Table 2. Contribution of the modules composing our proposal, considering only the generation phase of our training procedure.

Method	Transf. ABS	Token Memory	Hann	Window	Dice
PosPadUNet3D	✗	✗	0	✗	0.797 ± 0.006
TransPosPadUNet3D	✓	✗	0	✗	0.796 ± 0.009
TransPosPadUNet3D	✓	✓	0	✗	0.801 ± 0.005
TransPosPadUNet3D	✓	✗	128	✗	0.800 ± 0.011
TransPosPadUNet3D	✓	✓	128	✗	0.802 ± 0.004
Ours (Complete)	✓	✓	128	✓	0.809 ± 0.004

application of the Hann Windows function as a post-processing strategy. Ultimately, the implementation of the [ABS] token results in a halved standard deviation, thereby supporting the robustness of the proposed model.

5.4 Qualitative Evaluation

A qualitative evaluation of the predictions obtained using our proposed model is provided in Fig. 6, where five pairs of automatic segmentations are coupled with

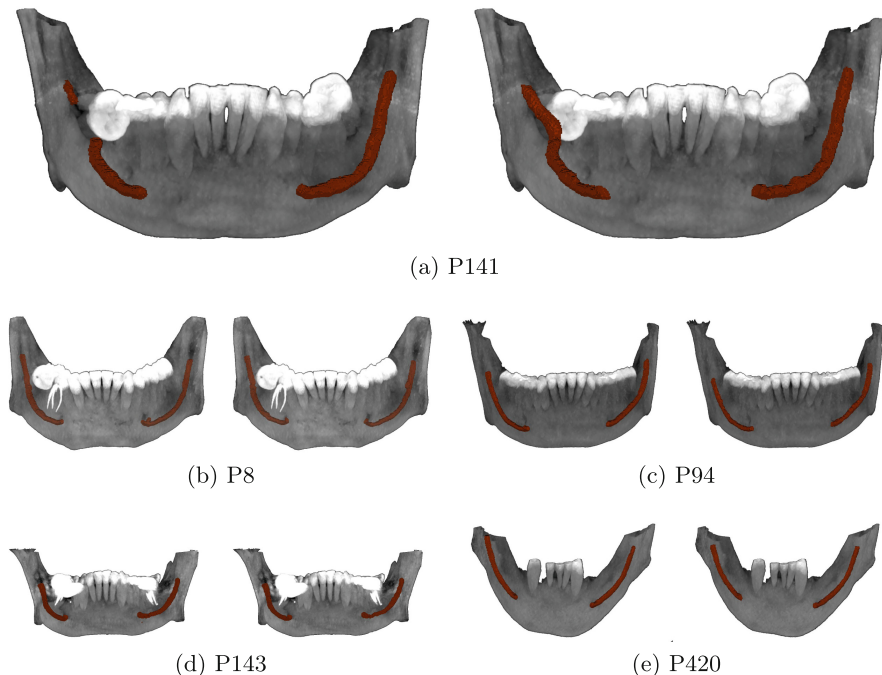


Fig. 6. Segmentation predictions proposed by our model (left) and corresponding ground-truths annotation (right) on examples taken from the ToothFairy public test set. The jaws face the camera view, thus the canal on the left side is the right IAC.

their corresponding ground-truth annotations. Sample data are taken from the public test case of the ToothFairy dataset.

While the majority of the predictions are exceptionally accurate and worth to be integrated in the daily clinical practice, a notable edge case is observed in the sample P141, where the canal on the left is heavily affected by the presence of a wisdom tooth, making it one of the hardest to be predicted. In this instance, our model’s prediction resulted in a non-continuous canal. Further improvements to our model may involve techniques to deal with such a kind of issues.

6 Discussion and Conclusion

One of the primary challenges associated with patch-based learning is the limited context available when modeling patches extracted from the original objects. In order to address this limitation, we propose an innovative approach by incorporating a transformer encoder with memory into the U-Net architecture, along with the introduction of the [ABS] token. Specifically, the [ABS] token is designed to embed the absolute position information of the processed patch within the original volume. By sharing this positional information with other elements within the bottleneck of the U-Net architecture, we are able to enhance the contextual understanding of the patches during the segmentation process and improve overall performance.

Moreover, our transformer encoder is equipped with memory tokens, which serve to store essential and generalized information pertaining to all patches. This stored information can be particularly valuable for the segmentation task, as it may be difficult to be directly retrieved from each patch singularly. By leveraging the transformer encoder with memory and the [ABS] token, our proposed method seeks to address the contextual information challenge in patch-based learning, improving the segmentation performance within the U-Net architecture.

To ensure the reproducibility of our experiments, we have made the described pipelines openly accessible to the scientific community as an open-source project. Furthermore, we conducted our experiments on public datasets, encouraging the broader scientific community to further enhance the results in the context of inferior alveolar canal segmentation and letting anyone reproduce the obtained results and verify our claims. Such a collaborative effort is crucial in critical medical domains to foster progress and innovation.

Future Work. While the suggested approach has proven effective in refining IAC segmentation, it could be adapted and potentially applied to any tasks where feeding an entire sample into the network is impractical, but having a global context is important. Future works will focus on studying the versatility of our proposed method, which will open doors to a broad range of applications beyond IAC segmentation. This will offer a promising research direction for further investigation into its performance across diverse neural networks, datasets, and data modalities.

Acknowledgements. This work was supported by the University of Modena and Reggio Emilia and Fondazione di Modena, through the FAR 2023 and FARD-2024 funds (Fondo di Ateneo per la Ricerca).

References







1. Abdolali, F., Zoroofi, R.A., Abdolali, M., Yokota, F., Otake, Y., Sato, Y.: Automatic segmentation of mandibular canal in cone beam CT images using conditional statistical shape model and fast marching. *Int. J. Comput. Assist. Radiol. Surg.* **12**(4), 581–593 (2017)
2. Blacher, J., Van DaHuvel, S., Parashar, V., Mitchell, J.C.: Variation in Location of the Mandibular Foramen/Inferior Alveolar Nerve Complex Given Anatomic Landmarks Using Cone-beam Computed Tomographic Scans. *J. Endodontics* **42**(3), 393–396 (2016)
3. Bontempo, G., Porrello, A., Bolelli, F., Calderara, S., Ficarra, E.: DAS-MIL: distilling Across Scales for MIL classification of histological WSIs. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 248–258. Springer (2023)
4. Cipriano, M., Allegretti, S., Bolelli, F., Di Bartolomeo, M., Pollastri, F., Pellacani, A., Minafra, P., Anesi, A., Grana, C.: Deep segmentation of the mandibular canal: a new 3D annotated dataset of CBCT volumes. *IEEE Access* **10**, 11500–11510 (2022)
5. Cipriano, M., Allegretti, S., Bolelli, F., Pollastri, F., Grana, C.: Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21137–21146. IEEE (2022)
6. Cornia, M., Baraldi, L., Cucchiara, R.: Explaining transformer-based image captioning models: an empirical analysis. *AI Commun.* **35**(2), 111–129 (2022)
7. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10578–10587 (2020)
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)* (2020)
9. Hatamizadeh, A., et al.: UNETR: transformers for 3D Medical Image Segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 574–584 (2022)
10. Hattab, J., et al.: Scoring Enzootic Pneumonia-like Lesions in Slaughtered Pigs: Traditional vs. Artificial-Intelligence-Based Methods. *Pathogens* **12**(12), 1460 (2023)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
12. Jaskari, J., et al.: Deep learning method for mandibular canal segmentation in dental cone beam computed tomography volumes. *Sci. Rep.* **10**(1), 1–8 (2020)
13. Kainmueller, D., Lamecker, H., Seim, H., Zinser, M., Zachow, S.: Automatic extraction of mandibular nerve and bone from cone-beam CT data. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 76–83. Springer (2009)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems* (2012)
15. Landi, F., Baraldi, L., Corsini, M., Cucchiara, R.: Embodied vision-and-language navigation with dynamic convolutional filters. In: *Proceedings of the 30th British Machine Vision Conference* (2019)

16. Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C.: Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. *IEEE Access* (2024)
17. Lumetti, L., Pipoli, V., Bolelli, F., Grana, C.: Annotating the inferior alveolar canal: the ultimate tool. In: *International Conference on Image Analysis and Processing*, pp. 525–536. Springer (2023)
18. Mercadante, C., Cipriano, M., Bolelli, F., Pollastri, F., Di Bartolomeo, M., Anesi, A., Grana, C.: A cone beam computed tomography annotation tool for automatic detection of the inferior alveolar nerve canal. In: *16th International Conference on Computer Vision Theory and Applications-VISAPP 2021*, vol. 4, pp. 724–731. SciTePress (2021)
19. Moris, B., Claesen, L.J.M., Sun, Y., Politis, C.: Automated tracking of the mandibular canal in CBCT images using matching and multiple hypotheses methods. *2012 Fourth International Conference on Communications and Electronics (ICCE)*, pp. 327–332 (2012)
20. Pielawski, N., Wählby, C.: Introducing Hann windows for reducing edge-effects in patch-based image segmentation. *PLoS ONE* **15**(3), e0229839 (2020)
21. Pipoli, V., Cappelli, M., Palladini, A., Peluso, C., Lovino, M., Ficarra, E.: Predicting gene expression levels from dna sequences and post-transcriptional information with transformers. *Comput. Methods Programs Biomed.* **225**, 107035 (2022)
22. Pollastri, F., Cipriano, M., Bolelli, F., Grana, C.: Long-range 3D self-attention for MRI prostate segmentation. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE (2022)
23. Pollastri, F., Maroñas, J., Bolelli, F., Ligabue, G., Paredes, R., Magistroni, R., Grana, C.: Confidence calibration for deep renal biopsy immunofluorescence image classification. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE (2021)
24. Pollastri, F., Parreño, M., Maroñas, J., Bolelli, F., Paredes, R., Ramos, D., Grana, C.: A deep analysis on high resolution dermoscopic image classification. *IET Comput. Vision* **15**(7), 514–526 (2021)
25. Porrello, A., et al.: Spotting insects from satellites: modeling the presence of *Culicoides imicola* through deep CNNs. In: *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pp. 159–166. IEEE (2019)
26. Roberti, I., Lovino, M., Di Cataldo, S., Ficarra, E., Urgese, G.: Exploiting gene expression profiles for the automated prediction of connectivity between brain regions. *Int. J. Mol. Sci.* **20**(8), 2035 (2019)
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer Assisted Intervention*. vol. 9351, pp. 234–241 (2015)
28. Stefanini, M., Lovino, M., Cucchiara, R., Ficarra, E.: Predicting gene and protein expression levels from DNA and protein sequences with Perceiver. *Comput. Methods Programs Biomed.* **234**, 107504 (2023)
29. Tang, Y., et al.: Self-supervised pre-training of swin transformers for 3D medical image analysis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740 (2022)
30. Usman, M., et al.: Dual-stage deeply supervised attention-based convolutional neural networks for mandibular canal segmentation in CBCT Scans. *Sensors* **22**(24), 9877 (2022)
31. Vaswani, A., et al.: Attention Is All You Need. *Advances in Neural Information Processing Systems (NIPS)* **30** (2017)

32. Wei, X., Wang, Y.: Inferior alveolar canal segmentation based on cone-beam computed tomography. *Medical Physics* (2021)
33. Worthington, P.: Injury of the inferior alveolar nerve during implant placement: a literature review. *Int. J. Oral Maxillofacial Implants* **19**(5) (2004)
34. Zhao, H., Chen, J., Yun, Z., Feng, Q., Zhong, L., Yang, W.: Whole mandibular canal segmentation using transformed dental CBCT volume in Frenet frame. *Heliyon* **9**(7) (2023)



Adaptive Class Learning to Screen Diabetic Disorders in Fundus Images of Eye

Shramana Dey¹(✉) , Pallabi Dutta¹ , Riddhasree Bhattacharyya¹ ,
Surochita Pal Das¹ , Sushmita Mitra¹ , and Rajiv Raman² 

¹ Machine Intelligence Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, West Bengal, India

shramanadey96@gmail.com

² Shri Bhagwan Mahavir Vitreoretinal Services, Sankara Nethralaya, Chennai, India

Abstract. The prevalence of ocular illnesses is growing globally, presenting a substantial public health challenge. Early detection and timely intervention are crucial for averting visual impairment and enhancing patient prognosis. This research introduces a new framework called *Class Extension with Limited Data (CELD)* to train a classifier to categorize retinal fundus images. The classifier is initially trained to identify relevant features concerning *Healthy* and *Diabetic Retinopathy (DR)* classes and later fine-tuned to adapt to the task of classifying the input images into three classes, *viz. Healthy, DR and Glaucoma*. This strategy allows the model to gradually enhance its classification capabilities, which is beneficial in situations where there are only a limited number of labeled datasets available. Perturbation methods are also used to identify the input image characteristics responsible for influencing the model's decision-making process. We achieve an overall accuracy of 91% on publicly available datasets.

Keywords: Fundus image · Class Extension · Data scarcity · Explainability

1 Introduction

With the rapid growth in the global population, the number of individuals diagnosed with diabetes is increasing at an alarming rate. Diabetes, a metabolic disorder characterized by high blood sugar levels, often leads to various visual impairments. Diabetic individuals must undergo regular eye screening due to the strong correlation between diabetes and eye abnormalities. A significant

This research was supported by the J. C. Bose National Fellowship, grant no. JCB/2020/000033 of S. Mitra.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15328, pp. 124–137, 2025.
https://doi.org/10.1007/978-3-031-78104-9_9

challenge lies in the shortage of trained eye care professionals, which hampers effective screening and treatment [1]. Diabetes is a precursor for several vision-threatening diseases, notably diabetic retinopathy (DR) and Glaucoma [3]. Approximately one-third of diabetics are likely to develop DR, and those with diabetes are twice as likely to be afflicted by Glaucoma as compared to non-diabetic individuals [15]. Both DR and Glaucoma often progress silently until significant vision loss occurs, potentially leading to irreversible blindness.

Regular, automated, non-invasive screening is crucial for early detection. This helps prevent the progression of these diseases in the preliminary stage, especially in remote regions with limited access to trained professionals. These technologies can help identify at-risk individuals early, allowing timely intervention [5]. Deep learning (DL) [12] is a popular choice in smart healthcare, for automating the screening process. Feature extraction with minimal human assistance, scalability, and high output efficiency, are some of the key factors attributed to the acceptability of deep learning methodologies in healthcare. This helps make the screening process efficient in terms of time, while coping with the limited number of trained professionals.

Several studies have utilized deep learning models for DR classification. For instance, the multi-resolution convolutional attention network (MuR-CAN) [14] emphasizes discriminative features using a multi-dilation attention block, with depth-wise convolution layers at various dilation rates to capture multi-scale spatial information. The DRNet13 [18] has been developed for automated DR stage classification. The Modified Generative Adversarial-based Crossover Salp Grasshopper (MGA-CSG) [16] approach predicts and classifies diabetic retinal diseases using fundus image datasets. Research has also focused on analyzing multiple pathologies from eye fundus images [7], using various CNN-based models like LeNet, AlexNet, Inception, VGG, and ResNet, for diagnosing Glaucoma and DR. Vision transformers [4] have also been employed for ocular disease detection and classification using fundus images.

Deep learning models require large volumes of annotated data to learn features for accurate prediction. Given the limited number of trained professionals, this becomes a major challenge. Consequently, automated detection of DR and Glaucoma, from eye fundus images, gets constrained by data scarcity issues. Most publicly available datasets provide retinal images without detailed labeling of the affected region(s). Transfer learning has been employed [2] to handle the scarcity of data by adapting weights from models trained on larger datasets. However, this often faces challenges like catastrophic forgetting and degraded performance due to domain shifts.

This paper proposes a framework - Class Extension with Limited Data (CELD), which trains classifiers to recognize additional (new) classes over time without forgetting relevant features from the previously learned classes. This framework is particularly useful in scenarios where new data classes get gradually incorporated. In real-life scenarios, as new and rarer ocular diseases are discovered or become more prevalent, it becomes necessary to update the diagnostic model to recognize these new conditions while still retaining the ability to

diagnose previously known diseases. The CELD framework addresses the data scarcity and imbalance issues prevalent in DR and Glaucoma classification when compared to healthy samples. Unlike transfer learning, which requires a substantial volume of data to fine-tune the model, the proposed framework updates the model as new data becomes available; without having to retrain from scratch. This approach allows the network to continuously learn from smaller, progressive batches; thereby, making it resource-efficient and scalable for dynamic environments. Several controlled data-perturbation techniques are incorporated to analyze the decision-making process of our model. This adds explainability to address the significance of each input attribute towards model behavior. The key contributions of the research are listed below.

- The class adaptation in CELD progresses incrementally. A deep neural network is first trained to classify fundus images into healthy and DR classes. It is then extended to additionally classify Glaucoma, to transform to a three-class learning model.
- The CELD framework prevents catastrophic forgetting of previous learning, while leveraging existing knowledge to learn new classes in the presence of limited data.
- Detailed empirical study and analysis of the CELD framework establishes its robustness to data and use of fewer computational resources.
- Feature relevance is explored, through data perturbation, to analytically observe changes in model performance.

The remaining sections of the paper are organized as follows. Section 2 describes the CELD framework. Section 3 outlines the experimental results to study the performance of our proposed framework CELD. Finally Sect. 4 concludes the article.

2 Methodology

A significant challenge in the task of retinal image classification is the limited availability of annotated data, which constrains the generalizability of the model. Specifically, there is a disproportionate ratio of healthy and DR data, with Glaucoma data being even scarcer. This imbalance complicates the task of improving classification accuracy. While data augmentation might intuitively address this issue, it risks overfitting, resulting in an inefficient model [6]. To effectively manage this data imbalance, the proposed CELD framework helps to classify fundus images as healthy, DR-affected, or Glaucoma-affected. Additionally, we evaluate the model’s decision-making process using explainability methods based on perturbation techniques. A schematic workflow is provided in Fig. 1.

The objective of this study is to develop a detection system for retinal color fundus images of three classes: healthy, DR, and Glaucoma. The subsequent parts of this section provide a detailed explanation of the classifier architecture, the proposed CELD framework, and the perturbation-based explainable methods used to gain insights regarding the decision-making process of the model.

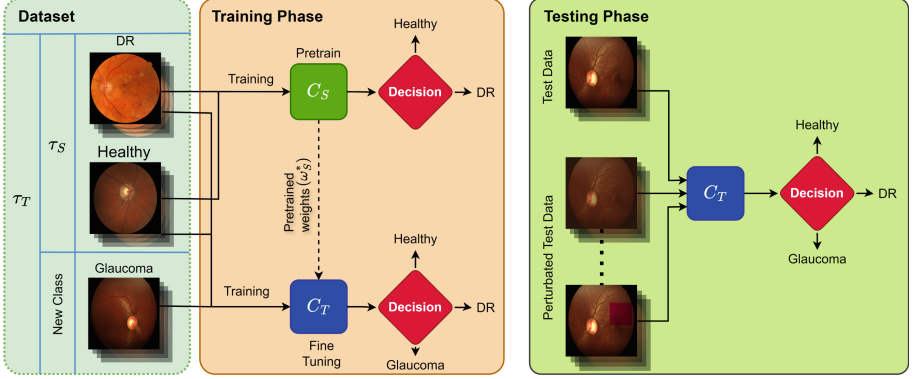


Fig. 1. The basic workflow of the proposed CELD-Framework

2.1 Class Extension with Limited Data (CELD)

A deep learning model tends to experience “catastrophic forgetting”, losing previously learned patterns when trained on new data distributions [19]. In contrast, humans have an inherent ability to learn new skill over the time without forgetting prior knowledge. In this work, the proposed CELD framework exploits this notion of natural learning ability by retaining the knowledge acquired from previously learned classes to enable the network to adapt to new class. This reduces the requirement for extensive datasets for each incoming new class. This makes it highly suitable for real-world scenarios characterized by limited data availability, such as the classification of retinal fundus images, where obtaining extensive labeled datasets is often a significant challenge. Employing models pre-trained on the ImageNet dataset and subsequently fine-tuning them for adapting to new tasks in the medical domain may lead to suboptimal performance due to the inherent differences in data distribution between natural images and medical images [8]. CELD framework mitigates the issue of performance degradation caused by domain shift since the datasets for source and target tasks belong to the same domain. In this work the source task is to train the classifier for categorizing healthy and DR images and the target task is defined by adapting the classifier from the source task to categorize the input fundus images into DR, Glaucoma and healthy category.

Formally, τ_S and τ_T represent the datasets for source and target task respectively, and $\tau_S \subset \tau_T \subset \tau$, where τ represents the universal domain of retinal fundus images. A classifier C_S , parameterized by a set of parameters ω_S , is initially trained on source data $(x_i, y_i) \in \tau_S$. Here $x_i \in \chi_S$ represent the input images and the corresponding labels $y_i \in \mathcal{Y}_S$ where $\mathcal{Y}_S = \{Healthy, DR\}$.

$$C_S : \chi_S \mapsto \mathcal{Y}_S \quad (1)$$

$$\omega_S^* = \arg \min_{\omega_S} \sum_{i=1}^M \mathcal{L}(\hat{y}_i = C_S(x_i; \omega_S), y_i) \quad (2)$$

Here \mathcal{L} is the loss function to be minimized and $M = |\tau_S|$ during training the classifier C_S . The optimized weights from the trained classifier C_S , ω_S^* are then used to initialize a new classifier C_T which classifies $(x_k, y_k) \in \tau_T$ where $x_k \in \chi_T$ represent the input images and the corresponding labels $y_k \in \mathcal{Y}_T$ where $\mathcal{Y}_T = \{Healthy, DR, Glaucoma\}$ with $N = |\tau_T|$. Subsequently, C_T is fine-tuned on the extended dataset.

$$C_T : \chi_T \mapsto \mathcal{Y}_T \quad (3)$$

$$\omega_T^* = \arg \min_{\omega_T} \sum_{i=1}^N \mathcal{L}(\hat{y}_k = C_T(x_k; \omega_T), y_k) \quad (4)$$

Here, ω_T^* is the updated weight of C_T after training on τ_T . The loss function during the incremental learning phase can be defined as:

$$\mathcal{L}(\tau_T; \omega_T) = \mathbb{E}_{(x,y) \sim \tau_T} \left[- \sum_{c=1}^C y_c \log \hat{y}_c \right] \quad (5)$$

where $(x, y) \sim \tau_T$ indicates that the input data x and the corresponding label y is drawn from the expanded dataset τ_T . y_c is the true label for class c and \hat{y}_c is the predicted label. $C = |\mathcal{Y}_T|$ represents the total number of classes in τ_T . The expectation \mathbb{E} denotes averaging over all samples in the dataset. This approach allows $C_T(\cdot)$ to retain patterns learned from τ_S while learning features relevant to the new class, thus improving the retention of previously learned knowledge and avoiding overfitting.

2.2 Classifier

This paper adapts DenseNet121 [10] as the backbone classifier based on the experimental results shown in Sect. 3. DenseNet121 is characterized by a dense connectivity pattern, where each convolutional layer receives inputs from all preceding layers within a dense block, thereby promoting efficient feature reuse and robust gradient flow. This ensures strong gradient signals even for the earliest layers during backpropagation [10]. The architecture consists of 121 convolutional layers, organized into four dense blocks and separated by three transition layers. The transition layers apply normalization, followed by a convolution and pooling operations to downsample the feature maps. Finally, the intermediate feature map obtained undergoes global average pooling and is fed to fully connected layers for classification. The dense connections reduce redundant parameters which lower model complexity and enable faster training of deeper networks [10]. This improved information flow helps to reduce overfitting, which is crucial for handling imbalanced data.

2.3 Perturbation Methods for Explainability

The black-box nature of deep neural networks hinders the understandability of how predictions are made by the model. This limits the usability of the AI

algorithms in critical scenarios like healthcare where the rationale behind the decision-making process of the model must align with the characteristics taken into account by the healthcare professionals. In the realm of deep neural networks, explainability is not just a desirable feature—it is a necessity. Without a clear understanding of how and why these complex models make decisions, particularly in medicine, we risk compromising trust and safety. To make the model more trustworthy and transparent, our framework uses perturbation techniques to identify the relevant characteristics of input data that influence the decision-making process of the model. An efficient model should learn from salient features rather than spurious information or noise that is present in the training data. Perturbation methods, being model-agnostic, allow dynamic analysis without requiring access to the model’s internal details. Techniques like applying occlusion masks or adding noise to image patches or pixels help in querying the model and developing test hypotheses on the fly. The main challenge is selecting appropriate perturbation techniques to analyze the model’s performance effectively.

The detection of DR requires identifying pathologies spread across various quadrants of the eye fundus image. In contrast, diagnosing Glaucoma necessitates a precise analysis of the optic disc, focusing on the cup-to-disc ratio. The formation of red and bright lesions are the two most common symptoms of DR. Research shows that the identification of red and bright lesions is most effectively done using the green channel of color fundus images [20]. Additionally, as the condition worsens, neovascularization, which involves the formation of new blood vessels, takes place. Neovascularization in advanced stages of DR can also impact the optic disc area, resulting in the formation of new blood vessels in the optic disc.

Based on the above insights gained regarding the relevant clinical features for DR and Glaucoma, we have designed perturbation techniques to further investigate the model’s decision-making process. Multiple controlled perturbations are applied to the test dataset and the performance of classifier C_T on this perturbed data was compared with the one obtained from the unperturbed test dataset. Two techniques were used to assess the influence of the green channel in the decision-making process: Reduce green (RG) which reduces the overall green channel weightage in comparison to the red and blue channels of color fundus images and Random green removal (RGR) which randomly removes segments of the green channel. Additional techniques like reducing image contrast (RC), adding Gaussian noise (GN) and applying edge sharpening (ES) were also used to study the impact of image quality on the model’s inference. A strategy namely Optic disk occlusion (ODC) was used to evaluate the relevance of the optic disc in the classification of Glaucoma and DR images. Figure 2 shows the different kind of perturbed images.

3 Experiments and Results

This section provides a comprehensive overview of the datasets used, the metric used to assess the classifier’s performance, the experimental setup, and the resulting experimental outcomes.

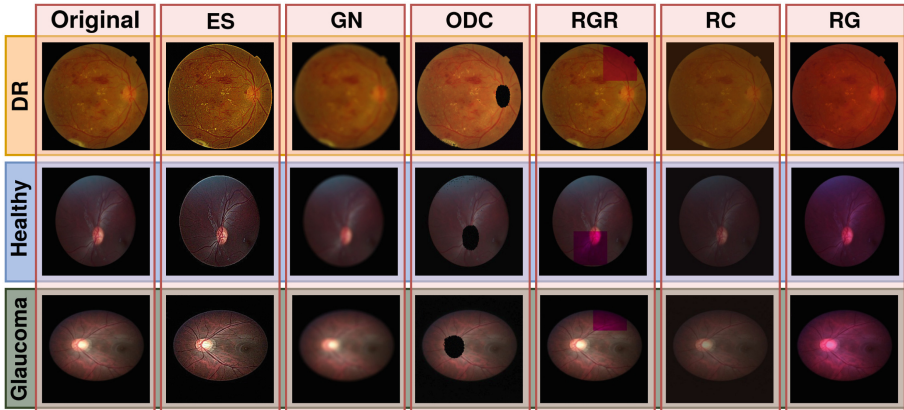


Fig. 2. The original image and its perturbed versions for each of the classes: DR, Healthy and Glaucoma.

3.1 Dataset

A total of 3,111 retinal color fundus images were obtained from three publicly available datasets: Messidor2¹, Chaksu [11], and LES-AV [17]. The Messidor2 dataset has 1,744 macula-centered RGB images. There are 1017 images belonging to the healthy class and 727 images belonging to the DR category. The Chaksu dataset comprises 1,345 images, with 188 images classified as Glaucoma and 1,157 images classified as healthy. These images were captured using three devices, including two non-mydratiac fundus cameras: the Remido non-mydratiac Fundus-on-Phone (FoP) and the Forus 3Nethra Classic non-mydratiac fundus camera and the Bosch handheld fundus camera. These images are Optic Disc-centered for Optic Disc assessment and Glaucoma detection. The LES-AV dataset has 22 images with 11 images categorized into Glaucoma and the remaining 11 images categorized into healthy category. The data details are given in Table 1.

Table 1. Pooled Dataset Details

Data	DR	Healthy	Glaucoma	Total
Messidor2	727	1017	-	1744
LES-AV	-	11	11	22
Chaksu	-	1157	188	1345
Overall Samples per Class	727	2185	199	3111

Data Pooling: To address data scarcity, data from three sources were combined to create a more diverse dataset, enabling the model to learn salient features

¹ <https://www.kaggle.com/datasets/mariaherrerot/messidor2preprocess>.

and generalize better. The inclusion of images from different populations and imaging conditions, along with an increased chance of capturing rare medical conditions, reduces model bias toward irrelevant features. After pooling from the three datasets, there are 2185 healthy, 727 DR, and 199 Glaucoma samples. All images were resized to dimensions of 256×256 from each dataset. The combined dataset was split into 80% for training, 10% for testing, and 10% for validation. To ensure accurate splits, the data was initially divided within each dataset before combining, ensuring that each train-test-validation split contained data from all sources. This approach was applied separately for healthy, DR, and Glaucoma-affected fundus images from each dataset within its split.

3.2 Experimental Setup

CELD is developed using Pytorch² and Monai³ on python 3.9 as the platform. All experiments were performed on a 12 GB NVIDIA Titan XP GPU. The initial learning rate was set to 10^{-5} . Early stopping is used to avoid over-fitting along with the AdamW optimizer [13]. A batch size of 8 is used in training.

3.3 Metric

The performance of the proposed framework was evaluated using accuracy, precision, recall, and F1-score. The mathematical definition of the listed metrics in terms of True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) is defined below.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.4 Result

The state-of-the-art (SOTA) models such as SeResNet101 [9], DenseNet121, and ViT were employed for classifying retinal images into Healthy, DR and Glaucoma, with their performance summarized in Fig. 3. DenseNet121 achieved the highest accuracy at 0.7910. However, all models exhibited poor performance in classifying Glaucoma and DR due to significant class imbalance.

² <https://pytorch.org/>.

³ <https://monai.io/>.

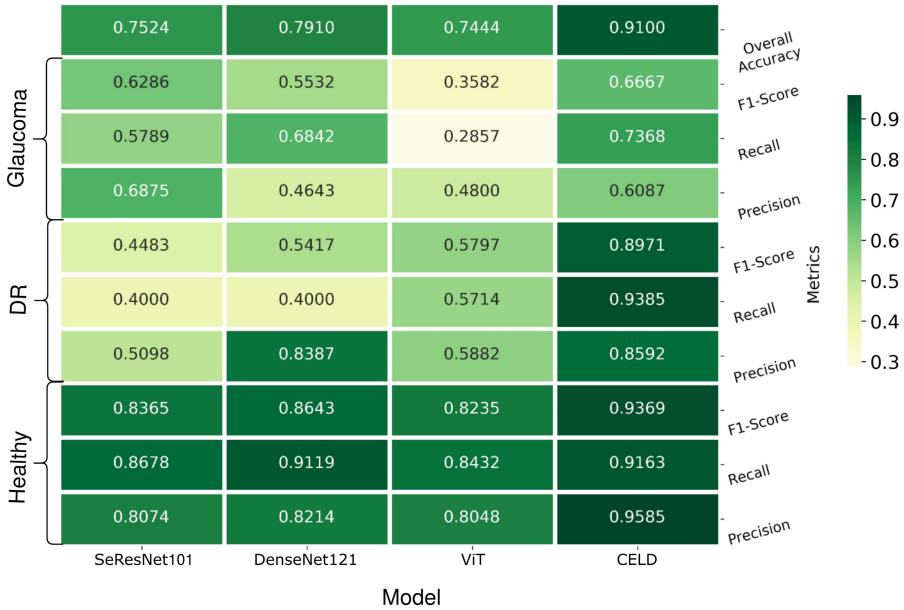


Fig. 3. Quantitative Result for 3 Class Classification.

Table 2. Quantitative Result for 2 Class Classification

Model	Classes						Overall Accuracy
	Healthy			DR			
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
SeResNet101	0.8264	0.8772	0.8511	0.7667	0.6866	0.7244	0.8066
DenseNet121	0.9027	0.8947	0.8987	0.8235	0.8358	0.8296	0.8729
ViT	0.8103	0.8246	0.8174	0.6923	0.6716	0.6818	0.7680

Further, the same models were tested, with results summarized in Table 2 where the classifiers are trained to categorize images into Healthy and DR only. Significant performance improvements were observed, particularly in the DR class, as indicated by balanced precision-recall scores. DenseNet121 outperformed the other models, achieving an overall accuracy of 0.8729 and F1-scores of 0.8987 for healthy and 0.8296 for DR, leading to its selection as the backbone architecture for the CELD framework.

For this three-class classification problem, the proposed CELD framework outperformed the state-of-the-art (SOTA) models, achieving an overall accuracy of 0.9100. The performance of the CELD framework has been listed in Fig. 3. It demonstrated significant improvement in the F1-scores for all classes, particularly for DR and Glaucoma. While the ViT model achieved a maximum F1-score of 0.5797 for DR, the CELD framework substantially improved this to 0.8971,

with a high yet balanced precision-recall. Similarly, SeResNet’s highest F1-score of 0.6286 for Glaucoma was improved to 0.6667 by the CELD framework. In medical image analysis, it is crucial for models to accurately detect positive cases, even if it occasionally results in false alarms. The CELD framework showed significant improvement in recall, albeit with a slight drop in precision for Glaucoma. Overall, the CELD framework significantly outperformed other models across various parameters.

Explaining CELD Framework with Data Perturbation: The performance of the proposed framework was evaluated using perturbed data and compared to unperturbed data, as summarized in Fig. 4. The corresponding confusion matrix is represented in Fig. 5. Reducing the weight of the green channel significantly decreased performance for DR and Glaucoma classifications, while the classification of healthy samples remained mostly unaffected. The confusion matrix shows increased mis-classification of DR and Glaucoma images as healthy when the green channel’s weight is negatively altered or partially removed. Notably, reducing the green channel’s weight across the entire image leads to higher mis-classification rates than randomly removing segments of the green channel. When random patches of green channel are removed, the classifier’s decision for the DR class is influenced by the remaining unperturbed data. The performance drop for the Glaucoma class is less significant, as the optic disc region often remains unperturbed in many images.

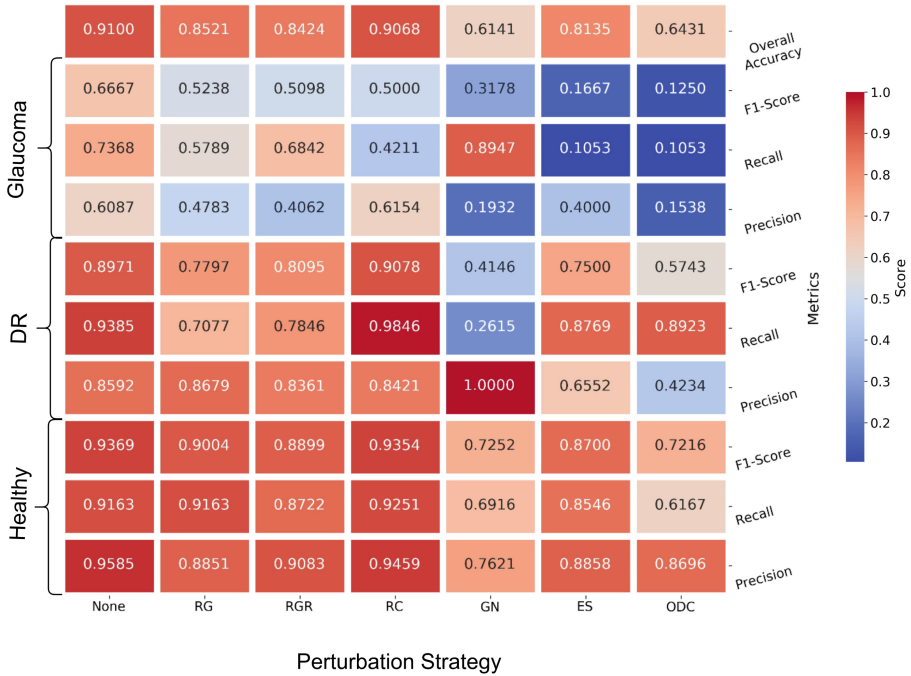


Fig. 4. Quantitative Result over CELD framework with Data Perturbation.

Reducing contrast does not significantly impact the classification of DR or healthy categories, as shown in the confusion matrix, but it does impair Glaucoma classification. Visually, this perturbation blurs the optic disc region, thus,

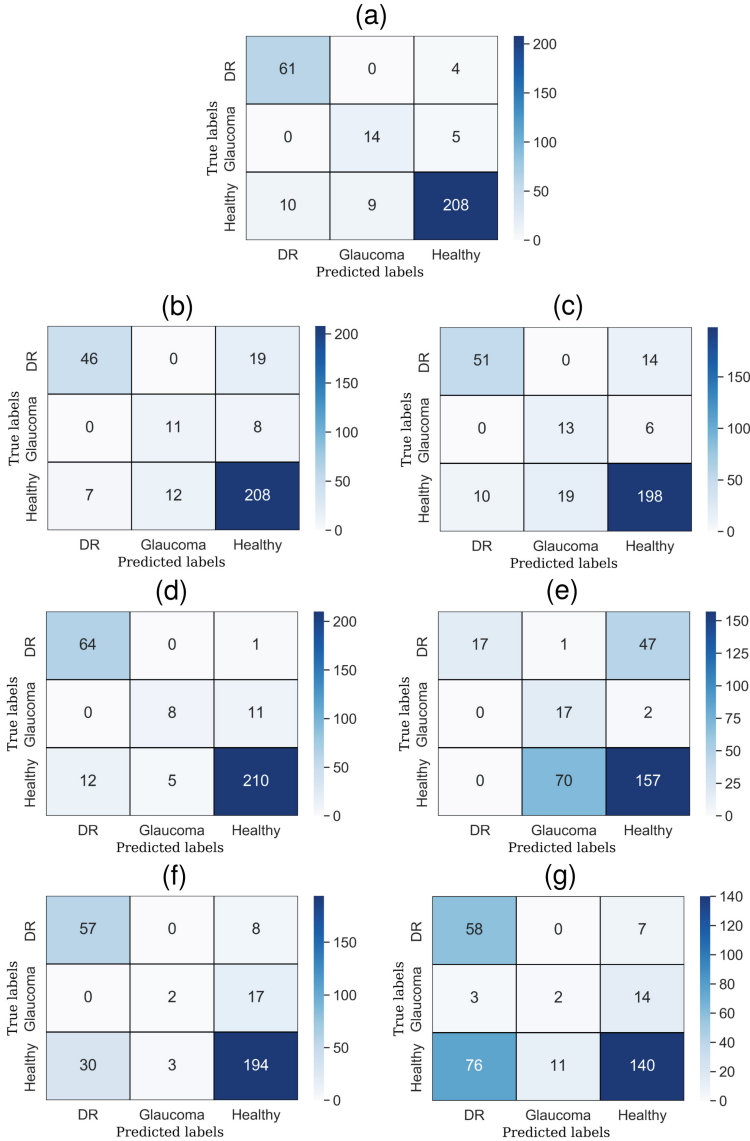


Fig. 5. Confusion matrix over CELD framework with Data Perturbation. The matrix shows performance of CELD (a) With no perturbation, (b) Reduce green (RG), (c) Random green removal (RGR), (d) Reducing image contrast (RC), (e) Gaussian noise (GN), (f) Edge sharpening (ES), (g) Optic disk occlusion (ODC) (Color figure online)

obscuring key features. Edge sharpening, which enhances pixels having high-intensity changes w.r.t its neighborhood, leads to a high mis-classification rate of Glaucoma as healthy, as reflected by the F1-score. Adding random Gaussian noise causes the model to falsely classify only one DR image as Glaucoma and most other DR images as healthy, resulting in high precision but a low F1-score for the DR class. It is important to note that in this study this the first perturbation strategy that generates ambiguous decisions between the two disease classes. The model's decision for Glaucoma is less affected by noise but becomes prone to classifying healthy images as Glaucoma, leading to high recall and low precision for the Glaucoma class. In summary, DR identification is challenging in poor-quality images, while Glaucoma can be diagnosed in noisy, low-quality images but not in those with poor contrast or excessive sharpening.

Observing the optic disc occlusion strategy reveals the model's high dependency on the optic disc for classifying Glaucoma and healthy eyes. These two classes exhibit high mis-classification rates, which is understandable since eye experts often diagnose Glaucoma by examining the optic disc region. For Glaucoma, the absence of relevant features leads to mis-classification, as shown in the confusion matrix and Fig. 4. Most mis-classified images are labeled as healthy, indicating the model relies on this feature for Glaucoma identification, thereby increasing the precision score for the normal class. The optic disc's features are crucial for determining eye health, reflected in the lower F1-score compared to unperturbed data for healthy eyes. For DR, performance is less affected since neovascularization in the optic disc is not always present in DR-affected images. However, there is an increase in false positives for the DR class due to the model's insufficient features for reliable decisions, resulting in high recall and low precision for the DR class. In conclusion, optic disc occlusion significantly impacts overall model accuracy, highlighting its importance as an input feature.

To conclude, the perturbation techniques revealed that the model heavily relies on the green channel and image quality for accurate classification, especially for DR and Glaucoma. Occlusion of the optic disc significantly impacts Glaucoma detection, emphasizing its critical role in the model's decision-making process.

4 Conclusion and Discussion

This research demonstrates the potential of deep neural networks to improve medical image classification, particularly for identifying conditions like diabetic retinopathy (DR) and Glaucoma from fundus images. Initially, we trained the network to differentiate between healthy and DR-affected images. Using the Class Extension with Limited Data (CELD) framework, we fine-tuned the model also to classify Glaucoma, transforming it into a three-class classifier. The CELD framework enables the model to maintain its performance on previously learned tasks while adapting to new classes while efficiently tackling data imbalance with minimal computational overhead and data requirements. Consequently, the model retains its proficiency in identifying DR while learning to classify Glaucoma, ensuring efficiency and resource-friendliness.

Our extensive empirical analysis compared the performance of two-class and three-class classifiers. The results highlighted that the Densenet121 architecture significantly improves classification accuracy, proving its suitability for this application. We conducted various experiments to assess accuracy and robustness, confirming the model's effectiveness. Additionally, we explored feature relevance through explainability using perturbed data. These studies provided insights into how changes in input data affect model performance, identifying the most critical features for accurate classification. The perturbation analysis summarized the robustness of the CELD framework. This approach represents a significant advancement in medical imaging and deep learning, providing an efficient method to expand model capabilities with limited data and computational resources. The CELD framework has the potential to be applied to diagnose a variety of other ocular diseases common in diabetic eyes.

References

1. Basu, S., Mitra, S.: Segmentation in diabetic retinopathy using deeply-supervised multiscalar attention. In: Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2614–2617. IEEE (2021)
2. Basu, S., Mitra, S., Saha, N.: Deep learning for screening covid-19 using chest x-ray images. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 2521–2527. IEEE (2020)
3. Bourne, R.R., Stevens, G.A., et al.: Causes of vision loss worldwide, 1990–2010: a systematic analysis. *Lancet Glob. Health* **1**, e339–e349 (2013)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
5. Fatima, M., Pachauri, P., et al.: Enhancing retinal disease diagnosis through AI: evaluating performance, ethical considerations, and clinical implementation. *Inf. Health* **1**, 57–69 (2024)
6. Garcea, F., Serra, A., et al.: Data augmentation for medical imaging: a systematic literature review. *Comput. Biol. Med.* **152**, 106391 (2023)
7. Grover, K.S., Kapoor, N.: Detection of glaucoma and diabetic retinopathy using fundus images and deep learning. In: Proceedings of the IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICC-CMLA), pp. 407–412. IEEE (2023)
8. He, K., Girshick, R., et al.: Rethinking ImageNet pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4918–4927 (2019)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
10. Huang, G., Liu, Z., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708 (2017)
11. Kumar, J.H., Seelamantula, C.S., et al.: Chákṣu: a glaucoma specific fundus image database. *Sci. Data* **10**, 70 (2023)
12. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**, 436–444 (2015)

13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019). <https://openreview.net/forum?id=Bkg6RiCqY7>
14. Madarapu, S., Ari, S., et al.: A multi-resolution convolutional attention network for efficient diabetic retinopathy classification. *Comput. Electr. Eng.* **117**, 109243 (2024)
15. Morya, A.K., Ramesh, P.V., et al.: Diabetes more than retinopathy, it's effect on the anterior segment of eye. *World J. Clin. Cases* **11**, 3736 (2023)
16. Navaneethan, R., Devarajan, H.: Enhancing diabetic retinopathy detection through preprocessing and feature extraction with MGA-CSG algorithm. *Expert Syst. Appl.* **249**, 123418 (2024)
17. Orlando, J.I., Barbosa Breda, J., et al.: Towards a glaucoma risk index based on simulated hemodynamics from fundus images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. pp. 65–73. Springer (2018)
18. Shamrat, F.J.M., Shakil, R., et al.: An advanced deep neural network for fundus image analysis and enhancing diabetic retinopathy detection. *Healthcare Analytics* **5**, 100303 (2024)
19. Van de Ven, G.M., Tuytelaars, T., et al.: Three types of incremental learning. *Nature Mach. Intell.* **4**, 1185–1197 (2022)
20. Walter, T., Massin, P., et al.: Automatic detection of microaneurysms in color fundus images. *Med. Image Anal.* **11**, 555–566 (2007)



EDB-Net: An Edge-Guided Dual-Branch Neural Network for Skin Cancer Classification

Amartya Ray¹, Soumyajit Gayen¹, Dmitrii Kaplun^{2,3},
and Ram Sarkar¹

¹ Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

² Artificial Intelligence Research Institute, China University of Mining and Technology, Xuzhou, China

³ Department of Automation and Control Processes, Saint Petersburg Electrotechnical University “LETI”, Saint Petersburg, Russian Federation
dikaplun@etu.ru

Abstract. Skin cancer poses a serious global health challenge, where timely and precise diagnosis is essential to improve patient outcomes. Recently, neural networks have proven to be highly effective tools for automated skin cancer classification, significantly advancing the field of dermatology. This paper introduces a novel approach to generate edge maps from dermoscopic images using a holistically nested edge detector model. These edge maps enhance the detection of shape and symmetry irregularities, which are key indicators of malignancy, and improve the focus on relevant regions of interest. We then propose an edge-guided dual-branch neural network, called EDB-Net, for the classification task. Branch 1 handles edge maps, while Branch 2 processes original dermoscopic images. To highlight significant regions and focus on specific lesion areas, we incorporate a novel channel-spatial synergistic attention block within Branch 2. Additionally, we introduce a unique strategy to modulate the generated attention maps using edge features extracted from the edge maps in Branch 1, creating edge-guided features that refine the overall feature representation. In the final stage, both edge-guided and attention-aided features are combined, producing more distinct and contextually relevant outputs, thereby significantly enhancing classification performance. Our model achieves accuracies of 0.927 and 0.848 on the challenging HAM10000 and ISIC 2016 datasets, respectively, without employing any data augmentation. The source code of the proposed model is available at: https://github.com/Cmatermedicalimageanalysis/EDB_Net.

Keywords: Skin Cancer · Lesion Classification · Neural Network · Dermoscopic Image · Edge Map · Attention Mechanism

1 Introduction

Skin cancer is among the most dangerous types of cancer worldwide, with its incidence increasing due to increased exposure to ultraviolet radiation from the sun. Periodic and intense exposure can cause sunburns, raising the risk of skin cancer. Additional risk factors include abnormal moles and a family history of the disease. The risk of developing skin cancer increases with age, making older individuals more vulnerable. Skin cancer is broadly categorized into melanoma and non-melanoma types. Despite being less common, melanoma is more likely to metastasize and is a significant cause of skin cancer-related deaths. Melanoma can appear on any part of the skin, but is often found in extensive sun-exposed areas such as the face, hands, and neck. Early detection and diagnosis are vital for effectively treating melanoma, as it spreads quickly and can result in severe and fatal consequences. Therefore, the artificial intelligence community in medical image analysis is dedicated to developing techniques for the early diagnosis of skin cancer.

Related Work: Classifying skin cancer is a complex task due to the diverse types of skin lesions. Visual features like color, pattern, shape, and texture can overlap between benign and malignant samples, making it difficult to distinguish between harmless moles and potentially cancerous ones. Furthermore, even among malignant lesions, there can be significant differences in appearance. From the literature survey, it has been observed that earlier machine learning-based techniques were predominantly used in the field of skin cancer classification, achieving limited success. However, the introduction of deep neural networks brought about rapid advancements in this area.

Kalouche et al. [1] employed a pre-trained deep convolutional neural network (CNN) architecture, VGG16, with the last three layers fine-tuned. They used a stochastic gradient descent (SGD) optimizer with a low learning rate for fine-tuning. Emara et al. [2] utilized the InceptionV4 backbone and enhanced it by incorporating feature reuse through a residual connection. This modification, which integrated features from earlier layers with those from higher-level layers, significantly improved classification performance. Iqbal et al. [3] proposed a deep CNN comprising 63 convolutional layers aimed at multi-class skin cancer classification. Although they accounted for inter-class similarities and intra-class variances, their model was not effective in addressing these challenges. Datta et al. [4] proposed a skin cancer classification model using InceptionResNetV2 as the backbone. They enhanced it with a soft attention unit that included a bilinear attention layer, which helped focus on small lesion areas and ignore artifacts by computing weighted feature maps.

Skin cancer datasets are often heavily imbalanced and to address this challenge, Shen et al. [5] employed a cost-effective and high-performance data augmentation strategy. They combined this approach with an EfficientNetB7 architecture to enhance automatic skin cancer screening in rural communities. Sarkar et al. [6] presented a novel classifier combination technique using the Dempster-Shafer theory for skin cancer classification. This approach significantly improved the recall rate for melanoma, the deadliest form of skin cancer. Leveraging deep

learning, researchers have made significant strides in addressing many challenges associated with skin cancer classification. However, despite these encouraging advancements, existing approaches still struggle to consistently deliver robust performance in real-world settings.

Motivation and Contributions: In dermatology, skin cancer classification heavily relies on evaluating asymmetry, border irregularity, color variation, and lesion diameter—commonly known as the ABCD rules [7]. Dermatologists utilize these criteria to assess skin lesions for signs of malignancy. Edge information of lesions can effectively capture aspects of the “A” (asymmetry) and “B” (border irregularity) components of the ABCD rules by delineating lesion boundaries and contours. In response to this, we introduce an edge-guided dual-branch neural network (EDB-Net), which focuses on recognizing lesion boundaries and intricate edge patterns within these boundaries. The main contributions of this paper are as follows:

1. We introduce an innovative pre-processing method for generating edge maps from dermoscopic images.
2. We propose the EDB-Net model, which features a novel dual-branch approach that incorporates both dermoscopic images and their corresponding edge maps. Branch 1 processes the edge maps, while Branch 2 handles the original dermoscopic images.
3. We integrate a novel channel-spatial synergistic attention (CSSA) block within Branch 2 to highlight the most distinct regions within the lesion.
4. We introduce an attention modulation block that implements a unique strategy to modulate the attention maps generated from the CSSA block. This modulation uses edge features extracted from the input edge maps in Branch 1, resulting in the creation of edge-guided features.
5. To assess the performance of our model, we conduct extensive experiments using the publicly available benchmark datasets for skin cancer: HAM10000 and ISIC 2016.

2 Proposed Method

In this section, we introduce our novel algorithm for generating edge maps from dermoscopic images and elaborate on our proposed model, EDB-Net. Additionally, we propose the CSSA block and elucidate the strategy used to modulate the attention maps generated by the CSSA block.

2.1 Edge Map Generation

In this section, a detailed explanation of the edge map generation from dermoscopic images is provided. The block diagram, as illustrated in Fig. 1, provides a comprehensive overview of the method.

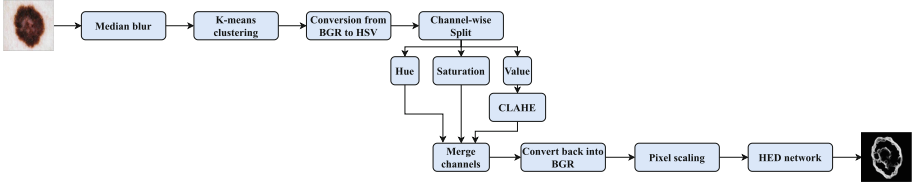


Fig. 1. Overview of the proposed edge map generation approach

Noise Removal: Median blur is applied to mitigate noise from the dermoscopic image in the blue-green-red (BGR) color space. This operation involves computing the median value for each pixel within a 5×5 square neighborhood centred around its position. For a pixel at position (x, y) , the operation can be represented mathematically as $Output(x, y) = Median(Neighborhood(x, y))$. This process enhances the overall quality of the image for subsequent analysis.

K-Means Clustering: K-means clustering, an unsupervised learning algorithm, partitions data points into K clusters. The objective function of this algorithm, denoted in Eq. 1, measures the within-cluster sum of squares (WCSS).

$$WCSS = \sum_{i=1}^K \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

In Eq. 1, K represents the number of clusters, C_i denotes the i^{th} cluster and x signifies a data point in C_i . μ_i and $\|\cdot\|$ denote the centroid of C_i and Euclidean distance respectively. The algorithm iteratively updates cluster assignments and centroids to minimize this objective function. In our algorithm, K-means clustering is applied to the enhanced image with $K=8$ clusters. The exact value for K has been determined through empirical testing. The cluster centers are computed, and each pixel is assigned to its nearest center, allowing the image to be reconstructed into a segmented image with similarly attributed pixels.

Adaptive Histogram Equalization: The K-means segmented image undergoes a conversion from BGR color space to hue-saturation-value (HSV) color space. This transformation facilitates the independent processing of color and intensity information. Subsequently, contrast limited adaptive histogram equalization (CLAHE) is applied exclusively to the value channel, representing image intensity or brightness. The hue and saturation channels remain unaltered to preserve the color information. CLAHE restricts contrast enhancement to localized regions by dividing the image into small tiles and applying histogram equalization individually to each tile. The threshold parameter for contrast limiting is specified as 3, with a tile grid size of 8×8 for CLAHE initialization. Then, the individual hue, saturation, and value channels are merged together to produce the enhanced image in the lightness-A-B (LAB) color space. Finally, the image

is converted from LAB to BGR color space, thereby restoring the image to its original color representation while incorporating enhanced contrast in the value channel, due to the CLAHE operation.

Edge Detection: The BGR image is processed in a way suitable for input to a holistically nested edge detector (HED) [8] model for predicting edge maps. This processing typically involves scaling pixel values and performing mean subtraction. Following this, the resultant image is utilized as input for the pre-trained HED model. Unlike conventional edge detectors, which rely on low-level features such as gradients or filters, HED employs deep learning to acquire rich hierarchical representations of edges from the input image. The fundamental concept underlying HED involves conducting edge detection across multiple scales, followed by fusing the multi-scale information to generate a holistic edge map.

HED consists of a stem network, side output layers, and a weighted fusion layer. A pre-trained CNN model like VGGNet serves as the backbone for hierarchical feature extraction from the input image. The feature maps from the stem network are fed into the side output layers across different stages. Each side output layer produces an edge map at a specific scale, capturing edges at varying levels of granularity. These edge maps from all the side output layers are then fused together using the weighted fusion layer. The weighted fusion layer assigns learned weights to the edge maps from each side output layer and combines them, thereby allowing selective emphasis or de-emphasis based on the significance for edge detection.

In this work, HED is utilized solely for generating edge maps, without altering the model’s parameters or performing any training or optimization. The pre-trained weights are directly employed to predict edge maps for the input image. During the forward pass of an input image through the HED model, the image traverses through the layers of the network, and edge features are computed at each layer. Following this process, the HED model generates a set of edge maps from the input image. These edge maps encompass edges identified at diverse scales and levels of detail, capturing both fine and coarse edges within the image. For the input image, edge map predictions from the side output layers ($\hat{Y}_{\text{side}}^{(1)}, \dots, \hat{Y}_{\text{side}}^{(M)}$) and the weighted-fusion layer (\hat{Y}_{wfuse}) are obtained. Here, M denotes the number of side-output layers. The final consolidated edge map (\hat{Y}_{HED}) is achieved by aggregating these generated edge maps denoted in Eq. 2.

$$\hat{Y}_{\text{HED}} = \text{Average}(\hat{Y}_{\text{wfuse}}, \hat{Y}_{\text{side}}^{(1)}, \dots, \hat{Y}_{\text{side}}^{(M)}) \quad (2)$$

The resulting edge map undergoes rescaling to ensure that edge intensities fall within the appropriate range for visualization, typically between 0 and 255. Figure 2 illustrates two complex cases where the original dermoscopic images are processed using the HED model to generate edge maps. Additionally, the images processed through our proposed algorithm are also fed to the HED model for edge map generation. HED struggles to accurately delineate lesion boundaries in the original images and often detects edges based on artifacts. In contrast,

when using the images processed by our algorithm, the HED model successfully distinguishes even small infected areas, despite the presence of artifacts.

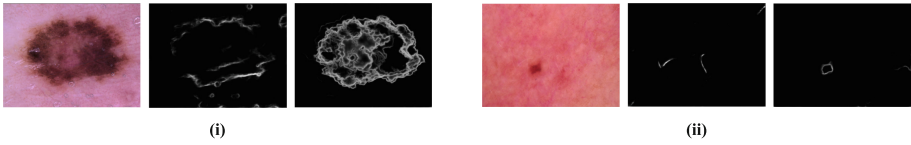


Fig. 2. Comparison of edge maps obtained from the original image with the processed image generated using our proposed algorithm

2.2 Edge-Guided Dual-Branch Neural Network

EDB-Net consists of two branches: Branch 1 is dedicated to processing the edge maps, while Branch 2 focuses on the original dermoscopic images. Both branches employ a DenseNet-based backbone [9] to extract features from their respective inputs. This section deals with the detailed exploration of EDB-Net, elucidating the rationale behind each architectural choice and discussing the integration of edge features into the novel CSSA block for improved classification performance. Figure 3 illustrates the architecture of EDB-Net.

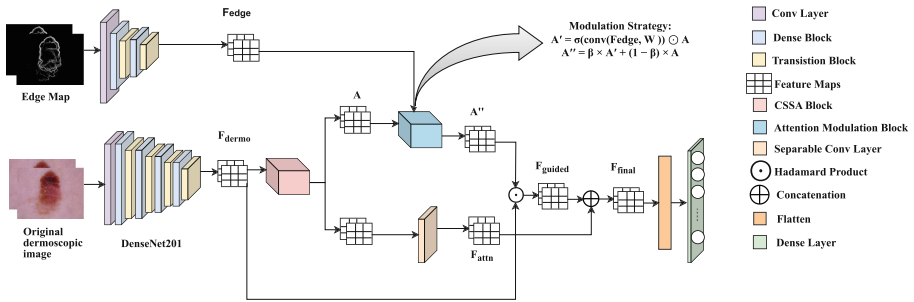


Fig. 3. Architecture of the proposed EDB-Net model

Branch 1 Extraction of Edge Features: Here, a set of dense and transition blocks is used for extracting the edge features, F_{edge} , from the input edge maps. The architectural sequence commences with an initial convolution operation to extract basic features from the input images. This convolutional layer is followed by subsequent batch normalization and rectified linear unit (ReLU) activation to normalize and introduce non-linearity into the feature maps. A sequence of dense

blocks and transition blocks is then iteratively applied to capture increasingly complex edge features. Within this iterative process, two dense-transition iterations are specified, each comprising a dense block with 4 layers and a growth rate of 12, followed by a transition block. The decision to limit the number of dense-transition iterations underscores the aim to prioritize the extraction of edge features while minimizing the likelihood of the network acquiring extraneous information. This strategic approach aligns with our research goal of developing a specialized model tailored specifically for the edge feature extraction task.

Branch 2 Dermoscopic Feature Extraction: Here, the DenseNet201 architecture is used to extract features, F_{dermo} , from dermoscopic images. This architecture is customized by excluding the last 28 layers and integrating our channel-spatial synergistic attention block into the dense block of the network architecture, where the size of image feature maps, F_{dermo} , is 7×7 . Figure 4 depicts the architecture of this block.

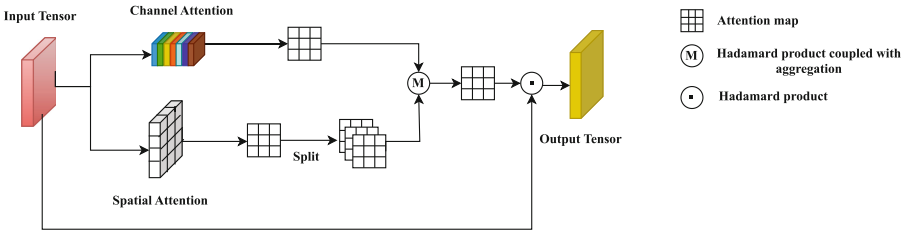


Fig. 4. Architecture of the proposed CSSA block

The design of this CSSA block is inspired by the architecture of the convolutional block attention module (CBAM) [10] and involves the computation of both channel and spatial attention mechanisms. However, like CBAM, we do not apply the attention mechanisms sequentially. In our CSSA block, initially, channel attention is computed by globally pooling the input tensor, F_{dermo} , along the spatial dimensions, using both average and max pool operations. These activations are then fused through dense layers to generate the channel attention map, $M_c \in \mathbb{R}^{C \times 1 \times 1}$. Subsequently, spatial attention is computed by averaging and max-pooling activations along the channel dimension, followed by convolutional operations to derive the spatial attention map, $M_s \in \mathbb{R}^{H \times W}$. This map is then split into five smaller maps, $M_s^{(i)} \in \mathbb{R}^{H \times W}$, each corresponding to different heads. The decision to use five distinct maps is based on empirical testing and is intended to enhance the focus on specific and localized areas. This strategy allows for a more detailed and precise analysis of the diverse characteristics of skin lesions, which differ significantly in terms of shape, size, and color. Each of these smaller spatial attention maps $M_s^{(i)}$ is combined with the channel attention map M_c using the Hadamard product. The outputs from each of the five heads

are aggregated to yield a unified attention map, $A \in \mathbb{R}^{C \times H \times W}$, which integrates all localized attentions into a global context. This refined attention map, A , is subsequently applied to the input features, F_{dermo} , through a Hadamard product. Equations 3, 4, 5, and 6 represent the key formulations within the CSSA block. Thereafter, a SeparableConv2D layer is employed, resulting in enhanced feature representations, F_{attn} . These refined representations encapsulate both channel-wise and spatially relevant information, facilitating improved performance in the subsequent classification task.

$$M_s^{(i)} = \text{Split}(M_s, \text{parts} = 5) \quad \text{for } i = 1, \dots, 5 \quad (3)$$

$$A^{(i)} = M_s^{(i)} \odot M_c \quad \text{for } i = 1, \dots, 5 \quad (4)$$

$$A = \sum_{i=1}^5 A^{(i)} \quad (5)$$

$$F_{attn} = F_{dermo} \odot A \quad (6)$$

Attention Modulation Block: Here, the attention map A is refined through the integration of edge features, F_{edge} . These features are instrumental in modulating the attention map, generated using our CSSA block, enhancing the model’s ability to focus on the regions of interest defined by prominent edges within the input images. This modulation process is governed by a gating mechanism equipped with learnable convolutional weights W , which adaptively capture contextually relevant edge patterns. The gating mechanism is formulated in Eq. 7.

$$A' = \sigma(\text{conv}(F_{edge}, W)) \odot A \quad (7)$$

Here, in Eq. 7, σ represents the sigmoid activation function, that normalizes the convolution output, ensuring that the values lie within the range $[0, 1]$. This normalization is crucial as it conditions the modulation to be conducive for multiplicative scaling. The operation \odot represents the Hadamard product. This operation allows the edge features to selectively enhance or suppress features within the attention map A , based on the edge information. This process aligns the model’s focus with the most salient regions marked by significant edge features.

Further, to introduce flexibility and maintain a balance between the original and edge-enhanced attention maps, a hybrid modulation strategy is applied to generate a refined attention map A'' , defined in Eq. 8.

$$A'' = \beta \times A' + (1 - \beta) \times A \quad (8)$$

Here, in Eq. 8, β is a trainable parameter that determines the extent to which the edge-modulated attention map A' influences the final attention representation A'' . Such a hybrid approach not only enriches the attention mechanism with precise and edge-based contextual information but also preserves the integrity of the initial attention cues. This strategy ensures that the network remains robust to variations in edge relevance across different contexts.

Integration of Edge-Enhanced Attention: This stage is pivotal for incorporating the final attention map A'' into the feature maps to enhance the model’s feature representations. Initially, the Hadamard product is employed to fuse the features extracted during the dermoscopic feature extraction phase, F_{dermo} , with the refined attention map A'' . This operation, denoted in Eq. 9, produces edge-guided feature maps, F_{guided} , which are enhanced by the attentional focus driven by edge-specific information.

$$F_{guided} = F_{dermo} \odot A'' \quad (9)$$

Subsequently, to ensure a comprehensive feature representation, these edge-guided features F_{guided} are concatenated with F_{attn} , the features generated using the CSSA block. This concatenation, shown in Eq. 10, helps preserve both vital edge information and high-level semantic information, culminating in an enriched feature set F_{final} .

$$F_{final} = F_{guided} \oplus F_{attn} \quad (10)$$

Finally, the enriched feature set F_{final} is flattened and fed into the final dense layers to perform the classification task.

3 Results

3.1 Datasets and Experimental Protocols

Datasets: We have used the challenging HAM10000 [11] and ISIC 2016 [12] datasets to assess the performance of EDB-Net. The HAM10000 dataset includes 10,015 images categorized into seven skin disease groups: 327 actinic keratosis and intraepithelial carcinoma images, 514 basal cell carcinoma images, 1,099 benign keratosis images, 115 dermatofibroma images, 1,113 melanoma images, 6,705 melanocytic nevus images, and 142 vascular malformation images. Since there is no official split for this dataset, we have applied an 80:20 train-test split, consistent with the approach used by many other state-of-the-art methods. For the ISIC 2016 dataset, we have adhered to the official training and test sets. The training set consists of 900 images, while the test set contains 379 images, categorized into malignant melanomas and benign nevi. We have utilized 20% of the training set from each dataset for validation. Despite significant class imbalances in these datasets, we do not augment the training sets, aiming to demonstrate that the integration of edge information in our proposed model effectively distinguishes various classes, highlighting its robustness and potential for accurate skin cancer classification. All images and their corresponding edge maps have been resized to 224×224 dimensions. Our proposed model has been retrained on the training sets by fine-tuning all layers for 60 epochs, with an early stopping patience of 30. For all training tasks, we have used a learning rate of 0.001, the Adam optimizer with an epsilon value of 0.1, and cross-entropy loss.

Evaluation Metrics: To evaluate our model, we have employed the following metrics: $Accuracy(Acc) = \frac{TP+TN}{TP+TN+FP+FN}$, $Precision(Pre) = \frac{TP}{TP+FP}$,

$Recall(Rec) = \frac{TP}{TP+FN}$, and $F1 - score(F1) = 2 \times \frac{(Precision \times Recall)}{(Precision+Recall)}$. Here TP , TN , FP , and FN stand for true positives, true negatives, false positives, and false negatives, respectively.

3.2 Quantitative Analysis

To determine the optimal backbone network for our proposed model, several CNNs have been trained on the HAM10000 dataset. These include MobileNetV2, ResNet50, InceptionResNetV2, and DenseNet201, with DenseNet201 demonstrating the highest performance. The outcomes of these evaluations are summarized in Table 1a. Although DenseNet201 and InceptionResNetV2 perform similarly, DenseNet201 has significantly fewer parameters. Therefore, we decided to further investigate various architectural configurations using DenseNet201 as the backbone network. The different configurations are as follows:

- (i) DenseNet201 + CBAM
- (ii) DenseNet201 + CSSA module
- (iii) EDB-Net (DenseNet201 + CSSA module + edge guidance)

The detailed analysis of these configurations is provided in Table 1b.

Table 1. Ablation results for selecting the best-performing baseline model and the best architectural setup for the HAM10000 dataset

(a) Baseline model selection					(b) Best setup selection				
Model	Acc	Pre	Rec	F1	Model	Acc	Pre	Rec	F1
MobileNetV2	0.858	0.852	0.858	0.849	(i)	0.908	0.910	0.908	0.909
ResNet50	0.872	0.877	0.868	0.874	(ii)	0.913	0.910	0.913	0.911
InceptionResNetV2	0.892	0.888	0.892	0.890	(iii)	0.927	0.924	0.927	0.926
DenseNet201	0.894	0.900	0.897	0.900					

From the ablation analysis presented in Table 1b, it is evident that our novel CSSA module exhibits an enhancement over CBAM for skin cancer classification. Additionally, the integration of edge information contributes to improving accuracy by a minimum of 1.40% compared to using attention mechanisms alone. Figure 5a, Fig. 5b, and Fig. 5c depict the confusion matrix, receiver operating characteristic (ROC) curve, and the loss curve of EDB-Net on the HAM10000 dataset, respectively. Additionally, to assess robustness, we have evaluated the performance of EDB-Net on the ISIC 2016 dataset using the best setup that was obtained for the HAM10000 dataset. For ISIC 2016, Fig. 6a, Fig. 6b, and Fig. 6c present the confusion matrix, ROC curve, and the loss curve of EDB-Net, respectively. The results summarized in Table 2 illustrate EDB-Net’s superior performance compared to most existing methods for skin cancer classification on both the HAM10000 and the ISIC 2016 datasets. Additionally, Table 2 indicates that although the model proposed by Gajera et al. [13] achieves the highest accuracy on the ISIC 2016 dataset, its relatively low recall rate is a notable flaw. In

contrast, EDB-Net offers a balanced performance in terms of both accuracy and recall.

Table 2. Performance comparison of the proposed model with some recent methods on HAM10000 and ISIC 2016 datasets

Dataset	Work Ref.	Acc	Pre	Rec	F1
HAM10000	Bajwa et al. [14], 2020	0.815	0.728	0.782	0.783
	Iqbal et al. [3], 2021	0.888	0.905	0.888	0.891
	Datta et al. [4], 2021	0.916	0.915	0.916	0.914
	Sai Charan et al. [15], 2022	0.886	-	-	-
	Gururaj et al. [16], 2023	0.912	-	-	0.917
	Roy et al. [17], 2023	-	0.723	0.706	0.702
	Khan et al. [18], 2024	0.870	-	0.869	-
	Gairola et al. [19], 2024	0.920	0.690	0.920	0.730
	Roy et al. [20], 2024	0.908	0.908	0.908	0.912
	Kumar et al. [21], 2024	0.886	0.888	0.883	0.880
	Ours, 2024	0.927	0.924	0.927	0.926
ISIC 2016	Yu et al. [22], 2016	0.855	-	0.507	-
	Saba et al. [23], 2019	0.786	-	0.667	-
	Gajera et al. [13], 2022	0.871	0.783	0.480	0.595
	Gajera et al. [24], 2023	0.805	0.506	0.560	0.532
	Sahoo et al. [25], 2024	0.781	-	0.780	-
	Ours, 2024	0.848	0.846	0.848	0.847

3.3 Discussion

Accurate classification of skin cancer poses significant challenges, particularly with unbalanced datasets like HAM10000 and ISIC 2016. As demonstrated in Fig. 5a, EDB-Net achieves a high level of accuracy in identifying various skin cancer classes, even without the use of data augmentation. Figure 7 illustrates samples of seven types of dermoscopic lesion images, their corresponding edge maps generated using our proposed algorithm, and edge feature maps obtained from Branch 1 of EDB-Net. Notably, Fig. 7 showcases EDB-Net’s ability to delineate intricate lesion borders, affirming its efficiency.

Figure 8 shows the t-distributed stochastic neighbor embedding (t-SNE) plots for our model on the HAM10000 and the ISIC 2016 datasets. In HAM10000, the plot shows distinct clusters for most classes, indicating effective model learning. Class overlap occurs for some classes (akiec and bkl) suggesting room for potential improvements. For ISIC 2016, the malignant (mel) class forms a relatively distinct cluster towards the right side of the plot, isolated from the benign (nv)

class. However, the significant overlap between the two classes in the plot’s central region suggests potential difficulties in the model’s ability to differentiate between them.

Figure 9 showcases gradient-weighted class activation mapping (Grad-CAM) heatmaps for sample images of each lesion type in the HAM10000 test set, visually indicating the areas within the images that most influence our model’s predictions. These heatmaps highlight regions of elevated diagnostic significance, with warmer colors (red and yellow) denoting higher relevance to the model’s decisions. Notably, the heatmaps effectively focus on the primary lesion areas while ignoring uninfected skin and hair.

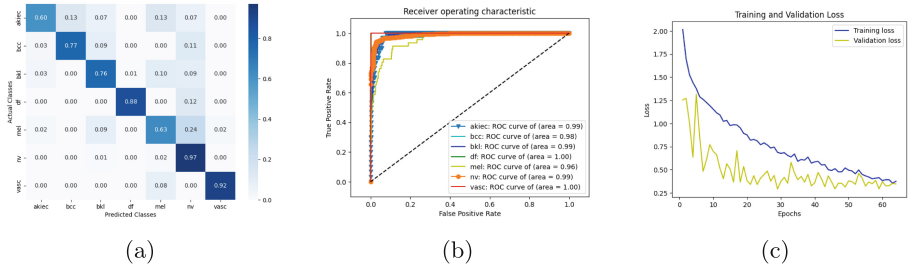


Fig. 5. (a) Confusion matrix; (b) ROC curve; (c) Loss curve for HAM10000

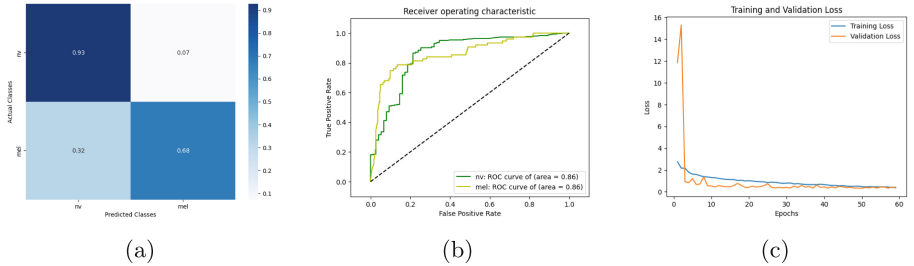


Fig. 6. (a) Confusion matrix; (b) ROC curve; (c) Loss curve for ISIC 2016

To offer a balanced view, we acknowledge some limitations of the proposed method. Although EDB-Net presents a novel approach, its efficiency heavily relies on the quality of the generated edge maps, which significantly influences the model’s overall performance. Additionally, identifying the optimal hyperparameters for the CSSA module and the edge feature modulation strategy poses challenges, necessitating extensive experimentation.

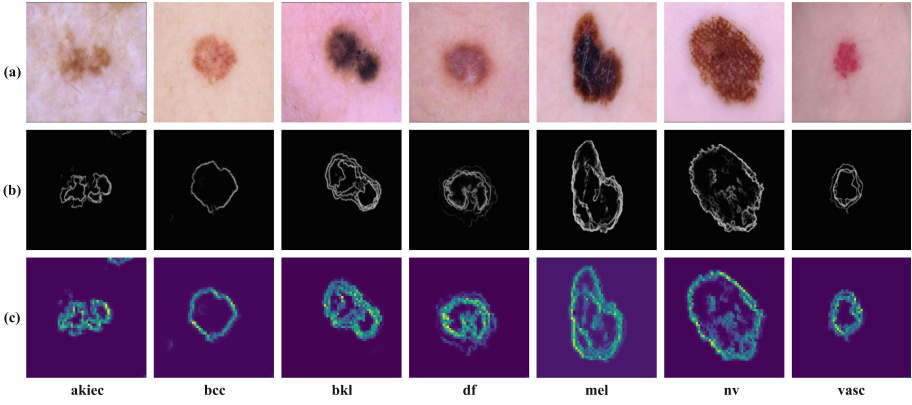


Fig. 7. (a) Dermoscopic images; (b) Edge maps; (c) Edge feature maps

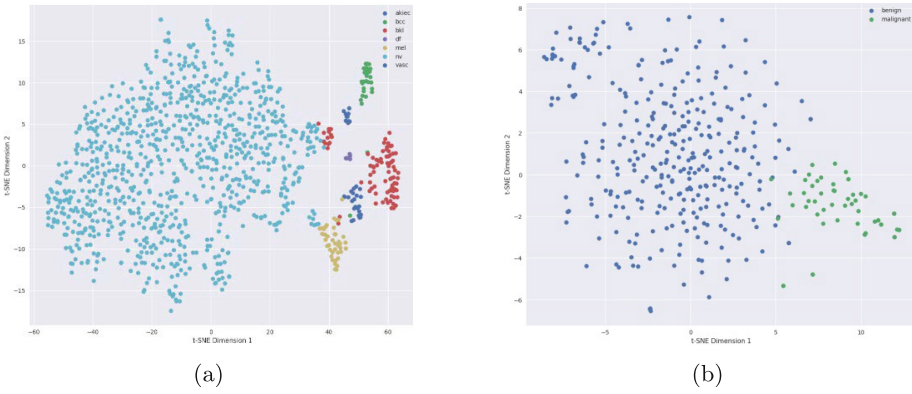


Fig. 8. t-SNE plots for (a) HAM10000 and (b) ISIC 2016 datasets

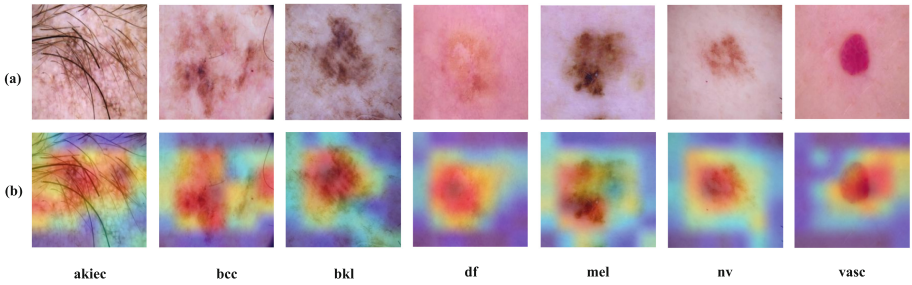


Fig. 9. (a) Dermoscopic images; (b) Grad-CAM heatmaps

4 Conclusion

Skin cancer is responsible for a significant number of fatalities worldwide each year, making prompt and accurate diagnosis crucial for improving survival rates and treatment efficacy. Skin cancer classification is a critically important and challenging research topic. In this work, we draw inspiration from the fact that the edge information of lesions is a crucial indicator of malignancy. Consequently, we introduce an innovative algorithm for generating edge maps from dermoscopic images. Subsequently, we propose an edge-guided dual-branch network aided by our CSSA block and attention modulation block. This approach focuses on intricate edge patterns and enhances the emphasis on regions of interest within the images, producing more distinctive features and significantly boosting classification performance. Our model achieves accuracies of 0.927 on the HAM10000 dataset and 0.848 on the ISIC 2016 dataset. Notably, our model operates without data augmentation, demonstrating its efficiency and effectiveness.

While our proposed model showcases promising results, further enhancements are required to improve the recall rates of melanoma and a few other minority classes to ensure real-world usability. Our next attempts involve a focused investigation into specific loss functions tailored to address this issue. Additionally, we aim to integrate contrastive learning techniques to refine feature extraction, thereby enhancing overall classification performance.

Acknowledgments. We appreciate the essential infrastructure support provided by the [CMATER Research Lab](#) of the Department of Computer Science and Engineering at Jadavpur University, India.

References

1. Kalouche, S., Ng, A., Duchi, J.: Vision-based classification of skin cancer using deep learning. 2015, conducted on Stanfords Machine Learning course (CS 229) taught (2016)
2. Emara, T., Afify, H.M., Ismail, F.H., Hassanien, A.E.: A modified inception-v4 for imbalanced skin cancer classification dataset. In: 2019 14th International Conference on Computer Engineering and Systems (ICCES), pp. 28–33. IEEE (2019)
3. Iqbal, I., Younus, M., Walayat, K., Kakar, M.U., Ma, J.: Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized Med. Imaging Graph.* **88**, 101843 (2021)
4. Datta, S.K., Shaikh, M.A., Srihari, S.N., Gao, M.: Soft attention improves skin cancer classification performance. In: Reyes, M., Henriques Abreu, P., Cardoso, J., Hajj, M., Zamzmi, G., Rahul, P., Thakur, L. (eds.) *IMIMIC/TDA4MedicalData-2021*. LNCS, vol. 12929, pp. 13–23. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87444-5_2
5. Shen, S., et al.: A low-cost high-performance data augmentation for deep learning-based skin lesion classification. *BME Frontiers* (2022)

6. Sarkar, S., Ray, A., Kaplun, D., Sarkar, R.: A combination of soft attention-aided cnn models using dempster-shafer theory for skin cancer classification. In: International Conference on Current Problems of Applied Mathematics and Computer Systems, CPAMCS-2023. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-64010-0_38
7. Weigert, U., Burgdorf, W.H.C., Stolz, W.: c abcd rule. In: An Atlas of Dermoscopy, pp. 123–127. CRC Press (2012)
8. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403 (2015)
9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
10. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
11. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**(1), 1–9 (2018)
12. Gutman, D., et al.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv preprint [arXiv:1605.01397](https://arxiv.org/abs/1605.01397) (2016)
13. Gajera, H.K., Nayak, D.R., Zaveri, M.A.: Fusion of local and global feature representation with sparse autoencoder for improved melanoma classification. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 5051–5054. IEEE (2022)
14. Bajwa, M.N., et al.: Computer-aided diagnosis of skin diseases using deep neural networks. *Appl. Sci.* **10**(7), 2488 (2020)
15. Charan, D.S., Nadipineni, H., Sahayam, S., Jayaraman, U.: Method to classify skin lesions using dermoscopic images. arXiv preprint [arXiv:2008.09418](https://arxiv.org/abs/2008.09418) (2020)
16. Gururaj, H.L., Manju, N., Nagarjun, A., Manjunath Aradhya, V.N., Flammini, F.: A deep learning approach for skin cancer classification. *IEEE Access, Deepskin* (2023)
17. Roy, D., Pramanik, R., Sarkar, R.: Margin-aware adaptive-weighted-loss for deep learning based imbalanced data classification. *IEEE Trans. Artif. Intell.* (2023)
18. Khan, M.A., Muhammad, K., Sharif, M., Akram, T., Kadry, S.: Intelligent fusion-assisted skin lesion localization and classification for smart healthcare. *Neural Comput. Appl.* **36**(1), 37–52 (2024)
19. Gairola, A.K., Kumar, V., Sahoo, A.K., Diwakar, M., Singh, P., Garg, D.: Multi-feature fusion deep network for skin disease diagnosis. *Multimed. Tools Appl.*, 1–26 (2024)
20. Roy, A., Sarkar, S., Ghosal, S., Kaplun, D., Lyanova, A., Sarkar, R.: A wavelet guided attention module for skin cancer classification with gradient-based feature fusion. arXiv preprint [arXiv:2406.15128](https://arxiv.org/abs/2406.15128) (2024)
21. Kumar, A., Vishwakarma, A., Bajaj, V., Mishra, S.: Novel mixed domain hand-crafted features for skin disease recognition using multi-headed cnn. *IEEE Trans. Instrum. Measur.* (2024)
22. Lequan, Yu., Chen, H., Dou, Q., Qin, J., Heng, P.-A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans. Med. Imaging* **36**(4), 994–1004 (2016)

23. Tanzila Saba, Muhammad Attique Khan, Amjad Rehman, and Souad Larabi Marie-Sainte. Region extraction and classification of skin cancer: A heterogeneous framework of deep cnn features fusion and reduction. *Journal of medical systems*, 43(9):289, 2019
24. Gajera, H.K., Nayak, D.R., Zaveri, M.A.: A comprehensive analysis of dermoscopy images for melanoma detection via deep cnn features. *Biomed. Signal Process. Control* **79**, 104186 (2023)
25. Sahoo, S.R., Dash, R., Mohapatra, R.K.: Fusion of deep and wavelet feature representation for improved melanoma classification. *Multimed. Tools Appl.* 1–27 (2024)



PSIVUS: Atherosclerotic Plaque Segmentation in Intravascular Ultrasound Images via Active Learning

Anuradha Mahato^{1,2}(✉), Paromita Banerjee³, Rutvik Narendrabhai Jethava², Bhanu Duggal³, Angshuman Paul², Mayank Vatsa², and Richa Singh²

¹ Indian Institute of Science Education and Research Bhopal, Bhopal, India
anuradha19@iiserb.ac.in

² Indian Institute of Technology Jodhpur, Jodhpur, India

³ All India Institute of Medical Sciences Rishikesh, Rishikesh, India

Abstract. Atherosclerosis, characterized by the deposition of fats, cholesterol, and other substances along arterial walls, poses a significant risk to cardiovascular health, leading to arterial narrowing and potentially fatal events such as heart attacks and strokes. Intravascular ultrasound (IVUS) imaging plays a crucial role in cardiovascular medicine, offering high-resolution views of arterial cross-sections. Accurate segmentation of IVUS images is essential for quantifying pathological features such as atherosclerotic plaque, which is necessary for assessing disease burden, planning therapeutic procedures, and evaluating responses to medications. This paper introduces a novel approach leveraging machine learning and deep learning techniques to segment atherosclerotic plaques in IVUS images. The proposed methodology incorporates active learning techniques into the segmentation pipeline to strategically select the most informative data points for training, thereby enhancing model performance and mitigating data dependency. Experimental results demonstrate promising outcomes, achieving comparable segmentation performance measured by mean Intersection over Union (IoU) using a significantly smaller portion of the dataset. This highlights the efficacy of our methodology in optimizing segmentation performance while reducing reliance on extensive data. We will release the dataset on <https://iab-rubric.org/resources>.

Keywords: Intravascular Ultrasound · Segmentation · Active Learning

1 Introduction

Cardiovascular diseases are a leading cause of morbidity and mortality worldwide [25]. Therefore, imaging techniques for an accurate diagnosis of such diseases are highly valued by clinicians. Intravascular ultrasound (IVUS) [18] is a powerful modality that provides high-resolution cross-sectional images of the coronary arteries. The information obtained from IVUS, particularly in the evaluation of

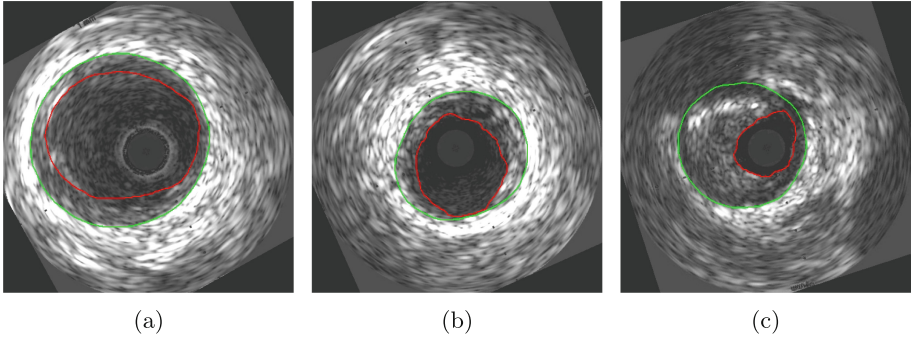


Fig. 1. Annotated IVUS images showing cross-sectional view of coronary artery depicting the plaque (fibro/fatty plaque in (a) and (b), calcified plaque in (c) deposition of three different individuals having varying degrees of plaque marked between red (inner) and green (outer) boundaries for healthy and diseased patients with (a) low plaque, (b) moderate plaque, and (c) high plaque.

vessel wall morphology and the identification of atherosclerotic plaques, makes it extremely useful in interventional cardiology. While IVUS provides detailed images, the manual analysis of these images is a time-consuming and challenging task, requiring the intervention of experts. Image segmentation, the process of partitioning images into distinct regions or structures, is fundamental to extracting meaningful information from IVUS images (Fig. 1).

Accurate segmentation of plaques from IVUS images assist clinicians in making precise diagnoses and treatment plans for cardiovascular diseases. However, the manual analysis of IVUS images is challenging because of several factors, including subjectivity, interobserver variability, and the time-consuming nature of the process [12]. This shows the need for automated and accurate image segmentation techniques to help healthcare professionals obtain clinically relevant information efficiently. Deep learning-based segmentation methods may address these challenges and contribute to ongoing efforts to enhance the utility of IVUS in cardiovascular medicine. Traditional segmentation approaches face challenges posed by the intricate nature of vascular structures, morphology variations, and the presence of speckle noise.

Modern deep learning techniques often require vast amounts of training data; however, annotated datasets for plaque segmentation in IVUS images remain scarce. The challenge lies in the labor-intensive nature of manual annotation, compounded by potential interobserver variability. To address this gap, we have developed one of the first annotated IVUS image datasets, named *PSIVUS*, specifically for plaque segmentation. This dataset has the potential to significantly advance research in the automated and efficient diagnosis of cardiovascular diseases. We will release *PSIVUS* dataset to further catalyze progress in this critical area of research.

Given the elusiveness of annotated IVUS datasets, there is a pressing need to develop automated plaque segmentation methods that can be trained with a limited number of training images. At the same time, it is important to recognize that unannotated images are often available in much larger quantities. To address these challenges, we introduce PSIVUS-Net, an active learning method for plaque segmentation from IVUS images. Our approach minimizes the reliance on extensive annotated datasets by leveraging unannotated images during training. PSIVUS-Net utilizes the underlying data distribution to generate pseudo labels for unannotated images. We then identify a subset of images where the pseudo labels are predicted with high confidence and incorporate these images, along with the annotated ones, into the training process. This iterative approach continues based on the available annotation budget, optimizing the model’s performance with minimal manual annotation. In this paper, our key contributions are as follows:

1. We have developed an annotated dataset specifically for plaque segmentation from Intravascular Ultrasound (IVUS) images.
2. We propose an active learning framework tailored for the segmentation of IVUS images.
3. We provide a comprehensive benchmark by evaluating the performance of various state-of-the-art image segmentation techniques on our dataset.

2 Related Works

Since this research focuses on segmentation for IVUS and active learning, we summarize here the literature in these two areas.

2.1 Segmentation

Recent studies in IVUS image segmentation employ various machine learning approaches in conjunction with deep learning. These methods include active contour-based techniques such as snake and level set methods [28], probability-based methods [16], and other machine learning-based strategies [4]. Recent work focused on fully automated segmentation that mimics human expert procedures, with specific applications in assessing coronary artery dimensions, balloon sizing, and the automatic extraction of lumen and vessel boundaries [15].

These advancements indicate the growing role of machine learning in enhancing the accuracy and efficiency of IVUS image segmentation tasks. Deep learning techniques, especially convolutional neural networks (CNNs), are used in various medical image segmentation tasks [10]. It has shown remarkable success in various image segmentation tasks because it can automatically learn hierarchical features from raw data [20]. Nishi et al. [17] used a DL-based segmentation system built using a fully convolutional neural network (CNN) with DeepLabv3 architecture, incorporating a ResNet34 encoder and Adam optimizer. Bajaj et al. [2] employed a ResNet-based convolutional neural network with a Pix2pix

[11] conditional generative adversarial network (GAN). Dong et al. [6] used an approach using an 8-layer U-Net to segment the coronary artery lumen and EEM.

These investigations have shown promising results, showcasing the potential of automated approaches to streamline the analysis workflow and improve diagnostic accuracy. However, a comprehensive evaluation of custom IVUS image segmentation deep learning techniques remains an active research area. This study aims to contribute to the existing body of knowledge by presenting a detailed exploration of the application of CNNs for IVUS image segmentation. By building upon the foundation laid by previous studies, we seek to address the limitations of traditional segmentation methods and provide insights into the potential of deep learning to revolutionize the analysis of IVUS images. Through rigorous experimentation and evaluation, our objective is to establish the efficacy and reliability of our proposed methodology to enhance the clinical utility of IVUS for the diagnosis and planning of cardiovascular treatment.

2.2 Active Learning in Medical Image Segmentation

Active learning has become increasingly important in medical image segmentation due to its ability to reduce the manual annotation burden by strategically selecting the most informative samples for labeling. One notable recent method is the integration of Bayesian active learning with deep learning models. Billah et al. [3] proposed a Bayesian Convolutional Neural Network (BCNN) framework that employs Monte Carlo dropout to estimate the model uncertainty. Their approach focuses on selecting samples with the highest uncertainty, thus providing a significant boost in performance with fewer labeled instances. This method has been applied to tasks such as brain tumor segmentation, demonstrating enhanced efficiency in labeling efforts [1].

The suggestion annotation [27], one of the initial deep AL frameworks, used bootstrapping to estimate the uncertainty of the sample, and used a greedy measure of cosine similarity to evaluate the similarity between the candidate set and the unlabeled pool. In contrast to using multiple models, [19] utilized a Monte Carlo dropout Bayesian network to compute the prediction variance and adopted a Borda-count-based sampling strategy to identify the candidates who ranked the best in terms of uncertainty and representativeness. An extension of this approach computed representativeness with an infoVAE [29] for maximum-likelihood sampling in the latent space [19]. Mahapatra et al. [14] used a conditional generative adversarial network (cGAN) to generate realistic chest X-ray images conditioned on real images, and a Bayesian neural network to select the most informative samples for training.

To enhance model performance with a smaller amount of annotated data during training, two methodologies have emerged to harness the potential of unlabeled data: active learning and semi-supervised learning [8]. Active learning (AL) focuses on selecting informative samples for labeling and inclusion in training. Semi-supervised learning aims to enhance the learned representation from data by leveraging unlabeled samples alongside the limited labeled ones.

Nonetheless, the challenge persists in selecting the most suitable samples for the labeled set, emphasizing the significance of active learning in this context.

3 PSIVUS Dataset

Due to the lack of multiple publicly available datasets in this domain, we prepared a dataset of IVUS imaging at All India Institute of Medical Sciences Rishikesh (AIIMS) Rishikesh. For this, appropriate IRB approvals were taken. IVUS procedures were performed according to established protocols, ensuring standardization and reliability of image acquisition. An automated transducer pullback, operating at a velocity of 0.5 mm/s, is utilized in conjunction with a commercially available IVUS imaging system under the expertise of a experienced cardiologists. The acquired IVUS image data are initially stored in the “.iq” format and then processed using the cardiac imaging software QIVUS Research Edition 3.1 by Medis Medical Imaging to convert it into the more accessible “.jpeg” format. The resulting dataset includes a collection of cross-sectional artery images annotated by domain experts. These annotations delineated the vessel and lumen with distinct red and green boundaries, ensuring clarity and precision in subsequent segmentation tasks, as shown in Fig. 2 and the dataset statistics are summarized in Table 1.

To train a segmentation model effectively, it is essential to have not only the images, both also the annotations for a corresponding mask. Therefore, the dataset is further annotated by domain experts which will provide a significant support towards the development of segmentation algorithms. These annotations primarily involved delineating boundaries that mark the plaque regions within IVUS images, serving as ground-truth labels for mask generation. The careful annotation process ensured the accuracy and reliability of the dataset, thereby providing a strong foundation for developing and evaluating segmentation methods to identify plaque regions in IVUS images. Annotated images that show delineations in green and red hues are processed for mask development. The dataset and the annotation will be released publicly to the research community.

The binary mask features the white plaque region (plaque) and the black background. Green and red boundaries are used to generate a binary mask. Initially, a black background is created to match the size of the image. Then, the presence of green pixels in the annotated image is identified using the Suzuki algorithm [24]. This algorithm allows for the precise tracing of continuous white (foreground) regions by scanning pixels within a defined range of green values. It employs a pixel-following technique, navigating through adjacent pixels to determine contour boundaries with a non-recursive method. Following this approach, red contours are also generated. Then, area under the red contour is subtracted from the area under green contour on the background black image, resulting in the final plaque region. This process yields a binary mask that is used in model training.

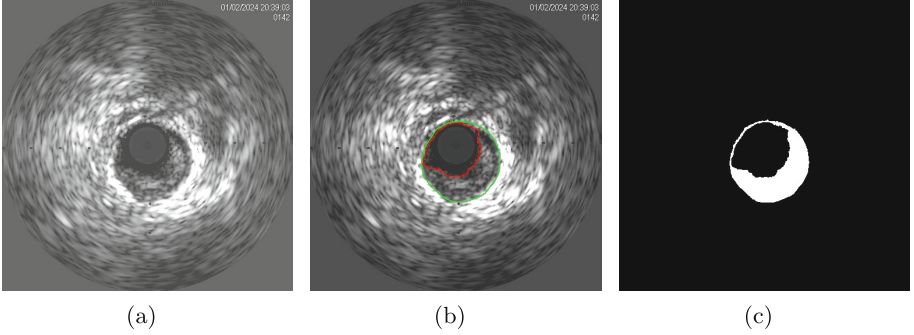


Fig. 2. (a): Raw image from QIVUS, (b): After annotation from a medical expert, (c): Ground truth mask for plaque segmentation.

Table 1. Description of the dataset split for training and evaluation.

	Train	Validation	Test	Total
Images	8606	1385	4706	14697

4 PSIVUS-Net

We propose PSIVUS-Net, an active learning method for plaque segmentation from IVUS images. It utilizes a Variational Autoencoder (VAE) [13] with a shared encoder and two decoders. While one decoder branch reconstructs the input image, the other decoder branch performs plaque segmentation. The latent space of the VAE helps in extracting salient features from the input IVUS images. Our design is motivated by the Variational Adversarial Active Learning (VAAL) method [23]. We also follow the training protocol of [23]. The proposed PSIVUS-Net architecture is illustrated in Fig. 3. We perform active learning utilizing the latent space representation. The operation of the segmentation branch and the reconstruction branch is discussed next followed by the description of the active learning method.

The integration of the segmentation task with a variational autoencoder (VAE) for the extraction of latent features is done where the network has an encoder-decoder structure. The encoder maps the input image to a latent space, and the decoder reconstructs the image from this latent representation. The training procedure involves training the segmentation model, VAE, and discriminator models. Specifically, the VAE transforms the input image x into a distribution over the latent variables z .

4.1 The Segmentation Decoder

The segmentation decoder operates in conjunction with the encoder, using the same encoded features for segmentation tasks. The encoder $E_\phi(x)$ processes

the input image x to produce the latent space feature map f . The segmentation branch (decoder) implements an upsampling function $U(f)$ that takes the encoded features f and increases their dimensions to match those required by the segmentation task. This is done with transposed convolutional layers. Then, the segmentation head S_ψ processes these upsampled features to generate pixel-wise segmentation predictions for an input image x .

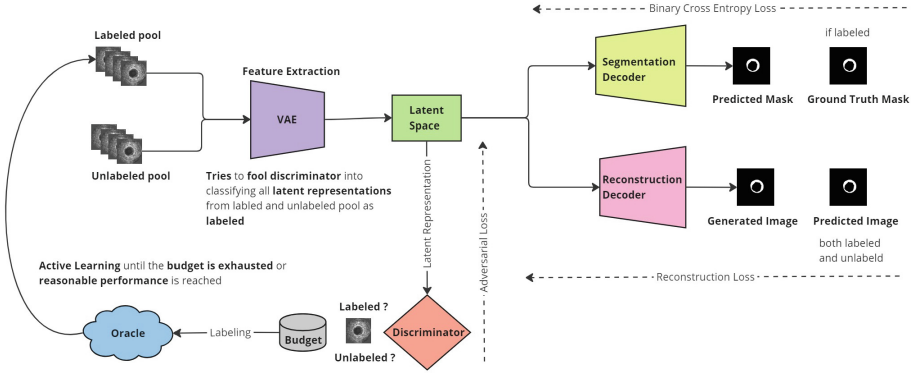


Fig. 3. Proposed PSIVUS-Net Architecture: Both labeled and unlabeled images are passed through an encoder for feature extraction then latent space representations are used for different purposes. Discriminator: tries to differentiate the latent space representation between labeled and unlabeled and according to the budget the unlabeled samples identified by discriminator are given to oracle. Reconstruction decoder: helps capture important features from these representations. Segmentation decoder: creates the mask, leveraging the refined features for accurate segmentation results.

The segmentation branch employs instance normalization and dropout for regularization along with LeakyReLU and ReLU activations. For each pixel of an input image x , this branch outputs probabilities of belonging to the foreground (and background). These probabilities are used to calculate the cross-entropy loss \mathcal{L}_{seg} between the predicted and ground truth segmentation masks. This loss is used alongside other losses to train our model.

4.2 The Reconstruction Decoder

The reconstruction decoder outputs the reconstructed version of the input image using the latent space representation of the VAE. Reconstruction of images from the latent space is achieved through decoding layers, involving upsampling to the original image size. In each epoch, the model iterates through batches of labeled and unlabeled data, computing losses for segmentation, VAE transductive and adversarial loss, and discriminator loss. Let x_L represent the labeled data samples and x_U represent the unlabeled input data samples. The encoder

q_ϕ processes these inputs to produce the latent representations z_L and z_U , respectively. The VAE learns to encode input data x into a compressed latent space z and then decode it back to reconstruct the input. The reconstruction should be as close to the original as possible, which is encouraged through the reconstruction loss, measured as the mean squared error (MSE) between the input and the reconstructed output. The objective function of the VAE aims to minimize the variational lower bound on the marginal likelihood of a given sample is given by,

$$\begin{aligned} \mathcal{L}_{\text{trd}}^{\text{VAE}} = & \mathbb{E}[\log p_\theta(x_L|z_L)] - \beta KL(q_\phi(z_L|x_L)|p(z)) \\ & + \mathbb{E}[\log p_\theta(x_U|z_U)] - \beta KL(q_\phi(z_U|x_U)|p(z)), \end{aligned} \quad (1)$$

where $\mathcal{L}_{\text{trd}}^{\text{VAE}}$ is the transductive loss for the VAE; $\mathbb{E}[\log p_\theta(x_L|z_L)]$ and $\mathbb{E}[\log p_\theta(x_U|z_U)]$ are the expected log probabilities of reconstructed labeled x_L and unlabeled x_U data from their corresponding latent representations z_L and z_U , respectively; β is a hyperparameter that balances the reconstruction loss and the KL divergence in the VAE loss function; $KL(q_\phi(z_L|x_L)|p(z))$ and $KL(q_\phi(z_U|x_U)|p(z))$ are the Kullback-Leibler divergences between the approximate posterior distributions of the latent variables given the labeled and unlabeled data, respectively, and the prior distribution $p(z)$.

4.3 Discriminator

The discriminator distinguishes between the latent representations derived from the labeled data and those from the unlabeled data. If the VAE creates different latent space representations for the labeled and the unlabeled data, the discriminator penalizes the VAE by increasing the adversarial loss given by

$$\mathcal{L}_{\text{adv}}^{\text{VAE}} = -\mathbb{E}[\log D(q_\phi(z_L|x_L))] - \mathbb{E}[\log D(q_\phi(z_U|x_U))], \quad (2)$$

where $\mathcal{L}_{\text{adv}}^{\text{VAE}}$ is the adversarial loss for VAE; \mathbb{E} denotes the expectation over the specified distributions; $\log D(q_\phi(z_L|x_L))$ and $\log D(q_\phi(z_U|x_U))$ are the logarithms of the discriminator outputs for latent representations of labeled z_L and unlabeled z_U data samples, respectively, conditioned on their input data x_L and x_U . Through the minimization of this loss, the discriminator enforces the creation of similar latent space representations for the labeled and the unlabeled data. Similar representations of the labeled and unlabeled data makes the VAE model robust to the variation in the data.

When the discriminator identifies a sample as unlabeled, the sample is likely to have a significant difference in distribution compared to the labeled samples. Hence, such a sample should be labeled by an oracle and used in the training process to accommodate such varieties in the distribution. To that end, such a sample is forwarded to an oracle for labeling within the constraints of a predefined budget. The budget is a parameter that optimizes the labeling process by ensuring that only a limited number of samples are labeled in each iteration to maintain efficiency and cost-effectiveness. Once the oracle labels these samples,

the models are subsequently re-trained using the updated sets of labeled and unlabeled data. This iterative re-training process aims to improve the model’s accuracy and performance.

4.4 Sampling for Active Learning

Our sampling strategy, as shown in Fig. 3 of PSIVUS-Net, uses the probability scores of the discriminator’s predictions. We collect a batch of b samples with the lowest confidence scores predicted as “unlabeled” and send them to the oracle for labeling. The closer the probability is to zero, the higher the likelihood that the sample comes from the unlabeled pool. The core idea behind our approach is to prioritize the representativeness of samples, rather than depending solely on the training algorithm’s performance on the main task, which tends to be unreliable, especially in the early stages. Thus, we select samples based on their likelihood of belonging to the unlabeled pool according to the discriminator.

Our experiments start with an initial labeled pool comprising 100 images of the training set. For each batch, the budget size is set to 100 samples from the unlabeled pool. The remaining portion of the training set forms the pool of unlabeled data, from which samples are selected for annotation by the oracle and also used for producing better segmentation results. Once these samples are labeled, they are incorporated into the initial training set, and the training process is repeated on this augmented dataset.

4.5 Implementation Details

For the purpose of model training, validation and evaluation, the dataset is divided into three subsets where fine-tuning is performed with 6 patients where a subset (1 vessel data) of 1 patient is held from these 6 patients and is used for validation, curated to ensure representative sampling in various clinical scenarios and anatomical presentations. Subsequently, the remaining 4 patients’ data are in the test set, held separate from model training and validation processes that provide an unbiased evaluation of algorithmic performance.

The segmentation models UNet, UNet++, DeepLabV3+, and Multi-Scale Attention Net are initialized with ImageNet weights, which proves to be particularly advantageous given the limited size of the initial dataset. Using pre-trained weights allows the models to use the vast knowledge acquired from large-scale datasets. For the Variational Adversarial Active Learning (VAAL) framework, kaiming initialization is used to address the challenges associated with initializing neural network weights, ensuring more stable and efficient training. The combination of pre-trained weights for the segmentation models and kaiming initialization for VAAL optimizes the overall model initialization process.

The training process for the segmentation models is further optimized using the Adam optimizer and configured with a learning rate determined through grid search testing various batch sizes from 4 to 64 depending on model size and learning rates between 0.1 and 0.00001. Additionally, a learning rate scheduler,

specifically the ReduceLROnPlateau scheduler, adjusts the learning rate dynamically based on validation loss. If the validation loss stagnates for a set number of epochs, the learning rate is reduced to enhance convergence. An early stopping mechanism with patience of 10 epochs is implemented to prevent overfitting and halt training when no improvement in validation loss is observed within the specified period, thus ensuring the development of a robust and generalized segmentation model. Training is carried out on an A100 GPU, ensuring efficient handling of computational demands.

5 Results and Discussion

The experiments are performed on the proposed PSIVUS dataset using the protocol summarized in Table 1. In order to evaluate the effectiveness of the proposed model, we have compared the performance with state-of-the-art segmentation algorithms including UNet [21], UNet++ [30], DeepLabV3+ [5], MANet [7] along with different backbones: ResNet18 [9], VGG19 [22], MiT-b2 [26]. In addition, we have also compared the performance with active learning methodology, Variational Adversarial Active Learning (VAAL), and the results are evaluated in terms of Jaccard index, F1 score, precision and recall. Active learning methods are evaluated with incrementally adding a subset of the training images. Therefore, starting with an initial training set of 100 images, we have incrementally used 10%, 20%, 30%, and the full training dataset, while the testing subset remains the same throughout. The results obtained with the best encoder-decoder combinations are summarized in Table 2.

Some visual results of the proposed and existing algorithms are shown in Fig. 4. The first row displays the original input images that are grayscale and show cross-sectional views of coronary artery. The second row provides the ground truth segmentations, binary masks indicating the plaque in white region. The third to seventh row indicate the outputs from different segmentation models. Here, UNet results closely follow the ground truth but sometimes have more FPs in comparison to others as in Sample 1 and more FNs as in Sample 4. UNet++ provides more refined boundaries and captures details better. However, there are still some areas where the segmentation could be improved like in Sample 5. MANet’s performance is similar to UNet++, with well-defined boundaries but some segmentations show slight over-segmentation, where the predicted mask is larger than the ground truth as visible in Sample 5. DeepLabV3+ tends to produce smoother boundaries and captures the overall shape well. The final row presents the results from the proposed model and is very close to the ground truth, with well-defined boundaries and accurate representations of the regions of interest. It seems to handle both the overall shape and finer details better than the other models but sometimes has worse performance than others.

Across both existing and proposed active learning algorithms, we observe that training with only 20–30% of the entire dataset yields results comparable to those obtained by using the entire dataset. PSIVUS-Net exhibits promising performance across various data splits, achieving competitive mean IoU scores.

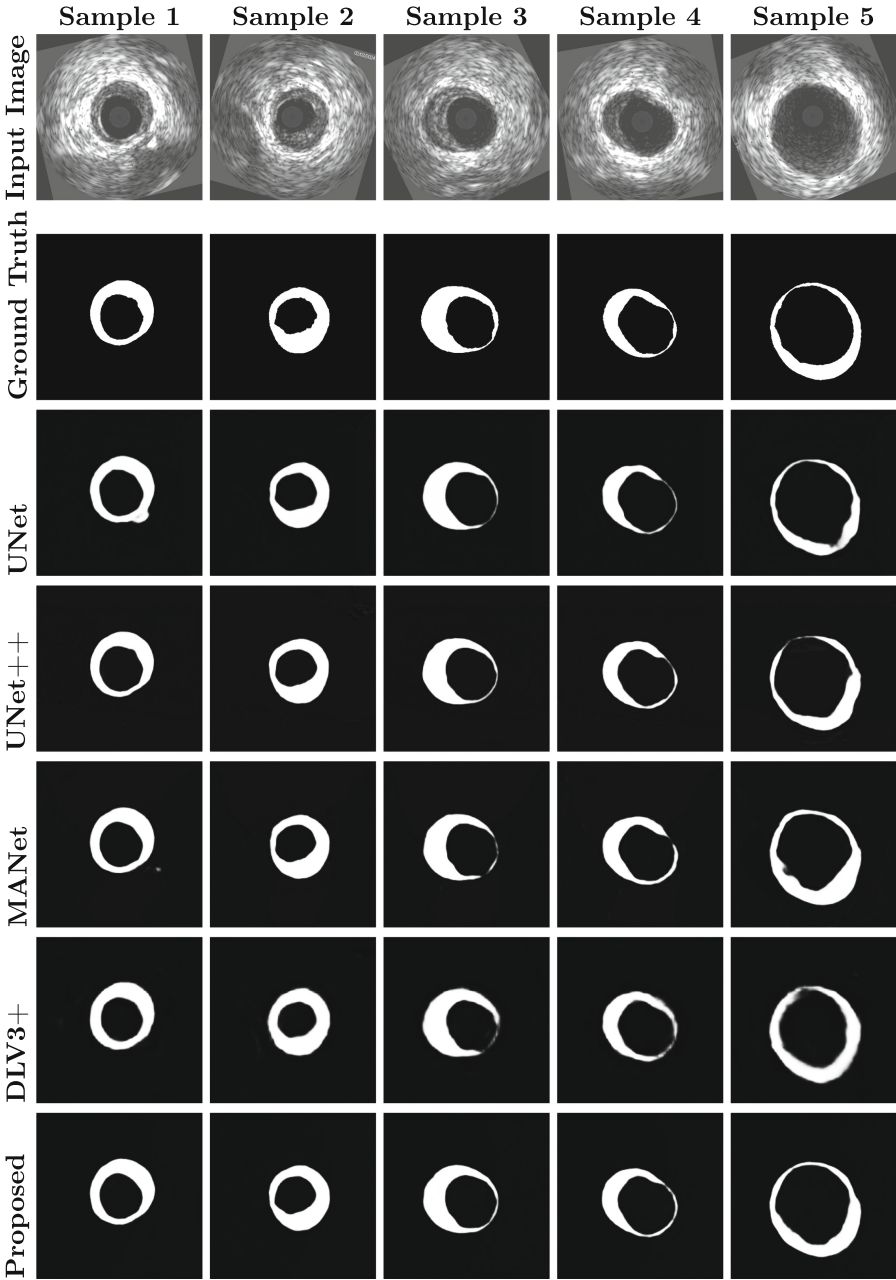


Fig. 4. The visual results after training with 30% of the training data for different models are shown here, where the various columns denote the different sample images used and each row has a different model; Row 1: Input Image, Row 2: Ground Truth Mask, Row 3: UNet, Row 4: UNet++, Row 5: MANet, Row 6: DeepLabV3+, Row 7: Proposed PSIVUS-Net.

Table 2. Performance of different models with Active Learning (AL) on 10%, 20%, 30% and 100% of the training data.

AL	Model	Ratio	Jaccard	F1	Recall	Precision
VAAL [23]	UNet [21]	0.10	0.77	0.85	0.81	0.92
		0.20	0.77	0.86	0.82	0.92
		0.30	0.78	0.87	0.84	0.92
		1.00	0.79	0.88	0.85	0.91
	UNet++ [30]	0.10	0.76	0.85	0.81	0.90
		0.20	0.77	0.86	0.83	0.92
		0.30	0.78	0.86	0.83	0.92
		1.00	0.79	0.87	0.84	0.91
	MANet [7]	0.10	0.63	0.73	0.67	0.93
		0.20	0.76	0.85	0.81	0.92
		0.30	0.77	0.86	0.82	0.92
		1.00	0.78	0.92	0.83	0.87
	DeepLabV3+ [5]	0.10	0.73	0.82	0.78	0.89
		0.20	0.74	0.84	0.80	0.89
		0.30	0.77	0.86	0.86	0.88
		1.00	0.77	0.86	0.83	0.89
PSIVUS-Net (Proposed)	0.10	0.77	0.84	0.81	0.90	
	0.20	0.77	0.86	0.82	0.91	
	0.30	0.78	0.86	0.82	0.91	
	1.00	0.79	0.88	0.85	0.92	

This range highlights its robust segmentation capabilities. Furthermore, precision scores indicate high accuracy in identifying positive instances, while recall values reflect its effectiveness in capturing relevant instances from the dataset. PSIVUS-Net shows comparable performance to VAAL with UNet model and has better performance than the other models.

The results indicate that PSIVUS-Net demonstrates the ability to select the most representative samples in each iteration that are equivalent to those obtained with the entire dataset. The proposed architecture can be further explored to accommodate a balance between segmentation and sample selection with more adaptability that could help improve its performance. The other models present competitive metrics across various evaluation criteria, indicating their effectiveness in reducing data annotation costs and accelerating learning. These findings highlight the potential of these techniques to improve the efficiency of segmentation tasks in various applications. More experimentation can be done to explore their performance under different scenarios and datasets and identify potential areas for improvement and optimization.

6 Conclusion

This research makes contributions to the field of medical image analysis by addressing the challenges associated with plaque segmentation in intravascular ultrasound images. Through the development of the PSIVUS dataset and the introduction of PSIVUS-Net, an active learning framework, we have demonstrated that deep learning models can achieve high performance with substantially fewer annotated samples. This efficiency not only reduces the burden of manual annotation but also accelerates the adoption of automated diagnostic tools in clinical settings. Our work offers a promising pathway towards more accurate, reliable, and scalable solutions for cardiovascular disease diagnosis.

Acknowledgement. This research is funded through the grant from iHub Drishti Foundation, TIH on CV & ARVR under NM-ICPS, DST, Government of India.

References

1. Alshehhi, R., Alshehhi, A.: Quantification of uncertainty in brain tumor segmentation using generative network and bayesian active learning. In: VISIGRAPP, pp. 701–709 (2021)
2. Bajaj, R., et al.: Advanced deep learning methodology for accurate, real-time segmentation of high-resolution intravascular ultrasound images. *Int. J. Cardiol.* **339**, 185–191 (2021)
3. Billah, M.E., Javed, F.: Bayesian convolutional neural network-based models for diagnosis of blood cancer. *Appl. Artif. Intell.* **36**(1), 2011688 (2022)
4. Blanco, P.J., Ziemer, P.G., Bulant, C.A., Ueki, Y., Bass, R., Räber, L., Lemos, P.A., García-García, H.M.: Fully automated lumen and vessel contour segmentation in intravascular ultrasound datasets. *Med. Image Anal.* **75**, 102262 (2022)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision, pp. 801–818 (2018)
6. Dong, L., Jiang, W., Lu, W., Jiang, J., Zhao, Y., Song, X., Leng, X., Zhao, H., Wang, J., Li, C., et al.: Automatic segmentation of coronary lumen and external elastic membrane in intravascular ultrasound images using 8-layer u-net. *Biomed. Eng. Online* **20**(1), 1–9 (2021)
7. Fan, T., Wang, G., Li, Y., Wang, H.: Ma-net: a multi-scale attention network for liver and tumor segmentation. *IEEE Access* **8**, 179656–179665 (2020)
8. Gaillochet, M., Desrosiers, C., Lombaert, H.: Active learning for medical image segmentation with stochastic batches. *Med. Image Anal.* **90**, 102958 (2023)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Hesamian, M.H., Jia, W., He, X., Kennedy, P.: Deep learning techniques for medical image segmentation: achievements and challenges. *J. Digit. Imaging* **32**, 582–596 (2019)
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)

12. Katouzian, A., Sathyanarayana, S., Baseri, B., Konofagou, E.E., Carlier, S.G.: Challenges in atherosclerotic plaque characterization with intravascular ultrasound (ivus): from data collection to classification. *IEEE Trans. Inf. Technol. Biomed.* **12**(3), 315–327 (2008)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114)* (2013)
14. Mahapatra, D., Poellinger, A., Shao, L., Reyes, M.: Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. *IEEE Trans. Med. Imaging* **40**(10), 2548–2562 (2021)
15. Matsumura, M., et al.: Accuracy of ivus-based machine learning segmentation assessment of coronary artery dimensions and balloon sizing. *JACC: Advances* **2**(7), 100564 (2023)
16. Mendizabal-Ruiz, G., Rivera, M., Kakadiaris, I.A.: A probabilistic segmentation method for the identification of luminal borders in intravascular ultrasound images. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. *IEEE* (2008)
17. Nishi, T., et al.: Deep learning-based intravascular ultrasound segmentation for the assessment of coronary artery disease. *Int. J. Cardiol.* **333**, 55–59 (2021)
18. Nissen, S.E., Yock, P.: Intravascular ultrasound: novel pathophysiological insights and current clinical applications. *Circulation* **103**(4), 604–616 (2001)
19. Ozdemir, F., Peng, Z., Fuernstahl, P., Tanner, C., Goksel, O.: Active learning for segmentation based on bayesian sample queries. *Knowl.-Based Syst.* **214**, 106531 (2021)
20. Razzak, M.I., Naz, S., Zaib, A.: Deep learning for medical image processing: overview, challenges and the future. *Classification in BioApps: Automation of Decision Making*, pp. 323–350 (2018)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241 (2015)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)* (2014)
23. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: *IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981 (2019)
24. Suzuki, S., et al.: Topological structural analysis of digitized binary images by border following. *Comput. Vision Graph. Image Process.* **30**(1), 32–46 (1985)
25. Wilkins, E., et al.: European cardiovascular disease statistics 2017 (2017)
26. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: simple and efficient design for semantic segmentation with transformers. *Adv. Neural. Inf. Process. Syst.* **34**, 12077–12090 (2021)
27. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention*, pp. 399–407 (2017)
28. Zakeri, F.S., Setarehdan, S.K., Norouzi, S.: Automatic media-adventitia ivus image segmentation based on sparse representation framework and dynamic directional active contour model. *Comput. Biol. Med.* **89**, 561–572 (2017)
29. Zhao, S., Song, J., Ermon, S.: Infovae: information maximizing variational autoencoders. *arXiv preprint [arXiv:1706.02262](https://arxiv.org/abs/1706.02262)* (2017)
30. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: a nested u-net architecture for medical image segmentation. In: *MICCAI Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11. Springer (2018)



Generalist Segmentation Algorithm for Photoreceptors Analysis in Adaptive Optics Imaging

Mikhail Kulyabin¹(✉), Aline Sindel¹, Hilde R. Pedersen², Stuart Gilson²,
Rigmor Baraas², and Andreas Maier¹

¹ Pattern Recognition Lab, Department of Computer Science,
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
mikhail.kulyabin@fau.de

² National Centre for Optics, Vision and Eye Care, Faculty of Health and Social
Sciences, University of South-Eastern Norway, Kongsberg, Norway

Abstract. Analyzing the cone photoreceptor pattern in images obtained from the living human retina using quantitative methods can be crucial for the early detection and management of various eye conditions. Confocal adaptive optics scanning light ophthalmoscope (AOSLO) imaging enables visualization of the cones from reflections of waveguiding cone photoreceptors. While there have been significant improvements in automated algorithms for segmenting cones in confocal AOSLO images, the process of labeling data remains labor-intensive and manual. This paper introduces a method based on deep learning (DL) for detecting and segmenting cones in AOSLO images. The models were trained on a semi-automatically labeled dataset of 20 AOSLO batches of images of 18 participants for 0° , 1° , and 2° from the foveal center. F1 scores were 0.968, 0.958, and 0.954 for 0° , 1° , and 2° , respectively, which is better than previously reported DL approaches. Our method minimizes the need for labeled data by only necessitating a fraction of labeled cones, which is especially beneficial in the field of ophthalmology, where labeled data can often be limited.

Keywords: AOSLO · cones · photoreceptors · segmentation · detection

1 Introduction

Adaptive optics scanning light ophthalmoscopy (AOSLO) [15] offers a noninvasive approach to achieve high-resolution, in vivo imaging of the cone photoreceptors (cones) mosaic in both healthy and diseased retinas [23]. The AOSLO technique integrates an adaptive optics (AO) system within a scanning light ophthalmoscope (SLO) [17]. The AO system employs a wavefront sensor and an actuated mirror to measure and dynamically compensate for wavefront aberrations caused by the eye's inhomogeneous medium. While AO can be utilized with any ophthalmic imaging device requiring light passage into or out of the

eye, it is predominantly used with SLOs due to its superior contrast and resolution capabilities. Multimodal AOSLO imaging captures three channels simultaneously (confocal, split-detection, and dark-field), each emphasizing different retinal structures. The confocal modality of AOSLO facilitates relatively explicit imaging of cones and rods [17], presenting clinicians and researchers with quantifiable but complex retinal structural information [12]. Using this technology, one can obtain various quantitative measures of the cone mosaic from AOSLO images, such as cone density, spacing, and pattern regularity [2, 8]. Such quantities are useful for developing sensitive biomarkers for early diagnosis and monitoring of ocular and systemic disease progression.

Considering just the cones, peak foveal density can approach 200,000 cones per mm^2 [5], making manual labeling impractical. On the other hand, existing automatic labeling techniques may not consistently enable the automatic identification of every cone within an image, particularly in the presence of blood vessels or when the image clarity is compromised. Furthermore, the challenge intensifies when examining retinal locations that are more eccentric from the fovea.

Using the Voronoi algorithm, we cover the area from center-to-center of a cone detected in the confocal image [8]. As we move out from the foveal center, we move from an area with only cones and where the Voronoi cell is equal to the cone’s size to areas with rods in between cones. This has already happened about 0.5° – 1° from the foveal center. Thus, in areas with rods and cones, the Voronoi represents distances between cones but not their size. Classical methods, such as presented in work by Li and Roorda [7], which are currently used in contemporary works, rely on the optical fiber properties of cone photoreceptors. In practice, the algorithm can mislabel rods as cones. Therefore, it needs to be revised by a human expert. New algorithms should take this into account. Several algorithms have been previously developed to detect inner segments in split-detection images [4, 18]. In general, AOSLO split-detection images are semi-automatically analyzed to extract the location of cone photoreceptor cells within the images, with compulsory refinement by a medical expert. Creating a fully automatic method for the segmentation and detection of cones will significantly increase the possibilities of retinal research and reduce the workload of retinal researchers. This paper introduces a deep learning (DL) –based method for automatically detecting and segmenting the cones.

2 Related Works

Cellpose [20] is a versatile, generalist algorithm for cell segmentation in microscopy images, regardless of the imaging modality or the type of cells being analyzed. It employs a DL model to identify cell boundaries, enabling automated and accurate segmentation of individual cells or nuclei across various applications. The algorithm uses a novel approach based on the concept of “flows” to capture cells’ complex shapes and sizes, making it highly effective in different

biological contexts. The term “flows” refers to the vector field that is generated for each pixel in the image, pointing towards the center of the cell to which that pixel belongs. Cellpose 2.0 [11] is an updated version with a manual correction step for training custom models. However, it requires considerable effort to manually correct the detected polygons of multiple cells, which is significant for the number of receptor cells in AOSLO images (up to 200,000 cells per mm^2). Another segmentation method, PolarMask, is a single-shot, anchor-free convolutional neural network (CNN) framework designed for instance segmentation [24]. Unlike traditional instance segmentation methods that rely on bounding boxes or complex, multi-stage processes, PolarMask simplifies this by utilizing a polar representation to capture the shape of each object. It generates a center point for the object and then defines the segmentation boundary through a set of rays emanating from the center to the boundary in polar coordinates. This approach allows PolarMask to perform instance segmentation efficiently and accurately without the need for anchor boxes, reducing the complexity and computational demands of the task. StarDist is a novel image segmentation method optimized for microscopy images, particularly those of nuclei and cells, leveraging a shape-based approach to outline individual objects’ boundaries [16, 22]. The core innovation of StarDist lies in its use of star-convex shapes for segmentation, where it predicts the distances from the center of an object to its boundary in a fixed set of directions, effectively capturing the often complex and irregular shapes of biological cells. This method is implemented through a DL framework, allowing it to learn from annotated training data and generalize well to new, unseen images. StarDist stands out for its ability to handle overlapping structures and varying shapes, making it highly effective for tasks where segmenting closely packed or irregularly shaped cells is critical. Its performance and efficiency make it a valuable tool for biomedical image analysis, facilitating advanced quantitative studies of cellular structures.

These methods are versatile and efficient computational tools for segmentation, demonstrating significant performance in various biological imaging contexts [19, 21]; however, they are designed to generalize across different types of cells and imaging modalities by leveraging a unique representation of cell shapes. Despite its robustness and adaptability, applying them to segment structures derived from Voronoi diagrams may require modifications. Cunefare et al. [4] applied CNN to confocal AOSLO images to detect the cones, extracting the training patches using the Voronoi algorithm. However, the method does not involve the segmentation of the cells. The AOSLO images belonged to patients with achromatopsia disease –which have many inactive cones – and are, in fact, black regions on the images and are very different from the active cones in our dataset.

3 Methods

3.1 Dataset

In this study, we employed a semi-automatically labeled dataset of 20 AOSLO batches of images of 18 healthy participants with normal vision from a wide age range, 14–65 years, representing a wide scope of healthy retinas. Each batch consists of approximately 40 confocal images. The cone centers were first automatically identified with the classical method [7]. Then, missed cones were manually added by a human expert or removed if they were mislabeled by the automatic algorithm, representing approximately 5% of all cones across the dataset images. The data was split on the participant’s level so that images belonging to one participant appeared only in one subset, with a train: test split ratio of 70:30. Therefore, we had 14 batches (540 images) of AOSLO images for training and 6 batches (240 images) for testing. Each image was cropped to 550×550 pixel resolution with 350 labeled cones on each image on average, for a total of 190k segmented cells in the training subset at the starting point. Figure 1 shows examples of confocal images with labeled cone centers.

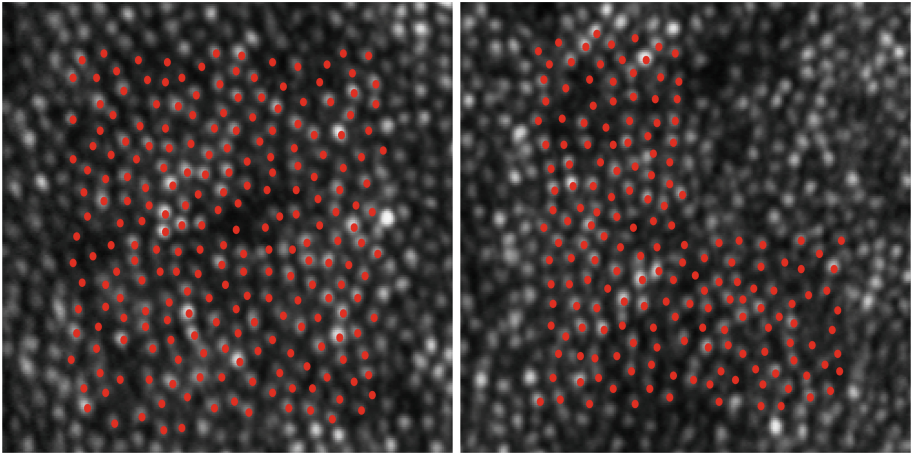


Fig. 1. Two examples of confocal AOSLO images with labeled cone centers using the existing semi-automatic segmentation method [7] followed by refinement by a medical expert.

3.2 Human-In-The-Loop Approach

Figure 2 illustrates the overall pipeline of the proposed method. AOSLO images were labeled and split into the test and train subsets as described in the dataset Sect. 3.1. To all labeled areas, we applied the Voronoi algorithm to obtain the

masks of the cones. Then, the human-in-the-loop step was applied: on the initially labeled AOSLO images, we trained DL-based models to generate semantic masks on unlabeled images. Then, from semantic masks, we calculated the center of mass for each segment (cone), which is basically the center of the cone we manually labeled. Therefore, we could evaluate the models by comparing the obtained centers with ground truth labels. Given the potential for initial inaccuracies, manual correction is a crucial step in the method. This step ensures the precision of the model’s output by allowing the expert to review and adjust the segmented cone centers, mitigating the risk of errors in the initial automated segmentation. Adding new annotations and the manual correction step are not involved in the initial zero iteration, they are only applied starting from the first iteration step.

The Voronoi algorithm is reapplied to the refined data after correction of the centers of the cones, which was done in EXACT [9]. This iterative process refines the segmentation accuracy and enriches the training dataset with additional, corrected instances. Thus, the next 15% of the total number of images is labeled at each iteration, increasing the training dataset. Each iteration concludes with an evaluation step on a test dataset to quantify the improvements. Additionally, at this stage, we apply the K -means algorithm for clustering the cones by the mean brightness of the center part (reflection). Therefore, we obtain the percentage of reflecting and non-reflecting cones, which is also a priori information for diagnostics. This cyclical process, encompassing both automated segmentation and expert review, ensures the development of a robust model capable of high-precision cone segmentation.

3.3 Voronoi Algorithm

The Voronoi algorithm is one of the more useful geometrical constructions to study point patterns since it provides all the information needed to study proximity relations between points [10]. Connecting surrounding cones and characterizing the number of sides, the Voronoi diagram allows assessment of the degree of hexagonality, and it is often used to show how disease and aging can affect this aspect of packing geometry [1]. In a healthy retina, cones are packed in the most efficient manner possible, which is a hexagonal (honeycomb) arrangement. The degree of hexagonality, therefore, can be used as a proxy for general retinal health. Applying the Voronoi algorithm, we can obtain a reasonably accurate approximation of photoreceptor segmentation by labeling only the centers of cones (Fig. 3).

3.4 Attention-Augmented U-Net

Figure 4 shows the overview of the model. In our model, we applied the concept of flows (vector gradient fields) [20]. This means that we trained a neural network to predict the horizontal and vertical gradients of the topological maps. Additionally, the network predicts a binary map to indicate if a given pixel is inside or outside of regions of interest. Our model was based on the general U-Net

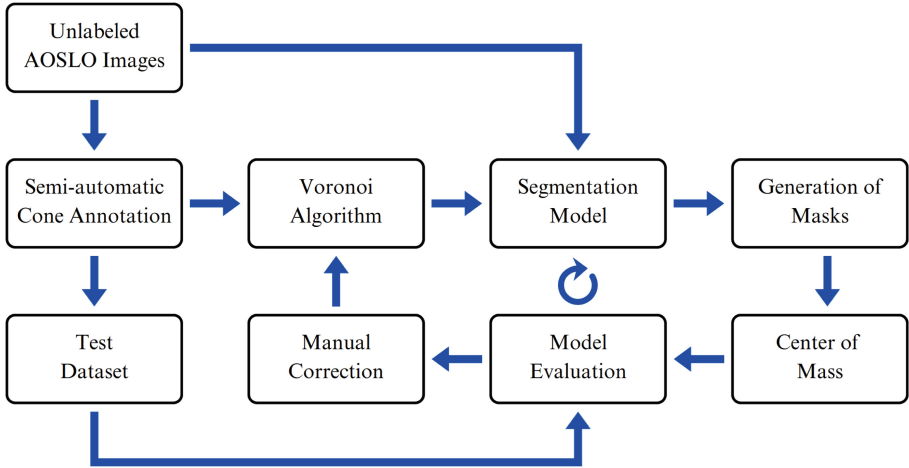


Fig. 2. Pipeline of the method. Voronoi algorithm is applied to initially semi-automatically annotated cones to obtain the masks. Then, a segmentation model was trained, which generates segmentation masks for unlabeled AOSLO images. The center of mass function is applied to get the centers of the cells from segmentation masks. After each iteration step, the model was evaluated using the test subset. A manual correction step is involved in the pipeline to improve the annotations of the segmentation model of initially unlabeled images.

architecture [14] with an additional attention-augmented module (AA module) [13]. This module dynamically adjusts the importance of different spatial regions and channels in the input data, enabling the network to prioritize more relevant features for improved segmentation accuracy. Such augmentation facilitates precise localization and detailed segmentation in complex image datasets, which is particularly beneficial in medical imaging applications where accuracy is essential. Attention mechanisms can help the model to focus on relevant features and ignore distractions, therefore, improving segmentation accuracy.

The AA module improves the performance of overlapping or docked objects. The nature of the Voronoi algorithm ensures the cells are always tightly packed, with no possibility of spaces in between. The AA module helps to distinguish between adjacent objects by prioritizing spatial features that define boundaries, enhancing the model’s ability to separate and accurately segment individual cells.

3.5 Center of Mass

We calculated the center of mass to extract the centers of the cones. The center of mass is a point that corresponds to the average position of all the mass in a

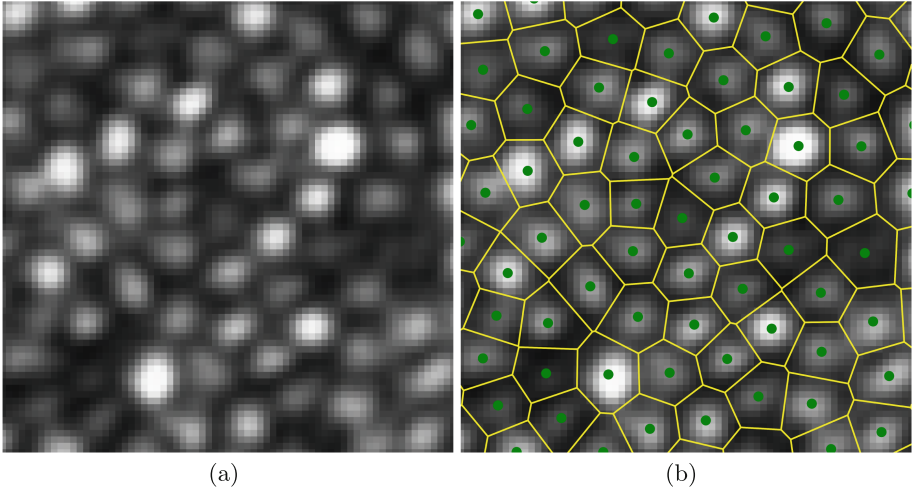


Fig. 3. Application of Voronoi algorithm on the labeled AOSLO images: (a) example of the original image; (b) segmented image.

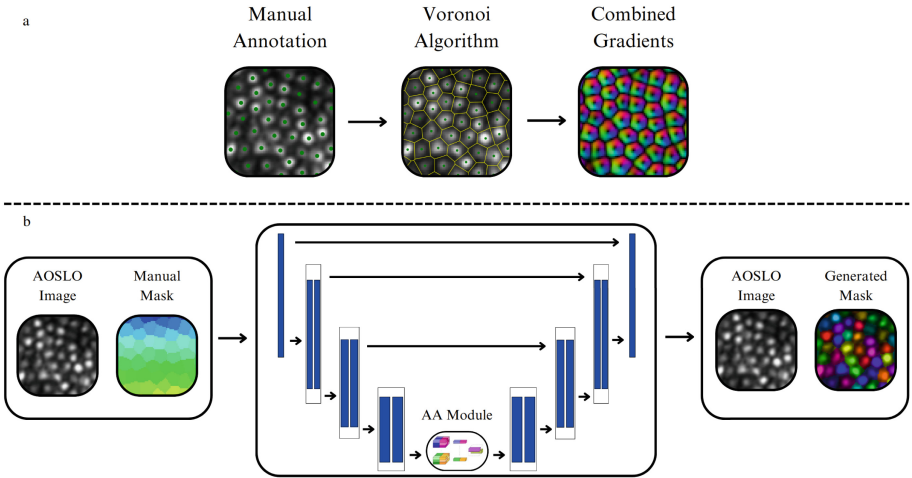


Fig. 4. Model overview. (a) Transformation from the center of the cell to a gradient vector field using the Voronoi algorithm. (b) U-Net model with additional Attention-augmented module.

system. For discrete systems, the center of mass can be considered the weighted average of the positions of all elements, where the weights are the values of those elements. The cone may have several pixels corresponding to the brightest color: using the center of mass, we get the center of the brightest area.

4 Experiments

4.1 Training Setup

The model was trained for 500 epochs on each iteration with stochastic gradient descent with a learning rate of 0.001, a momentum of 0.9, a batch size of 16 images, and a weight decay of 0.0001. All the models were trained on a single NVIDIA A100 graphics processing unit on a machine with two Intel Xeon Gold 6134 3.2 GHz and 96 GB RAM. One training iteration on this setup lasts 30 min, with about 10s for the further inference of one image from a batch.

To predict the horizontal and vertical gradients, we used the MSE loss function. We applied the cross-entropy loss function to predict the probability that a pixel was inside or outside a cell.

4.2 Metrics

To match predicted points and ground truth, we applied the KDTree algorithm [6]. Each predicted cell center matched with a ground truth pair was True Positive (TP), a predicted cone without a ground truth pair was False Positive (FP), and when nothing was detected where ground truth indicates a cone was a False Negative (FN) case. The L_2 distance (D_{L_2}) between pairs of points was calculated using the following formula:

$$D_{L_2} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}, \quad (1)$$

where a and b are predicted and ground truth centers, respectively. Detected cones were evaluated using Recall, Precision, and F1-score:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (2)$$

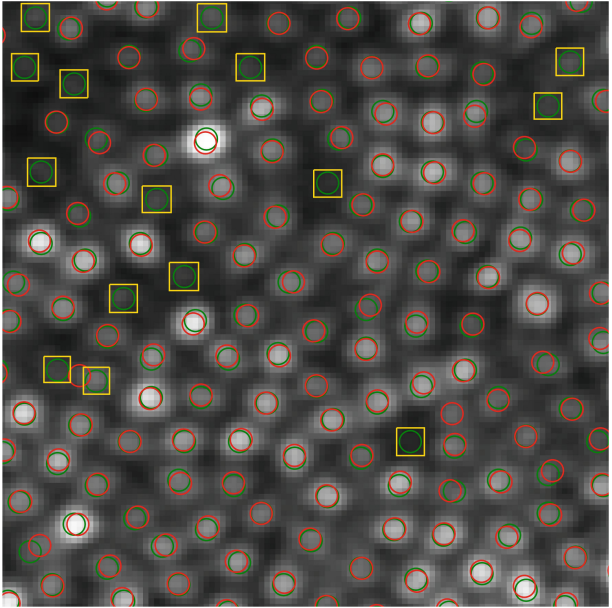
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (3)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (4)$$

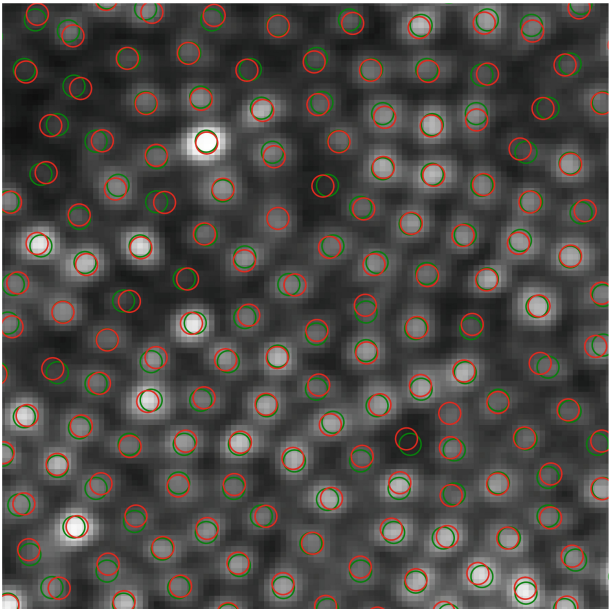
4.3 Results

Figure 5 shows an example of the application of the algorithm on the first (a) and second (b) iterations. Green and red circles correspond to ground truth and predicted cell centers, respectively. Yellow squares on the first iteration show unlabeled cells (FNs); on the second iteration, they are correctly labeled.

Figure 6a shows an example of the predicted semantic mask by our model. For this mask, the center of mass was calculated, obtaining the centers of cones that are shown in Fig. 6b. Predicted centers (red) are matched with ground truth (green), and the distance is shown with blue connection lines.



(a) First iteration



(b) Second iteration

Fig. 5. Evaluation of the proposed method on first (a) and second (b) iterations on a 0° test sample. Green circles correspond to the ground truth cone centers, and red to the predicted centers. Yellow squares show the False Negative predictions of the model. (Color figure online)

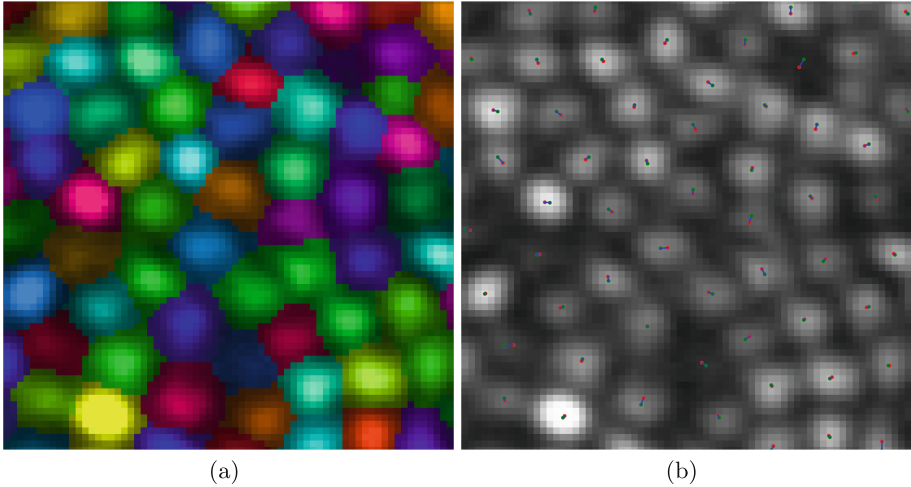


Fig. 6. Example of a predicted segmentation mask (a); example of the matching of predicted (red) and ground truth (green) centers (b). Blue lines show the L_2 distance. (Color figure online)

Table 1 presents the comparative performance of the StarDist, Cellpose, and our models. Recall, Precision, and F1 score were computed separately for 0° , 1° , and 2° from the fovea. D_{L_2} was calculated for all three degrees together. All models are characterized by a fundamental improvement after the second correction: on average, the F1 score improved by 7%. We also see a tendency for the scores to deteriorate slightly at higher eccentricities.

Table 1. Evaluation metrics of the trained models. Best results are marked in bold.

Model	It.	0°			1°			2°			All D_{L_2}
		Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	
StarDist	1	0.833	0.930	0.879	0.824	0.919	0.869	0.810	0.920	0.861	7.653
Cellpose	1	0.843	0.952	0.895	0.843	0.941	0.890	0.811	0.942	0.872	7.641
Ours	1	0.854	0.953	0.901	0.846	0.954	0.897	0.841	0.955	0.894	7.637
StarDist	2	0.927	0.946	0.936	0.927	0.936	0.931	0.891	0.937	0.913	7.539
Cellpose	2	0.937	0.967	0.952	0.937	0.957	0.947	0.902	0.948	0.925	7.534
Ours	2	0.958	0.978	0.968	0.948	0.968	0.958	0.940	0.969	0.954	7.529

The obtained centers were clustered in terms of brightness to monitor the distribution of light-reflecting and non-reflecting (dark) photoreceptors during training. The K -means algorithm is a popular unsupervised machine learning technique for clustering data into a specified number of clusters, denoted by K .

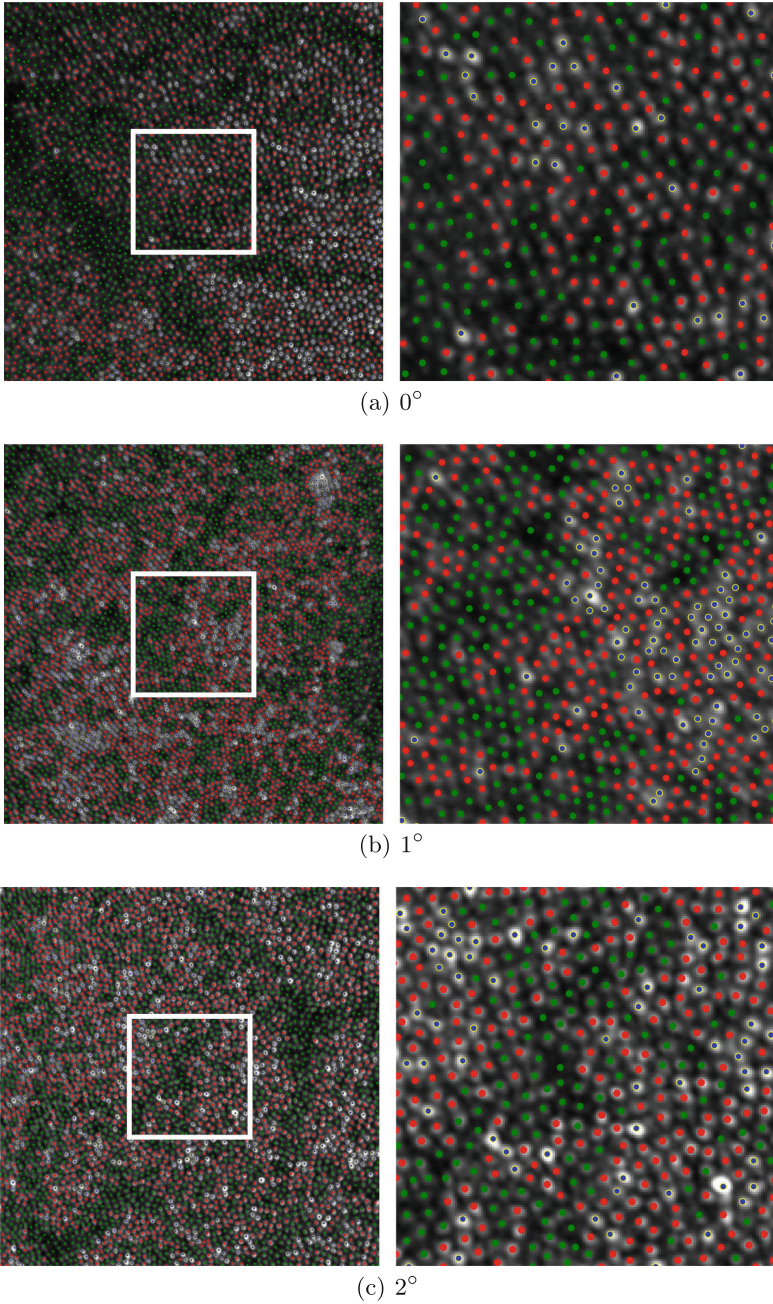


Fig. 7. Examples of the method performance on the test images located at 0° , 1° , and 2° from the fovea with applied *K*-Means clustering algorithm. Blue-marked labels correspond to the cones with the highest reflection, green to the lowest, and red to the middle. White boxes show the location of the zoomed area in the right column.

We applied the K -means algorithm with three clusters ($K = 3$) during each iteration for additional distribution of reflecting cone control.

Clustering the cones based on their brightness level is particularly useful in retinal imaging for understanding differences between healthy and diseased retinas. Figure 7 shows examples of clustering on images of 0° , 1° , and 2° .

Figure 8 plots the cumulative average number of corrected cone center identifications made for all three models. The values show that our proposed improvement will decrease the number of corrections for each image, compared with using Cellpose or StarDist models for the human-in-the-loop approach, potentially saving the human expert’s time cost for the AOSLO cells segmentation.

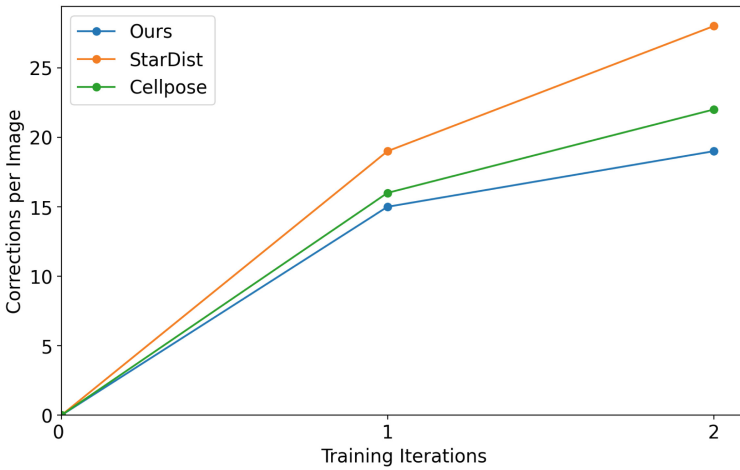


Fig. 8. The cumulative average number of corrections of cone centers on the first and second iterations per image for Cellpose, StarDist, and the proposed model. The initial training iteration was done with the original labeled dataset; therefore, the number of corrections was equal to zero.

5 Conclusion

This work describes and evaluates a method for the identification and segmentation of cone photoreceptors from AOSLO confocal images. Models were trained and tested on images covering a more extensive range of images of 18 participants with only 5% labeled cones. Our proposed method received an overall F1 score of 0.968 for cones for 0° , 0.958 for 1° , and 0.954 for 2° , which is better than previously reported DL approaches [3,4]. Our method can reduce the labeling effort by requiring only a portion of labeled cones and is particularly advantageous in the ophthalmology field, where labeled data can be scarce. The

work is limited to the range of eccentricities from the center of the fovea - 0° , 1° , and 2° . Rods are already present at 1° but peaks in density at around 15° ; thus, rods become more and more visible on images between cones, which also require detection. Therefore, a potential improvement of the method could be to add the annotations for rods for implementing rods detection for eccentricities more than 2° . This could be done using the calculated modality of the AOSLO images.

The method can be extended to the automatic identification of areas that are not cones, enabling these regions to estimate rod density. Incorporating automatic detection of inner segments in split-detection images could help to confirm that the reflected light and/or dark areas in confocal images correspond to cones. This would allow for an estimation of the number of dark cones. Furthermore, identifying retinal pigment epithelium (RPE) cells as part of this process would significantly enhance the methods' utility for clinical work and research, which leads us to future work.

6 Code Availability

The code used to generate the results in this paper will be available at github.com/MikhailKulyabin/AOSLO

Acknowledgments. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).

References

1. Baraas, R.C., Carroll, J., Gunther, K.L., Chung, M., Williams, D.R., Foster, D.H., Neitz, M.: Adaptive optics retinal imaging reveals s-cone dystrophy in tritan color-vision deficiency. *JOSA A* **24**(5), 1438–1447 (2007)
2. Cooper, R.F., Wilk, M.A., Tarima, S., Carroll, J.: Evaluating descriptive metrics of the human cone mosaic. *Investigative Ophthalmology Visual Sci.* **57**(7), 2992–3001 (2016)
3. Cunefare, D., Huckenpahler, A.L., Patterson, E.J., Dubra, A., Carroll, J., Farsiu, S.: Rac-cnn: multimodal deep learning based automatic detection and classification of rod and cone photoreceptors in adaptive optics scanning light ophthalmoscope images. *Biomed. Opt. Express* **10**(8), 3815–3832 (2019)
4. Cunefare, D., et al.: Deep learning based detection of cone photoreceptors with multimodal adaptive optics scanning light ophthalmoscope images of achromatopsia. *Biomed. Opt. Express* **9**(8), 3740–3756 (2018)
5. Curcio, C.A., Sloan, K.R., Kalina, R.E., Hendrickson, A.E.: Human photoreceptor topography. *J. Comparative Neurol.* **292**(4), 497–523 (1990)
6. Friedman, J.H., Bentley, J.L., Finkel, R.A.: An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw. (TOMS)* **3**(3), 209–226 (1977)
7. Li, K.Y., Roorda, A.: Automated identification of cone photoreceptors in adaptive optics retinal images. *JOSA A* **24**(5), 1358–1363 (2007)

8. Litts, K.M., Cooper, R.F., Duncan, J.L., Carroll, J.: Photoreceptor-based biomarkers in aoslo retinal imaging. *Investigative Ophthalmology Visual Sci.* **58**(6), BIO255–BIO267 (2017)
9. Marzahl, C., Aubreville, M., Bertram, C.A., Maier, J., Bergler, C., Kröger, C., Voigt, J., Breininger, K., Klopffleisch, R., Maier, A.: Exact: a collaboration toolset for algorithm-aided annotation of images with annotation version control. *Sci. Rep.* **11**(1), 4343 (2021)
10. Mozos, O.M., Bolea, J.A., Ferrandez, J.M., Ahnelt, P.K., Fernandez, E.: V-proportion: a method based on the voronoi diagram to study spatial relations in neuronal mosaics of the retina. *Neurocomputing* **74**(1–3), 418–427 (2010)
11. Pachitariu, M., Stringer, C.: Cellpose 2.0: how to train your own model. *Nature Methods* **19**(12), 1634–1641 (2022)
12. Pedersen, H.R., Gilson, S., Hagen, L.A., Holtan, J.P., Bragadottir, R., Baraas, R.C.: Multimodal in-vivo maps as a tool to characterize retinal structural biomarkers for progression in adult-onset stargardt disease. *Front. Ophthalmol.* **4**, 1384473 (2024)
13. Rajamani, K.T., Rani, P., Siebert, H., ElagiriRamalingam, R., Heinrich, M.P.: Attention-augmented u-net (aa-u-net) for semantic segmentation. *SIViP* **17**(4), 981–989 (2023)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer (2015)
15. Roorda, A., Romero-Borja, F., Donnelly, W.J., III, Queener, H., Hebert, T.J., Campbell, M.C.: Adaptive optics scanning laser ophthalmoscopy. *Opt. Express* **10**(9), 405–412 (2002)
16. Schmidt, U., Weigert, M., Broaddus, C., Myers, G.: Cell detection with star-convex polygons. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II*, pp. 265–273 (2018). https://doi.org/10.1007/978-3-030-00934-2_30
17. Scoles, D., Sulai, Y.N., Langlo, C.S., Fishman, G.A., Curcio, C.A., Carroll, J., Dubra, A.: In vivo imaging of human cone photoreceptor inner segments. *Investigative Ophthalmol. Visual Sci* **55**(7), 4244–4251 (2014)
18. Sredar, N., Razeen, M., Kowalski, B., Carroll, J., Dubra, A.: Comparison of confocal and non-confocal split-detection cone photoreceptor imaging. *Biomed. Opt. Express* **12**(2), 737–755 (2021)
19. Stevens, M., Nanou, A., Terstappen, L.W., Driemel, C., Stoecklein, N.H., Coumans, F.A.: Stardist image segmentation improves circulating tumor cell detection. *Cancers* **14**(12), 2916 (2022)
20. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**(1), 100–106 (2021)
21. Waisman, A., Norris, A.M., Elías Costa, M., Kopinke, D.: Automatic and unbiased segmentation and quantification of myofibers in skeletal muscle. *Sci. Rep.* **11**(1), 11793 (2021)
22. Weigert, M., Schmidt, U., Haase, R., Sugawara, K., Myers, G.: Star-convex polyhedra for 3d object detection and segmentation in microscopy. In: *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. <https://doi.org/10.1109/WACV45572.2020.9093435>

23. Wynne, N., Carroll, J., Duncan, J.L.: Promises and pitfalls of evaluating photoreceptor-based retinal disease with adaptive optics scanning light ophthalmoscopy (aoslo). *Prog. Retin. Eye Res.* **83**, 100920 (2021)
24. Xie, E., et al.: Polarmask: single shot instance segmentation with polar representation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12193–12202 (2020)



UNeXt++: A Serial-Parallel Hybrid UNeXt for Rapid Medical Image Segmentation

Yan Li¹(✉), Juelin Wang¹, Yunteng Deng², Binyang Li¹, and Junlin Hu³

¹ School of Cyber Science and Engineering, University of International Relations, Beijing, China

liyan@uir.edu.cn

² School of Cyberspace Security, Beijing University of Posts and Telecommunication, Beijing, China

³ School of Software, Beihang University, Beijing, China

Abstract. Recently, a growing interest has been seen in rapid medical image segmentation for point-of-care applications. UNeXt, a convolutional multilayer perceptron (MLP)-based rapid medical image segmentation network has shown an outstanding performance in single-organ segmentation. However, there is still a large room for improvement in multi-organ segmentation by exploring sufficient information from a global view. To this end, we propose UNeXt++, a more powerful framework that adds a lightweight Serial-Parallel Hybrid Attention module named SPHTension to the UNeXt. The proposed SPHTension is designed to assist in the detection and localization of lesion tissue by extracting image local features and global semantic context through parallel structures, respectively. These structures play distinct roles in instance segmentation. Furthermore, we introduce the Attentional Feature Fusion (AFF) approach, which simultaneously cascades learning blocks to fuse and optimize the feature representation. The proposed hybrid architecture is capable of simultaneously focusing on local and global features in different regions, effectively integrating them to sense the location and edges of lesion tissues, and performing accurate segmentation. It is noteworthy that UNeXt++ is capable of efficiently aggregating global representations by adding only a very small number of parameters. Experimental results demonstrate that our UNeXt++ outperforms UNeXt in terms of segmentation performance on the multi-organ segmentation dataset Synapse and three single-organ segmentation datasets. This improvement is observed to be between 5% and 18%, while the computational cost is reduced by 17% and the amount of parameters is reduced by 19%.

Keywords: Medical image segmentation · UNeXt · point-of-care · Serial-Parallel

1 Introduction

Medical image segmentation is a vital auxiliary tool in computer-aided diagnosis, image-guided surgical systems, and advanced medical care. Owing to the

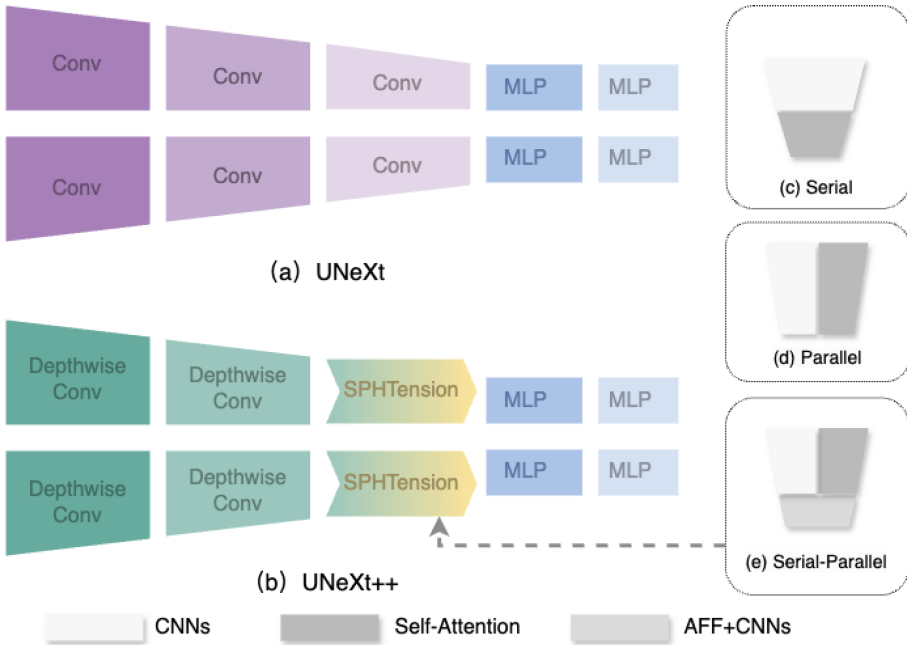


Fig. 1. (a) Network structure of UNeXt, (b) Network framework for UNeXt++ frameworks, (c) Serial Structure, (d) Parallel Structure, and (e) Serial-Parallel Structure.

powerful capabilities demonstrated by deep learning methods in image processing tasks, based on Convolution (CNN) U-shaped network UNet [27] and its variants like TransUnet [6], UNet++ [38], and 3D UNet [34] have become the cutting-edge work in the field of medical image segmentation in recent years. However, the above networks mainly focus on amplifying lab performance, frequently neglecting their practical application scene like point-of-care (PoC) [19], which is characterized by fast testing and strong individualization, demonstrating significant advantages in many clinical tests. Limited by hardware computational resources, PoC devices cannot run complex networks with large parameters. Therefore, rapid medical image segmentation aimed at reducing parameter size and maintaining good performance has become a hot research topic. Following the goal of speeding up inference, UNeXt [32], whose design is centered on reducing model parameters by decreasing the number of filters in convolutional blocks and replacing traditional transformer blocks with tokenized MLP [32] blocks, is a groundbreaking model due to its strong segmentation performance, as shown in Fig. 1(a).

Although UNeXt has seen substantial success, its performance for multi-organ segmentation lags behind the transformer-based TransUNet model by 10% on the Synapse dataset. For instance, the UNeXt model in the experiments only utilizes CNN and MLP modules. Due to the limitations of CNNs-based oper-

ations in handling long-distance dependency information, some crucial global information is lost in image processing, which causes performance bottlenecks for organ segmentation tasks. Transformer, relying purely on attention mechanisms to model global dependencies, should have been an alternative architecture with better performance. However, the large-scale parameters of transformer blocks limit the application for rapid medical image segmentation. Therefore, the motivation of our work is to design a lower computational architecture with global attention mechanisms and incorporate it into the UNeXt framework for effective representation learning of latent spaces.

In this paper, we propose a novel U-shaped hybrid framework, UNeXt++, which integrates the strengths of Depthwise Convolution neural networks, attention mechanisms, and multi-layer perceptrons, while maintaining a minimal parameter cost for efficient medical image segmentation tasks, as shown in Fig. 1(b). Specifically, our framework primarily uses depthwise convolution and MLP to hierarchically extract local intensity features, optimizing computational speed and avoiding parameter loss large-scale pretraining of self-attention. At the same time, it incorporates serial-parallel hybrid attention (SPHTension) to enhance complex spatial transformations and long-range feature dependencies, as shown in Fig. 1(e). The SPHTension interspersed between hierarchical convolution and MLP, extracts global context and models local features concurrently via two heads, and feeds the adapted fused features into the learning block for further enhancement and Refinement. Through the local connection and weight sharing of the learning blocks, the noise in the features can be effectively suppressed and the quality of the features can be improved. The overall framework follows a lightweight design, and the serial-parallel structure further reduces the influence of the attention mechanism on the number of parameters of the model, which is 19% less than that of the UNeXt network. Extensive experiments on both multi-organ and single-organ segmentation show the better performance of UNeXt++ compared to UNeXt. The main contributions of this paper include:

- We propose the UNeXt++ framework for rapid medical image segmentation with excellent performance.
- We propose the lightweight SPHTension module, which effectively retains both local and global image features with only a minimal increase in the number of network parameters.
- We perform extensive experiments both on multi-organ segmentation and single-organ segmentation datasets, all of which demonstrate the effectiveness of our proposed method.

2 Related Work

2.1 Rapid Medical Image Segmentation

Rapid medical image segmentation has transformed from conventional approaches to deep learning-based methods [17]. The majority of conventional approaches rely on low-level features of images [3, 14, 20, 26], including pixel intensity, color, edge,

and texture. All of them are simple, intuitive, high interpretability, and may be effective in some specific applications of simple scenarios. However, constrained by the sensitivity to noise variations [24], they cannot adapt to complex scenes and result in poor segmentation performance. The deep learning widely used in the development of the increasing popularity of CNN has opened up new opportunities for medical image segmentation. Among them, UNet and its variants have shown excellent performance and become a pioneering architecture.

Since then, with the development of portable medical devices and bedside care, UNet and its variants have started to explore lightweight and fast personalized modifications to the model. For example, ShuffleNet [37] is a CNN architecture designed for mobile applications. It uses pointwise group convolution and channel shuffling operations and effectively reduces the computational complexity associated with 1×1 convolution. SegNet [4] achieves rapid semantic segmentation by using a deep fully convolutional neural network architecture and efficiently leveraging pooling indices for non-linear upsampling, resulting in reduced memory usage compared to competing models. Ping et al. [11] achieve fast model performance through a fast spatial attention mechanism and additional spatial reduction in intermediate feature stages, enhancing computational speed. SHFormer [29] employs a shallow hierarchical Transformer architecture and a spatial-channel connection module to reduce model complexity and achieve lightweight design. Super-BPD [33] achieves efficient image segmentation by employing a novel super boundary-to-pixel direction method, effectively partitioning images into information-rich superpixels with directional similarity, thus enhancing segmentation accuracy and efficiency. These models are either optimised based on the CNN structure that is not sufficient to capture the global information of the image or based on the attentional mechanism which takes too much time for model inference, despite significant progress in the field of image segmentation. Therefore, a comprehensive plan for addressing the aforementioned issues needs to be taken into consideration.

2.2 Hybrid Models of CNN and Self-attention

The CNN structure excels at extracting local features with lower computational costs, while Self-Attention captures global dependencies at the expense of increased complexity. Combining these methods is advantageous for rapid medical image segmentation. To improve Self-Attention's efficiency in handling high-resolution images, Wang et al. [35] integrated a multi-scale pyramid CNN into the Vision Transformer (ViT), marking one of the first hybrid models. The ViT [9] enhances input adaptability and feature extraction with residual links between transformer layers. Touvron et al. [30] introduced CNN's inductive bias into Self-Attention using knowledge distillation. Another hybrid approach combines CNN and Self-Attention in series or parallel. DETR [8] uses CNN as a feature extractor followed by Self-Attention for end-to-end object detection, while Beal et al. [5] connect Faster R-CNN after ViT. Conformer [23] introduces a

Feature Coupling Unit (FCU) for parallel feature fusion, and Yoo et al. [36] use bidirectional bridging for parallel fusion via synchronized blocks.

Recently, several models have been proposed to explore the potential of creating networks using a mixture of CNN and Self-Attention in various ways. EdgeViTs [21] and CMT [10] employ a multi-layer structure similar to ResNet. In EdgeViTs, the model is organized into sequential stages, which consist of a local aggregation module based on deep convolution and point-wise convolution, a global sparse self-attention module, and a local propagation module based on transposed convolutional local propagation module. The CMT consists of three parts: a CNN, a lightweight multi-head self-attention mechanism, and an inverse residual feed-forward network. On the other hand, the SegFormer [23] model utilizes a CNN to extract features, the structure is shown in Fig. 1(c). Finally, it fuses the features through an ALL-MLP layer with multilevel features. PHTrans [16] is a similar approach to our work. It innovatively constructs CNN and the Swin Transformer in parallel to model hierarchical representations of local and global information, the structure is shown in Fig. 1(d). As standard convolutional kernels are input-independent and unable to adapt to different inputs, the performance improvement in these hybrid networks is limited.

Consequently, in light of the existing literature, we employ the more lightweight and efficient depthwise convolution [7] and optimized sparse attention [18] to process image features and generate fused features with the adaptation of local and global feature information of images processed in parallel according to the task characteristics, the structure is shown in Fig. 1(d). Further enhancement and optimization of the fused features by the learning block improves the stability and quality of the features, resulting in an improvement in the model’s segmentation performance.

3 Method

In this paper, we propose a new three-stage framework UNeXt++, as shown in Fig. 2. The DSC stage consists of two depthwise convolution layers for generating a local feature map of a given image. The SPHTension stage aggregates the global information by the proposed SPHTension module, which learns the associations between different local features. The SPHTension stage uses SPHTension parallel depthwise convolution and sparse attention serial-parallel convolution to learn the association between different local features. The MLP stage applies tokenized MLP to efficiently tokenize and project the image features.

3.1 Depthwise Convolution Stage

The UNeXt++ algorithm employs a lighter and more efficient Depth Convolution to extract shallow image features. Each convolution block is divided into two steps. Depthwise convolution and pointwise convolution are employed, whereby each input channel is spatially convolved independently, and the output of depthwise convolution is channel fused. Depthwise Convolution performs a spatial

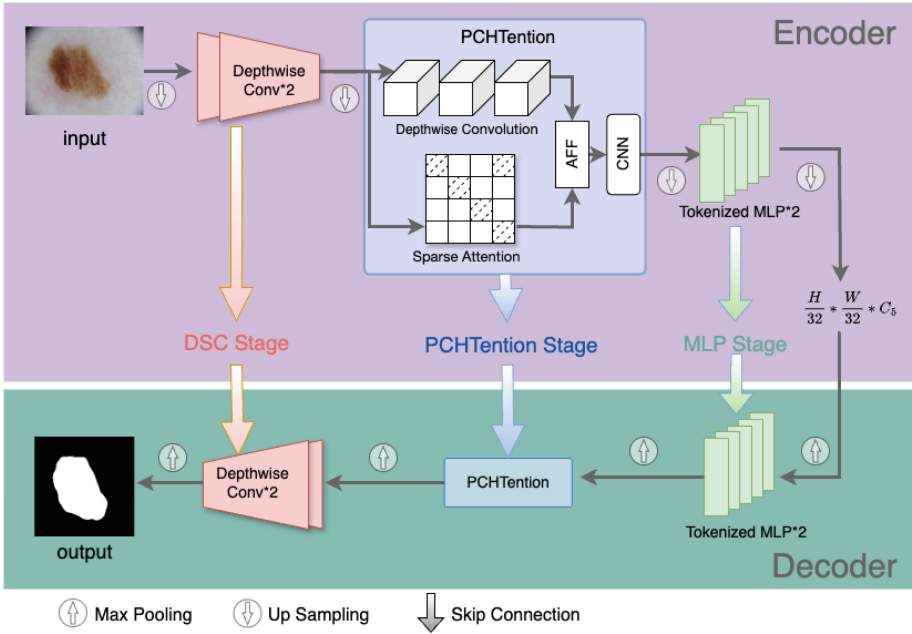


Fig. 2. Overall of the UNeXt++ framework. In the encoder, the input image will go through the Depthwise Convolution stage, the SPHTension stage, and the MLP stage sequentially. This architecture is also used in the decoder stage to ensure consistency throughout the model. Each stage uses jump connections to fuse the feature maps at corresponding positions during encoding and decoding.

convolution operation independently for each input channel, while Pointwise Convolution performs channel fusion for the output of Depthwise Convolution. In comparison to conventional convolution, depthwise convolution employs a smaller number of parameters. This reduction not only reduces the storage requirements of the model but also minimizes the risk of overfitting, thereby enhancing the model’s generalisability. Depthwise Convolution is applied in the first stage of the encoder and the last stage of the decoder, which can significantly improve the computational efficiency of the model.

3.2 SPHTension Stage

As illustrated in Fig. 2, the image features extracted in the Depthwise Convolution stage are copied into two copies and transferred to the parallel part of the SPHTension module. The parallel part employs depthwise convolution and sparse attention to obtain long-range dependent and local representations of image features, respectively. The attention feature fusion (AFF) method is used to adaptively fuse the features of the two parallel branches. In addition, sparse attention can more effectively filter out unimportant information and focus on

the key regions in the image. The selective attention mechanism has the potential to enhance the model’s sensitivity to crucial features, thereby optimizing the efficacy of image feature extraction. In comparison to the global attention mechanism employed in conventional Transformers, sparse attention significantly reduces computational complexity and memory usage. Assuming that the input features are N , the feature processing and fusion process of the SPHTension module can be expressed as follows.

Sparse Attention. The input feature maps are transformed in a linear manner to obtain query, key, and value vectors as

$$Q, K, V = NW_{qkv}, \quad (1)$$

where $N \in \mathbb{R}^{B \times (H \times W) \times C}$ is the input feature map with B samples, each with $H \times W$ spatial locations and a 3D tensor of the C channels, $W_{qkv} \in \mathbb{R}^{C \times 3D_r}$ is the shared weight matrix, and $D_r = \frac{C}{2}$. The sparse attention mechanism enhances computational efficiency by adjusting the parameter W_{qkv} to focus on the relationship between the image elements related to the lesion region, thereby ignoring those elements that have little effect on the segmentation result. The sparse attention calculates the attention weights and is subsequently employed to weight and sum the value vectors, yielding the final output feature X as

$$A = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) \quad (2)$$

$$X = \text{Dropout}(A)VW_{\text{proj}} \quad (3)$$

where A is the attention weight, d_k is the dimension of the key vector, and $W_{\text{proj}} \in \mathbb{R}^{D_r \times C}$ is the projection matrix.

Depthwise Convolution. In the context of depthwise convolution processing, the input feature map N is initially convolved in a depthwise and pointwise manner, resulting in the generation of a convolved feature map S_{conv} as

$$S_{\text{conv}} = (N * K_d) * K_p \quad (4)$$

where the parameters of the convolution kernel for deep convolution are K_d , and point-by-point convolution are K_p . The convoluted feature map is then subjected to normalization, Dropout, and non-linear activation to obtain the final output feature map S by

$$S = \text{NonLin}(\text{Dropout}(\text{Norm}(S_{\text{conv}}))) \quad (5)$$

Subsequently, X and S are merged with features utilizing AFF as

$$F_{\text{fused}} = \alpha \cdot X + \beta \cdot S \quad (6)$$

among them, α and β are fusion weights, which are usually generated dynamically by the attention mechanism. Furthermore, the local connectivity and weight-sharing properties in the serialized convolutional block are exploited to spatially reprocess and enhance the features, effectively suppressing noise in the input features to improve the quality and stability of the features.

3.3 MLP Stage

This stage is located at the bottom of the whole framework, where two tokenized MLP blocks [32] are used in the encoder and decoder, with the same structure as the tokenized MLP in UNeXt [32]. Unlike UNet, which uses two convolutions for input raw images to extract feature maps, UNeXt uses one convolution layer in the encoder and subsequently downsampling the input image by pooling. The local connectivity of convolutional operations and downsampling through pooling operations restrict neurons from perceiving only local information from the image. The image after the SPHTension stage can compensate for this. The image features are then fed into the Tokenized MLP module through two cleverly designed components, PatchEmbed and ShiftedBlock, to maintain segmentation performance within the overall lightweight framework of the model.

4 Experiments

4.1 Datasets

The datasets used in this paper fall into two main categories: multi-label and single-label medical image segmentation datasets.

Multi-label Medical Image Segmentation Dataset. We select Synapse multi-organ segmentation dataset [15] for the multi-label segmentation task, which contains 30 abdominal computed tomography (CT) scans totaling 3,779 slices. Each CT scan contains 85 to 198 slices that were manually labeled with 13 abdominal organs by two experienced undergraduate students. The labeling is validated at the volume level by a radiologist using MIPAV software. The dataset is divided into a training set and a test set containing 18 and 12 instances, respectively. The subset covers eight abdominal organs, including the aorta, gallbladder, right and left kidneys, liver, pancreas, spleen, and stomach. It is mainly used to compare the results with the UNeXt model, evaluated in terms of DSC [28] and HD [2].

Single-Label Medical Image Segmentation Dataset. The single-label medical image segmentation datasets include the Breast Ultrasound Image (BUSI) [1] dataset, the Database of Thyroid Ultrasound Images (DDTI) [22], and the Dermoscopy image classification dataset (ISIC-2020 [25], ISIC-2018 [31]). The BUSI dataset contains ultrasound images of normal, benign, and malignant breast cancers, along with their corresponding segmentation maps. In this paper, we selected benign and malignant tumor images, totaling 647, all resized to 224×224 . The DDTI dataset contains lesion types such as thyroiditis, cystic nodules, adenomas, and thyroid cancers, providing accurate lesion segmentation stored in XML format. The ISIC series of datasets are derived from the world’s largest skin image analysis challenge hosted by the International Skin Imaging Collaboration (ISIC), with the ISIC-2018 dataset comprising more than 12,500 images and the ISIC-2020 dataset expanding to 33,126 dermoscopy training images.

4.2 Evaluation Metrics

We evaluate the proposed UNeXt++ model from two perspectives. For the segmentation accuracy of the model, we use three commonly used standard evaluation metrics in image segmentation and compare them with other SOTA methods. These metrics contain the Dice coefficient (Dice) [28], the Intersection over Union (IoU) [12], and the 95th percentile of Hausdorff distance (HD95) [2]. For the efficiency of the model, we also consider model computational complexity and inference time as evaluation metrics. The computational complexity is determined by both model parameters and computational cost (GFLOPs) [13]. A lower computational complexity results in shorter inference time, leading to better real-time performance and response speed in practical applications.

The Dice coefficient is used to measure the similarity between the predicted result ‘pred’ and the real annotation ‘label’:

$$Dice = \frac{2|pred \cap label|}{|pred| + |label|}, \quad (7)$$

Here, ‘pred’ represents the model’s predicted result, and ‘label’ represents the real annotation. The Dice coefficient value ranges from 0-1. The larger the value, the greater the similarity of the segmented result to the actual label, meaning better segmentation performance.

The Intersection over Union (IoU) is a metric for evaluating the performance of image segmentation models. IoU is used to calculate the overlap between the model’s predicted results and the actual annotations. With a range from 0 to 1, a larger IoU score implies heightened segmentation accuracy. Specifically, IoU measures the accuracy of segmentation by calculating the ratio of the intersection area to the union area between the predicted results and annotations. The formula is shown below:

$$IoU = \frac{|pred \cap label|}{|pred \cup label|}. \quad (8)$$

The Hausdorff distance (HD95) represents the maximum distance (Hausdorff distance) between the predicted and actual results, retaining the average distance of the top 95% data points after distance sorting. A lower HD95 suggests a greater resemblance between the predicted and true results, pointing to enhanced algorithmic effectiveness.

4.3 Experimental Settings

The UNeXt++ model runs on the PyTorch framework. A combination of Binary Cross Entropy (BCE) loss and Dice Loss was used to train UNeXt++. The loss L between the predicted result y and the labeled result \hat{y} is expressed as

$$L = 0.5BCE(y, \hat{y}) + Dice(y, \hat{y}). \quad (9)$$

For all training samples, image enhancement operations such as rotation, flipping, transposition, adding Gaussian noise, and adjusting image saturation were

Table 1. Comparison on the Synapse multiple organ segmentation dataset. Note that the \uparrow indicates that UNeXt++ is better than UNeXt’s Dice score.

Network structure	Inference time (s/it)	Aorta		Gallbladder		Left kidney		Right kidney		Liver		Pancreas		Splenic organ		Stomach		Average	
		DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD
UNet [27]	246.08	85.32	61.38	82.62	2.64	93.07	47.04	82.33	65.10	73.69	30.08								
UNet++ [38]	286.98	87.10	62.52	81.21	73.99	93.08	50.62	83.90	67.57	75.00	28.82								
TransUNet [6]	212.41	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62	77.48	31.69								
UNeXt [32]	79.54	76.43	51.64	74.54	67.94	91.11	34.95	79.20	60.70	67.07	40.47								
UNeXt++	83.36	79.45 \uparrow	51.38	77.60 \uparrow	64.41	91.18 \uparrow	44.79 \uparrow	82.39 \uparrow	65.94 \uparrow	69.64 \uparrow	41.17								

Table 2. Comparison on the single-label datasets.

Network	Params (M)	GFLOPs	DDTI		BUSI		ISIC-2020	
			IoU (%)	Inference time (s/it)	IoU (%)	Inference time (s/it)	IoU (%)	Inference time (s/it)
UNet [27]	31.13	55.84	77.42	0.24	64.26	0.45	75.51	0.08
UNet++ [38]	9.16	34.65	82.00	0.22	65.04	0.74	75.77	0.13
TransUNet [6]	105.32	38.52	81.69	0.26	65.46	0.62	81.23	0.20
UNeXt [32]	1.47	0.57	81.22	0.02	64.29	0.01	83.23	0.01
UNeXt++	1.19	0.47	85.63	0.01	75.62	0.01	89.86	0.01

used to enhance the diversity of the data. The learning rate of Adam’s optimizer was set to 0.001, and the momentum parameter was set to 0.9. Additionally, a cosine annealing learning rate scheduler was used, with the minimum learning rate set to 0.00001. The factor was set to 0.1, and patience was set to 2. The batch size of the dataset was set to 16. For the Synapse dataset, a total of 250 calendar events were trained. For the BUSI, DDTI, and ISIC datasets, the training and test sets were randomly divided in an 8:2 ratio. To ensure the accuracy of the experimental results, the number of epochs for all experiments was set to 300, and the experiments were conducted using a 32 GB NVIDIA V100.

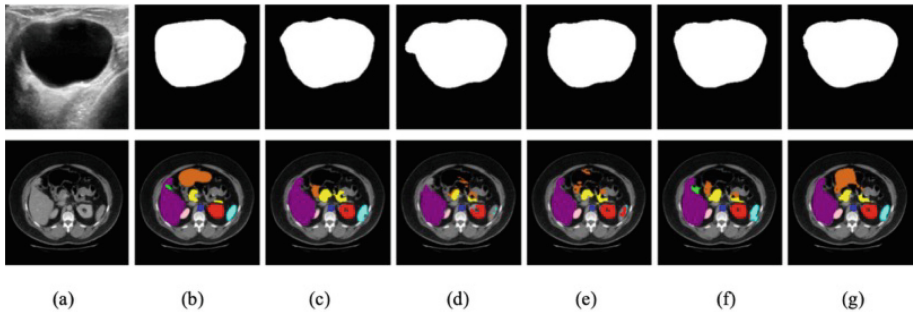
4.4 Experimental Results

To evaluate the segmentation performance of the proposed method in real-world application scenarios, this paper compares UNeXt++ with widely used medical image segmentation frameworks. It is compared with the convolution-based UNet [27] and its variant UNet++ [38]. It is also compared with the Transformer-based TransUNet [6] and the MLP-based UNeXt. All comparison experiments were conducted using the same equipment and parameter settings.

Results on Multi-label Medical Image Segmentation Dataset. As shown in Table 1, experimental results demonstrate that compared to UNeXt, the proposed UNeXt++ model achieves a 2.57% improvement in average DSC, with only a 4.6% increase in inference time. Due to the fewer attention blocks included in UNeXt++, its inference time on the Synapse dataset is only 29-36% of complex models such as TransUNet. These findings underscore the effectiveness of our model in the direction of lightweight design.

Table 3. Comparison on the ISIC-2018 dataset.

Network	Params (M)	Dice (%)	IoU (%)	Inf-T(CPU) (ms)	Inf-T(GPU) (ms)
UNet [27]	31.13	86.05	78.42	498.98	13.79
UNet++ [38]	9.16	87.09	79.77	1559.06	38.29
TransUNet [6]	105.32	90.02	83.23	1871.92	66.93
UNeXt [32]	1.47	89.45	83.83	59.99	9.02
SHFormer [29]	1.94	90.73	84.41	68.76	10.30
UNeXt++	1.19	93.27	87.44	60.09	5.79

**Fig. 3.** Qualitative segmentation results of two images from BUSI dataset and Synapse dataset. (a) Input image, (b) Ground truth, (c) TransUNet, (d) UNeXt, (e) UNeXt with encoder SPHTension, (f) UNeXt with decoder SPHTension, and (g) UNeXt++.

Results on Single-Label Medical Image Segmentation Datasets. It is noteworthy that the UNeXt++ model demonstrates satisfactory segmentation performance when applied to single-label datasets, such as BUSI, DDTI, and ISIC. The results of the experiment are presented in Table 2 and Table 3. Furthermore, we evaluated the performance with the latest lightweight segmentation method on the commonly used ISIC-2018 dataset. The results show that the IoU metric improves by 3.5% and the Dice metric improves by 2.7% when compared to the state-of-the-art SHFormer network. From the table, it can be observed that compared to existing techniques, the proposed UNeXt++ model demonstrates significant improvements in model lightweight, efficiency, and segmentation performance. The reduction in model parameters ranges from 79.81% to 98.24%, while the acceleration in inference time ranges from 91.82% to 98.65%. The IoU performance is enhanced by a margin of 5% to 18%. Furthermore, UNeXt++ outperforms UNeXt by more than three percentage points in terms of segmentation performance (as measured by IoU), while maintaining a lightweight model. In comparison to the CNN-based UNet, UNet++, and the Transformer-based TransUNet, UNeXt++ demonstrates a relatively high degree of performance improvement, particularly because of the significant reduction in model parameters and accelerated inference times. This modification does not compromise

Table 4. Ablation experiments on the Synapse multi-organ segmentation dataset.

Network structure		Inference	Aorta	Gallbladder	Left kidney	Right kidney	Liver	Pancreas	Splenic organ	Stomach	Average
encoder	decoder	time (s/it)									
UNeXt	UNeXt	79.54	76.43	51.64	74.54	67.94	91.11	34.95	79.20	60.70	67.07
UNeXt+SPHTension	UNeXt	82.87	78.56	52.44	77.07	68.70	90.88	40.53	74.63	62.71	68.19
UNeXt	UNeXt+SPHTension	83.74	77.78	51.65	73.12	65.93	89.76	38.96	73.36	62.25	66.60
UNeXt++		83.36	79.45	51.38	77.60	64.41	91.18	44.79	82.39	65.94	69.64

Table 5. Ablation experiments on the BUSI dataset.

Network structure		Params (M)	Inference	IoU (%)
Encoder	Decoder		time (it/s)	
UNeXt+Serial CNN&Self-Attention		1.67	9.10	73.26
UNeXt+parallel CNN&Self-Attention		1.74	9.10	72.94
UNeXt++		1.19	9.65	75.62

computational memory or power consumption, aligning more closely with the practical application requirements. Furthermore, Fig. 3 presents a comparison of the qualitative segmentation outcomes of the distinct methodologies on two images, namely the BUSI dataset and the Synapse dataset.

4.5 Ablation Experiments

To validate the effectiveness of the SPHTension stage, ablation experiments were conducted on the BUSI and Synapse datasets, focusing on two aspects. Firstly, the study examined the necessity of using the SPHTension module in both the encoder and decoder of the model. This involved observing segmentation performance with and without the SPHTension module in both the encoder and decoder. The results are presented in Table 4, revealing that adding the SPHTension module in both the encoder and decoder stages outperforms adding it only in the decoder or encoder stages, demonstrating the superiority of SPHTension in learning-rich semantic features. Secondly, the necessity of the SPHTension module in the serial-parallel structure was evaluated. The SPHTension module in the second stage was replaced with parallelly concatenated depthwise convolution and Sparse Attention modules, as well as parallelly concatenated depthwise convolution and sparse attention modules, on the BUSI dataset to observe the results. As shown in Table 5, models using only serial or parallel structures exhibited segmentation performance 3.6% and 4.1% lower, respectively, compared to models using both serial and parallel structures with the SPHTension module. This difference persists even when the parameter count increases negligibly.

5 Conclusion

We present UNeXt++, an efficient U-shaped hybrid framework designed for rapid and lightweight medical image segmentation. By integrating depthwise

convolution, MLP, and SPHTension modules, UNeXt++ effectively balances local feature extraction with global context modeling, all while maintaining a minimal parameter cost. This design results in fast inference and improved segmentation performance across various medical imaging tasks. However, UNeXt++ does have certain limitations. Although the SPHTension module enhances the model's ability to capture global dependencies, it may still fall short in scenarios requiring highly intricate global information processing, such as in complex multi-organ segmentation tasks. Compared to transformer-based models like TransUNet, which excel in modeling long-range dependencies due to their pure attention mechanisms, UNeXt++ might struggle to maintain the same level of accuracy in capturing fine-grained details across distant regions of an image. Additionally, while UNeXt++ is designed to be lightweight, this efficiency comes with trade-offs. The reduction in parameters, although beneficial for speed, may result in performance bottlenecks when dealing with highly variable or complex anatomical structures that require a more nuanced understanding of spatial relationships. This could limit its effectiveness in certain clinical scenarios where precision is critical. These limitations highlight areas where further research could enhance the model's capabilities. We hope that UNeXt++ will inspire ongoing efforts to develop models that not only offer efficiency but also robustness in handling the diverse challenges of medical image segmentation.

Acknowledgement. This work was supported by National Natural Science Foundation of China (Grant number: 61976066), Research Funds for NSD Construction, University of International Relations (Grant number: 2024GA07).

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. *Data Brief* **28**, 104863 (2020)
2. Ali, N.A., Subki, L., Alwee, R., Amin, M.M.: A review on medical image segmentation: techniques and its efficiency. *PERINTIS eJournal* **7**(2), 59–82 (2017)
3. Arthur, D., Vassilvitskii, S.: K-means++ the advantages of careful seeding. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035 (2007)
4. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017)
5. Beal, J., Kim, E., Tzeng, E., Park, D.H., Zhai, A., Kislyuk, D.: Toward transformer-based object detection. *arXiv preprint [arXiv:2012.09958](https://arxiv.org/abs/2012.09958)* (2020)
6. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. *arXiv preprint [arXiv:2102.04306](https://arxiv.org/abs/2102.04306)* (2021)
7. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258 (2017)
8. Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L.: Dynamic detr: end-to-end object detection with dynamic attention. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2988–2997 (2021)

9. Dong, X., et al.: Cswin transformer: a general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12124–12134 (2022)
10. 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)
11. Hu, P., et al.: Real-time semantic segmentation with fast attention. *IEEE Robot. Autom. Lett.* **6**(1), 263–270 (2020)
12. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat* **37**, 547–579 (1901)
13. Kollár, J.: Flops. *Nagoya Math. J.* **113**, 15–36 (1989)
14. Kolmogorov, V., Zabini, R.: What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(2), 147–159 (2004)
15. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proceedings of MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge, vol. 5, p. 12 (2015)
16. Liu, W., et al.: Phtrans: parallelly aggregating global and local representations for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 235–244 (2022)
17. Liu, X., Song, L., Liu, S., Zhang, Y.: A review of deep-learning-based medical image segmentation methods. *Sustainability* **13**(3), 1224 (2021)
18. Lu, M., et al.: Smile: sparse-attention based multiple instance contrastive learning for glioma sub-type classification using pathological images. In: MICCAI Workshop on Computational Pathology, pp. 159–169 (2021)
19. Lippa, P.B., Müller, C., Schlichtiger, A., Schlebusch, H.: Point-of-care testing (POCT): current techniques and future perspectives. *TrAC, Trends Anal. Chem.* **30**(6), 887–898 (2011)
20. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)
21. Pan, J., et al.: Edgevit: competing light-weight CNNs on mobile devices with vision transformers. In: European Conference on Computer Vision, pp. 294–311 (2022)
22. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: 10th International Symposium on Medical Information Processing and Analysis, vol. 9287, pp. 188–193 (2015)
23. Peng, Z., et al.: Conformer: local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 367–376 (2021)
24. Ramesh, K., Kumar, G.K., Swapna, K., Datta, D., Rajest, S.S.: A review of medical image segmentation algorithms. *EAI Endorsed Trans. Pervasive Health Technol.* **7**(27), e6 (2021)
25. Rotemberg, V., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci. Data* **8**(1), 34 (2021)
26. Sethian, J.A., et al.: Level Set Methods and Fast Marching Methods, vol. 98. Cambridge Cambridge UP (1999)
27. Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V.: U-net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access* **9**, 82031–82057 (2021)

28. Sorensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter* **5**, 1–34 (1948)
29. Su, D., Luo, J., Fei, C.: An efficient and rapid medical image segmentation network. *IEEE J. Biomed. Health Inform.* (2024)
30. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: *International Conference on Machine Learning*, pp. 10347–10357 (2021)
31. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**(1), 1–9 (2018)
32. Valanarasu, J.M.J., Patel, V.M.: Unext: MLP-based rapid medical image segmentation network. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 23–33 (2022)
33. Wan, J., Liu, Y., Wei, D., Bai, X., Xu, Y.: Super-BPD: super boundary-to-pixel direction for fast image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9253–9262 (2020)
34. Wang, C., MacGillivray, T., Macnaught, G., Yang, G., Newby, D.: A two-stage 3D unet framework for multi-class segmentation on full resolution image. *arXiv preprint [arXiv:1804.04341](https://arxiv.org/abs/1804.04341)* (2018)
35. 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)
36. Yoo, J., Kim, T., Lee, S., Kim, S., Lee, H., Kim, T.: Rich CNN-transformer feature aggregation networks for super-resolution. *arXiv preprint [arXiv:2203.07682](https://arxiv.org/abs/2203.07682)* (2022)
37. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856 (2018)
38. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **39**(6), 1856–1867 (2019)



Efficient Adapter on Pre-trained Visual Feature Reliance in Medical Visual Question Answering

Aakansha Mishra^(✉), Prateek Keserwani, Vikram N. Rajendiran,
and Ashok K. Senapati

Samsung R&D Institute India-Bangalore, Bengaluru, India
{a11.mishra,k.prateek,vikram.nr,a.ashokkumar}@samsung.com

Abstract. Medical Visual Question Answering (MedVQA) is crucial for medical data analysis and patient diagnosis, aiding medical practitioners with fast and accurate answers. The recent deep learning model requires a huge amount of data to train; however, collecting large samples and annotations in medical domain is challenging. Therefore, training the model from scratch with small samples easily leads to overfitting. To overcome this problem, pre-trained models can be leveraged and transfer prior knowledge to the medical domain. Efficiently transferring knowledge from pre-trained models with limited data to different domains remains challenging. To address this issue, an efficient convolution-based adapter (EC-Adapter) is introduced, which is versatile and applicable to any pre-trained architecture. The proposed adapter leverages the depth-wise and point-wise convolution operation and add this as parallel layer to the model. The proposed EC-Adapter is simple, lightweight and effective as compared to state-of-the-art low-rank adapters, potentially benefiting large language or vision models. It achieves superior performance while requiring significantly fewer parameters than existing complex methods. In the era of increasingly large and diverse medical datasets, EC-Adapter offers a promising solution to enhance the adaptability and efficiency of pre-trained models in medical applications. The efficacy of the model is demonstrated through extensive experiments and analysis on two publicly available MedVQA datasets: *SLAKE* and *PathVQA*.

Keywords: Medical visual question answering · Parameter efficient adapter · Less training data

1 Introduction

Medical Visual Question Answering [5, 10, 12, 19] is the task where a natural language question is asked about the content of the medical images and the objective is to predict the answer in natural language [18]. As deep learning has revolutionized the domain of medical image analysis in the last decade [41] which includes the development of efficient methods for disease diagnosis based on various non-invasive sensory data such as X-ray [31], structural MRI [3], and fMRI [37] via

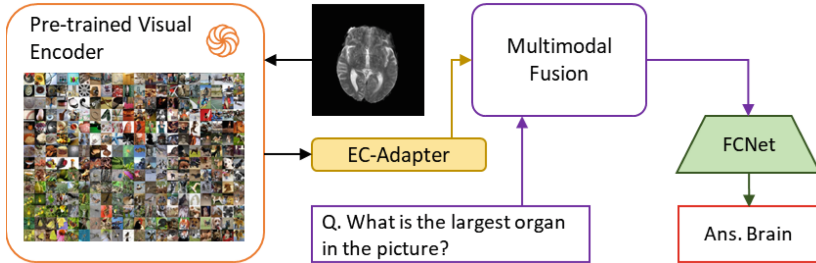


Fig. 1. Overview of proposed work. Visual encoder is trained on real image dataset for multiple tasks. This knowledge is exploited for MedVQA task where it is expensive to train large model due to unavailability of huge amount of data. The medical image features achieved from visual encoder are adapted through the proposed adapter which is based on convolution and consists of small number of parameters.

classification and segmentation approaches. These methods provide a second opinion for the pathologists and doctors to investigate the patient’s medical condition. However, this assistance is mostly disease, modality or task specific. Generic assistance can be achieved with the help of developing a system that provides answers to questions or raised concerns for captured visual sensory data. This question-answering system that covers various modalities of visual sensory has extended the visual question answering (VQA) [2] in the medical domain [18].

VQA has been investigated extensively in the past [2, 24, 25], however, despite its crucial importance, the domain of Medical Visual Question Answering has not received significant attention from researchers. Recently, there has been a growing interest in the field of MedVQA [5, 10, 12, 19]. This task has its own challenges, which includes the lack of availability of well-annotated datasets by medical experts. The annotation of medical images is more time-consuming and complicated than that of natural images. Most medical datasets for VQA contain only a few hundred of medical images, such as SLAKE [19], and PathVQA [12]. The annotation of medical images requires the domain expert which is rare therefore annotation is error prone and time consuming. Additionally, the questions of the medical domains are more complex, and the system’s answer needs to possess high accuracy since it belongs to the area of health and safety.

Recently, a massive effort has been put by the vision and language research community by introducing various foundation models such as CLIP [29]. These methods shows the promising performance due to the usage of high-volume dataset in the model training. A huge amount of effort and time has been invested in collecting massive well annotated datasets. The data used to train these model are highly diverse and collected to the various domains like social media, general web, medical etc. Therefore, these model shows the high generalization ability across the various task. It is observed that these model can be used to adapt the medical domain with the helps of only few available samples. However full training of these model losses its generalization ability and on the small data

easily overfit to the training samples. Hence, the concept of adapter has been coined in the past. The adapter is some additional parameter which need to trained over the foundation/pre-trained models which has a capability to get trained with very small amount of training data. In adapter there are two fundamental approaches, namely, add some new parameters in different part of the pre-trained network [15, 22], and low-rank factorization update of the weights of the pre-trained network [14].

However, these adapter does not fully utilize the parameters and works well only if the adaption scenario is close to the base model’s task. For a significantly different task it requires to increase the adapter parameters which again poses a problem of overfitting. Also, in the today’s large models (vision and language) scenarios a parameter efficient model is required which can easily fit to the GPU memory. To overcome the above challenges, this work proposed an efficient convolution-based adapter (EC-Adapter) which requires the significantly less parameter as compared to the recent state-of-the-art adapter [14]. The proposed EC-Adapter leverages the efficient depthwise and pointwise convolution operation and added parallel to transformer layer. The proposed model achieves the promising result without extensive pre-training on big medical domain data. The proposed adapter exploits the knowledge of the generic pre-trained visual model, and transfer the knowledge to the medical domain. The end-to-end model is highly efficient which requires less parameters and computing resources. The extensive experiments over the publicly available MedVQA dataset *SLAKE* [19] and *PathVQA* [12] shows that EC-Adapter outperform the recent state-of-the-adapter by a significant margin, which requiring very few parameters. An overview of proposed workflow is presented in Fig. 1. The key contributions of proposed work are summarized as:

1. Proposed a lightweight efficient convolution-based adapter for transforming the visual features from natural image domain to the medical image domain.
2. The proposed EC-Adapter is easily plug-and-play on any existing efficient visual backbone for natural images.
3. Extensive experimentation has been done on two publicly available dataset *SLAKE* [19] and *PathVQA* [12] for MedVQA and established the efficiency of EC-Adapter by producing similar results compared to existing SOTA methods.

The rest of the work is organized into five sections. In Sect. 2, the work that is closely related to the proposed method has been described. The proposed EC-Adapter has been detailed in Sect. 3. Finally, the experimental details, results, and conclusion are presented in Sect. 4, 5, and 6, respectively.

2 Related Work

In the following section, the existing works relevant to proposed work are discussed. These works can be broadly grouped into three categories, based on their methods: Traditional Method; Medical Vision-Language Pre-training Method and Adapter-based Method.

Traditional Method. MedVQA framework [18] basically consist of three steps include, visual & text feature extraction, feature fusion and answer reasoning, that generate answer for the asked question on some medical image. For visual and text feature extraction the pre-trained visual models such as VGGNet [33] and textual feature extraction such as LSTM [13], combining Glove [28] with LSTM, BERT [6], BioBERT [16] has been used. These visual and language encoding contains information from two different domains, hence to establish a relationship between features of two different modality, the feature fusion stage is required. The basic methods to conduct the feature fusion are attention mechanism [36] and pooling module [9]. In [1], impact of stacked attention network [36] and multimodal compact bilinear pooling [9] has been shows for MedVQA. In [32], multi-mode decomposition bilinear pools [38] has been used. After feature fusion, the answer for the question is generated either by using classifiers [7, 20, 27] or by generative methods [34, 40].

Medical Vision-Language Pre-training Method. Medical vision-language pre-training aims to learn generic representations from large-scale medical image-text data, which can be transferred to various medical vision-and-language downstream tasks. Chen et al. [5] used a self-supervised learning (SSL) method to learn representations from medical images and text. The method learned cross-modal domain knowledge via the reconstruction of missing pixels and tokens from randomly masked images and text. In [39], three pre-training tasks are taken. It includes image reconstruction, report reconstruction, and Global and Local Alignment. Li et al. [17], has exploited an SSL method that applies masked image modeling, masked language modeling, image text matching, and image text alignment via contrastive learning.

Adapter-Based Method. Recently, the MedVQA has start aligning with the adapter based approaches. In [20], a parameter-efficient way to transfer the knowledge from pre-trained CLIP model [29] to the medical domain by introducing a lightweight adapter. The proposed adapter is a stack of two linear layers, which first downscale the features and then upscale the features for learning the new features over the CLIP features. Additionally, Liu et al. [20] has also employed a denoise auto-encoder and label smoothing to boost the performance of the method. The denoise autoencoder’s encoder feature has been concatenated with the adapter features. However, the usage of such complicated pipeline shows that the knowledge transfer from natural image training based CLIP is not easy.

In comparison with the existing methods of adapter for medical visual question answering, the proposed method has advocate that a simple convolution based adapter on generic VQA pipeline is sufficient to obtain a comparative results as compared to complicated adapter based MedVQA methods. Further, it is observed that a convolution based feature projection of pre-trained vision transformer model [21, 35] with basic language encoding such as Glove with LSTM is producing competitive results as [11, 29].

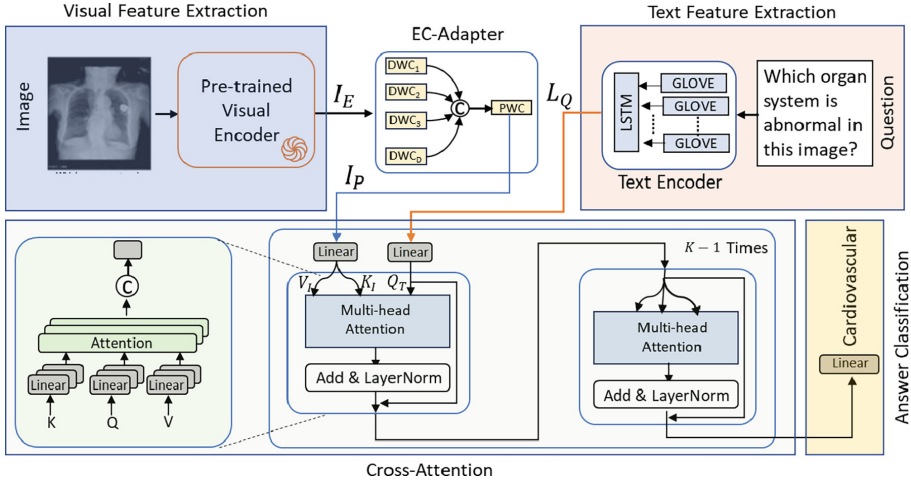


Fig. 2. The schematic diagram for the proposed method. The method consist of five components, includes, visual feature extraction (pre-trained frozen weights), text feature extraction, cross-attention, EC-Adapter, and answer classification.

3 Proposed Method

With the development of large models which are heavily pre-trained on huge amount of data for multiple tasks, it is realised to use their capability for further downstream task. In this work, the ability of large vision model which is pre-trained on general images is exploited to accomplish MedVQA by using it as a visual feature extractor. As the domain of data visual encoder is pre-trained and the medical images are significantly different, it is required to adapt the obtained features for medical images.

3.1 Problem Formulation

In the proposed formulation, MedVQA task is solved as classification problem. Given a set of medical image (\mathcal{I}), related question (\mathcal{Q}) and answer (\mathcal{A}) VQA system starts with feature extraction for visual and text modality. Further the parameters are learned to adapt the medical image features. These visual features are fed to cross modal interaction module to extracts the attended visual feature in context of question. Later attended features from visual and text modality are fused to obtain the unified embedding. With this embedding answer prediction is performed by feeding to fully connected network. The overall framework of proposed method is demonstrated in Fig. 2.

3.2 Feature Encoder

For MedVQA, let’s consider the image, question, answer triplet samples as: $\{I, Q, A\}$. Here, $I \in \mathcal{R}^{M \times N \times 3}$, $Q \in \mathcal{R}^{n_w}$, $A \in \mathcal{R}^{n_c}$. Here, M, N represents the

height, width of image, n_w is the number of words in the question, and n_c is the number of answer (class labels) in the dataset.

Visual Encoder: For encoding the visual representation image (I) is fed to a pre-trained vision transformer [21] to obtain the features as:

$$I_E = \text{ImgEnc}_\theta(I) \quad (1)$$

where $I_E \in \mathcal{R}^{p \times d_p}$. p is the number of patches and d_p is the dimension of each patch.

Text Encoder: Each question is trimmed or padded to a uniform length of n_w words. These words are embedded with GloVe embeddings [28] to obtain $W = [w_1, \dots, w_{n_w}] \in \mathcal{R}^{n_w \times d_w}$. Further, to keep the contextual information, W are fed to LSTM to obtain $L_Q = [q_1, \dots, q_{n_w}] \in \mathcal{R}^{n_w \times d_q}$.

3.3 EC-Adapter

The pre-trained parameters (θ) are obtained by training the encoder model over the natural images. The general natural image collection are very different from the medical data (MRI, CTScan etc.), therefore the encoded embedding is not well suited to the medical domain.

To address the same problem, an efficient adapter is proposed which leverages the efficient (depthwise and pointwise) convolutional filters. Say, $I_E \in \mathcal{R}^{p \times d_p}$ is the image encoder output, obtained with the pre-trained parameters θ . The image encoder output I_E as a tensor of dimension one i.e. $I_E \in \mathcal{R}^{p \times d_p \times 1}$. Let DWC is the depthwise convolution filter of size $L \times L \times 1$, we apply the D number of DWC over the I_E and concatenated all the output as:

$$I_D = \text{Concat}([\text{DWC}_1(I_E), \dots, \text{DWC}_D(I_E)]) \quad (2)$$

where $I_D \in \mathcal{R}^{p \times d_p \times D}$ and **Concat** is the concatenation operation. Further to combine the I_D to a single feature map the pointwise convolution operation (**PWC**) which is of size $1 \times 1 \times D$. The **PWC** operation over the I_D is defined as:

$$I_P = \text{PWC}(I_D) \quad (3)$$

The $I_P \in \mathcal{R}^{p \times d_p}$ feature embeddings are used as the final adapted visual features. Say, ϕ be the total parameter in the adapter i.e. combined **DWC** and **PWC** convolutional parameter. Here we can observe that the ϕ contains the total parameter $|\phi| = L \times L \times D + D$.

3.4 Cross-Attention

The adapted visual feature obtained from EC-Adapter is fed to transformer [35] based multi-head cross attention in the context of the question. To accomplish this, visual ($I_P \in \mathbb{R}^{p \times d_p}$) and textual ($L_Q \in \mathbb{R}^{n_w \times d_q}$) features are projected to a shared hidden dimension of d_h as $I_{PH} \in \mathcal{R}^{p \times d_h}$ and $L_{QH} \in \mathcal{R}^{n_w \times d_h}$. First,

the text is encoded through self attention by generating *key* (K_T), *query* (Q_T) and *value* (V_T) through linear projections $W_K^T \in \mathcal{R}^{d_h \times d_h}$, $W_Q^T \in \mathcal{R}^{d_h \times d_h}$, and $W_V^T \in \mathcal{R}^{d_h \times d_h} \in \mathcal{R}^{d_h \times d_h}$ from L_{QH} as:

$$K_T = L_{QH}W_K^T \quad (4)$$

$$Q_T = L_{QH}W_Q^T \quad (5)$$

$$V_T = L_{QH}W_V^T \quad (6)$$

where $W_K^T, W_Q^T, W_V^T \in \mathcal{R}^{d_h \times d_h}$

$$\begin{aligned} MHA(Q_T, K_T, V_T) &= \text{Concat}[\text{head}_1, \dots, \text{head}_h]W_O^T \\ \text{head}_i &= \text{Attention}(Q_TW_{Qi}, K_TW_{Ki}, V_TW_{Vi}) \\ \text{Attention}(Q_T, K_T, V_T) &= \text{softmax}\left(\frac{Q_TK_T^T}{\sqrt{d_k}}\right)V_T \end{aligned} \quad (7)$$

Say, L_S be the multihead self attended question representation. With this self attended question, cross attention is given on the image by generating *key* (K_I), *query* (Q_I) and *value* (V_I) from I_{PH} similar to Eq. 4–6. For cross attention on image, query would be *query* (Q_T). Further the multi-head attended visual representation (I_C) is obtained from Eq. 7. This multi-head attention is applied in k blocks where each block take as an input the attended representation from previous block and connected through skip connection and LayerNorm as proposed by [35].

3.5 Model Learning and Answer Classification

The obtained text and visual features from self and cross attention are fused via pointwise multiplication to obtain the final unified multimodal embedding (say, U). This embedding is fed to a fully connected network for answer classification.

$$\hat{a} = \text{FCNet}_{\theta_c}(U) \quad (8)$$

$\hat{a} \in \mathcal{R}^{n_c}$ is the predicted answer vector with n_c as the number answer categories in the dataset.

The model is trained with parameters in EC-Adapter ϕ , multihead attention block parameters, classification network parameters θ_c in end-to-end manner. The loss employed is binary cross entropy between predicted answer vector \hat{a} and ground truth answer vector $a \in A$.

4 Experimental Details

In the following discussion, the details of experimental setup are presented for end-to-end model learning.

4.1 Implementation Details

The size of image is $(M, N) = (256, 256)$, $n_w = 14$, GloVe dimension (d_w) is 300. The number of patches (p) is 64, dimension of each patch (d_p) is 768. The LSTM hidden state output (d_q) is 512. The visual encoder used in proposed method is SwinV2 [21]. The dimension of hidden shared space (d_h) for vision and text is 512. Number of attention blocks (k) is 2. The size of kernel (L) for EC-adapter is 17 and number of kernel (D) is 32. Number of answers i.e. class labels (n_c) for SLAKE [19] and for PathVQA [12] are 507 and 957 respectively. All experimentation has done on A100 GPU with 40 GB GPU memory. The experiments was conducted with batch size 128 and trained for 30 epoch with cross-entropy loss. Adam optimizer are used with initial learning rate of 0.0001.

4.2 Datasets

To validate the efficacy of proposed model, experiments are performed on two public datasets namely, PathVQA [12] and SLAKE [19]. The details of these datasets are as follows:

(1) *PathVQA* [12] is the first dataset for pathology VQA. The data is collected from digital library and pathology textbooks. The question can be divided into seven categories consist of how, where, when, what, whose, (how many, how much), and (yes,no). First six categories belongs to open ended whereas last one is for closed ended questions. In totality it consist of 32799 question-answer pair from 4998 pathology images. Among them 16465 is open ended questions and rest are close ended.

(2) *SLAKE* [19] is a bilingual MedVQA dataset consist of semantic labels and medical domain knowledge. The semantic label consist of mask and bounding boxes, whereas medical knowledge base is provided in the form of knowledge graph. The dataset is collected from three datasets and annotated by physicians. It consists of 14028 question-answer pairs from 642 images. It covers a wide range of human body parts, including the chest, brain, pelvic cavity, neck, and abdomen.

4.3 Evaluation Metric

Following the dataset and existing work [10, 12, 19, 39] the model is evaluated in terms of *overall accuracy* as the metric. Alongwith the overall accuracy for total number of samples in the dataset, evaluation is also performed on the category of question i.e., *open* and *close*.

5 Results and Analysis

In this section, quantitative and qualitative results are presented for proposed model on Path-VQA and SLAKE dataset. Further, the ablation studies conducted to evaluate the impact of different components on the efficacy of overall model.

Table 1. Performance Comparison of Proposed Method with SOTA methods on PathVQA and SLAKE-VQA datasets.

Model	Dataset						Visual Encoder Additional Training*
	PathVQA			SLAKE-VQA			
	Open	Close	All	Open	Close	All	
MPMA [39]	16.40	86.80	50.20	–	–	–	✓
M3AE [5]	–	–	–	80.31	87.82	83.25	✓
VQAMix [10]	13.40	83.50	48.60	–	–	–	✓
PEFA [20]	–	–	–	–	–	81.90	✓
M2I2 [17]	36.30	88.00	62.20	74.70	91.10	81.20	✓
Proposed	37.98	87.70	62.92	85.29	84.45	84.95	✗

*An additional training of visual encoder on medical data, which further finetune on MedVQA datasets.

5.1 Quantitative Results

In Table 1, a comparative analysis is presented between the proposed method and recent approaches on the Path-VQA [12] and SLAKE [19] datasets. The accuracy metrics are provided for both the ‘Open’ and ‘Close’ categories of questions, along with the ‘Overall Accuracy’. In comparison to the best-performing M2I2 model [17], the proposed method demonstrates a superior performance with a $\sim 1.7\%$ increase in overall accuracy on the SLAKE dataset. Notably, for open-category questions, the proposed approach exhibits a significant improvement of $\sim 5\%$, while for close-category questions, a slight decrease of $\sim 6\%$ is observed. For the Path-VQA [12], the proposed method achieves a gain of $\sim 0.7\%$ in overall accuracy compared to M2I2 [17]. It’s important to mention that M2I2 [17], M3AE [5] is pre-trained on a medical image caption dataset and fine-tuned for the MedVQA task. In contrast, the proposed method doesn’t require pre-training on medical data which is crucial to obtain.

5.2 Qualitative Results

For the qualitative results, the well-known Grad-Cam [30] visualization method is applied to the medical images. As shown in Fig. 3, the first column represents the medical image. The second and third column shows the Grad-Cam of the medical image without adapter and EC-Adapter (proposed method). In last columns the questions are mentioned with their corresponding answer and predicted answers. The Grad-Cam activation visually explains what parts of an image the model focused on to make its prediction. In the first row, an MRI image of the brain is provided, the Grad-Cam activation of the proposed method (EC-Adapter) highlights the location useful for identification of the organ i.e. “brain” whereas without the adapter the activation is misaligned, and produces a wrong prediction. In the second row, the Grad-Cam maps tried to highlight the

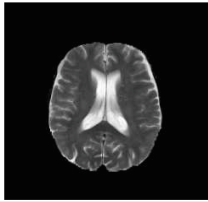
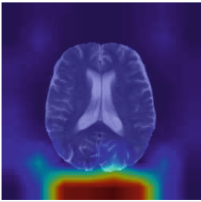
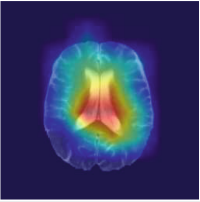

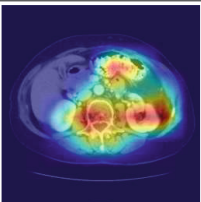
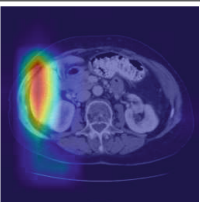

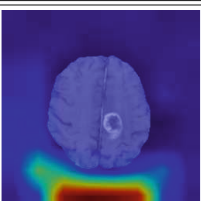
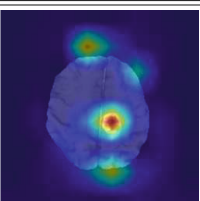
Image	Without Adapter	EC-Adapter	Question, Answer, Prediction
			Question: Is the brain healthy? Answer : Yes Without Adapter : No EC-Adapter : Yes
			Question: Which is bigger in this image, kidney or liver? Answer : Liver Without Adapter : Kidney EC- Adapter : Liver
			Question: What diseases are included in the picture? Ground Truth : Brain Edema, Brain Enhancing Tumor, Brain Non-enhancing Tumor Without Adapter : Right Lobe EC- Adapter : Brain Edema, Brain Enhancing Tumor, Brain Non-enhancing Tumor

Fig. 3. Grad-Cam visualization for the MedVQA. The first column contains the medical image, the second and third column shows the Grad-cam maps without adapter and with adapter (EC-Adapter). The fourth column provides the questions, ground truth, and the predicted answer without adapter and with adapter (EC-Adapter). [Red color is used to represents wrong prediction whereas green color is sued for correct prediction]. (Color figure online)

regions that help in making the decision for the biggest organ. Without the EC-Adapter, the non-liver region major contributed to the decision and predicted the biggest organ in the image as “kidney”, whereas after adding the adapter the selected region is the liver, which is the correct prediction. In the third row, the Grad-Cam highlighted the tumor regions with high activation scores. This qualitative analysis helps to conclude that the EC-Adapter helps improve the selection of image regions for getting better results.

5.3 Ablation Analysis

In this section, ablation studies conducted to analyze the impact of various model’s component.

Exploring Visual Encoder Variants: An analysis is performed by experimenting with different visual encoders pre-trained on general images and results are summarized in Table 2. Initially, ViT [8] was employed, resulting in performance scores of 57.96% and 82.23% on the Path-VQA and SLAKE

Table 2. Performance analysis with different visual encoders on PathVQA and SLAKE-VQA datasets

Visual Encoder	Dataset	
	PathVQA	SLAKE-VQA
ViT [8]	57.96	82.23
DiNOV2 [4]	53.47	82.32
SwinV2 [21]	62.92	84.95

Table 3. Performance analysis with different training settings of Swin Transformer on PathVQA and SLAKE-VQA datasets

Variants	Dataset	
	PathVQA	SLAKE-VQA
Scratch Training	53.37	78.12
FineTuned	61.09	83.19
Finetune with EC-Adapter	62.92	84.95

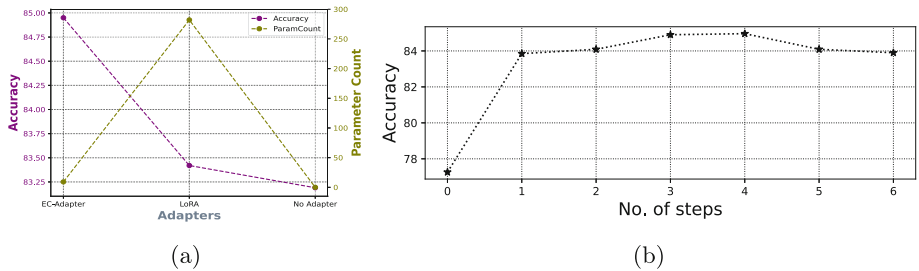
datasets, respectively. However, when using one of the latest vision transformers, DiNOV2 [4], a lower performance of 53.47% was observed on the Path-VQA dataset compared to the ViT transformer. While on SLAKE a comparable performance was achieved. We observed that SwinV2 [21] visual encoder perform consistently and produce superior results.

Impact of Swin Transformer: The superiority of SwinV2 as a visual encoder is evident from Table 3. To further analyse, an analysis is conducted to understand the effects of training on pre-trained SwinV2. The exploration began by training SwinV2 from scratch, as vision transformers requires a huge amount of training data to converge. As anticipated, the model’s learning on smaller MedVQA datasets was limited, resulting in a comparatively lower performance (Table 3, row 1). However, notable improvements were achieved when SwinV2 was fine-tuned on MedVQA images, showcasing gains of $\sim 8\%$ and $\sim 5\%$ on the Path-VQA and SLAKE datasets, respectively. Furthermore, applying the proposed EC-Adapter on SwinV2 features facilitated effective and efficient knowledge transfer from general images to medical images, enhancing performance.

Impact of EC-Adapter: To analyze the influence of the proposed adapter, experiments are conducted with various adapter configurations. Initially, the model was tested without any adapter, utilizing pre-trained SwinV2-encoded visual representations as features. This resulted in an accuracy of 61.09% and 83.19% on the Path-VQA and SLAKE datasets, respectively (Table 4, row 1). Subsequently, LoRA [14] was employed as an adapter on all sets of key and query parameters of the Swin model. Leveraging the pre-trained model’s learning capability with respect to medical images, a performance improvement was observed (Table 4, row 2). Another experiment focused on adapting only the key

Table 4. Performance Analysis with Different Adapters on PathVQA and SLAKE-VQA datasets

Adapter	Dataset	
	PathVQA	SLAKE-VQA
No Adapter	61.09	83.19
LoRA	61.25	83.42
LoRA-SwinV2 [21]	60.96	83.22
EC-Adapter	62.92	84.95

**Fig. 4.** (a) Parameter Count with respect to different Adapters Vs. Test Accuracy, (b) The impact of varying the no. of cross modal attention steps on accuracy on SLAKE dataset

and value parameters of the last block of SwinV2, limiting the model’s adaptability to significant information from the last block (Table 4, row 3). Finally, the EC-adapter was introduced on the pooled features obtained from SwinV2 [21]. The best performance was achieved on both datasets when the lightweight and efficient EC-Adapter was applied to SwinV2-encoded features (Table 4, row 4).

Efficacy of EC-Adapter: Additionally, the effectiveness of the proposed EC-adapter is evaluated by comparing it to a no-adapter setup and the LoRA adapter on the SLAKE dataset. From Fig. 4(a) it is observed that in no-adapter setup, the model achieved a performance of 83.19% without any additional parameters. Introducing the LoRA adapter led to a slight improvement in performance, reaching 83.42% with an increase in parameters to 282K. In contrast, the EC-adapter outperformed both setups, achieving the best performance of 84.95% with significantly fewer parameters, only 9.2K. This showcases the efficiency and effectiveness of the proposed EC-adapter in enhancing performance with a minimal increase in model complexity. The number of parameters required for EC-adapter ϕ is $\sim 9K$ parameter. While the total number of parameters in full model is 18.1M. Hence, the number of parameters for proposed adapter is 0.0005% of total model parameters.

Number of Cross Modal Attention Steps: Figure 4(b) illustrates the effects of adjusting the number of attention steps. Starting with 0 steps, no cross-modal

attention is applied; instead, the two modalities are fused directly. Attention is crucial in VQA tasks [23, 26], and its absence significantly reduces model accuracy. Introducing one attention step enhances modality interaction, resulting in a more enriched contextual feature representation. Performance continues to improve up to 4 attention steps. However, increasing the number of attention steps beyond this does not yield further positive results.

6 Conclusion

In the proposed approach, a lightweight adapter is designed to facilitate the transfer of knowledge acquired from general image-related tasks to the specific context of MedVQA. This adapter acts as a connector, allowing the model to leverage insights from broader visual tasks and apply them effectively to the more specialized domain of medical question answering. To assess the effectiveness of the proposed method, thorough experiments and analysis are conducted on two extensively studied datasets in the domain of MedVQA. The results demonstrate the capability of the proposed EC-Adapter to get trained on a small set of data and produce better results as compared to various SOTA methods.

References

1. Abacha, A.B., Gayen, S., Lau, J.J., Rajaraman, S., Demner-Fushman, D.: NLM at imageclef 2018 visual question answering in the medical domain. In: CLEF (Working Notes) (2018)
2. Antol, S., et al.: VQA: visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015)
3. Bhatele, K.R., Bhadauria, S.S.: Brain structural disorders detection and classification approaches: a review. *Artif. Intell. Rev.* **53**, 3349–3401 (2020)
4. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021)
5. Chen, Z., et al.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 679–689. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_65
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
7. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple meta-model quantifying for medical visual question answering. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 64–74. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_7
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
9. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multi-modal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint [arXiv:1606.01847](https://arxiv.org/abs/1606.01847) (2016)

10. Gong, H., Chen, G., Mao, M., Li, Z., Li, G.: Vqamix: conditional triplet mixup for medical visual question answering. *IEEE Trans. Med. Imaging* **41**(11), 3332–3343 (2022)
11. He, J., Li, P., Liu, G., Zhao, Z., Zhong, S.: Pefomed: parameter efficient fine-tuning on multimodal large language models for medical visual question answering. arXiv preprint [arXiv:2401.02797](https://arxiv.org/abs/2401.02797) (2024)
12. He, X., et al.: Pathological visual question answering. arXiv preprint [arXiv:2010.12435](https://arxiv.org/abs/2010.12435) (2020)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
14. Hu, E.J., et al.: LoRA: low-rank adaptation of large language models. arXiv preprint [arXiv:2106.09685](https://arxiv.org/abs/2106.09685) (2021)
15. Karimi Mahabadi, R., Henderson, J., Ruder, S.: Compacter: efficient low-rank hypercomplex adapter layers. *Adv. Neural. Inf. Process. Syst.* **34**, 1022–1035 (2021)
16. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**(4), 1234–1240 (2020)
17. Li, P., Liu, G., Tan, L., Liao, J., Zhong, S.: Self-supervised vision-language pretraining for medial visual question answering. In: *IEEE 20th International Symposium on Biomedical Imaging*, pp. 1–5. IEEE (2023)
18. Lin, Z., et al.: Medical visual question answering: a survey. *Artif. Intell. Med.* 102611 (2023)
19. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: *2021 IEEE 18th International Symposium on Biomedical Imaging*, pp. 1650–1654. IEEE (2021)
20. Liu, J., et al.: Parameter-efficient transfer learning for medical visual question answering. *IEEE Trans. Emerg. Top. Comput. Intell.* (2023)
21. Liu, Z., et al.: Swin transformer V2: scaling up capacity and resolution. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019 (2022)
22. Mahabadi, R.K., Ruder, S., Dehghani, M., Henderson, J.: Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. arXiv preprint [arXiv:2106.04489](https://arxiv.org/abs/2106.04489) (2021)
23. Mishra, A., Anand, A., Guha, P.: CQ-VQA: visual question answering on categorized questions. In: *2020 International Joint Conference on Neural Networks*, pp. 1–8. IEEE (2020)
24. Mishra, A., Anand, A., Guha, P.: Multi-stage attention based visual question answering. In: *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 9407–9414. IEEE (2021)
25. Mishra, A., Anand, A., Guha, P.: Dual attention and question categorization-based visual question answering. *IEEE Trans. Artif. Intell.* **4**(1), 81–91 (2022)
26. Mishra, A., Anand, A., Guha, P.: Aggregated co-attention based visual question answering. In: *Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 1–10 (2023)
27. Nguyen, B.D., Do, T.-T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: Shen, D., et al. (eds.) *MICCAI 2019. LNCS*, vol. 11767, pp. 522–530. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_57
28. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543 (2014)

29. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
31. Seyfi, G., Esme, E., Yilmaz, M., Kiran, M.S.: A literature review on deep learning algorithms for analysis of x-ray images. *Int. J. Mach. Learn. Cybern.* 1–17 (2023)
32. Shi, L., Liu, F., Rosen, M.P.: Deep multimodal learning for medical visual question answering. In: CLEF (working notes) (2019)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
34. van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., Snoek, C.G., Worring, M.: Open-ended medical visual question answering through prefix tuning of language models. arXiv preprint [arXiv:2303.05977](https://arxiv.org/abs/2303.05977) (2023)
35. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
36. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–29 (2016)
37. Yin, W., Li, L., Wu, F.X.: Deep learning for brain disorder diagnosis based on FMRI images. *Neurocomputing* **469**, 332–345 (2022)
38. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1821–1830 (2017)
39. Zhang, K., et al.: Multi-task paired masking with alignment modeling for medical vision-language pre-training. *IEEE Trans. Multimedia* (2023)
40. Zhang, X., et al.: PMC-VQA: visual instruction tuning for medical visual question answering. arXiv preprint [arXiv:2305.10415](https://arxiv.org/abs/2305.10415) (2023)
41. Zhou, S.K., Greenspan, H., Shen, D.: Deep Learning for Medical Image Analysis. Academic Press (2023)



MUMR: Mask-UnMask Regions Framework for AMD Grades Classification Based on Inter-regional Interactions

Ibrahim Abdelhalim¹, Mohamed Elsharkawy¹, Namuunaa Nadmid¹,
Mohammed Ghazal², Ali Mahmoud¹, and Ayman El-Baz¹(✉)

¹ Department of Bioengineering, University of Louisville, Louisville, KY, USA
ayman.elbaz@louisville.edu

² Electrical and Computer Engineering Department, Abu Dhabi University,
Abu Dhabi, UAE

Abstract. The early diagnosis and effective treatment of age-related macular degeneration (AMD), a leading cause of vision impairment, is contingent upon accurate grading. This paper introduces a novel framework, named Mask-UnMask Regions (MUMR), designed to distinguish between normal retina, intermediate AMD, geographic atrophy (GA), and wet AMD using retinal fundus images, with the input resolution standardized to 1024×1024 pixels. The framework begins by downscaling images to a quarter of their size using a Preserving High-Frequency Information (PHFI) module, which maintains critical details essential for further analysis. Furthermore, we developed a simple, lightweight, yet effective ResNet-like network for efficient feature extraction and introduced a Region Interaction (RI) module, which consists of Adaptive Mask and UnMask Sub-Modules. This module identifies significant regions while reconstructing the insignificant ones using a direction-constrained self-attention mechanism to ensure the learning of global structural cues of AMD grades. The proposed method was evaluated on a dataset of 864 retinal fundus images. Our model consistently achieves superior results compared to other state-of-the-art models, with mean accuracy, mean F1-score, and mean Cohen's Kappa of 92.55%, 92.59%, and 89.97%, respectively. Additionally, we demonstrate that these results are statistically significant compared to other models based on F1-score, indicating that our proposed framework achieves robust and improved AMD grading performance.

Keywords: Age-related Macular Degeneration · Retinal Diseases · Deep Learning · Transformer

1 Introduction

The sense of vision is crucial for humans, offering vital visual data necessary for numerous activities. Retinal disorders, including age-related macular

I. Abdelhalim and M. Elsharkawy—These authors contributed equally to this work.

degeneration (AMD), diabetic retinopathy (DR), and glaucoma, are the primary conditions leading to visual impairment and blindness globally [17]. AMD is a persistent retinal disease predominantly affecting the macular region of the retina, typically observed in older adults.

The progression of AMD is marked by the formation of drusen, with the quantity and size of these deposits determining the disease's stage. AMD is categorized into dry AMD and wet AMD. Dry AMD, the more prevalent type, is further subdivided into early, intermediate, and late stages. Conversely, wet AMD, regarded as the advanced stage of the disease, is classified into inactive and active stages [7]. In wet AMD, vision loss occurs due to the abnormal growth of blood vessels beneath the retina. The shift from dry to wet AMD can happen abruptly, making early detection essential for preventing disease progression. Receiving treatment at this stage is vital for maintaining vision and potentially stopping the further advancement of the disease [9].

Recent advances in deep learning (DL), transfer learning (TL), and vision transformers (ViTs) have demonstrated significant promise in medical image analysis applications, particularly in the diagnosis and grading of retinal diseases using fundus and optical coherence tomography (OCT) images. Many researchers nowadays use vision transformers (ViTs), an innovative architecture derived from transformers originally developed for natural language processing [22], which adapt self-attention mechanisms to interpret image patches as sequences. This method enhances the model's ability to handle complex visual tasks. Recent studies have shown that models incorporating the ViT architecture are particularly effective in identifying and differentiating between various stages of AMD and healthy retinas [4, 6, 10, 18, 19]. For example, Chakraborty et al. [4] implemented their deep convolution neural network (DCNN) model for diagnosing AMD using two public datasets utilizing fundus images, namely, ARIA and iChallenge-AMD [14]. Their method is used to discriminate between AMD and healthy retinas. Pečiulis et al. [18] used the MobileNetV3 pre-trained model trained on a private dataset. They performed a binary classification to distinguish between normal eyes and AMD. Nevertheless, the main problem in [4, 18] is that they didn't distinguish between dry and wet AMD, which requires different treatments for each type. Kumar et al. [15] proposed an ensemble approach combining EfficientNet-B0, VGG16, and ResNet152 pre-trained models to differentiate between dry, wet AMD, and other retinal diseases using public datasets. A deep learning model was implemented by Bhuiyan *et al.* [3] for binary classification, achieving 99.2% accuracy, 98.9% sensitivity, and 99.5% specificity in distinguishing between normal/early and intermediate/late AMD stages. Additionally, a four-class classification was performed to differentiate between normal, early, intermediate, and advanced AMD, with an accuracy of 96.1%. The model was also utilized for predicting disease progression, with accuracy rates of 66.79% for dry AMD and 68.15% for wet AMD over a one-year period. AMDNet23 was introduced by Ali *et al.* [2] to diagnose three different retinal disorders and distinguish between normal retina, AMD, cataract, and DR. An accuracy of 96.50%, specificity of 99.32%, sensitivity of 96.5%, precision of 96.51%, and an F1-score of 96.49% were recorded. The model was trained on 2000 high-quality fundus

images from six public databases: ODIR, Eye Diseases Classification from Kaggle, DR-200, Fundus Dataset, RFMiD, and ARIA. Furthermore, Gour et al. [10] used a VGG16 pre-trained model trained on a private dataset composed of 5K fundus images for eight different retinal diseases. They succeeded in performing multi-label disease detection and differentiating between a normal retina, AMD, and other diseases, recording an accuracy of 84.93%. However, AMD can manifest in dry or wet forms, each carrying a different prognosis. Domínguez et al. [5] evaluated and compared the performance of different ViT-based models to DL-based pre-trained models for AMD disease classification, concluding that working with convolutional-based architectures is better than using transformer-based models for AMD detection and classification using fundus images.

Upon reviewing the existing literature, our findings indicate that many studies extensively resize images to dimensions such as 224×224 or 256×256 to accommodate pretrained models. However, such resizing practices often result in significant information loss critical for accurate classification of AMD grades. Furthermore, these approaches typically fail to effectively identify informative regions corresponding to AMD pathologies within input images. As a consequence of these limitations, this paper proposes a novel framework called Mask-UnMask Regions (MUMR). This framework aims to comprehensively capture the distinctive characteristics of AMD grades to enhance the accuracy of AMD grading using high-resolution retinal fundus images (1024×1024 pixels). Accordingly, our contributions are as follows:

- We introduce the Preserving High-Frequency Information (PHFI) module, designed to retain crucial details while resizing inputs to a quarter of their resolution.
- Developing a simple, lightweight, yet effective ResNet-like network for efficient feature extraction.
- Introducing the Region Interaction (RI) module, comprising Adaptive Mask Sub-Module (AMSM) and Adaptive UnMask Sub-Module (AUMSM). AMSM identifies significant regions while masking irrelevant ones, whereas AUMSM unmaskes masked regions to enhance the understanding of crucial areas in the input, thereby ensuring the acquisition of salient semantic cues relevant to AMD pathology.

2 Methodology

The workflow of the proposed framework (i.e., MUMR) for AMD grading is illustrated in Fig. 1. The dataset consists of colored retinal fundus images varying in resolution from 500×500 to 3152×3000 pixels. To standardize the input, all images are resized to a resolution of 1024×1024 pixels. These resized images are then input into our model, which consists of several steps. The first step in our model is the Preserving High-Frequency Information (PHFI) module, designed to maintain critical high-frequency details crucial for learning, especially in medical images (i.e., retinal fundus images). The resulting feature maps from the

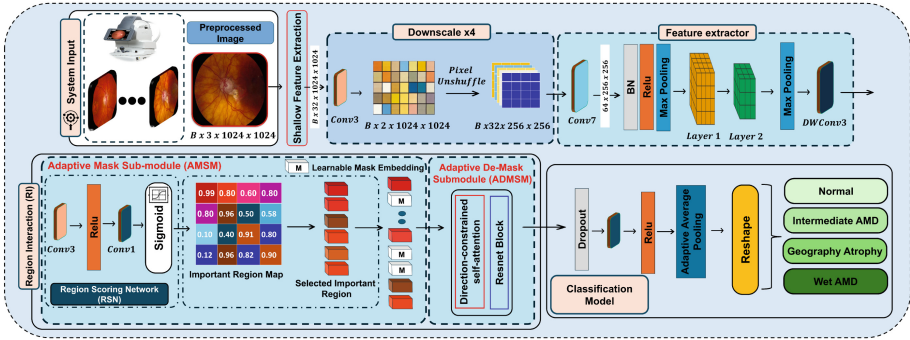


Fig. 1. An overview of the proposed framework for predicting AMD grades.

PHFI module are then passed to the second step, where ResNet-like network is developed for efficient feature extraction. To distinguish between important and unimportant regions effectively, a Region Interaction (RI) model is developed. Finally, the output from the RI module is fed into a classification head to predict the AMD grades.

2.1 Preserving High-Frequency Information (PHFI)

Fundus images contain intricate details essential for accurate grading. Accordingly, preserving these details becomes inevitable for effectively predicting AMD grades. Thus, maintaining high-frequency details is crucial. While maintaining high resolution throughout the network structure is intuitive, it comes at the cost of significantly increased computational demands. Conversely, downsampling via convolution with striding or using pooling mechanisms inevitably leads to information loss and degraded performance. To mitigate these challenges, we employ Pixel Unshuffle for downsampling the image to one-quarter of its original size while expanding the channels without losing high-frequency details. Specifically, the input image $x \in \mathbb{R}^{3 \times W \times H}$ undergoes processing through a 3×3 convolutional layer to derive shallow features $s \in \mathbb{R}^{C \times W \times H}$, where $C = 32$ in our experiments. Subsequently, another 3×3 convolutional layer reduces the channel dimensions to $s \in \mathbb{R}^{C/r^2 \times W \times H}$, which are then expanded back using Pixel Unshuffle to $s \in \mathbb{R}^{C \times W/r \times H/r}$, with $r = 4$.

2.2 Feature Extraction

In the second stage of our pipeline, we employ feature maps generated by PHFI to extract meaningful features for subsequent steps. To achieve this, we developed a modified version of ResNet18 [12], where the initial convolutional layer was adapted to process input tensors with 32 channels instead of the standard 3, with corresponding adjustments made for batch normalization. Additionally, the first two layers (i.e., Layer1 and Layer2) of ResNet18 were adapted, followed by

a MaxPooling layer and a depth-wise convolution layer. The output from this stage, denoted as $F \in \mathbb{R}^{128 \times 32 \times 32}$, is subsequently fed into the third stage, where the model captures inter-regional interactions. It's worth mentioning that Layer1 and Layer2 are initialized with the ImageNet-pretrained weights of ResNet-18.

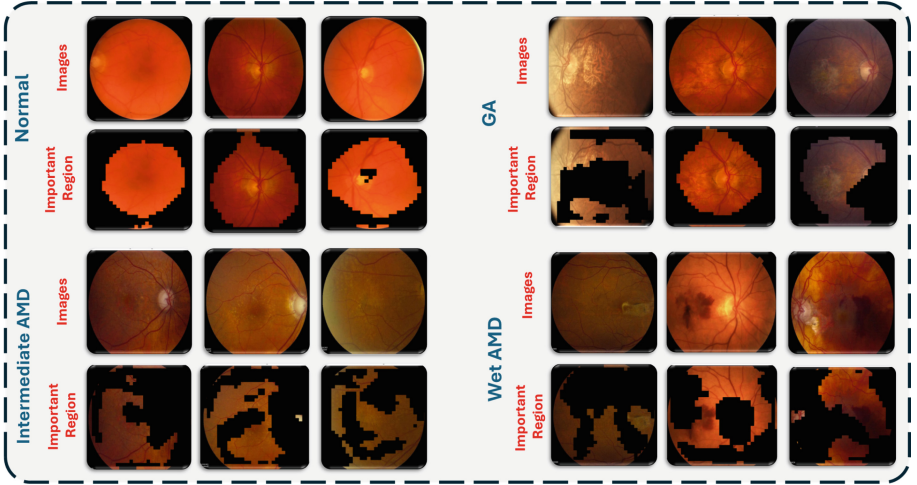


Fig. 2. This figure demonstrates examples of salient regions of AMD grades from the MUMR perspective.

2.3 Region Interaction (RI)

In this step of our pipeline, the RI module captures varying perceptual importance across input feature maps. Specifically, the module comprises two sub-modules: the Adaptive Mask Sub-Module (AMSM) and the Adaptive UnMask Sub-Module (AUMSM). Within AMSM, each region is assigned a score based on its relevance, with high scores indicating significant shape and structural information being retained, while low-scoring regions are masked. Subsequently, AUMSM restores the masked regions to ensure the model incorporates finer details relative to the characteristics of AMD grades, alongside learning global structural cues. This approach enhances the model's ability to discern inter-regional correlations, prioritizing target shapes and structures, thereby improving understanding of AMD grades compared to neighboring regions in fundus images. Such enhancement proves particularly advantageous in medical imaging, where emphasizing structure is critical.

Adaptive Mask Sub-module (AMSM). It operates by passing the feature map F through a lightweight Region Scoring Network (RSN) to evaluate the significance of each feature region. This network comprises two convolutional layers:

$$RSN(F) = \sigma(\Omega_{1 \times 1}(ReLU(\Omega_{3 \times 3}(F)))) \quad (1)$$

where $\Omega_{1 \times 1}$ and $\Omega_{3 \times 3}$ denote convolutional layers with kernel sizes 1×1 and 3×3 , respectively. σ represents the sigmoid function, while ReLU indicates the rectified linear unit. The output of RSN is reshaped to yield scores s_l for each region r_l , $l = 1, \dots, L$, where L represents the number of regions. Regions are then sorted in descending order based on their scores, and the top K scores α_l along with their corresponding regions r_l are selected. These scores are multiplied with normalized region features as modulating factors:

$$\begin{aligned} \Phi &= \{\phi_l \mid \phi_l = \text{LayerNorm}(r_l) \cdot \alpha_l\}, \quad l = 1, \dots, K \\ POS &= \{p_{\phi_l} \mid p_{\phi_l} \in \{0, \dots, L\}\}, \quad l = 1, \dots, K \end{aligned}$$

Here, Φ denotes the selected set of important region features, and POS denotes their respective positions in the original 2D feature map. The parameter K is set to $\beta \times L$, where β is a constant fractional value, and the mask ratio is defined as $1 - \beta$. In our experiments, $\beta = 0.5$. Refer to Fig. 2 for examples illustrating the selected important region from the MUMR perspective.

Adaptive UnMask Sub-Module (AUMSM). Following the masking in AMSM, AUMSM reconstructs these masked regions to enhance the MUMR’s understanding of the characteristics of AMD grades. In other words, this process improves the MUMR’s ability to holistically understand the inter-regional interactions. AUMSM initially filling masked regions with uniformly initialized learnable mask code embeddings (see Fig. 1). Afterwards, it employs a novel direction-constrained self-attention mechanism [13] to facilitate information flow from unmasked to masked regions while preventing reverse flow. This design leverages unmasked region features to infer masked ones without negative impact. The mathematical formulation of direction-constrained self-attention is:

$$Q, K, V = W_Q \Phi, W_K \Phi, W_V \Phi \quad (2)$$

$$A = \left(\text{SoftMax} \left(\frac{QK^T}{\sqrt{C}} \right) \right) \odot M \quad (3)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{C \times C}$ are learnable parameters, M denotes the attention mask, and C is the number of channels in the input. To improve the learning process by enhancing the flow of gradients, a skip connection is used that adds the input F to the output of the RI module:

$$A = A + F \quad (4)$$

It’s worth mentioning that the direction-constrained self-attention mechanism was proposed in [13] for image generation, but here we adapted it to capture inter-regional interactions for AMD grades classification. Finally, A is passed through the Classification Head (CH):

$$\text{CH}(A) = D(F_{1 \times 1}(\text{ReLU}(\text{AAP}(A)))) \quad (5)$$

where D represents a Dropout layer with a probability $p = 0.5$, and AAP represents Adaptive Average Pooling.

Table 1. A comparison between our MUMR and other state-of-the-art models reveals that our model consistently achieves superior results, as indicated by the values highlighted in bold. However, Res2NeXt-DLA-60 [8] achieved the best Recall for 'wet' and the best Precision for 'intermediate'. Results are presented as percentages, with mean and standard deviation for each experiment, each repeated three times with data shuffled each time. T indicates the tiny version of that model.

Model	Metric	GA(%)	Wet (%)	Normal (%)	Intermediate (%)
ConvNext-T [16]	Precision	75.27 ± 6.59	44.53 ± 4.98	58.38 ± 10.89	75.29 ± 1.03
	Recall	77.21 ± 7.28	53.24 ± 17.98	62.62 ± 4.38	60.56 ± 8.58
	F1 Score	76.12 ± 6.45	47.57 ± 8.8	60.22 ± 7.74	66.86 ± 5.62
GFNet-T [21]	Precision	89.16 ± 2.84	73.87 ± 14.66	89.99 ± 3.37	85.02 ± 2.49
	Recall	87.99 ± 7.09	85.30 ± 9.86	91.43 ± 6.17	75.07 ± 8.92
	F1 Score	88.50 ± 4.75	78.76 ± 11.36	90.60 ± 3.78	79.30 ± 4.21
ALOFT-T [11]	Precision	91.72 ± 2.55	74.45 ± 12.24	92.22 ± 8.75	87.29 ± 5.34
	Recall	87.62 ± 6.88	86.85 ± 4.87	84.76 ± 1.35	83.21 ± 9.31
	F1 Score	89.41 ± 2.94	79.63 ± 7.41	88.10 ± 3.98	84.56 ± 2.56
Res2NeXt-DLA-60 [8]	Precision	87.59 ± 8.85	81.78 ± 10.85	93.53 ± 1.68	97.53 ± 3.49
	Recall	96.44 ± 2.74	94.70 ± 2.69	90.71 ± 6.31	79.55 ± 3.56
	F1 Score	91.68 ± 5.99	87.22 ± 4.84	91.92 ± 2.53	87.58 ± 2.90
Ours	Precision	91.87 ± 4.87	88.76 ± 8.13	93.74 ± 2.51	95.47 ± 1.51
	Recall	97.33 ± 3.77	88.93 ± 6.02	97.62 ± 3.37	88.42 ± 4.56
	F1 Score	94.51 ± 4.32	88.47 ± 4.05	95.56 ± 1.18	91.71 ± 1.78

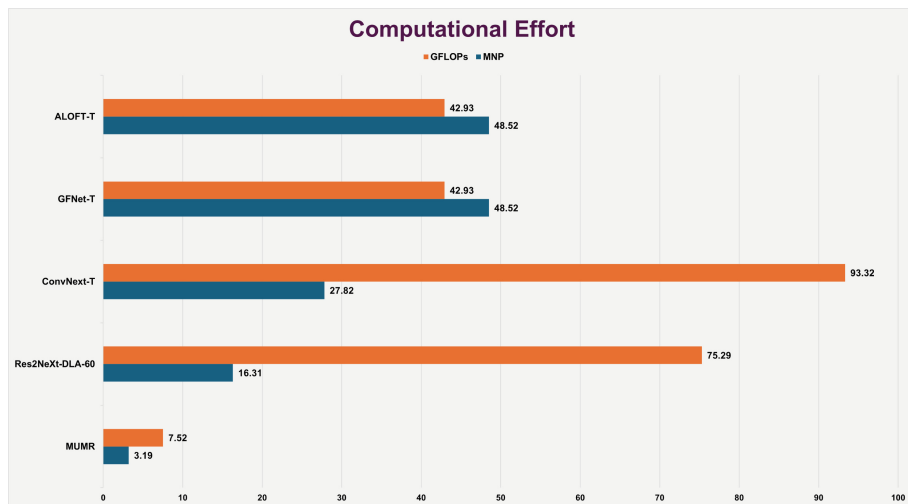


Fig. 3. This figure demonstrates the computational effort required for each model based on MNP and GFLOPs.

3 Experiments and Results

Dataset: The proposed approach is evaluated on a dataset consisting of 864 fundus images, with 216 images for each category: normal retina, intermediate AMD, geographic atrophy (GA), and wet AMD. These images were gathered by the Comparisons of Age-Related Macular Degeneration Treatments Trials (CATT), a study group sponsored by the University of Pennsylvania [1].

Table 2. A comparison between MUMR and other state-of-the-art models using mean accuracy, mean F1-score, and mean Cohen’s Kappa. Bootstrap resampling was employed to establish 95% confidence intervals, confirming statistically significant differences in performance based on mean F1-score. T indicates the tiny version of that model.

Model	Accuracy (%) (95% CI)	F1-Score (%) (95% CI)	Cohen Kappa (%) (95% CI)	Diff
ConvNext-T [16]	62.33 (55.38, 69.23)	62.49 (55.81, 69.02)	49.62 (40.44, 58.74)	30.1 (29.99, 30.21)
GFNet-T [21]	83.86 (78.46, 88.72)	84.24 (79.08, 89.07)	78.31 (71.13, 84.91)	8.35 (8.26, 8.44)
ALOFT-T [11]	85.12 (80.26, 90.07)	85.31 (80.26, 86.27)	79.95 (73.04, 86.27)	7.27 (7.18, 7.36)
Res2NeXt-DLA-60 [8]	89.52 (85.13, 93.85)	89.6 (85.07, 93.67)	85.95 (79.95, 91.64)	2.99 (2.91, 3.07)
Ours	92.55 (88.72, 95.9)	92.59 (88.75, 96.00)	89.97 (84.82, 94.51)	–

Setting: The proposed system was trained using the AdamW optimizer with a learning rate of 0.0001 and a cosine annealing scheduler. The dataset was divided into train (80%), validation (10%), and test (10%) sets for the purpose of training and testing. Each experiment was repeated three times, ensuring the dataset was shuffled each time. Additionally, cross-entropy loss was employed. The implementation was carried out using PyTorch, utilizing a single NVIDIA Quadro P5000 GPU with 16 GB of memory.

Results and Analysis. As shown in Tables 1 and 2, MUMR consistently outperformed other models in mean Accuracy, mean F1-score, and mean Cohen’s kappa, as well as across all classes according to Precision, Recall, and F1-score. However, Res2NeXt-DLA-60 [8] achieved the best recall for ‘wet’ and the best precision for ‘intermediate’. Furthermore, Res2NeXt-DLA-60 [8] achieved the second-best results among the models compared, as shown in Table 2. In Table 2, we employed bootstrap resampling with 10,000 samples from the test set, using the three models from the three experiments mentioned earlier, to construct 95% confidence intervals (CIs) for these metrics. The 95% CIs for the mean accuracy, mean F1-score, and mean Cohen’s Kappa were derived from the 2.5th to the 97.5th percentiles of the bootstrap distribution. To assess the statistical significance of these findings, we followed Rajpurkar et al. [20] in determining the statistical significance of the proposed approach. This involved computing the difference in mean F1-score between MUMR and other models on the same bootstrap samples. The absence of zero within the 95% CI of this difference, as shown in Table 2 under the column labeled ‘Diff,’ indicates a statistically significant superiority of MUMR’s performance over the others. In addition, as illustrated in Fig. 3, we compared all models based on the Million Number of

Parameters (MNP) and Giga Floating-point Operations Per second (GFLOPs) required. As observed from the figure, MUMR requires fewer MNP and GFLOPs, thus outperforming other models in terms of computational efficiency. Like all models, as shown in Fig. 4, MUMR experienced limitations inherently related to the images. As illustrated in the figure, images with illumination issues due to scanning problems caused MUMR to incorrectly identify non-AMD grade regions as important.

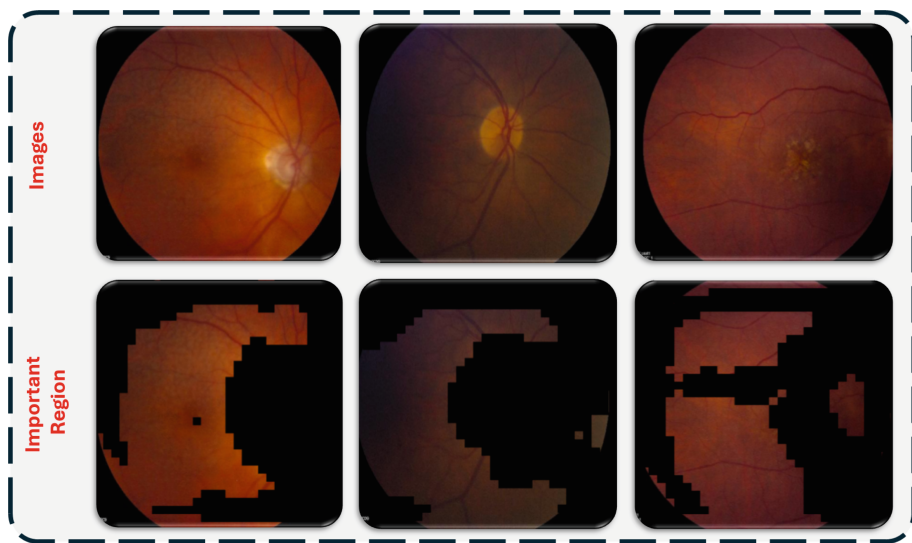


Fig. 4. This figure shows examples of regions obtained incorrectly by MUMR, identified as crucial regions related to AMD grades.

4 Conclusion and Future Work

This paper introduces the MUMR framework for predicting AMD diseases using color retinal fundus images. As detailed in Sect. 2, MUMR comprises several key steps. Initially, high-frequency information is preserved using PHFI to mitigate information loss, while MUMR downscales the image to one-quarter of its original size. Next, a lightweight variant of ResNet-18 is employed to efficiently extract features. Thirdly, MUMR models inter-regional relationships using the RI module to better understand the characteristics of AMD grades regions. Experimental results demonstrate that MUMR achieves superior performance compared to state-of-the-art models in terms of mean Accuracy, mean F1-score, and mean Cohen’s Kappa. Additionally, MUMR achieves statistically significant results in terms of mean F1-score compared to other models. Based on these findings, we posit that this work has the potential to advance clinical decision-making and enhance patient treatment outcomes. In future work, we plan to

explore additional state-of-the-art models and compare them against MUMR, including larger model variants. Furthermore, we intend to evaluate MUMR on additional medical datasets and extend MUMR's capabilities to more complex tasks such as detection and segmentation within the medical domain.

References

1. The Comparisons of Age-Related Macular Degeneration Treatments Trials (CATT). <https://www.med.upenn.edu/cpob/catt.html>
2. Ali, M.A., Hossain, M.S., Hossain, M.K., Sikder, S.S., Khushbu, S.A., Islam, M.: Amdnet23: hybrid CNN-LSTM deep learning approach with enhanced preprocessing for age-related macular degeneration (AMD) detection. *Intell. Syst. Appl.* **21**, 200334 (2024)
3. Bhuiyan, A., Wong, T.Y., Ting, D.S.W., et al.: Artificial intelligence to stratify severity of age-related macular degeneration (AMD) and predict risk of progression to late AMD. *Transl. Vision Sci. Technol.* **9**(2), 25–25 (2020)
4. Chakraborty, R., et al.: DCNN-based prediction model for detection of age-related macular degeneration from color fundus images. *Med. Biol. Eng. Comput.* **60**(5), 1431–1448 (2022)
5. Domínguez, C., et al.: Binary and multi-class automated detection of age-related macular degeneration using convolutional-and transformer-based architectures. *Comput. Methods Programs Biomed.* **229**, 107302 (2023)
6. El-Den, N.N., et al.: Scale-adaptive model for detection and grading of age-related macular degeneration from color retinal fundus images. *Sci. Rep.* **13**(1), 9590 (2023)
7. Elsharkawy, M., et al.: Role of optical coherence tomography imaging in predicting progression of age-related macular disease: a survey. *Diagnostics* **11**(12), 2313 (2021)
8. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: a new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(2), 652–662 (2019)
9. Gehrs, K.M., et al.: Age-related macular degeneration-emerging pathogenetic and therapeutic concepts. *Ann. Med.* **38**(7), 450–471 (2006)
10. Gour, N., et al.: Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network. *Biomed. Signal Process. Control* **66**, 102329 (2021)
11. Guo, J., Wang, N., Qi, L., Shi, Y.: Aloft: a lightweight MLP-like architecture with dynamic low-frequency transform for domain generalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24132–24141 (2023)
12. 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)
13. Huang, M., Mao, Z., Wang, Q., Zhang, Y.: Not all image regions matter: masked vector quantization for autoregressive image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011 (2023)
14. iChallenge: Broad (Baidu Research Open-Access Dataset) iChallenge-AMD dataset (2019). <http://ai.baidu.com/broad/subordinate?dataset=amd>. Accessed 10 May 2023

15. Kumar, K.S., et al.: Retinal disease prediction through blood vessel segmentation and classification using ensemble-based deep learning approaches. *Neural Comput. Appl.* **35**(17), 12495–12511 (2023)
16. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986 (2022)
17. Mitchell, P., et al.: Age-related macular degeneration. *The Lancet* **392**(10153), 1147–1159 (2018)
18. Pečiulis, R., et al.: Automated age-related macular degeneration area estimation—first results. *arXiv preprint [arXiv:2107.02211](https://arxiv.org/abs/2107.02211)* (2021)
19. Philippi, D., et al.: A vision transformer architecture for the automated segmentation of retinal lesions in spectral domain optical coherence tomography images. *Sci. Rep.* **13**(1), 517 (2023)
20. Rajpurkar, P., et al.: Chexnet: radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint [arXiv:1711.05225](https://arxiv.org/abs/1711.05225)* (2017)
21. Rao, Y., Zhao, W., Zhu, Z., Zhou, J., Lu, J.: Gfnet: global filter networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(9), 10960–10973 (2023)
22. Vaswani, A., et al.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, vol. 30 (2017)



A New Attention Based UNet and Gated Edge Attention Network for Retinal Vessel Segmentation

Ayush Roy¹, Shivakumara Palaiahnakote^{2(✉)}, Umapada Pal³, and Sukalpa Chanda⁴

¹ Department of Electrical Engineering, Jadavpur University, Kolkata, India

² School of Science, Engineering and Environment, University of Salford, Salford, UK
s.palaiahnakote@salford.ac.uk

³ Computer Vision and Pattern Recognition Unit, Indian Statistical Institute,
Kolkata, Kolkata, India
umapada@isical.ac.in

⁴ Østfold University College, Halden, Norway
sukalpa@ieee.org

Abstract. Early diagnosis of retinal diseases is crucial for preventing blindness. However, due to background variations and degradation in the images, retinal vessel segmentation has become challenging. As a result, accurate segmentation of the retinal vessels is essential for enhancing diagnosis to identify the disease. To achieve this, inspired by the special ability of the attention mechanism, which detects vital regions, and UNet, which segments the vital region, we propose a combination of a new attention mechanism and modified UNet for segmenting vessels in retina images. In the proposed segmentation model, the convolutional blocks have been modified to capture multiscale spatial information by varying the convolution dilation rates. Similarly, the Trainable tanh activation (T-Tanh) is adapted in a new way to identify changes in the flow of the feature gradients to differentiate between the retinal vessel pixels and the background. Furthermore, to make the segmentation robust, the Gated Edge Attention (GEA) network is proposed. The effectiveness of the segmentation is demonstrated by testing on two benchmark datasets, namely, STARE and CHASE. The results show that the performance of the proposed method is superior to the state-of-the-art methods.

Keywords: Retinal vessel segmentation · UNet · Attention mechanism · Diabetic retinopathy

1 Introduction

Deformations within the retina's internal structure can lead to various ocular diseases, making early detection crucial [1]. It is true that retinal vessels are the key indicators for identifying diseases such as Diabetic Retinopathy (DR), Glaucoma, age-related Macular Degeneration, and cardiovascular diseases. Therefore, accurate analysis of retinal vessels is vital for effective disease detection. Thus, it is necessary to segment vessels

accurately to make them visible. Color Fundus Photography (CFP) is the most commonly used imaging modality due to its non-invasive nature, avoiding the inconvenience and hospitalization associated with invasive methods. Retinal vessel segmentation is not a new problem, we can find several methods in the literature [2–4]. However, the past methods are not robust and effective for the images affected by adverse factors, such as low resolution, background color changes, etc. As a result, segmenting blood vessels in retina images presents several challenges:

- Low contrast in fundus images makes differentiation between vessels and background difficult.
- Pathological features like exudates and hemorrhages can be mistaken for vessels.
- The complex morphology of retinal vasculature varies in orientation and scale.

Our aim is to address these challenges effectively, as illustrated in Fig. 1. It is noted that the traditional medical image segmentation methods, which usually rely on conventional image processing steps utilize handcrafted features and domain knowledge [2–4]. To overcome these limitations, recent advancements in deep learning, particularly convolutional neural networks (CNNs), have transformed medical image segmentation [5]. In the same way, architectures like UNets [6] and their variants, including Attention UNet [7], ResUNet [8], and DeepLab V3 + [9], demonstrate exceptional performance. However, these architectures ignore boundary information for segmentation. Additionally, many existing methods overlook critical edge information necessary for accurate segmentation and vessel boundary adherence, leading to false positives. Transformers, although effective in capturing multiscale feature information, require extensive training data and involve a large number of trainable parameters.

This observation motivated us to introduce the method that explores the attention mechanism, modified UNet with a gated edge attention network. The proposed method can be seen in Fig. 1(a) and Fig. 1(b), where it can be seen that the predicted results obtained by the proposed model are almost the same as the ground truth for all the images with different complexities. The special characteristics of the attention mechanism, UNet, and gated edge attention network inspired us to propose a new model to integrate the strength of the above-modified model for accurate vessel segmentation in this work.

Therefore, the key contributions of the proposed method are as follows. (i) The chosen backbone is the Attention UNet, which integrates attention mechanisms into the U-Net framework. This enhances its ability to capture finer details and complex structures in retinal vessels. (ii) The convolution blocks of the encoder, decoder, and bottleneck layers of the Attention UNet are modified to capture multiscale spatial information. (iii) The Gated Edge Attention (GEA) captures edge information, highlights spatial regions, and provides boundary adherence for accurate segmentation mask generation. This reduces the false positive rate by streamlining the focus of the model and making it vessel boundary-aware.

The structure of the paper is as follows. A review of the different methods of retinal vessel segmentation is presented in Sect. 2. Section 3 discusses the architecture of the proposed segmentation method. The results and analysis to validate the performance of the proposed segmentation are presented in Sect. 4. Conclusion and Future work are listed in Sect. 5.

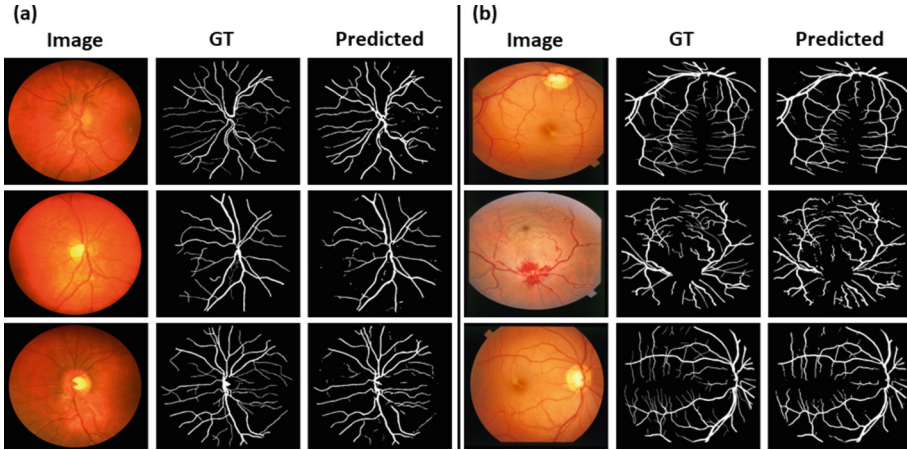


Fig. 1. Sample images, ground truth (GT), and the predicted mask of the proposed model for (a) CHASE and (b) STARE datasets.

2 Related Works

Recently, Biswas et al. [10] stated that the methods of medical imaging, especially the approach of segmentation play an important role for precise analysis and early detection of various diseases. This study presents a review of different deep-learning approaches for segmentation and classification. The methods used different architectures for vessel segmentation in the past. Anwar et al. [11] observed that accurate and efficient classification depends on the success of segmentation approaches. This study uses EfficientNet for the classification of different types of cancer, which include brain tumors, breast cancer, chest cancer, and skin cancer.

Since the aim of the work is to segment vessels from retina images, this work reviews different deep-learning models. For example, Iter-Net [12] enhances retinal vessel segmentation by cascading a UNet with mini UNets while reducing channel numbers for efficiency. The main objective of the approach is to find obscured details of the vessel from the segmented vessel image itself, rather than using the raw input image. Therefore, there are chances of missing sometimes vital details in the images. UNet++ [13] improves skip connections to fuse information across multiple scales. This model has the ability to reduce the effect of unknown network depth with an efficient ensemble of U-Nets. CE-Net [14] addresses spatial information loss in UNet with a multi-scale branch structure and dilated dense blocks. The CE-Net includes a feature encoder module, a context extractor module, and a feature decoder module for image segmentation. However, the model works well for 2D images. Genetic UNet [15] employs an evolutionary neural architecture search for retinal vessel segmentation, resulting in a streamlined network design. The key idea for the successful segmentation is that the model generates a U-shaped CNN which can segment vessels with high accuracy and few parameters.

DE-DCGCN-EE [16] features a dynamic-channel graph CNN with dual encoders for edge enhancement and topological relation utilization. The key step of the proposed work is to propose an edge detection-based dual encoder to preserve the edges of the

vessel. Sine-Net [17] introduces unconventional upsampling and down-sampling operations. The authors noted that deep learning achieves the best results when the models extract contextual features. However, it is not so easy to extract contextual features for all situations. LIOT [18] enhances Iter-Net's generalization through innovative preprocessing sensitive to curvilinear structures. The main idea to achieve generalization ability is that the approach transfers a grayscale image into contrast invariant channels based on pixel values and their neighboring values. SegR-Net [19] utilizes feature extraction and fusion for precise segmentation masks. While CNNs excel at exploiting translation symmetry, they often fall short in effectively handling rotation and scale symmetries, which are equally important for segmentation tasks. FRS-Net [20] addresses this issue by introducing FRS-Conv, a novel convolution operator equivariant to both rotation and scaling. However, predefined convolution operators cannot generalize across multiple datasets and depend on local information, ignoring global dependencies.

Overall, despite powerful deep learning models being proposed for vessel segmentation, the scope of the methods is limited to particular situations and applications to achieve the best results. In addition, as the accuracy improves, the number of computations increases due to heavy architectures. Therefore, since segmentation is a preprocessing step of detection, classification, and identification, it is necessary to develop a model that can work for any dataset and images. This is the major weakness of the existing methods. Thus, this work aims to develop a generalized and efficient model to achieve the best accuracy without additional computational burden.

3 The Proposed Methodology

The main objective of the proposed method is to segment vessels in the retina images. As discussed in the previous section, segmenting an accurate vessel is not easy due to background complexities and variations at the edges of the diseased images. Therefore, based on special characteristics of the attention network, UNet, and gated edge attention network, we propose a new model that integrates the strength of each modified component mentioned earlier for accurate segmentation, which results in a new model for segmentation.

The segmentation model we are using is based on the Attention UNet [21] architecture, which incorporates attention mechanisms into the U-Net framework. This enhances its ability to capture finer details and complex structures in retinal vessels. The traditional convolution layers in Attention UNet have been replaced by separable convolutional layers, reducing the number of trainable parameters without a decrease in performance. The convolution blocks of the encoder, decoder, and bottleneck layers have been modified to capture multiscale features by introducing spatial attention from the ASPP module to capture global and local dependencies. The parameterized tanh, i.e. T-Tanh, is used to capture the feature gradient flow of the image, adjusting the parameters accordingly to demarcate the flow changes in the images and spot the retinal vessels, thus ensuring differentiation between the foreground and background pixels.

The traditional gated attention aids the skip connections in the Attention UNet. To make the model adhere to the boundaries of the retinal vessels, we modified this attention by introducing the edge information to formulate the Gated Edge Attention (GEA)

module. GEA highlights the edge information for streamlining the model’s spatial focus. Furthermore, to differentiate between the foreground and background pixels, feature gradient flow is utilized for further spatial information enrichment. Feature gradient flow leverages the network’s ability to learn representations that identify the change in feature gradients along the retinal vessels as seen in Fig. 2. This gradient information showcases the feature flow specifically pointing in the direction of the retinal vessels.

Finally, a convolution layer with sigmoid activation produces the predicted mask. A detailed block diagram of the proposed model is shown in Fig. 3. As we go deeper into the network, i.e. from the bottleneck to the decoder layers, the spatial information becomes more and more prominent as shown in Fig. 4, where one can see the proposed method is effective and capable of segmenting vessels for the images of different complexities.

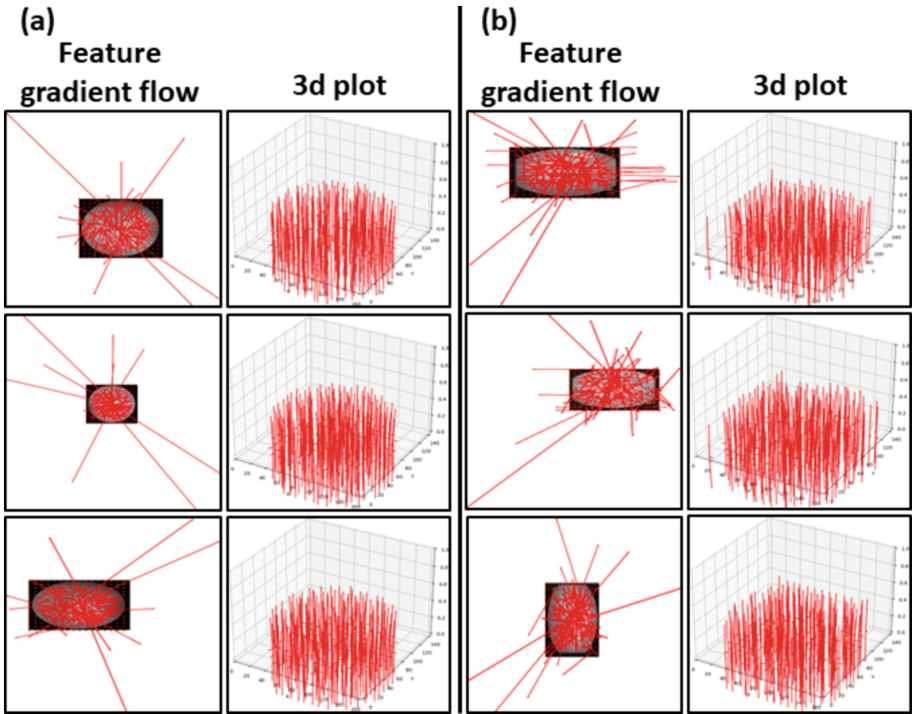


Fig. 2. Feature gradient flow of the images shown in Fig. 1. The arrow showcases the direction in which the feature gradient changes, i.e., the direction of the retinal vessels.

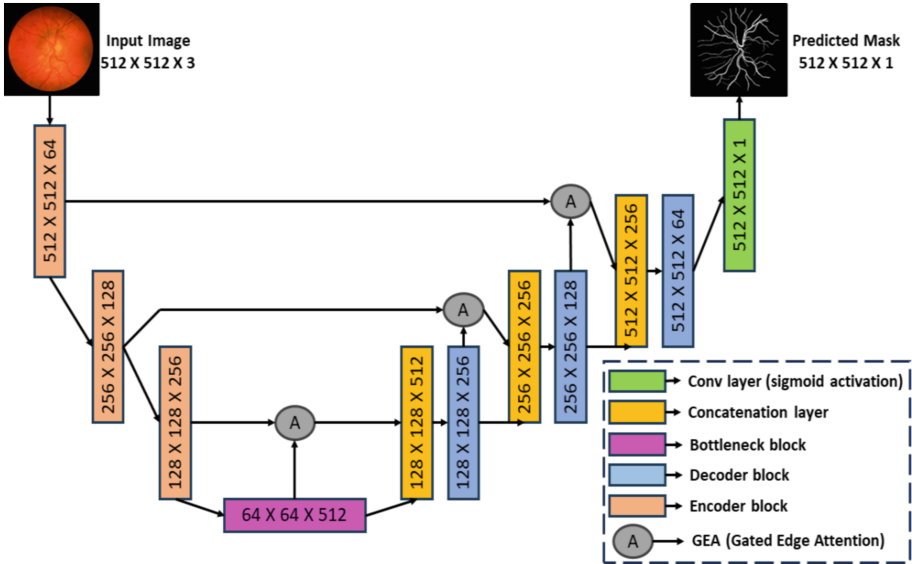


Fig. 3. A block diagram of the proposed model. Attention UNet is used as the baseline UNet. The gated attention is modified to Gated Edge Attention (GEA) to focus on the retinal vessel edges for boundary adherence. The convolution blocks of the encoder, decoder, and bottleneck layers are modified to capture both local and global features

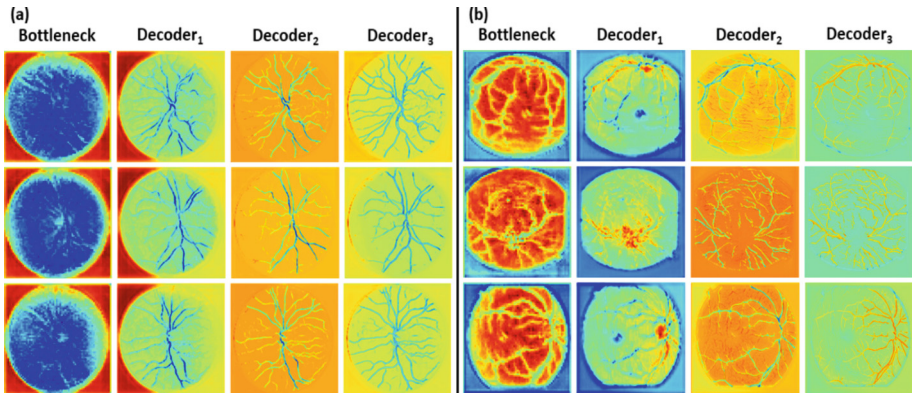


Fig. 4. Heatmaps of the decoder layers for images in Fig. 1.

3.1 Convolution Block

In the Attention UNet, convolution layers are the fundamental building blocks. Traditionally, these layers consist of convolution operations followed by batch normalization and ReLU activation. However, we propose a modification to these convolution layers in our work. Modified convolution layers aim to capture multiscale features to achieve scale invariance effectively. Specifically, we introduce a trainable Tanh (T-Tanh) activation

function to enhance spatial feature gradient flow and distinguish between foreground and background pixels.

The Tanh function is chosen for its steep gradient around the origin, making it highly sensitive to input changes, particularly in regions with values close to zero. In retinal vessel segmentation, such regions often correspond to areas with subtle transitions or gradients, such as vessel boundaries. To adapt the Tanh function to complex spatial information, we parameterize it and make it trainable. The equation for T-Tanh is represented in Eq. 1, where x is the input, and shift and slope are trainable parameters.

$$T - \text{Tanh}(x) = \text{Tanh}(\text{slope} \times (x - \text{shift})) \tag{1}$$

The shift parameter horizontally translates the Tanh function, enabling adaptation to varying background intensities or brightness levels in retinal vessel segmentation. The slope parameter adjusts the steepness of the transition from minimum to maximum values. By dynamically adjusting both parameters during training, the trainable Tanh function effectively captures spatial feature gradient flow, accommodating intensity variations and enhancing sensitivity to spatial gradients. Thus, with adjustable parameters, the trainable Tanh function accurately captures flow variations in retinal vessels, facilitating precise segmentation tasks such as retinal vessel segmentation.

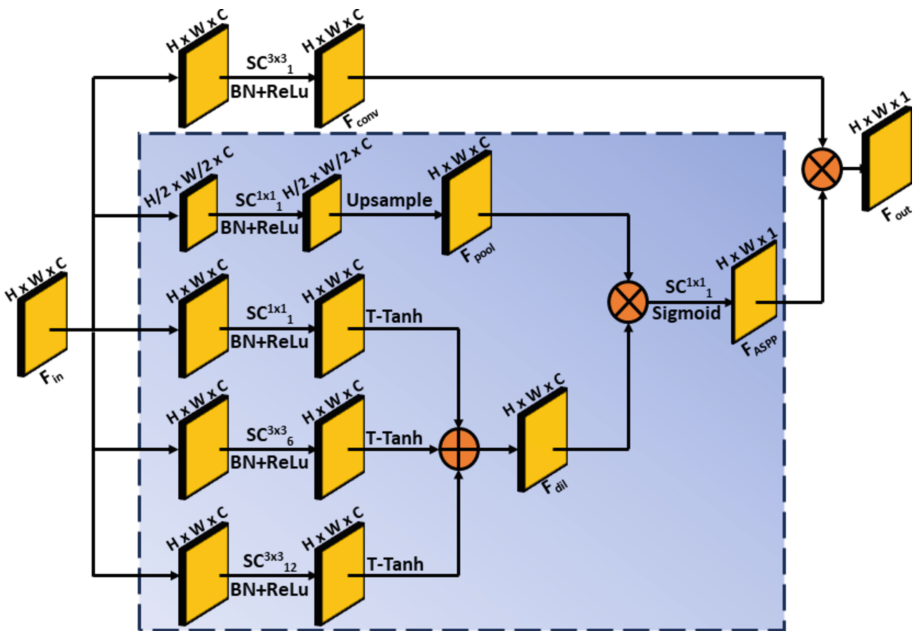


Fig. 5. The modified convolution block. The Atrous Spatial Pyramid Pooling (ASPP) captures multiscale features and the trainable Tanh (T-Tanh) captures the feature gradient flow to differentiate between the foreground and background pixels. BN stands for Batch Normalization and $SC_d^{k \times k}$ is a Separable Convolution layer with a dilation rate of d and kernel size of $k \times k$.

F_{in} is the input to the convolution block with dimensions $H \times W \times C$. It is treated by a separable convolution layer, batch normalization layer, and ReLU activation to generate F_{conv} of dimensions $H \times W \times C$. It is then subsequently treated by the ASPP module to generate multiscale features. It is a cascaded setup where F_{in} is treated by separable convolution layers with varying dilation rates ($d = 1, 6, 12$) to capture features of various receptive fields. The features extracted for different dilation rates are then treated by the T-Tanh activation to capture the spatial feature gradient flow and then added to produce F_{dil} of dimensions $H \times W \times C$. F_{in} is also max-pooled to capture the dominant spatial features and then subsequently treated by separable convolution layers and upsample layer to produce F_{pool} of dimensions $H \times W \times C$. F_{pool} and F_{dil} are multiplied element-wise and treated by a separable convolution layer with sigmoid activation to generate F_{ASPP} of dimensions $H \times W \times 1$. F_{ASPP} consists of the multiscale spatial feature gradient flow information to highlight the foreground regions, i.e. the retinal vessels. Finally, F_{ASPP} is element-wise multiplied with F_{conv} to generate F_{out} of dimensions $H \times W \times C$ as shown in Fig. 5.

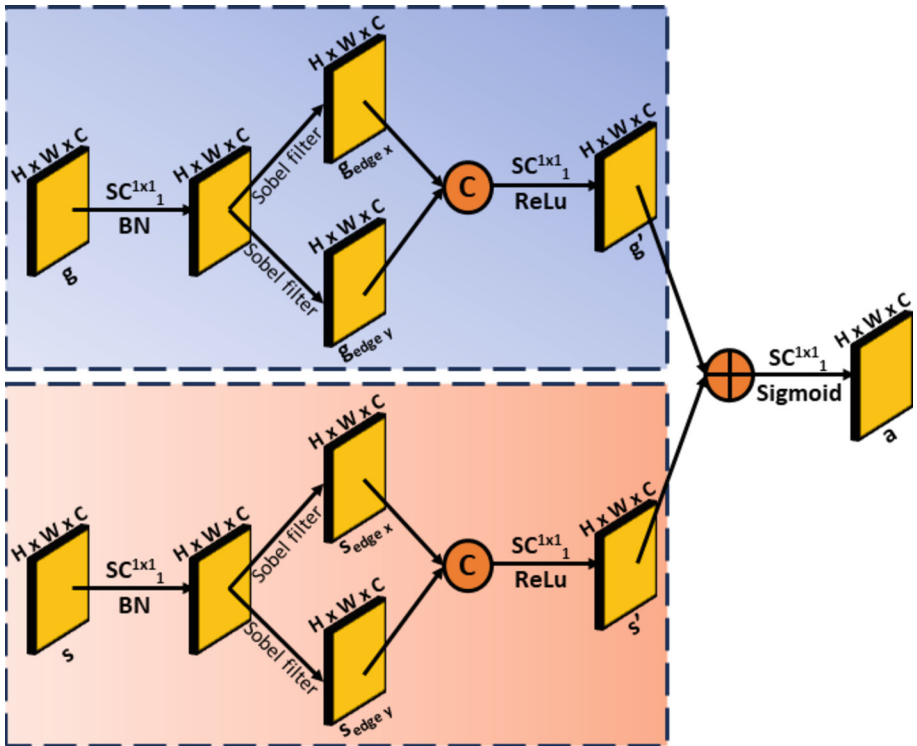


Fig. 6. An illustration of the Gated Edge Attention (GEA) module.

3.2 Gated Edge Attention (GEA)

The Gated Edge Attention (GEA) module is a modified version of the gated attention used in the Attention UNet. Its main purpose is to enhance the extracted spatial features with edge information. Segmentation models often smooth out the boundaries of the region of interest too much, which can be problematic for tasks like retinal vessel segmentation where precise demarcation of thin vessel boundaries is crucial. The input feature maps g and s are processed by separable convolution layers, batch normalization layers, and Sobel filters to capture edge information across the height and width directions. The resulting edge features $g_{\text{edge } x}$, $s_{\text{edge } x}$ (along the height direction), $g_{\text{edge } y}$, and $s_{\text{edge } y}$ (along the width direction) are concatenated and processed by a separable convolution layer with ReLU activation to generate g' and s' . These are then added together and processed by a separable convolution layer with sigmoid activation to generate 'a', which represents the spatial attention weight including the edge information. This helps the decoder layers produce a boundary-aware segmented output. A detailed block diagram of GEA is shown in Fig. 6. After the incorporation of both the GdAM and CSC modules.

4 Experimental Results

To evaluate the performance of the proposed model, two benchmark datasets, namely, STARE [22] and CHASEDB1 [23] are considered in this work. The STARE dataset comprises 20 digital retinal images from 10 subjects, each with a resolution of 700×605 pixels and 8-bit pixel depth. On the other hand, the CHASEDB1 dataset includes 28 digital retinal images from 28 subjects, with images of dimensions 999×960 pixels and 8-bit pixel depth. Expertly annotated retinal vessel masks accompany both datasets for evaluating segmentation performance. We utilized the original image size of 512×512 pixels for input and used a train-validation-test split of 70-10-20%. For the training set, we used data augmentation of horizontal flip, vertical flip, and rotation by 90° in both clockwise and anticlockwise directions. The details of the datasets are given in Table 1.

Standard metrics to evaluate segmentation performance like accuracy (Acc), Dice coefficient (Dc), Intersection over Union (IoU), Sensitivity (Se), and Specificity (Sp) were used. Accuracy measures the overall correctness of the segmentation results. The dice coefficient quantifies the overlap between the predicted segmentation and the ground truth. IoU measures the overlap between the predicted segmentation and the ground truth, normalized by the total area covered. Sensitivity measures the proportion of actual positives that are correctly identified. Specificity measures the proportion of actual negatives that are correctly identified. The measures are defined in Eq. (2)–Eq. (6), where TP, TN, FP, and FN represent True Positive pixels, True Negative pixels, False Positive pixels, and False Negative pixels respectively.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Dice Coefficients} = \frac{2TP}{2TP + FP + FN} \quad (3)$$

$$\text{Intersection Over Union} = \frac{TP}{TP + FP + FN} \quad (4)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (5)$$

$$Specificity = \frac{TN}{TN + FP} \quad (6)$$

More details about the measures can be found in [22, 23]. The training involves using a learning rate of 0.0001, the Adam optimizer, a batch size of 4, and training for 150 epochs. Implementation is carried out using TensorFlow on an NVIDIA TESLA P100 GPU, ensuring consistency throughout the ablation study.

Table 1. A summary of the number of images used for training and testing for the two datasets.

Dataset	Training	Training (augmented)	Testing
STARE [22]	16	80	4
CHASEDBI [23]	110	22	6

4.1 Ablation Study

To figure out the contribution and effectiveness of the key components used in the proposed method for segmenting vessels in retina images, we conducted ablation study experiments on the CHASEDBI dataset as listed in Table 1. (i) Attention UNet, which is the baseline architecture used to show that the baseline architecture is not capable of achieving the best segmentation results. (ii) Attention UNet (replacing traditional convolution layers in the convolution block with separable convolution layers to make it lightweight), which is to show the effectiveness of a modified attention network. (iii) Adding Gate Edge Attention (GEA) network to the model in (ii), which is to show the effectiveness of edge information for boundary adherence. (iv) Adding a modified convolution block to the (iii), which is to show the effectiveness of modified convolutional blocks to achieve the best segmentation, and (v) The proposed model with the T-Tanh activation function, which is to show the contribution of integrating all the above components to achieve high accuracy.

Table 2 showcases the significant influence of the GAE’s edge information. This information aids in enriching the features and highlighting the spatial boundaries of the retina. Moreover, the use of separable convolution layers instead of the traditional convolution layer reduces the number of trainable parameters by approximately 6.5 times with comparable performance. The use of a modified convolution block helps capture multiscale spatial feature gradient flow, thus boosting the performance of the model. Furthermore, the use of T-Tanh provides the necessary flexibility to adjust the threshold to capture the feature flow for retinal vessels. Overall, Table 2 shows that the performance of the method in terms of the number of parameters, accuracy, dice coefficients, sensitivity, specificity, and IoU is the best compared to the performance of individual components. This is because as key components and modifications are

Table 2. Performance of the key components of the proposed segmentation models on the CHASEDBI dataset (in %).

#	Steps	Parameters	Accuracy	Dice Coefficients	Sensitivity	Specificity	IoU
(i)	Baseline Attention UNet	8.14 M	95.05	75.20	76.17	98.39	57.87
(ii)	Attention UNet (Replacing traditional convolution layers with separable convolution layers)	1.26 M	95.65	74.78	76.84	98.05	58.72
(iii)	Modified Attention UNet + GEA	1.26 M	96.72	75.25	78.26	98.03	59.32
(iv)	Modified Attention UNet + GEA + Modified Convolutional Blocks	1.48 M	96.98	75.53	82.42	98.22	61.41
(v)	Proposed method + T-Tanh activation function	1.48 M	96.96	75.78	84.03	98.14	61.56

added to baseline architectures, the performance in terms of all the parameters improves. Therefore, one can infer that the proposed model integrates the strengths effectively to achieve the best segmentation results.

4.2 Comparison with the State-of-the-Art Methods

The proposed model performs better than the state-of-the-art (SOTA) models for both the STARE and CHASEDBI datasets, as demonstrated in Table 3 and Table 4 respectively. Our model shows lower specificity than LIOT Iter-Net [18] for the CHASEDBI dataset. However, LIOT Iter-Net [18] has poor sensitivity, resulting in a lower dice score. In contrast, our model demonstrates better sensitivity and consequently achieves a higher dice score. For the STARE dataset, SegR-Net [19] has the highest sensitivity, and Sine-Net [17] has the highest specificity, although both have lower specificity and sensitivity, respectively. Conversely, our proposed model exhibits higher specificity and sensitivity than SegR-Net [19] and Sine-Net [17], resulting in a higher dice score compared to the existing SOTA methods. It is evident from Table 2 that our proposed model has 1.48

million parameters, an improvement over standard UNets and comparable to state-of-the-art models. However, although our model has fewer parameters than most models, there are models such as Genetic U-Net [15] and SegR-Net [19] that have lower latency than our proposed model.

Table 3. Performance comparison of the proposed model with the SOTA models for the STARE dataset (in %)

Models	Parameters	Accuracy	Dice Coefficients	Sensitivity	Specificity	IoU
UNet [6]	31.03 M	97.05	71.24	80.49	97.41	55.33
UNet++ [13]	-	97.14	-	79.47	98.82	-
Iter-Net [12]	-	97.07	-	77.62	98.92	-
CE-Net [14]	39.35 M	96.90	-	79.85	98.52	-
Attention UNet [7]	35.60 M	-	72.00	82.01	97.23	56.25
ResUNet [8]	33.16 M	-	62.98	80.27	96.73	45.96
DeepLab V3+ [9]	36.89 M	-	65.58	77.71	95.19	48.79
Sine-Net [17]	6.9 M	97.11	-	67.76	99.46	-
DE-DCGCN-EE [16]	-	96.79	-	73.98	98.96	-
Genetic UNet [15]	0.27 M	97.19	-	79.94	98.83	-
LIOT Iter-Net [18]	-	96.94	-	78.53	98.69	-
FRS Iter-Net [20]	-	97.30	-	80.13	98.93	-
SegR-Net [19]	0.65 M	-	72.49	82.12	98.14	56.86
Ours	1.48 M	97.35	72.84	79.57	98.66	57.97

Cross-dataset evaluation involves training models on one dataset and testing them on another. This approach is essential in the medical field because models developed using a hospital's own datasets often need to be applied to different datasets. In addition, this experiment indicates the proposed model is robust to different datasets and has generalization ability. A strong cross-dataset performance indicates a model's reliability in practical settings. For instance, $CD_{STARE_CHASEDB1}$ refers to a model trained on STARE and tested on CHASEDB1, while $CD_{CHASEDB1_STARE}$ refers to a model trained on CHASEDB1 and tested on STARE. Cross-dataset evaluation is more challenging for a model's generalization and robustness than training and testing on the same dataset. Table 5 provides the numerical results for these cross-dataset experiments. The proposed model demonstrates superior accuracy and specificity compared to state-of-the-art (SOTA) models, though it does not show improvement in sensitivity. Despite this, the significant improvements in other metrics, particularly accuracy, suggest that the proposed method is superior and has promising potential for clinical applications. Note that the results of the existing methods presented in Tables 3, 4, and 5 are sourced directly from the reported findings of the respective authors of the cited papers.

Table 4. Performance comparison of the proposed model with the SOTA models for the CHASEDB1 dataset (in %)

Models	Parameters	Accuracy	Dice Coefficients	Sensitivity	Specificity	IoU
UNet [6]	31.03 M	95.71	70.56	82.72	98.05	54.51
UNet++ [13]	-	96.62	-	79.64	98.31	-
Iter-Net [12]	-	96.54	-	79.57	98.23	-
CE-Net [14]	39.35 M	96.43	-	77.90	98.27	-
Attention UNet [7]	35.60 M	-	71.25	81.72	98.35	55.34
ResUNet [8]	33.16 M	-	61.97	82.47	98.04	44.89
DeepLab V3+ [9]	36.89 M	-	65.50	76.35	97.69	48.70
Sine-Net [17]	6.9 M	96.78	-	80.11	98.15	-
DE-DCGCN-EE [16]	-	96.35	-	76.25	98.35	-
Genetic UNet [15]	0.27 M	96.58	-	79.85	98.25	-
LIOT Iter-Net [18]	-	96.37	-	75.66	98.43	-
FRS Iter-Net [20]	-	96.71	-	81.55	98.22	-
SegR-Net [19]	0.65 M	-	72.29	83.29	98.38	56.60
Ours	1.48 M	96.96	75.78	84.03	98.14	61.56

Table 5. Performance comparison of the proposed model with the SOTA models for cross-dataset validation (in %)

Model	CDCHASEDB1_STARE			CDSTARE_CHASEDB1		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
UNet [6]	95.84	56.04	99.61	94.66	67.87	97.32
IterNet [12]	95.65	54.19	99.59	94.49	58.36	98.08
UNet++ [13]	95.70	55.16	99.55	94.36	60.63	97.72
CE-Net [14]	96.07	63.18	99.19	94.63	69.43	97.14
DE-DCGCN-EE [16]	95.79	62.84	98.92	93.46	46.97	98.09
Genetic UNet [15]	96.59	70.83	99.04	94.29	59.20	97.78
LIOT Iter-Net [18]	96.44	69.37	99.01	94.36	64.32	97.35
FRS Iter-Net [20]	96.78	71.15	99.21	95.01	72.66	97.24
Ours	96.81	66.64	99.21	95.36	62.08	98.55

4.3 Error Analysis

Although the proposed segmentation model exhibits superior performance compared to state-of-the-art models, there are still areas for improvement. Figure 8 illustrates the

specific images that result in an erroneous segmentation output. The foreground-to-background feature similarity is the cause of over and under-segmentation. This can be avoided by introducing a loss to separate out the background and foreground feature distribution to enhance prominent foreground features. Also, discontinuous segmented retinal vessels are present in the predicted mask. This can be tackled by traditional techniques of comparing K different neighboring pixels to maintain the connectivity of the vessels. In this study, we focused on the boundary adherence problem of existing segmentation models and the aforementioned issues will be studied in our future research.

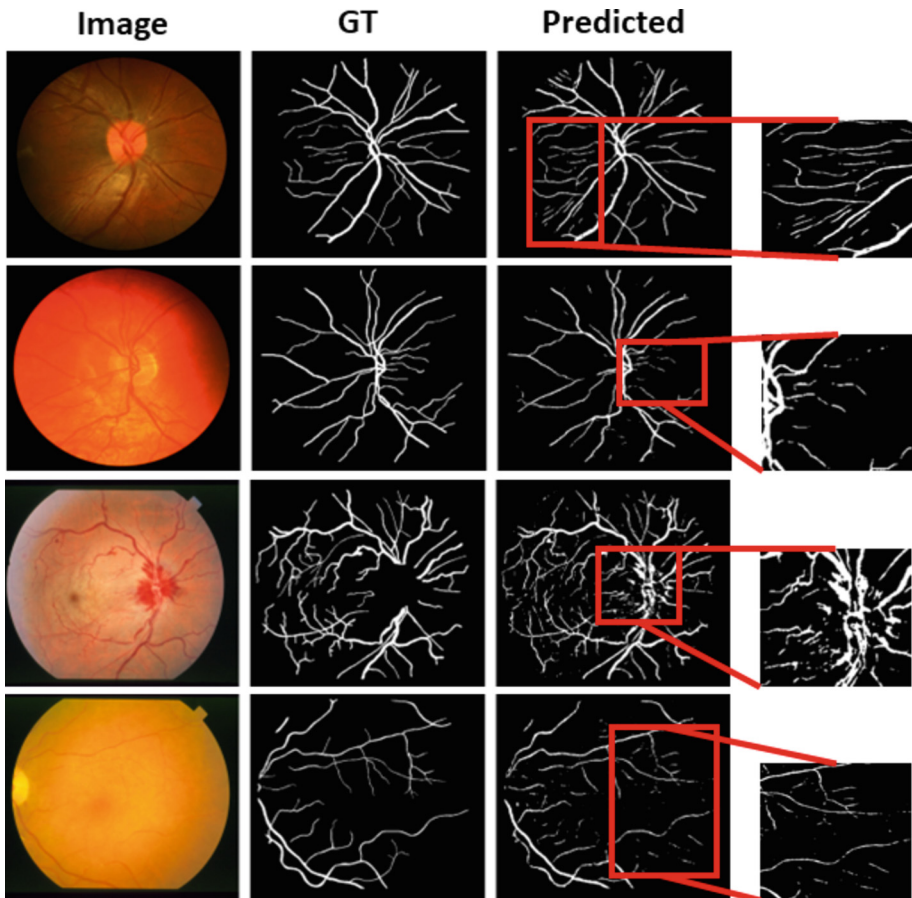


Fig. 8. Some error cases of the proposed model.

5 Conclusions and Future Work

We have introduced an innovative retinal vessel segmentation approach based on the modified attention UNet architecture. By adopting separable convolution layers instead of traditional ones, we achieve a streamlined model design without compromising performance, thus reducing computational demands. The convolution blocks of Attention UNet are modified to capture multiscale spatial information by varying convolution dilation rates. The trainable tanh activation function (T-Tanh) enriches spatial information by detecting feature gradient flow, improving segmentation accuracy by effectively distinguishing foreground and background pixels. Additionally, the Gated Edge Attention (GEA) mechanism enhances edge information extraction, highlights spatial regions of interest, and promotes boundary adherence for precise segmentation mask generation. By reducing false positives and enhancing vessel boundary awareness, GEA significantly improves overall performance. Our model surpasses existing ones on established retinal vessel segmentation benchmarks such as the CHASEDB1 and STARE datasets. As discussed in the experimental section, the proposed model may not work well when the contrast between the background and vessels is too low. To address this challenge, we plan to introduce a feedback mechanism to fine-tune the attention network and gated edge convolution network, which will be discussed in future work.

References

1. Mansour, R.F.: Evolutionary computing enriched computer-aided diagnosis system for diabetic retinopathy: a survey. *IEEE Rev. Biomed. Eng.* **10**, 334–349 (2017)
2. Mendonca, A.M., Campilho, A.: Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. *IEEE Trans. Med. Imaging* **25**(9), 1200–1213 (2006)
3. Chaudhuri, S., Chatterjee, S., Katz, N., Nelson, M., Goldbaum, M.: Detection of blood vessels in retinal images using two-dimensional matched filters. *IEEE Trans. Med. Imaging* **8**(3), 263–269 (1989)
4. Zana, F., Klein, J.-C.: Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation. *IEEE Trans. Image Process.* **10**(7), 1010–1019 (2001)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, vol. 25 (2012)
6. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) *MICCAI 2015. LNCS*, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
7. Oktay, O., et al.: Attention U-net: learning where to look for the pancreas, arXiv preprint [arXiv:1804.03999](https://arxiv.org/abs/1804.03999) (2018)
8. Li, D., Dharmawan, D.A., Ng, B.P., Rahardja, S.: Residual U-net for retinal vessel segmentation. In: *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1425–1429. IEEE (2019)
9. Baheti, B., Innani, S., Gajre, S., Talbar, S.: Semantic scene segmentation in unstructured environment with modified deeplabv3+. *Pattern Recogn. Lett.* **138**, 223–229 (2020)
10. Biswas, A., Banik, R.: Advancements in fundus image analysis: a comprehensive method of AI-based classification and segmentation technique. *Artif. Intell. Appl.* 1–11 (2023)

11. Anwar, R.S.S.: EfficientNet algorithm for classification of different type of cancers. *Artif. Intell. Appl.* 1–10 (2023)
12. Li, L., Verma, M., Nakashima, Y., Nagahara, H., Kawasaki, R.: Internet: retinal image segmentation utilizing structural redundancy in vessel networks. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3656–3665 (2020)
13. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **39**(6), 1856–1867 (2019)
14. Gu, Z., et al.: Ce-net: context encoder network for 2D medical image segmentation. *IEEE Trans. Med. Imaging* **38**(10), 2281–2292 (2019)
15. Wei, J., et al.: Genetic u-net: automatically designed deep networks for retinal vessel segmentation using a genetic algorithm. *IEEE Trans. Med. Imaging* **41**(2), 292–307 (2021)
16. Li, Y., Zhang, Y., Cui, W., Lei, B., Kuang, X., Zhang, T.: Dual encoder-based dynamic-channel graph convolutional network with edge enhancement for retinal vessel segmentation. *IEEE Trans. Med. Imaging* **41**(8), 1975–1989 (2022)
17. Atli, I., Gedik, O.S.: Sine-net: A fully convolutional deep learning architecture for retinal blood vessel segmentation. *Eng. Sci. Technol. Int. J.* **24**(2), 271–283 (2021)
18. Shi, T., Boutry, N., Xu, Y., Géraud, T.: Local intensity order transformation for robust curvilinear object segmentation. *IEEE Trans. Image Process.* **31**, 2557–2569 (2022)
19. Ryu, J., Rehman, M.U., Nizami, I.F., Chong, K.T.: Segr-net: a deep learning framework with multi-scale feature fusion for robust retinal vessel segmentation. *Comput. Biol. Med.* **163**, 107132 (2023)
20. Sun, Z., Xie, Q., Meng, D.: Frs-nets: fourier parameterized rotation and scale equivariant networks for retinal vessel segmentation, arXiv preprint [arXiv:2309.15638](https://arxiv.org/abs/2309.15638) (2023)
21. Oktay, O., et al.: Attention unet: learning where to look for the pancreas. arxiv 2018, arXiv preprint [arXiv:1804.03999](https://arxiv.org/abs/1804.03999) (2018)
22. Soares, J.V., Leandro, J.J., Cesar, R.M., Jelinek, H.F., Cree, M.J.: Retinal vessel segmentation using the 2-D gabor wavelet and supervised classification. *IEEE Trans. Med. Imaging* **25**(9), 1214–1222 (2006)
23. Fraz, M.M., et al.: An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Trans. Biomed. Eng.* **59**(9), 2538–2548 (2012)



TractoEmbed: Modular Multi-level Embedding Framework for White Matter Tract Segmentation

Anoushkrit Goel¹(✉), Bipanjit Singh¹, Ankita Joshi¹, Ranjeet Ranjan Jha²,
Chirag Ahuja³, Aditya Nigam¹, and Arnav Bhavsar¹

¹ Indian Institute of Technology (IIT) Mandi, Mandi, India
s22042@students.iitmandi.ac.in

² Indian Institute of Technology (IIT) Patna, Patna, India

³ Post-Graduate Institute of Medical Education and Research (PGIMER),
Chandigarh, India

Abstract. White matter tract segmentation is crucial for studying brain structural connectivity and neurosurgical planning. However, segmentation remains challenging due to issues like class imbalance between major and minor tracts, structural similarity, subject variability, symmetric streamlines between hemispheres etc. To address these challenges, we propose TractoEmbed, a modular multi-level embedding framework, that encodes localized representations through learning tasks in respective encoders. In this paper, TractoEmbed introduces a novel hierarchical streamline data representation that captures maximum spatial information at each level i.e. individual streamlines, clusters, and patches. Experiments show that TractoEmbed outperforms state-of-the-art methods in white matter tract segmentation across different datasets, and spanning various age groups. The modular framework directly allows the integration of additional embeddings in future works.

Keywords: Tract Segmentation · PointCloud · 3D Computer Vision · Tractography · Diffusion MRI

1 Introduction

Diffusion MRI (dMRI) [1, 2] facilitates the non-invasive examination of the brain's white matter (WM) microstructural organization. A crucial component of the dMRI analysis pipeline is fiber tractography [3, 22, 23], which tracks fibers or streamlines under anatomical constraints from the dMRI signal received from the scanner (refer to Sect. 3). Tract Segmentation involves dividing the streamlines into distinct, anatomically meaningful tracts, with each tract corresponding to a specific white matter pathway. These tracts can be broadly grouped into 3

Supported by SERB (Science and Engineering Research Board) of India.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78104-9_17.

types based on structural connectivity, i.e. Association, Commissural, and Projection Fibers. Each type is further subdivided based on its specific structural connectivity and function, allowing for more granular distinctions. Through the segmentation process, it becomes possible to conduct quantitative studies of white matter (WM), which is important in understanding neurological disorders such as Alzheimer’s, and Parkinson’s [16], the effect of tumors on segmenting fiber streamlines, etc.

In addition, tract segmentation is also crucial for preoperative neurosurgical planning [14], as it helps identify eloquent white matter areas and determine optimal surgical approaches that minimize post-operative damage. Tract segmentation is also extensively used to visualize particular segments for focused examination by clinicians. However, this process is typically performed by expert Neuroanatomists using their knowledge of brain anatomy to divide fibers into multiple bundles. As a result, it is very time-consuming and can vary between experts, affecting the consistency and reliability of the results.

Taking challenges associated with manual tract segmentation, various techniques have been developed over the years. These techniques range from classical methods to ATLAS-based and distance-based algorithms [8, 9, 21, 25] (refer to Sect. 2). An ATLAS refers to a standardized reference that allows spatial mapping of neuroimaging data from different studies (refer to Table 1) and modalities. They approximate the shape, location, and brain region boundaries in a common coordinate space, facilitating the comparison of brain structure and function across individuals.

These methods require significant manual intervention and are prone to age-related brain changes, also their effectiveness depends on the alignment and quality of the ATLAS. Considering the limitations of manual and classical methods, as well as the importance of tract segmentation, machine learning, and deep learning-based frameworks have been proposed for automatic tract segmentation [4, 28, 33]. Deep learning algorithms can learn information from shape, structure, relative location, fiber orientations, etc.

However, a notable drawback is that these models often fail in classifying streamlines that are linear in shape due to over-reliance on shape, such as striato-thalamo-pallido projection fibers, which in existing methods, require global reference along with streamlines [7]. Additionally, when neurosurgeons are concerned with segmenting only a specific set of streamlines, global tractography can become a computational overhead. Due to these complexities in streamline classification-based tract segmentation, each method inherently has a certain drawback.

To address this, we propose TractoEmbed, a modular framework that combines multi-level embeddings extracted from hierarchical data representations specifically at streamline, patch, and cluster levels (refer to Fig. 1). Our approach surpasses state-of-the-art (SOTA) results in tract segmentation. In this work, we present an approach with the following major contributions:

1. We introduce **TractoEmbed**, a **novel modular multi-embedding framework**, which leverages learning task-specific encoders to embed data representations, and generate embeddings. Where the encoders and their hyper-parameters are selected after rigorous experimentation.

2. We propose **novel hierarchical and descriptive streamline data representations**. These representations includes spatial information about regional patches, neighboring streamlines and the streamline itself, providing a comprehensive understanding of the streamline characteristics. In contrast to recent advances, our method leverages minimal neighbouring streamlines, hyperlocal streamlines, enhancing robustness to practical clinical settings
3. It is demonstrated that TractoEmbed framework, **generalizes across various datasets** encompassing different age groups (refer Table 1). Additionally, the framework is **modular** at the embedding level, allowing researchers to **integrate their own learnable embeddings** to achieve even richer representations of streamline data. This modularity enhances the flexibility and adaptability of the framework.

2 Related Work

In recent years, a plethora of classical and deep learning methods have been developed for tract segmentation, capable of performing in diverse conditions and data formats with minimal supervision from the skilled medical practitioners. Among these methods, clustering and distance-based methods, QuickBundles [9] and RecoBundles [8] are fast algorithms that utilize clustering approaches. QuickBundles is known for its speed in grouping streamlines based on their similarity, while RecoBundles excels in identifying parent anatomical bundles of streamlines. RecoBundles achieves this by recognizing and clustering similar streamlines based on their shape and spatial location, meanwhile leveraging a model of known white matter anatomy for accurate segmentation. Additionally, the Fast Streamline Search (FSS) [21] is a highly accurate distance-based search method. FSS indexes streamlines in a spatial data structure, enabling efficient retrieval of similar streamlines in tractography data.

Other notable methods include GeoLab [25], a tract segmentation framework for analyzing the geometry, topology, and structural connectivity of white matter fiber bundles. Classifyber [4] is a linear classifier that uses distance-based embeddings with local and global streamlines and regions of interest (ROIs) in the brain, concatenated into a weight vector, which serves as a hybrid of distance-based and learning-based algorithms. TractSeg [28], one of the seminal works, uses a 2D U-Net model that directly works on fODF peaks [23] to segment tracts, without the need for parcellation and registration. In DeepWMA [33], shape information of a single streamline is used to feed a FiberMap to a simple CNN model, preserving local information. BrainSegNet [10] employs bi-directional LSTMs, while FS2Net [11] uses an LSTM-based model to develop a rotation-invariant segmentation model. TRAFIC [12] uses geometry and 265 landmarks to accurately label, classify, and clean the traced paths of streamlines in streamline space.

Xue *et al.* use the PointNet model to classify streamlines using a local-global data representation, and Wang *et al.* [27] utilize a transformer encoder for fiber segmentation by incorporating features related to fiber shape and position. In [13], a graph convolution (GCNN)-based framework, Spectral GCNN extracts

geometry-invariant features. In FIESTA, [5, 34, 35] Dumais *et al.* segment tracts in latent space, via an autoencoder-based segmentation algorithm.

3 Diffusion MRI Data

In this section, we discuss how diffusion MRI data (refer to Table 1) is acquired and processed to generate input and labels for the proposed framework. Additionally, we explain how the data is divided for training and testing purposes and converted to variations of Point Cloud Data before feeding to encoders.

Table 1. Description of the publicly available dMRI Datasets containing a total of 1 million streamlines with (15,3) dimension each i.e. (1000000, 15, 3) using UKF Tractography and Parcellation (refer Sect. 3.1. [30])

Neuroimaging Datasets	N subs	b s/mm^2	N Volumes (mm^3)	TE/TR (ms)	Resolution (mm^3)
dHCP [6] developing Human Connectome Project	20	0	20 vol.	90/3800	$1.5 \times 1.5 \times 1.5 \text{ mm}^3$
		400	64 vol.		
		1000	88 vol.		
		2600	128 vol.		
ABCD [26] Adolescent Brain Cognitive Development	25	0	1 vol.	88/4100	$1.7 \times 1.7 \times 1.7$
		3000	60 vol.		
HCP [24] Human Connectome Project	25	0	18 vol.	89/5520	$1.25 \times 1.25 \times 1.25$
		3000	90 vol.		
PPMI [16] Parkinson’s Progression Markers Initiative	25	0	1 volume	88/7600	$2 \times 2 \times 2$
		1000	64 vol.		
BTP [33] Brigham’s Tumor Patient Data	25	0	1 volume	98/12700	$2.2 \times 2.2 \times 2.3$
		2000	30 vol.		

3.1 Data Preparation

Diffusion MRI data [2] (refer to Table 1) is acquired by applying magnetic diffusion gradients and measuring the resulting signal attenuation, which depends on the local tissue microstructure. This diffusion MRI is preprocessed using standardized algorithms [22, 23], followed by streamlines tracking using a tractography algorithm [3, 15]. ATLAS based labelling of streamlines is performed in the parcellation process. ATLAS registration on different brains can be inconsistent, non-scalable, knowledge intensive, dataset-specific and time-consuming because they are created by expert neuroanatomists. Hence there is a need for algorithms to automate tract segmentation.

For **Tractography**, we utilize the Unscented Kalman Filter (UKF) [15], which estimates microstructural parameters and fiber orientations from diffusion MRI data to track neuronal paths from multiple seed points. After tractography, the extracted streamlines are bundled into parcels or clusters, where these parcels are mapped to anatomically meaningful tracts using ATLAS, as discussed below.

Parcellation refers to the division of the brain into anatomical regions based on ATLAS and clustering techniques into parcels. Initially, division of hemisphere in streamline space, registration on ATLAS, and transformations are performed to align current brain with the ATLAS.

Through this process, we obtain 800 parcels that are appended to anatomical tracts and labeled along with quality control. These parcels are further refined using the ATLAS and diffusion measurements to separate outlier streamlines from each parcel, dividing each parcel into 2, resulting in 800 outlier parcels and 800 plausible parcels.

Using ATLAS, we club and label all 800 Outlier parcels to “Other” label, and 800 plausible parcels to 42 anatomical tracts. This entire procedure can be executed using *whitematteranalysis* package [17, 18, 32], which follows these steps from tractography streamlines to parcellation. This method ensures consistency across subjects and datasets. ATLAS used in this paper was derived from mean of *100 registered tractography of young healthy adults in the Human Connectome Project (HCP)* [24].

3.2 Training and Testing Data

After sequentially performing fiber tractography, parcellation, and labeling of each parcel, we obtain a total of 1 million labeled streamlines from 100 out of 120 subjects (refer to Table 1), where each subject contains 10,000 streamlines. And 20 subjects out of 120 subjects are kept aside for real world test cases, and not included in data splits.

From a total corpus of 100 subjects, we obtain an array consisting of 1M streamlines of shape **(1000000, 15, 3)** (refer Table 1), where (15,3) streamline array is derived from feature data in RAS (Right, Anterior, Superior) coordinate space (refer Supplementary Material). This dataset is subsequently partitioned into **train, validation, and test** sets in a ratio of 70 subjects for training, 10 subjects for validation, and 20 subjects for testing. Data is split subject-wise, where a subject will only belong to one data split at a time [30].

The dataset encompasses 43 tract classes: 42 anatomically significant tracts spanning the entire brain, and one category labeled as “other”, which includes anatomically implausible outlier streamlines identified during the parcellation process (refer Sect. 3.1). Here **PCD** is an acronym for Point Cloud Data.

3.3 Model Input Data

Training and Testing Data (refer Sect. 3.2), is in the form of a three-dimensional array that contains (number of streamlines, points per streamline, number of features) and is in unusable form for most encoders. Hence, to make it suitable

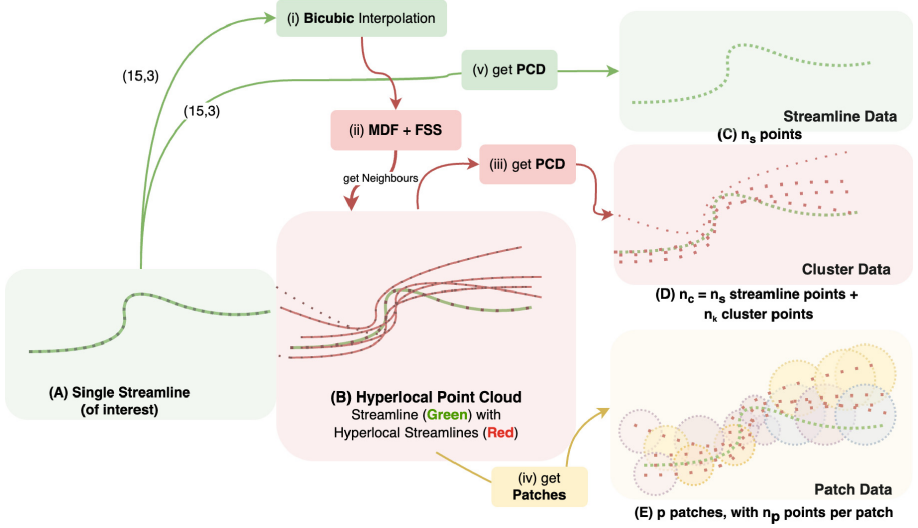


Fig. 1. Data Representations: Streamline, Patch, and Cluster. For (C) Streamline Data, (A) input streamline of shape (15,3), is (v) converted to a point cloud. For (B) Hyperlocal Streamlines, (refer Sect. 3.3) input streamline undergoes (i) bicubic interpolation to make a streamline of shape (40,3), on which k_{local} neighboring streamlines are sampled using (ii) MDF Distance. In k_{local} search space, FSS [21], (ii) Fast Streamline Search is used to get 5 (B) hyperlocal streamlines. (iii) *get PCD* converts hyperlocal streamlines to (D) Cluster Data with $(n_c, 3)$ points. For (E) Patch Data, (iv) p farthest points are sampled using FPS (refer Sect. 4.3), n_p points in each patch using kNN to find neighboring points.

for encoder-specific preprocessing methods, we represent data in 3 forms (as mentioned in Fig. 1) which would be utilised by encoders in Sects. 4.1, 4.2, 4.3 and in results Sect. 5.

1. **Streamline Data:** As mentioned in Fig. 1, Streamline Data is a (15, 3) array, which is created by undersampling actual streamlines of different lengths, containing RAS coordinates as features (also refer Sect. 4.1).
2. **Local Point Cloud (Local PCD):** Streamlines are interpolated using bicubic interpolation to approximate a smooth curve of the streamline, experimentally tested to be a (40,3) streamline. On the interpolated streamline, we use **MDF** (Mean Direct Flip) [9] distance to find k_{local} neighboring local streamlines. These k_{local} streamlines are then converted to **Local Point Cloud** (Local PCD) by merging and randomly sampling n_c (*number of points in a cluster*) points from $((k_{local} + 1) * 40, 3)$. Here interpolated streamlines give dense point clouds, giving richer representation.
3. **Hyperlocal Point Cloud (Hyperlocal PCD):** In the limited search space of k_{local} Local Streamlines, we employ **FSS** (Fast Streamline Search) [21] with a radius of 4mm-6mm to find 5 closest streamlines to the streamline of interest. This group of 5 hyperlocal streamlines is then converted to make a Hyperlocal

Point Cloud by merging and randomly sampling n_c points from total points of dimensions (240,3), from $((5 + 1) * 40, 3)$ which is (240,3), where 5 is no. of hyperlocal streamlines and 1 is the streamline itself.

Hyperlocal Point Cloud is a variation of Local Point Cloud which contains fewer spatially similar streamlines wherein local streamlines in Local PCD can range to higher numbers also, and may contain dissimilar streamlines with neighboring spatial information rather than structural shape information. Models we have used as respective encoders can utilize certain forms of data. **Streamline Encoder** can only process Streamline Data. **Patch Encoder** can process Hyperlocal and Local PCD. **Cluster Encoder** can process all forms of data mentioned above (refer Fig. 1).

4 Methodology

TractoEmbed utilizes a modular framework to fuse learnable embeddings trained on hierarchical streamline data representations, as detailed in the following Subjects. 4.1, 4.2, and 4.3.

In the Methodology section we describe pre-processing, training method, model architecture, and output embedding for each encoder.

1. **Streamline Encoder** essentially is any model that preserves intra-streamline information, its order of points, shape, and geometry amidst the random shuffling of data points in other models. To preserve intra-streamline spatial information we chose CNN-based method due to CNN's inherent capability of learning local and global features from a 2D array with channels. One can argue LSTM encoder for auto-regressive sequential information but LSTM struggles to encode spatial information (refer Sect. 5)
2. **Cluster Encoder**, should encode the shape, inter-streamline dependencies, and information of a cluster to resemble the target tract. Based on our evaluation PointNet is imperative in understanding spatial features from a cluster of points or point cloud. We did mild variations in kernel sizes and layers. We found the simple PointNet [19] model's ability to discern intricate patterns and dependencies better than others.
3. Objective of **Patch Encoder** is to learn regional information in a hyperlocal streamline point cloud, to embed origin and termination region information in the point cloud through regional patches. We chose a combination of minipointnet and Discrete Variational Autoencoder (dVAE) [20] to reconstruct point cloud patches and learn regional generative features. Patches are used to embed regional information as attention across only points fails due to minimal information in a single point and high compute requirements [31].

Broadly, three types of encoders are pre-trained or finetuned for classification downstream tasks. Embeddings from these encoders are combined to assist the classifier MLP (as illustrated in Fig. 2) in achieving accurate classification.

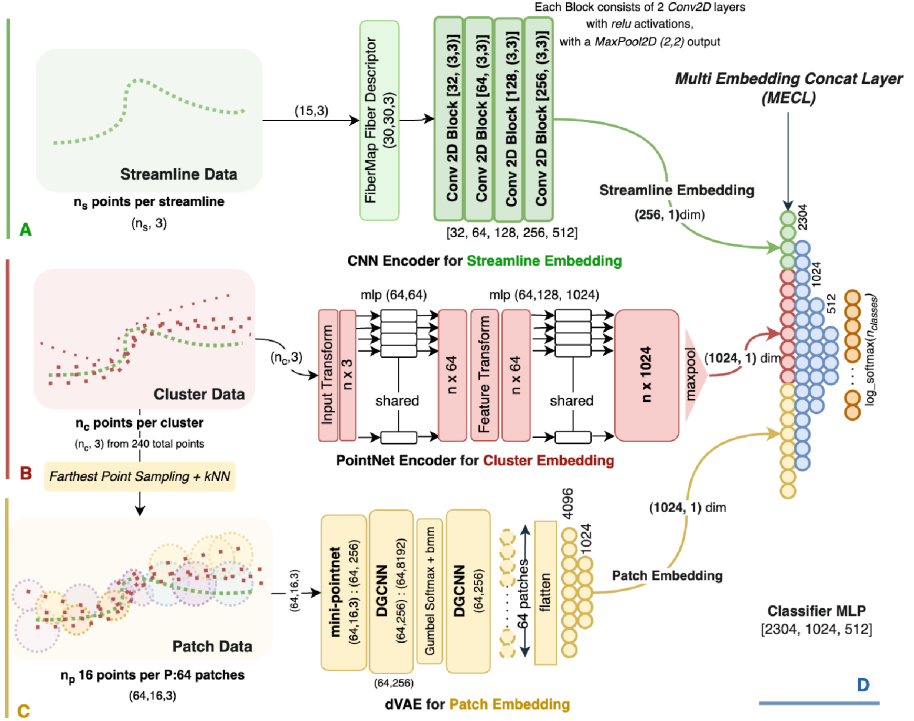


Fig. 2. Streamline, Patch, and Cluster data obtained through the processes illustrated in Fig. 1, are sent to respective encoders to generate embeddings. **(A) Streamline Data** of dimensions $(n_s, 3)$ serves as input to the Fiber Descriptor [33], producing an output of dimensions $(2 * n_s, 2 * n_s, 3)$, where n_s is number of points per streamline, which is fed to 4 CNN blocks, refer Table 2, to obtain a final embedding of dimensions $(256, 1)$ for the *MECL* (*Multi Embedding Concat Layer*). **(B) Cluster Data** $(n_c, 3)$ from either local or hyperlocal point cloud, refer to Sect. 3.3, is fed to the PointNet Encoder to give cluster embedding of 1024 dimensions. Patches on Cluster Data are created using Farthest Point Sampling to fetch 64 patches with 16 points each, resulting in $(64, 16, 3)$ dimensions. **(C) Patch data** is fed to a mini PointNet, which produces a $(64, 256)$ output, further input to dVAE, resulting in 64 patches each of dimension 256. This is flattened to be fed to $[4096, 1024]$ dense layers to give an output patch embedding of 1024 dimensions. **(D)** These multiple embeddings are concatenated at *MECL* to make $(256 + 1024 + 1024 = 2304)$ dimensional embedding. This is input to Classifier MLP resulting in a 512 dim classification embedding.

4.1 Streamline Encoder

Preprocessing: The streamline $(15, 3)$ is passed through the Fiber Descriptor to get streamline representation for streamline encoder. Fiber Descriptor is a streamline representation technique where concatenation of normal and flipped streamlines are stacked, so that a CNN kernel can learn local intra-streamline features.

In Fiber descriptor, the input streamline is flipped and concatenated horizontally with the original (15,3) input, creating a (30,3) row. In the second row, the flipped streamline is followed by the original, forming a 12–21 pattern (where 1 represents the original and 2 represents the flipped). This 12–21 pattern (representation of 2 rows separated by a dash '-') is vertically repeated 15 times, resulting in (30,30,3) representation from (15,3). This is input to our streamline encoder model (for more details refer [33]).

Streamline Embedding is the 256-dimensional output generated by the streamline encoder model (refer to Table 2). This CNN model is independently trained on the Fiber Descriptor representation of streamlines using cross-entropy loss for classification. The model is designed to learn streamline-specific discriminative spatial features, (as discussed [33]). We extract the streamline embedding from the output of **MaxPool2D** layer (refer Fig. 2 and Table 2).

Table 2. Streamline Encoder: Stacked CNNs with input data size (30,30,3), which we have compressed to represent as 4 Conv Blocks

Layer (type)	Output Shape	Kernel	Configuration
Conv Block 1	(None, 30, 60, 32)	(3, 3)	[Conv2D, Activation] x2
Max Pooling 2D	(None, 15, 30, 32)	(2, 2)	-
Conv Block 2	(None, 15, 30, 64)	(3, 3)	[Conv2D, Activation] x2
Max Pooling 2D	(None, 7, 15, 64)	(2, 2)	-
Dropout	(None, 7, 15, 64)	-	-
Conv Block 3	(None, 7, 15, 128)	(3, 3)	[Conv2D, Activation] x2
Max Pooling 2D	(None, 3, 7, 128)	(2, 2)	-
Conv Block 4	(None, 3, 7, 256)	(3, 3)	[Conv2D, Activation] x2
Max Pooling 2D	(None, 1, 3, 256)	(2, 2)	<i>Streamline Embedding</i>

4.2 Cluster Encoder

Preprocessing: From interpolated (40,3) streamlines we sample k_{local} local neighbour streamlines using MDF Distance from QuickBundles. After finding local streamlines for each streamline, we sample hyperlocal streamlines using **FSS** [21]. FSS uses barycenter of streamlines and distance parameters like radius to accurately find similar streamlines. Streamlines sampled from FSS are highly probable to belong to the same class which allow us to merge these streamlines creating hyperlocal streamline data (6,40,3) where $k_{hyperlocal} = 5$ and 1 streamline, resulting in (240,3) point cloud, from which n_c points are sampled for the input to PointNet Model (refer Table 3) for the Cluster Embedding.

Cluster Embedding is a 1024 dimensional output embedding of a PointNet Model (refer Table 3 [19]), which takes in Cluster data (refer Sect. 4.2) as input ($n_c, 3$), where n_c is the number of points in the total point cloud.

Table 3. Cluster Encoder: PointNet Model Architecture for Cluster Embedding, extracted at the last layer in this table.

Model	Module	C_{in}	C_{out}	Kernel Size
PointNet Encoder	Conv1D and BN	3	64	1
	STNkD	64	64	-
	Conv1D and BN	64	128	1
	Conv1D and BN	128	256	1
	Conv1D and BN	256	1024	1

Cluster Encoder architecture is mentioned in Table 3, all values are Xavier initialized, and the Cluster Embedding (1024 dimensional) is extracted at final MaxPool 1D. Cluster Encoder is trained along with Classifier MLP of TractoEmbed (discussed in Sect. 4.4). Effects on Classification Accuracy are observed by varying, n_c in Ablation Study Table 6

4.3 Patch Encoder

Patch Data refers to a 3D patch on the Cluster data. We use patch data to capture information across and among regional patches on a group of streamlines (Cluster). Enabling us to classify streamlines that are structurally linear in shape, and are very similar to other streamlines (like projection fibers). Regional Patches will learn to focus on Origin and Termination ROIs to help better segment difficult tracts.

Preprocessing: Patches are created by iteratively sampling, p_f farthest points using FPS, Farthest Point Sampling, over the point cloud. we then use kNN to sample p_{local} nearest points per patch. We get $p_f = 64$ patches, where every patch has $p_{local} = 16$ points making patch data dimensions to be (64, 16, 3) from an input cluster data of ($n_c, 3$), randomly sampled with replacement. (refer Ablation Study Table 6 to see effects on variation in n_c)

Patch Embedding is a 1024 dimensional output of dVAE decoder (refer: dVAE architecture used Table 4 and Subsect. 4.3) This dVAE model contains an encoder and a decoder trained to reconstruct patch data [31] using Chamfer and KL Divergence loss. The objective of the encoder is to create an 8192-dimensional token embedding for each patch, then the decoder scales down each token embedding to a 256-dimensional token embedding passing to the MLP to reconstruct the input patches. (For more details refer to Supplementary Material) We extract Patch Embedding at 1024 dimensional linear layer, and concat with other embeddings at *MECL*.

4.4 Training Strategy

TractoEmbed concatenates all three embeddings-Streamline, Cluster, and Patch- at the *MECL* (*Multi-Embedding Concat Layer*) to feed the classifier MLP.

Table 4. Patch Encoder: dVAE model architecture details [31]. The last layer from which Patch Embedding is extracted. Where C_{in} represents dimensions of input features, and C_{out} , dimensions of output features. N_{out} is the number of points in the query point cloud. K is the number of neighbors in kNN operation. C_{middle} is the dimension of the hidden layers for MLPs.

Model	Module	Block	C_{in}	C_{out}	K	N_{out}	C_{middle}
dVAE	Encoder	Linear	256	128	-	-	-
		DGCNN	128	256	4	64	-
		DGCNN	256	512	4	64	-
		DGCNN	512	512	4	64	-
		DGCNN	512	1024	4	64	-
		Linear	2304	8192	-	-	-
	Decoder	Linear	256	128	-	-	-
		DGCNN	128	256	4	64	-
		DGCNN	256	512	4	64	-
		DGCNN	512	512	4	64	1024
		DGCNN	512	1024	4	64	1024
		Linear	2304	256	-	-	-

During the training process, the Streamline and Patch Encoders, pre-trained on tasks mentioned above (refer to Sect. 4), are kept non-trainable, while the Cluster Encoder and Classifier MLP are trainable. This is trained for 40 epochs at an initial learning rate of 0.0001 with a Cosine Annealing Warm Restarts learning rate scheduler using *Focal Loss* to address class imbalance in major and minor tracts. Experimentally, we observed that concatenating the embeddings outperforms adding or merging them and focal loss performs better than cross entropy loss with these many classes. TractoEmbed extracts these multi-level embeddings to holistically represent streamlines and regional anatomy.

5 Results and Discussions

In this section, we present extensive ablation studies, and comparative results highlighting the effectiveness of our embeddings and data representations across different datasets. We evaluated all the results on the test split containing 20 subjects from a sample of 100 subjects. Classification Report with Accuracy and F1 scores for each class is described in the *Supplementary Material*.

We present a comparison with several models, including DeepWMA [33], DCNN++ [29], basic PointNet [19], and DGCNN, using Single Streamline data. In Local PCD, TractoEmbed outperforms both variations of TractCloud [30]. In Hyperlocal PCD, we see that with only similar streamlines in the point cloud, TractoEmbed performs better than its performance in Local PCD, due to extensive focus on learning shape information of streamlines, as shown in Table 6.

Table 5. Comparative Results: Model and Architecture performance across different Data Representations with a comparison of results with other state-of-the-art methods. Results for methods are sourced from [30], to eliminate discrepancies due to differences in training methods, except for *Hyperlocal PCD*. For TractCloud, hyperlocal PCD is made by setting k_{local} : 5 and k_{global} : 0. All the experiments are done with keeping with their prescribed configurations static and only changing the input data.

Data	Model: Type	Acc (%)	F1 (%)
Single Streamline	DeepWMA (CNN)	90.29	88.12
	DCNN++ (CNN)	91.26	89.14
	PointNet (PCD)	91.36	89.12
	DGCNN (Graph)	91.85	89.78
Local PCD (k = 20)	TractCloud: PointNet	91.51	89.25
	TractCloud: DGCNN (Graph)	91.91	90.03
	TractoEmbed (ours)	92.09	90.07
Hyperlocal PCD (k = 5)	TractCloud (PointNet)	91.12	88.66
	TractoEmbed (ours)	93.04	91.38
Local + Global Representation	TractCloud: PointNet	92.28	90.36
	TractCloud: DGCNN (Graph)	91.99	90.10

The comparative results in Table 5 highlight the efficacy and superior accuracy of our TractoEmbed framework for streamline classification and tract segmentation. Where TractCloud [30] relies on local and global streamlines using a PointNet model to achieve registration-free tract segmentation TractoEmbed performs better even with spatially sparse hyperlocal PCD. The ablation study presented in Table 6 reveals the effectiveness of combining multiple embeddings used by TractoEmbed. As the neighboring point cloud becomes sparser, the representations need to be denser, increasing the need for more embeddings.

Table 6. Ablation study across a combination of embeddings with varying input point cloud densities to study their effect on Model Performance and finding the optimal hyperparameters (also refer Fig. 1). Here, n_c number of points are randomly sampled from the total available points to make cluster data.

Multi Embeddings	Metric (%)	Hyperlocal PCD (k=5)			Local PCD (k=20)	
		$n_c = 190$ points	$n_c = 220$ points	$n_c = 240$ points	$n_c = 190$ points	$n_c = 240$ points
cluster + streamline	Acc	92.917	93.038	93.020	91.494	91.383
	F1	91.198	91.381	91.346	89.338	89.239
cluster + patch	Acc	92.078	91.90	90.94	81.502	80.451
	F1	89.891	89.574	88.489	74.765	72.799
streamline + patch	Acc	91.654	91.065	89.675	91.165	90.956
	F1	89.525	88.97	87.331	88.991	88.781
cluster + patch + streamline	Acc	92.946	92.837	92.876	91.431	91.409
	F1	91.284	91.091	91.164	89.239	89.36

Conversely, when the neighboring point cloud has a higher density of points, a pair of embeddings, cluster and streamline embedding, can achieve satisfactory performance. These findings ascertain the importance of incorporating dense streamline data representations from various perspectives/levels, including self, region, and neighbors.

Table 7. Ablation Study showcasing the performance of individual encoders on different data representations for the tract segmentation through streamline classification

Encoders only	TractoEmbed Models	(%)	Data		
			Streamline	Cluster	Patch
Streamline Encoder	CNN	Acc	91.024	-	-
		F1	88.894	-	-
Cluster Encoder	PointNet	Acc	91.36	90.54	-
		F1	89.12	88.31	-
Patch Encoder	dVAE	Acc	-	-	85.87
		F1	-	-	82.48

The efficacy of a combination of embeddings can further be proven vital in increasing streamline classification accuracy as individual encoders perform poorly when compared to a combination of these embeddings (see Tables 6 and 7).

Diving even further, there are slight improvements in F1 scores for projection fibers, striato-thalamo-pallido bundles, as observed in the Classification Report (refer to *Supplementary Material*), indicating that Patch Embedding can effectively make information-dense patches of an input point cloud. Having multiple embeddings decreases the over-reliance on one knowledge representation, and makes TractoEmbed robust to changes in either of the representations. Explicit addition of Streamline Embedding containing information on the order of points and intra-streamline spatial information makes TractoEmbed robust to point cloud perturbations in Cluster Encoder.

In summary, TractoEmbed demonstrates effectiveness in hyperlocal point clouds (regional examinations) and time-critical settings where a specific 3D brain segment is considered. It is also effective particularly for classifying structurally similar, minor, and projection fibers, achieving increased F1 scores and improved overall accuracy compared to LSTM-based approaches. TractoEmbed emphasizes the significance of fusing dense representations, incorporating various perspectives, including self, regional patches, and neighboring streamlines, which is crucial for extracting multiple types of information from low-fidelity streamline data. Future research can explore additional encoders, refined embedding combinations, optimal hyper-parameters, and different data representations. Also, there lies scope for improvement in finding unified models that can discriminate among highly similar streamlines or point clouds, with more classes.

6 Conclusion

With **TractoEmbed** we introduce an innovative method for Tract Segmentation, characterized by substantial accuracy, robustness, and modularity improvements. Our method integrates multiple embeddings from task-specific encoders to provide rich representations of streamlines, enabling a reduction in spatial input data requirements. It also demonstrates effectiveness in special cases, classifying structurally similar, minor, and projection fibers, by incorporating various data perspectives and minimal reliance on a single embedding. TractoEmbed also gives researchers the freedom to directly experiment with embeddings and data representations to get even better results. With its spatially minimal data requirements, TractoEmbed can be useful for focused ROI-specific, and time-sensitive clinical settings. *Code will be made available upon request.*

Acknowledgement. This research was supported by SERB Core Research Grant Project No: CRG/2020/005492, IIT Mandi.

References

1. Basser, P.J., Mattiello, J., LeBihan, D.: MR diffusion tensor spectroscopy and imaging. *Biophys. J.* **66**(1), 259–267 (1994)
2. Basser, P.J., et al.: In vivo fiber tractography using DT-MRI data. *Magn. Reson. Med.* **44**(4), 625–632 (2000)
3. Behrens, T.E., et al.: Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? *Neuroimage* **34**(1), 144–155 (2007)
4. Bertò, G., et al.: Classifyber, a robust streamline-based linear classifier for white matter bundle segmentation. *Neuroimage* **224**, 117402 (2021)
5. Dumais, F., et al.: Fiesta: autoencoders for accurate fiber segmentation in tractography. *Neuroimage* **279**, 120288 (2023)
6. Edwards, A.D., et al.: The developing human connectome project neonatal data release. *Front. Neurosci.* **16**, 886772 (2022)
7. Funk, A.T., et al.: In humans, striato-pallido-thalamic projections are largely segregated by their origin in either the striosome-like or matrix-like compartments. *Front. Neurosci.* **17**, 1178473 (2023)
8. Garyfallidis, E., et al.: Recognition of white matter bundles using local and global streamline-based registration and clustering. *Neuroimage* **170**, 283–295 (2018)
9. Garyfallidis, E., et al.: Quickbundles, a method for tractography simplification. *Front. Neurosci.* **6**, 175 (2012)
10. Gupta, T., Patil, S.M., Tailor, M., Thapar, D., Nigam, A.: Brainsegnet: a segmentation network for human brain fiber tractography data into anatomically meaningful clusters. arXiv preprint [arXiv:1710.05158](https://arxiv.org/abs/1710.05158) (2017)
11. Jha, R.R., Patil, S., Nigam, A., Bhavsar, A.: FS2Net: fiber structural similarity network (FS2Net) for rotation invariant brain tractography segmentation using stacked LSTM based siamese network. In: Vento, M., Percannella, G. (eds.) CAIP 2019. LNCS, vol. 11679, pp. 459–469. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29891-3_40
12. Lam, P.D.N., Belhomme, G., Ferrall, J., Patterson, B., Styner, M., Prieto, J.C.: Traffic: fiber tract classification using deep learning. In: Medical Imaging 2018: Image Processing, vol. 10574, pp. 257–265. SPIE (2018)

13. Liu, F., et al.: DeepBundle: fiber bundle parcellation with graph convolution neural networks. In: Zhang, D., Zhou, L., Jie, B., Liu, M. (eds.) GLMI 2019. LNCS, vol. 11849, pp. 88–95. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35817-4_11
14. Lucena, O., et al.: Informative and reliable tract segmentation for preoperative planning. *Front. Radiol.* **2**, 866974 (2022)
15. Malcolm, J.G., Shenton, M.E., Rathi, Y.: Neural tractography using an unscented kalman filter. In: Prince, J.L., Pham, D.L., Myers, K.J. (eds.) IPMI 2009. LNCS, vol. 5636, pp. 126–138. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02498-6_11
16. Marek, K., et al.: The parkinson progression marker initiative (PPMI). *Prog. Neurobiol.* **95**(4), 629–635 (2011)
17. O'Donnell, L.J., Westin, C.F.: Automatic tractography segmentation using a high-dimensional white matter atlas. *IEEE Trans. Med. Imaging* **26**(11), 1562–1575 (2007)
18. O'Donnell, L.J., Wells, W.M., Golby, A.J., Westin, C.-F.: Unbiased groupwise registration of white matter tractography. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7512, pp. 123–130. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33454-2_16
19. Qi, C., et al.: Pointnet: deep learning on point sets for 3D classification and segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660 (2017)
20. Rolfe, J.T.: Discrete variational autoencoders. arXiv preprint [arXiv:1609.02200](https://arxiv.org/abs/1609.02200) (2016)
21. St-Onge, E., et al.: Fast streamline search: an exact technique for diffusion MRI tractography. *Neuroinformatics* **20**(4), 1093–1104 (2022)
22. Tournier, J.D., Calamante, F., Connelly, A.: Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. *Neuroimage* **35**(4), 1459–1472 (2007)
23. Tournier, J.D., et al.: Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *Neuroimage* **23**(3), 1176–1185 (2004)
24. Van Essen, D.C., et al.: The WU-Minn human connectome project: an overview. *Neuroimage* **80**, 62–79 (2013)
25. Vindas, N., et al.: Geolab: geometry-based tractography parcellation of superficial white matter. In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE (2023)
26. Volkow, N.D., et al.: The conception of the ABCD study: from substance use to a broad NIH collaboration. *Dev. Cogn. Neurosci.* **32**, 4–7 (2018)
27. Wang, Z., et al.: Accurate corresponding fiber tract segmentation via FiberGeoMap learner. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13431, pp. 143–152. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16431-6_14
28. Wasserthal, J., Neher, P., Maier-Hein, K.H.: Tractseg-fast and accurate white matter tract segmentation. *Neuroimage* **183**, 239–253 (2018)
29. Xu, H., Dong, M., Lee, M.H., O'Hara, N., Asano, E., Jeong, J.W.: Objective detection of eloquent axonal pathways to minimize postoperative deficits in pediatric epilepsy surgery using diffusion tractography and convolutional neural networks. *IEEE Trans. Med. Imaging* **38**(8), 1910–1922 (2019)

30. Xue, T., et al.: Tractcloud: registration-free tractography parcellation with a novel local-global streamline point cloud representation. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14227, pp. 409–419. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43993-3_40
31. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: Point-bert: pre-training 3D point cloud transformers with masked point modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19313–19322 (2022)
32. Zhang, F., et al.: An anatomically curated fiber clustering white matter atlas for consistent white matter tract parcellation across the lifespan. *Neuroimage* **179**, 429–447 (2018)
33. Zhang, F., et al.: Deep white matter analysis (deepwma): fast and consistent tractography segmentation. *Med. Image Anal.* **65**, 101761 (2020)
34. Legarreta, J.H., et al.: Filtering in tractography using autoencoders (FINTA). *Med. Image Anal.* **72**, 102126 (2021)
35. Legarreta, J.H., Petit, L., Jodoin, P.M., Descoteaux, M.: Clustering in tractography using autoencoders (CINTA). In: International Workshop on Computational Diffusion MRI, pp. 125–136. Cham: Springer Nature Switzerland (2022)



Unsupervised Domain Adaptation for Cross-Device Iris Liveness Detection Model Transfer

Xiuying Wu¹, Chenxi Du^{2,3}, Hui Zhang^{1(✉)}, Jing Liu¹, Dexin Zhang¹,
and Hang Zou⁴

¹ Tianjin University of Science and Technology, Tianjin, China
zhanghui2022@tust.edu.cn

² SIAT, Chinese Academy of Sciences, Shenzhen, China

³ Southern University of Science and Technology, Shenzhen, China

⁴ China Telecom Research Institute (CTRI), Beijing, China

Abstract. Iris recognition is widely employed in unmanned detection environments, but it would suffer from a variety of attacks, such as, printed iris attack, wearing cosmetic contact lenses, glass eyes, prosthetic eyes, and so on. Although many previous methods have been proposed to resolve these problems, the generalization of models under cross-domain and cross-device scenarios is still need to be improved. To alleviate it, we propose an unsupervised domain adaptation transfer learning model with high detecting accuracy and generalization which is robust to attacks. The transfer learning is used to attain agile deployment in our model. Our model is mainly based on the CDD (Contrastive Domain Discrepancy) measurement method, which minimizes the intra-class difference and maximizes the inter-class difference, because of its satisfactory performance on liveness detection tasks. The transfer learning method based on MMD (Maximum Mean Discrepancy) measurement is proposed to attain agile deployment which selects the feature space alignment between the target domain and the source domain. Code is available at <https://github.com/Wuxiuying111/Cross-device-iris-liveness-detection.git>.

Keywords: Iris recognition · Liveness detection · Transfer Learning · CDD · MMD

1 Introduction

In recent years, iris recognition has been widely used in mines, prisons, banks, police, and entry and exit control. Iris recognition is one of the most important biometric recognition, it has stability, universality, and uniqueness. In 1993, Reference [1] successfully implemented an automated iris recognition system for the first time. In 1994, reference [2] developed an iris recognition system based on

This work is partially supported by National Natural Science Foundation of China (62076232, 62106015) and Beijing Nova Program (Z211100002121113).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15328, pp. 256–272, 2025.
https://doi.org/10.1007/978-3-031-78104-9_18

the Hough transform algorithm. A zero-crossing detection algorithm based on wavelet transform is proposed in reference [3] and applied to iris recognition. In practical application scenarios, iris recognition devices are mostly used in unattended situations. In this case, the importance of liveness detection is emerging and has attracted increasing attention. Typical fake iris attack methods include printed iris attacks, wearing cosmetic contact lenses, glass eye, ocular prosthesis, and screen replayed iris. Among these presentation attacks, wearing cosmetic contact lenses is the most common appearance attack for an iris recognition system in daily use. First, it may lead to false rejection. More importantly, it registers iris templates to the system may cause false identification when meeting similar cosmetic contact lenses, which is a high risk for the iris recognition system. Therefore, it is sorely needed for the classification of iris images with and without cosmetic contact lenses. The liveness detection methods based on deep learning have achieved significant improvement in accuracy and generalization [4–7]. However, a well-trained liveness detection classifier may completely fail when analyzing images from another iris capture device. Besides, the training data for the liveness detection classifier is a very small subset of the large cosmetic contact lens and end-user volume. When replacing recognition equipment or doing new user group adaptation, the iris-liveness detection module is the most likely to fail. The common operation is to collect a large number of images and retrain the classifier, which has a high cost in time, manual work, and price.

In traditional iris classification research, we often assume that the distribution of the training set and the test set is consistent. We train the model on the training set and test it on the test set. However, in practical scenarios, the test environment is often uncontrollable, leading to significant differences in the distribution between the test set and the training set. This discrepancy can cause overfitting and result in poor model performance on the test set. When the distributions of the training set and the test set are inconsistent—due to factors like changing devices, end users, or the presence of many unobserved contact lens types—agile deployment can be achieved through transfer learning technology. Domain adaptation, a representative method in transfer learning, involves using information-rich samples from the source domain to improve the performance of the target domain model. The source domain represents a different domain from the test sample but contains rich supervisory information, while the target domain represents the domain where the test sample is located, often with few or no labels. The source and target domains typically belong to the same task but have different distributions.

Domain adaptation is suitable for situations where there are multiple variations in images, such as changes in devices, collection targets, or environments. By using a small amount of both synthetic and real iris images collected from actual users with new devices, the existing model can be quickly optimized for agile and low-cost deployment. Specifically, when the iris image acquisition device is replaced or the near-infrared supplementary light source is adjusted, traditional liveness detection algorithms and other classification algorithms generally become completely ineffective, requiring the collection of a large number of

images for classifier retraining. Practical domain adaptation can optimize training with a small amount of new images to obtain a classification model suitable for the adjusted device. The paper conducts experiments and research on domain adaptation using liveness detection as an example, but the same method can also be applied to other classification, detection, and recognition algorithm transfers.

According to the different types of target and source domains, there are four different scenarios for domain adaptation problems: unsupervised, supervised, heterogeneous distribution, and multiple source domains. In this paper, we designed a classification transfer task based on practical applications: labeled source domain data with varying amounts of data, unlabeled target domain data with varying amounts of data, and various real-world scenarios through different combinations of source and target domains. An approach to feature space alignment was devised, leveraging the measures of Maximum Mean Discrepancy (MMD) and Contrastive Domain Discrepancy (CDD), thereby facilitating unsupervised model transfer learning within the target domain, consequently enhancing the efficacy of model transfer significantly.

To sum up, our contributions include:

- A transfer learning method based on MMD metrics is designed, which promotes the alignment of feature Spaces between source and target domains and emphasizes the differences within and between domains.
- A transfer learning method based on CDD measurement is designed. Based on MMD, it can minimize intra-class differences and maximize inter-class differences. Thus, the classification accuracy of the model is greatly improved.

In the rest of the paper, Sect. 2 provides the related work. In Sect. 3, the measurement method and implementation principle of MMD and CDD are introduced in detail. We also provide the proving and computing processes of them. Section 4 provides the detailed information of the dataset samples, the experimental results and the analyses of the results. Finally, the conclusion is drawn in Sect. 5.

2 Related Work

2.1 Iris Liveness Detection

Iris attack modes include printed iris attack, contact lens attack, and human fake eye attack. It is found that in response to printed iris attacks, Daugman [8] proposes to use the spectral characteristics of different positions of the human eye and four kinds of Purkinje reflections in the eye for printed iris deception. Reference [9] proposes that iris tremor or pupil response to sudden external light should be used to detect printed iris. Lee et al. [10] proposes an anti-deception method based on multispectral illumination to detect the difference in reflection characteristics between iris and sclera at different wavelengths. Lee et al. [11] then proposed an anti-deception method combining the reflectance characteristics of multi-spectral wavelength and the thickness characteristics of

corneosclera margin. He et al. [12] proposes an anti-spoofing method based on different positions using different bands of NIR illumination. Hoffman [6] proposes an anti-spoofing method to extract specific features of conjunctival blood vessels and iris texture from multispectral images. [13, 14] literature presents a method to obtain the fused gray samples of multiple images at different wavelengths. Lee et al. [15] presents a method for using near-infrared photodiodes at different positions in the image acquisition process. Park et al. [16] was distinguished by detecting the characteristics of iris tremor. Reference [17–20] applied the method of controlling pupil size fluctuation by external light intensity change to anti-iris deception. In response to contact lens attacks, Puhan et al. [21] proposes a method that combines pairwise pupil size and iris texture. Zhang et al. [22] proposes a weighted LBP iris texture analysis detection method. Reference [23, 24] proposes a detection method for printed photos after wearing contact lenses. A detection method based on in-depth information analysis is proposed in reference [25].

With the develop of iris recognition, some liveness detection algorithms have been proposed, which can be classified into two categories, e.g. algorithms based on the manual design features and the deep convolutional neural network. Early algorithms are based on the manual design features. Daugman [26] propose using the amplitude spectrum of Fourier transforms to distinguish iris images with and without cosmetic contact lens. Zhang et al. [27] develop the weighted LBP for iris lens detection. Lovish et al. [28] demonstrate a method based on Local Phase Quantization and Binary Gabor Patterns for detecting cosmetic lens. Daksha et al. [22] investigate the effects of texture lens on iris recognition by using variants of LBP. Sun et al. [29] propose an iris texture primitives encoding framework called Hierarchical Visual Codebook for iris liviness detection. In recent years, deep convolutional network based methods are gradually occupying the main status. Aimed at the three-class (cosmetic contact lens, soft contact lens and no contact lens) classification problem, Ragvendra et al. [4] propose an architecture based on the ContlensNet which is trained on image patches obtained from the segmented and normalized iris images. To solve the same three-class detection problem, a hierarchical network based on ResNet-50 is introduced in [5] called GHCLNet. It does not use any kind pre-processing and iris segmentation and performed well on most of the iris datasets except for some images that are illuminated to a large extent or highly occluded. Based on a shallow version of VGG net, Hoffman et al. [6] design an iris presentation attack detection method which takes a patch of iris image and the associated segmentation mask together as 2-channels inputs. While showing good cross-dataset generalization capability, this method needs high precision iris segmentation information in advance.

2.2 Domain Adaptation

Under the condition of consistent distribution between testing set and training set, these methods have achieved good results. The primary issue is that the accuracy of classification deteriorates significantly in cross-device scenarios, which refers to the non-homologous classification. In practical applications,

when the device changes, it is imperative to re-collect a substantial number of authentic human and prosthetic iris images for retraining purposes. This process can be time-consuming and costly, making the need for an efficient liveness detection adaptive deployment scheme more pressing. Furthermore, the collection of contact lenses used for training is merely a fraction of the vast space of contact lenses available. When new contact lenses are introduced, the standard practice involves gathering a substantial amount of data and retraining using these novel contacts, which demands considerable manpower and time. Zhang et al. [30] propose the Margin Disparity Discrepancy (MDD) which tailors to the distribution comparison with the asymmetric margin loss, and to the minimax optimization for easier training. Mingsheng Long et al. [31] introduced the architecture of Deep Adaptation Networks (DAN), employing the strategy of optimal multi-kernel selection to further mitigate the domain disparity through mean embedding alignment. Subsequently, the authors introduced the Joint Adaptation Network (JAN) [32], which is founded upon the Joint Maximum Mean Discrepancy (JMMD) criterion and employs adversarial network strategies to maximize the JMMD.

3 Methodology

Unsupervised domain adaptation (UDA) refers to the technique of transferring a model from a source domain to a target domain in the absence of annotated data in the target domain. In practice, the application of a model in a novel domain frequently necessitates a significant volume of annotated data. However, unsupervised domain adaptation methods expedite the adaptation to the new domain by transferring models from the source domain, thereby mitigating the expenses and labor associated with data annotation. As an unsupervised learning technique, the fundamental premise of unsupervised domain adaptation involves constructing a common space between the source and target domains, enabling domain transfer through feature alignment and domain adaptation within the target domain. The primary procedure entails training a shared feature processor on both the source and target domains, aligning the feature distributions of the source and target domains as closely as possible through feature alignment and adapting the model to the data distribution of the target domain through domain adaptation, while evaluating model performance through testing on the target domain.

Given a set of sample data $\mathcal{S} = \{(x_1^s, y_1^s), \dots, (x_{N_s}^s, y_{N_s}^s)\}$ from the source domain and sample data $\mathcal{T} = \{x_1^t, \dots, x_{N_t}^t\}$ from the target domain, where $y^s \in \{0, 1, \dots, M-1\}$ represents the labels of the source data for M classes and $y^t \in \{0, 1, \dots, M-1\}$ represents the labels of the target data for M classes, which are unknown. x^s and x^t denote the input data. Therefore, in the context of unsupervised domain adaptation, labeled source data \mathcal{S} and unlabeled target data \mathcal{T} can be employed to accurately predict $\{\hat{y}^t\}$ within \mathcal{T} . In deep neural networks, samples exhibit hierarchical features represented by activations at each layer $l \in \mathcal{L}$. In this paper, $\phi(\cdot)$ denotes the mapping defined by the deep neural

network from the input to a specific layer. $\phi_l(x)$ represents the output of layer l in deep neural network Φ_θ given input x .

3.1 Maximum Mean Discrepancy

The Maximum Mean Discrepancy (MMD) is a cornerstone among loss functions in the realm of transfer learning, notably in the domain of adaptation, and is predominantly utilized for gauging the divergence between two disparate yet interconnected distributions. Its fundamental premise lies in the integration of an adversarial element within the model, which is then refined through iterative training to enforce convergence between the feature distributions of the source and target domains. Through this mechanism, the transition from the source domain to the target domain can be facilitated.

The maximum mean discrepancy (MMD) serves as a metric for quantifying the disparity between two distributions within a Reproducing Kernel Hilbert Space (RKHS), constituting a kernel learning methodology. The Reproducing Kernel Hilbert Space (RKHS) is characterized by the property of reproduction, denoted by $\langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}} = K(x, y)$. In assessing the similarity between two distributions, diminution in the distance value of MMD indicates greater proximity between the respective data distributions.

Searching for a function f in the function set \mathcal{F} such that $MD = |\text{mean}(f(P)) - \text{mean}(f(Q))|$ achieves its maximum value. Therefore, the Maximum Mean Discrepancy (MMD) can be expressed as follows:

$$MMD[\mathcal{F}, P, Q] := \sup_{f \in \mathcal{F}} (E_{X^s} [f(X^s)] - E_{X^t} [f(X^t)]) \tag{1}$$

P and Q represent two datasets that conform to probability distributions, while x_i^s and x_i^t denote samples extracted from $P(X^s)$ and $P(X^t)$ respectively. Due to the preference for a more expansive dimensional space post-mapping, the function set should be as diverse as possible (in practice, it is infinite). When constitutes the unit ball on RHKS, it is as follows:

$$MMD[\mathcal{F}, P, Q] := \sup_{f \in \mathcal{F}, \|f\|_{\mathcal{H}} \leq 1} \left(\frac{1}{n_s} \sum_{i=1}^{n_s} f(x_i^s) - \frac{1}{n_t} \sum_{i=1}^{n_t} f(x_i^t) \right) \tag{2}$$

$\|f\|_{\mathcal{H}} \leq 1$ denotes the norm of f in RHKS, which ought to be less than 1, namely, any arbitrary vector within the unit ball. n_s and n_t respectively denote the sample sizes of the source and target domains. Additionally, $x^s \in \mathcal{S}' \subset \mathcal{S}$, $x^t \in \mathcal{T}' \subset \mathcal{T}$. Given that $f(x_i)$ is infinite-dimensional, the essence of its kernel trick lies in the avoidance of explicitly representing the mapping function to compute the inner product of two vectors. Therefore, we can square the MMD, simplify the result, and express it using a kernel function, namely:

$$MMD^2[\mathcal{F}, P, Q] = \left\| \begin{aligned} & \frac{1}{n_s^2} \sum_{i=1}^{n_s} \sum_{i=1}^{n_s} k_l(\phi_l(x_i^s), \phi_l(x_i^s)) \\ & - \frac{2}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} k_l(\phi_l(x_i^s), \phi_l(x_j^t)) \\ & + \frac{1}{n_t^2} \sum_{j=1}^{n_t} \sum_{j=1}^{n_t} k_l(\phi_l(x_j^t), \phi_l(x_j^t)) \end{aligned} \right\| \tag{3}$$

k_l represents the core selected by the first layer of the deep neural network.

3.2 Contrastive Domain Discrepancy

Contrastive Domain Discrepancy (CDD) is based on the MMD metric, but diverges from it in its criterion. CDD minimizes intra-class differences and maximizes inter-class differences. MMD employs the square of the kernel function, whereas CDD utilizes the kernel function directly. This variation results in different computational emphases. Consequently, MMD primarily focuses on source-target domain similarity, while CDD primarily focuses on contrastiveness.

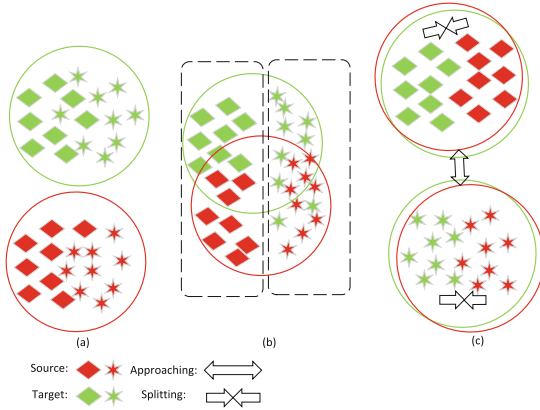


Fig. 1. This figure shows the comparison of the distribution of domain difference minimization methods, which minimizes the intra-class difference and maximizes the inter-class difference in our iris detection task.

As shown in the Fig. 1 (a) Before domain adaptation, it is evident that there exists a significant domain shift between the source and target data. (b) After conventional domain adaptation, the situation of ignoring class labels of samples often leads to poor generalization performance. (c) Our proposed method aims to minimize the intra-domain differences while maximizing the inter-domain variances, thereby significantly enhancing the classification accuracy.

With the detailed formulation of the Maximum Mean Discrepancy (MMD) method, the unsupervised domain adaptation approach based on the Conditional Distribution Discrepancy (CDD) metric is established upon the distinctions in the conditional data distributions across domains, where CDD refers to the marginal or conditional probability distributions under different domains.

Thus, the expression of CDD is:

$$CDD(P, Q) = CDD_{intra}(X_s, X_t) + CDD_{inter}(X_s, X_t) \tag{4}$$

Initially, consider how to minimize intra-domain differences. Intra-domain differences refer to the variances among samples within the same category. We aim for samples of the same category within the source and target domains to cluster as closely as possible. When we assume the domain differences of two

classes are respectively a and b, the intra-domain differences can be measured when $a = b$. Consequently, the definition of intra-domain difference is as follows:

$$\begin{aligned} CDD_{\text{intra}}(P, Q) &= \frac{1}{n_s^a} \sum_i^{n_s^a} \sum_j^{n_s^a} k(\phi(x_{s_i}^a), \phi(x_{s_j}^a)) + \frac{1}{n_s^b} \sum_i^{n_s^b} \sum_j^{n_s^b} k(\phi(x_{s_i}^b), \phi(x_{s_j}^b)) \\ &\quad + \frac{1}{n_t^a} \sum_i^{n_t^a} \sum_j^{n_t^a} k(\phi(x_{t_i}^a), \phi(x_{t_j}^a)) + \frac{1}{n_t^b} \sum_i^{n_t^b} \sum_j^{n_t^b} k(\phi(x_{t_i}^b), \phi(x_{t_j}^b)) \end{aligned} \quad (5)$$

Subsequently, deliberate on strategies to maximize inter-domain disparities. Inter-domain differences denote the distinctions observed among samples belonging to distinct categories. Our objective is for variances among samples from disparate categories between the source and target domains to be maximized. Under the same conditions as aforementioned, the measurement of inter-domain differences can be conducted when $a \neq b$. The definition of inter-domain difference can be formulated as follows:

$$CDD_{\text{inter}}(P, Q) = -\frac{2}{n_s^a n_s^b} \sum_i^{n_s^a} \sum_j^{n_s^b} k(\phi(x_{s_i}^a), \phi(x_{s_j}^b)) - \frac{2}{n_t^a n_t^b} \sum_i^{n_t^a} \sum_j^{n_t^b} k(\phi(x_{t_i}^a), \phi(x_{t_j}^b)) \quad (6)$$

The first term represents the differences between samples of category a from the source domain and samples of category b from the target domain, while the second term represents the differences between samples of category a from the target domain and samples of category b from the source domain. Estimate the target labels \hat{y}_i^t , assuming $a = b = m$. Therefore, the result of CDD calculation is:

$$CDD(P, Q) = \frac{1}{M} \sum_{m=1}^M CDD_{\text{intra}}(\hat{y}_{1:n_t}^t, \phi) - \frac{1}{M(M-1)} \sum_{a=1}^M \sum_{b=1, a \neq b}^M CDD_{\text{inter}}(\hat{y}_{1:n_t}^t, \phi) \quad (7)$$

Intra-domain difference is calculated by applying MMD to samples of the same category from both the source and target domains, while inter-domain difference is computed by applying MMD to samples of different categories from the source and target domains. This achieves the goal of minimizing intra-domain difference and maximizing inter-domain difference as much as possible. It is worth noting that since CDD is derived based on MMD, it exhibits robustness to noise. Therefore, the noise in label estimation can be disregarded.

Next, we leverage deep neural networks (CNNs) [33], a prevalent deep learning model commonly utilized for image recognition and computer vision tasks. These networks are composed of multiple convolutional and pooling layers for feature extraction from input images, followed by fully connected layers for classification or regression. The feature extractor consists of convolutional neural networks (CNNs), which are employed to extract feature representations from input data, mapping samples from both the source and target domains into a shared

feature space. Simultaneously, we leverage the concept of contrastive learning, using a contrastive adaptation module to compute contrastive losses, comprising intra-class contrastive loss and inter-class contrastive loss. Intra-class contrastive loss computes the contrastive loss between samples within the same category, encouraging samples of the same category to be closer in the feature space. Similarly, inter-class contrastive loss calculates the contrastive loss between samples from different categories, encouraging samples of different categories to be more dispersed in the feature space. We then employ the backpropagation algorithm to compute gradients of the loss function with respect to the parameters of the feature extractor.

Given a sample pair (x_i, x_j) , for the calculation of intra-class contrastive loss:

$$\mathcal{L}_{intra}^{loss} = \|f(x_i) - f(x_j)\|^2 \tag{8}$$

Here, $f(\cdot)$ denotes the feature extractor, and $\|\cdot\|$ represents the Euclidean distance between vectors.

Likewise, considering a sample pair (x_i, x_k) , for the calculation of inter-class contrastive loss:

$$\mathcal{L}_{inter}^{loss} = \max\left(0, m - \|f(x_i) - f(x_k)\|^2\right) \tag{9}$$

In this context, m represents a hyper parameter, often denoted as the margin. Consequently, the CDD loss can be expressed as follows:

$$\mathcal{L}_{CDD}^{loss} = \frac{1}{N} \sum_{(x_i, x_j)} \mathcal{L}_{intra}^{loss} + \frac{1}{M} \sum_{(x_i, x_k)} \mathcal{L}_{inter}^{loss} \tag{10}$$

Subsequently, we integrate a cross-entropy loss to augment the classification efficacy of the model. In this scenario, C represents the number of classes, y_c denotes the indicator function of the c^{th} class of true labels, and p_c signifies the probability of the c^{th} class predicted by the model. Therefore, the cross-entropy loss to be minimized is formulated as follows:

$$\mathcal{L}^{ce} = - \sum_c^C y_c \log(p_c) \tag{11}$$

Overall objective:

$$\mathcal{L}^{total} = \alpha \mathcal{L}^{CDD} + \beta \mathcal{L}^{ce} \tag{12}$$

In this context, α and β represent two weighting parameters of the losses, utilized to calibrate the impacts of the respective losses on the overarching objective. By minimizing both the CDD loss and the cross-entropy loss, it becomes feasible to proficiently learn feature representations across the source and target domains, thereby facilitating the optimization of unsupervised domain adaptation and classification tasks.

4 Experiment

We designed two sets of experiments to validate the effectiveness of the method. For the first set of experiments, we selected Clarkson2015LG, Clarkson20132015Dalsa and NDLiv-Det2017 datasets as dataset-1. Three data sets were collected from LG2200, Dalsa camera and LG4000 sensor respectively. For the second set of experiments, we selected ND-I, CASIA-IF and IF-VE datasets as dataset-2. Three data sets were collected from the LG4000, OKIIRISPASS-h and AI1000 sensors respectively. Details are presented in Sect. 4.1 and Sect. 4.2.

4.1 Experiment on Dataset-1

We conducted experiments on three publicly available iris cosmetic contact lens image datasets.

One public dataset is the Notre Dame Cosmetic Contact Lens Detection 2017 (NDLiv-Det2017) [34], collected by LG4000 and AD100 sensors. All genuine images conform to ISO/IEC 19794-6. Soft (or transparent) contact lenses are excluded from the data. We divided it into training subset and test subset, where the training subset includes 600 genuine iris images (without contact, soft, or make-up) and 600 images of cosmetic contact lenses from multiple manufacturers, and the test subset includes 900 textured contact lens images and 900 genuine iris images.

Table 1. This table shows the number of samples of different classes in dataset-1 and dataset-2

	Datasets	Train		Test	
		Genuine	Cosmetic	Genuine	Cosmetic
Dataset-1	NDLiv-Det2017	600	600	900	900
	Clarkson2015LG	450	540	379	577
	Clarkson20132015Dalsa	970	1275	625	1000
Dataset-2	ND-I	2000	1000	800	400
	CASIA-IF	4800	592	1200	148
	IF-VE	20000	20000	5000	5000

Another public dataset is Clarkson2015LG [34, 35]. We divided it into training subset and test subset, where the training subset includes 450 real-time images and 540 images of patterned cosmetic contact lenses, and the test subset includes 379 real-time images and 577 images of patterned cosmetic contact lenses.

Another public dataset is Clarkson20132015Dalsa [34, 35], and we also divided it into training subset and test subset. The training subset includes 2245 real-time images, images captured by Dalsa cameras, and images of patterned cosmetic contact lenses. The test subset includes 1625 real-time images, images captured by Dalsa cameras, and images of patterned cosmetic contact lenses.

In Fig. 2, an example of the Dataset-1 is presented. Table 1 presents the detailed breakdown of sample counts for training and testing subsets of each dataset.

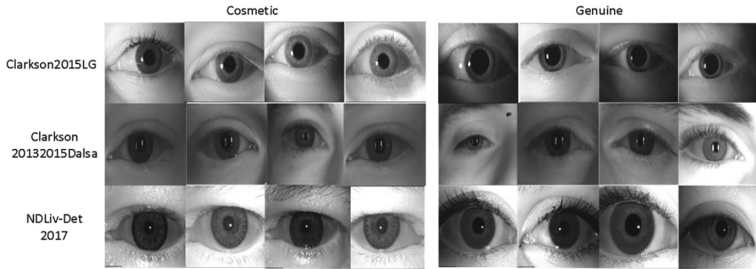


Fig. 2. Sample instances of genuine iris and cosmetic contact lens iris images in the first set of experiments. Among the three datasets, NDLiv-Det2017 was acquired using LG4000 and AD100 sensors, Clarkson2015LG was obtained using the LG2200 sensor, and Clarkson20132015Dalsa was captured using Dalsa brand camera sensors. As the three datasets were acquired using different equipment, to more effectively verify the performance of cross-device iris liveness detection, we amalgamated these datasets into a single group for experimentation.

4.2 Experiment on Dataset-2

We conduct experiments on two public iris cosmetic contact lens image datasets and a self collected cosmetic contact lens image dataset. One public dataset is the Notre Dame cosmetic contact lens 2013 I (ND-I for short) [36], which is captured by LG4000. Since our research emphasizes on cosmetic lens detection, merely genuine (non-lens and soft lens) and cosmetic iris images are considered. The other one public dataset is the CASIA-IrisFake(CASIA-IF for short) [37]. It contains three type attack means, including printed iris images, plastic eyes and cosmetic contacts. We use the cosmetic contact iris image and genuine iris image for experiments. We randomly split the dataset in train and test subsets of numbers referring to [36].

The self collected dataset is captured under various environment which is more closer to practical use. To test proposed methods on different quality images, we collect a large scale cross-sensor fake iris dataset under various environment (IF-VE for short). It contains 50000 images, with 40000 images from 80 volunteers for training and the other 10000 from other 20 volunteers for test. During the collection, different iris sensors are used, and volunteers are instructed to change angles, distances and positions, wear glasses, and squint, to enrich the quality types of iris images.

In Fig. 3, an example of the Dataset-2 is presented. Table 1 presents the detailed breakdown of sample counts for training and testing subsets of each dataset. Through literature review and analysis, it is evident that, for the purpose of facilitating comparison, we have employed a method of contrast by dividing the

data into the aforementioned two groups of datasets to validate the superiority of experimental results.

Implementation details: We used ResNet-50 and ResNet-101 [38,39] pre-trained on ImageNet [40] as our backbone networks. We fine-tuned the models using labeled source domain data and unlabeled target domain data and selected the labeled source data and unlabeled target data sets by jointly evaluating the test errors of the source classifier and the domain classifier.

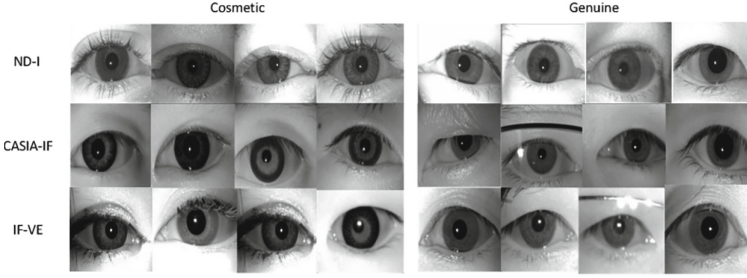


Fig. 3. Sample instances of genuine iris and cosmetic contact lens iris images in the second set of experiments. Similarly, among the three datasets, ND-I was acquired using the LG4000 sensor, CASIA-IF was obtained using the OKI IRISPASS-h sensor, and IF-VE was captured using the AI1000 sensor, with each dataset utilizing different sensors. Likewise, amalgamating these three datasets into a single group for experimentation can effectively verify the performance of cross-device iris liveness detection.

4.3 Performance Evaluation

In this part, we demonstrate the result of the experimental settings in Sect. 4.1 and Sect. 4.2, the accuracy of our model conducted on dataset-1 and dataset-2 respectively. For the dataset-1, we set six combinations of source and target domain data, as shown in Table 2. And for the second dataset, we set six permutations of source and target domain data, as shown in Table 3. We respectively employ the backbone networks as ResNet-50 and ResNet-101, and tested using two measurement methods, MMD and CDD. The results of the dataset-1 and dataset-2 are shown in the Table 2 and Table 3.

The classification accuracy of our model on dataset-1 is relatively high, especially when we tested using the CDD metric, the effect was more ideal. Concurrently, we employed the non-transfer method as a comparative baseline, clearly demonstrating that the MMD and CCD measurement methods significantly enhance accuracy and performance. To more accurately evaluate model predictions, the Receiver Operating Characteristic (ROC) curve is frequently utilized. The greater the distance of the ROC curve from the baseline, the superior the model's predictive performance. Figure 4 illustrates the ROC results from twelve experimental sets conducted on the ResNet50 and ResNet101 backbone networks, respectively. The solid line denotes the CDD algorithm, the long

dashed line signifies the MMD algorithm, and the short dashed line indicates the non-transfer method. Evidently, models employing the proposed CDD and MMD measurement methods demonstrate enhanced robustness and stability, consistently achieving high classification accuracy across various datasets. Furthermore, the CDD algorithm exhibits superior generalization capabilities across diverse datasets. Compared to non-transfer algorithms, it exhibits superior iris liveness detection characteristics.

Table 2. This table shows the sample instances of genuine iris and cosmetic contact lens iris images in the first dataset of experiments. The accuracy of our model performing on it is shown as well. RN50-NT and RN101-NT respectively denote the Non-Transfer of method with backbone Resnet-50 and Resnet-101

Source	Target/Test	ResNet50	RN50-ACC	RN50-NT	ResNet101	RN101-ACC	RN101-NT
Clarkson2015LG	Clarkson20132015Dalsa	MMD	71.7058	61.9943	MMD	76.0425	59.8818
		CDD	85.354		CDD	82.1147	
Clarkson2015LG	NDLiv-Det2017	MMD	89.4444	85.3889	MMD	91.5556	74.0556
		CDD	97.0566		CDD	93.5556	
Clarkson20132015Dalsa	Clarkson2015LG	MMD	88.8104	73.9542	MMD	86.1648	84.5941
		CDD	95.1348		CDD	90.563	
Clarkson20132015Dalsa	NDLiv-Det2017	MMD	79.3889	63.1111	MMD	77.4444	73.2778
		CDD	89.2778		CDD	83.6667	
NDLiv-Det2017	Clarkson2015LG	MMD	84.7222	81.7708	MMD	90.4473	80.0347
		CDD	90.7118		CDD	97.5198	
NDLiv-Det2017	Clarkson20132015Dalsa	MMD	66.6435	64.8798	MMD	69.1383	59.3186
		CDD	72.6096		CDD	71.0859	

Table 3. This table shows the sample instances of real iris and cosmetic contact lens iris images, in the second dataset of experiments. The accuracy of our model performing on this dataset is shown as well. RN50-NT and RN101-NT respectively denote the Non-Transfer of method with backbone Resnet-50 and Resnet-101

Source	Target/Test	ResNet50	RN50-ACC	RN50-NT	ResNet101	RN101-ACC	RN101-NT
ND-I	CASIA-IF	MMD	65.8615	51.0204	MMD	58.3856	53.4014
		CDD	78.4766		CDD	70.6764	
ND-I	IF-VE	MMD	24.6871	50.17	MMD	58.1913	50.82
		CDD	56.5526		CDD	77.0016	
CASIA-IF	ND-I	MMD	86.1875	85.3125	MMD	87.375	59.1875
		CDD	98.0625		CDD	93.6875	
CASIA-IF	IF-VE	MMD	67.2	62.77	MMD	67.1305	59.47
		CDD	98.52		CDD	86.5261	
IF-VE	ND-I	MMD	84.0625	77.1875	MMD	81.1875	80.6875
		CDD	91.8125		CDD	91.8125	
IF-VE	CASIA-IF	MMD	96.231	95.9184	MMD	95.5782	94.8563
		CDD	99.1948		CDD	98.6536	

By comparison with the non-transfer method conducted on dataset-2, it is observed that, except when ND-I is utilized as the training sample and IF-VE as the testing sample, the MMD algorithm exhibits lower classification accuracy than the non-transfer method. This discrepancy is attributable to the substantial difference in sample sizes between the ND and IFVE datasets. However, overall, our method demonstrates a significant enhancement in model classification accuracy. Similarly, in the second set of comparative experiments, we continue to utilize the ROC curve to evaluate model performance. As illustrated in Fig. 4, ResNet101 and ResNet50 are employed as the backbone networks, with the solid line denoting the CDD algorithm, the long dashed line signifying the MMD algorithm, and the short dashed line indicating the non-transfer method. According to the figures, the model trained on CASIA samples remains consistent with our conclusions. However, the ROC curve results derived from IFVE and ND as training samples are notably poor, potentially due to a substantial deviation in the number of training sample data. In the previous set of experiments, the sample size differences across the three datasets were minimal, leading to more precise model performance. In summary, across both sets of experiments, it is evident that the CDD and MMD algorithms substantially enhance the classification accuracy of iris liveness detection, as well as the robustness and stability of the model.

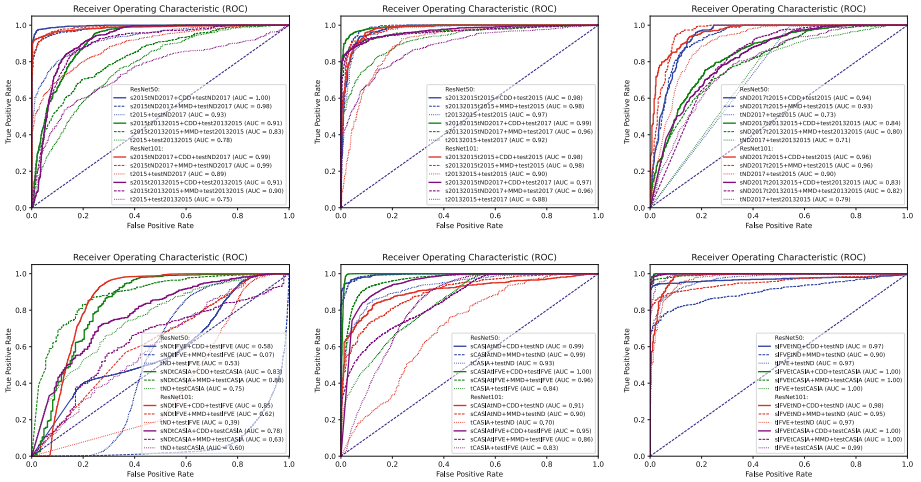


Fig. 4. Tree sets of diagrams on the first line: ROC curve obtained according to Clarkson2015, Clarkson20132015Dalsa and NDLiv-Det2017 training sample. According to the experimental grouping in Table 2, the MMD, CDD, and the Non-Transfer of method were visually compared using the ROC curve. The advantages of the CDD approach are clearly visible. Tree sets of diagrams on the second line: ROC curve obtained according to ND-I, CASIA-IF and IF-VE training sample. According to the experimental grouping in Table 3, the MMD, CDD, and the Non-Transfer of method were visually compared according to the ROC curve. The results are affected by the large gap in the number of data sets, but the advantages of the CDD method can still be clearly seen.

According to the results of the experiments on the two datasets, it is evident that both MMD and CDD metrics could improve detecting accuracy of our model on iris recognition task, which means the performance of our model has been enhanced notably. Moreover, compared with MMD metric under the same experimental conditions, the improvement of CDD metric method is significantly more established. So in the future work, we will further study on how to improve the establishment of our iris recognition model.

5 Conclusions

In this work, we propose an unsupervised domain adaptation transfer learning model based on MMD and CDD metrics for iris liveness detection through perceptual alignment. Modeling and optimizing intra-domain and inter-domain discrepancies, the CDD metric and MMD metric significantly improve the accuracy of our model when detecting. In addition, CDD is with a noticeable superiority in terms of classification accuracy and overall model improvement. We confirm the effectiveness of our unsupervised domain-adaptive transfer learning method from both theoretic proof and experimental results in cross-device iris liveness detection, which has obtained the capability of agile deployment.

The security of iris recognition technology is pertinent to every individual's life. The forgery of an individual's iris can have profound implications for personal, corporate, and national interests. The two transfer model methods we propose facilitate agile and cost-effective deployment when integrating new devices, thereby significantly advancing iris recognition technology.

References

1. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(11), 1148–1161 (1993)
2. Wildes, R.P., et al.: A system for automated iris recognition. In: *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pp. 121–128. IEEE (1994)
3. Boles, W.W., Boashash, B.: A human identification technique using images of the iris and wavelet transform. *IEEE Trans. Signal Process.* **46**(4), 1185–1188 (1998)
4. Raghavendra, R., Raja, K.B., Busch, C.: Contlensnet: robust iris contact lens detection using deep convolutional neural networks. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1160–1167. IEEE (2017)
5. Singh, A., Mistry, V., Yadav, D., Nigam, A.: Ghclnet: a generalized hierarchically tuned contact lens detection network. In: *2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, pp. 1–8. IEEE (2018)
6. Hoffman, S., Sharma, R., Ross, A.: Convolutional neural networks for iris presentation attack detection: toward cross-dataset and cross-sensor generalization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1620–1628 (2018)
7. Zhang, H., Bai, Y., Zhang, H., Liu, J., Li, X., He, Z.: Local attention and global representation collaborating for fine-grained classification. In: *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10658–10665. IEEE (2021)

8. Daugman, J.: Biometrics. personal identification in a networked society, chapter recognizing persons by their iris patterns (1999)
9. Daugman, J.: Iris recognition and anti-spoofing countermeasures. In: 7th International Biometrics Conference (2004)
10. Lee, S.J., Park, K.R., Kim, J.: Robust fake iris detection based on variation of the reflectance ratio between the iris and the sclera. In: Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference, pp. 1–6. IEEE (2006)
11. Lee, S.J., Park, K.R., Lee, Y.J., Bae, K., Kim, J.: Multifeature-based fake iris detection method. *Optical Eng.* **46**(12), 127204–127204 (2007)
12. He, Y., Hou, Y., Li, Y., Wang, Y.: Liveness iris detection method based on the eye's optical features. In: Optics and Photonics for Counterterrorism and Crime Fighting VI and Optical Materials in Defence Systems Technology VII, vol. 7838, pp. 236–243. SPIE (2010)
13. Park, J.H., Kang, M.G.: Iris recognition against counterfeit attack using gradient based fusion of multi-spectral images. In: Li, S.Z., Sun, Z., Tan, T., Pankanti, S., Chollet, G., Zhang, D. (eds.) IWBRIS 2005. LNCS, vol. 3781, pp. 150–156. Springer, Heidelberg (2005). https://doi.org/10.1007/11569947_19
14. Park, J.H., Kang, M.G.: Multispectral iris authentication system against counterfeit attack using gradient-based image fusion. *Opt. Eng.* **46**(11), 117003–117003 (2007)
15. Lee, E., Park, K., Kim, J.: Fake iris detection by using purkinje image. In: Proceedings of International Conference on Biometrics (ICB 2006) (2006)
16. Park, K.R.: Robust fake iris detection. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2006. LNCS, vol. 4069, pp. 10–18. Springer, Heidelberg (2006). https://doi.org/10.1007/11789239_2
17. Pacut, A., Czajka, A.: Aliveness detection for iris biometrics. In: Proceedings 40th Annual: International Carnahan Conference on Security Technology, pp. 122–129. IEEE (2006)
18. Bodade, R., Talbar, S.: Dynamic iris localisation: a novel approach suitable for fake iris detection. In: 2009 International Conference on Ultra Modern Telecommunications & Workshops, pp. 1–5. IEEE (2009)
19. Huang, X., Ti, C., Hou, Q.-Z., Tokuta, A., Yang, R.: An experimental study of pupil constriction for liveness detection. In: IEEE Workshop on Applications of Computer Vision (WACV), pp. 252–258. IEEE (2013)
20. Ruiz-Albacete, V., Tome-Gonzalez, P., Alonso-Fernandez, F., Galbally, J., Fierrez, J., Ortega-Garcia, J.: Direct attacks using fake images in iris verification. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M. (eds.) BioID 2008. LNCS, vol. 5372, pp. 181–190. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89991-4_19
21. Puhan, N.B., Sudha, N., Hegde, S.: A new iris liveness detection method against contact lens spoofing. In: IEEE 15th International Symposium on Consumer Electronics (ISCE), pp. 71–74. IEEE (2011)
22. Zhang, H., Sun, Z., Tan, T.: Contact lens detection based on weighted LBP. In: 2010 20th International Conference on Pattern Recognition, pp. 4279–4282. IEEE (2010)
23. Lee, E.C., Ko, Y.J., Park, K.R.: Fake iris detection method using purkinje images based on gaze position. *Opt. Eng.* **47**(6), 067204–067204 (2008)
24. Ghiani, L., et al.: Livdet: fingerprint liveness detection competition 2013. In: 2013 International Conference on Biometrics (ICB), pp. 1–6. IEEE (2013)

25. Lee, E.C., Park, K.R.: Fake iris detection based on 3D structure of iris pattern. *Int. J. Imaging Syst. Technol.* **20**(2), 162–166 (2010)
26. Zhang, Y., Liu, T., Long, M., Jordan, M.: Bridging theory and algorithm for domain adaptation. In: *International Conference on Machine Learning*, pp. 7404–7413. PMLR (2019)
27. Ren, M., Wang, Y., Sun, Z., Tan, T.: Dynamic graph representation for occlusion handling in biometrics. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 11940–11947 (2020)
28. Daugman, J.: Demodulation by complex-valued wavelets for stochastic pattern recognition. *Int. J. Wavelets Multiresolut. Inf. Process.* **1**(01), 1–17 (2003)
29. Lovish, Nigam, A., Kumar, B., Gupta, P.: Robust contact lens detection using local phase quantization and binary gabor pattern. In: *Azzopardi, G., Petkov, N. (eds.) CAIP 2015. LNCS*, vol. 9256, pp. 702–714. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23192-1_59
30. Yadav, D., Kohli, N., Doyle, J.S., Singh, R., Vatsa, M., Bowyer, K.W.: Unraveling the effect of textured contact lenses on iris recognition. *IEEE Trans. Inf. Forensics Secur.* **9**(5), 851–862 (2014)
31. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: *International Conference on Machine Learning*, pp. 97–105. PMLR (2015)
32. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: *International Conference on Machine Learning*, pp. 2208–2217. PMLR (2017)
33. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
34. Yambay, D., et al.: Livdet iris 2017-iris liveness detection competition 2017. In: *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 733–741. IEEE (2017)
35. Yambay, D., Doyle, J., Bowyer, K., Czajka, A., Schuckers, S.: Livdet-iris 2013-iris liveness detection competition 2013. In: *2014 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE (2014)
36. Sun, Z., Zhang, H., Tan, T., Wang, J.: Iris image classification based on hierarchical visual codebook. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(6), 1120–1133 (2013)
37. Doyle, J.S., Bowyer, K.W., Flynn, P.J.: Variation in accuracy of textured contact lens detection based on sensor and lens pattern. In: *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–7. IEEE (2013)
38. 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)
39. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: *Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS*, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
40. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE (2009)



A Collaborative Approach Using Ridge-Valley Minutiae for More Accurate Contactless Fingerprint Matching

Ritesh Vyas and Ajay Kumar^(✉)

Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong, China
ajay.kumar@polyu.edu.hk

Abstract. Contactless fingerprint identification has emerged as a reliable and user-friendly alternative for personal identification in a range of mobile and access control applications. This paper presents a systematic analysis of the extent of complimentary ridge-valley information in contactless fingerprint images and introduces a new approach to achieve significantly higher match accuracy over state-of-the-art fingerprint matchers commonly employed today. We also investigate the least explored methods for fingerprint color-space conversions, which can play a key role in more accurate contactless fingerprint matching from mobile sensors. We present the experimental results from different publicly available contactless fingerprint databases and incorporate the NBIS, MCC, and a commercial fingerprint matcher to ascertain the extent of performance enhancement. Our consistently outperforming results validate the effectiveness of the proposed approach for more accurate contactless fingerprint identification.

Keywords: Personal Identification · Biometrics · Fingerprint Identification

1 Introduction

Biometric patterns offer the most reliable signatures to conveniently and securely establish human identities. Among several physiological traits accessible from the human body, finger ridge patterns have been widely employed in law enforcement departments to establish the unique personal identity of suspects. The fingerprint features are formed before birth and are known for their high permanence, and today's high-computing machines make their use quite convenient and fast. The pervasiveness of fingerprint authentication can be conjectured by its use in the most ubiquitous devices like smartphones. Moreover, many e-business applications, like financial transactions or access to secured offices, have increasingly relied on fingerprint authentication.

The majority of the fingerprint-based systems deployed today still use contact-based sensors to acquire the fingerprint of any subject, i.e., these sensors require the subject to make contact with his or her finger with the platen or surface of the sensors. Such contact-based acquisition poses new challenges relating to user convenience, hygiene,

and security threats. The contact-based imaging requirements can be a serious threat to hygiene as there is a wide range of diseases, *e.g.* severe acute respiratory syndrome or coronavirus, which are known to transmit or spread from unintended contacts. Therefore, the user’s hygiene becomes vulnerable in such contact-based acquisition. The leftover or latent impressions on the surface of contact-based sensors not only interfere with the new acquisitions but are also known to pose a security threat as these can be lifted to reconstruct spoof fingerprints for the presentation attacks.

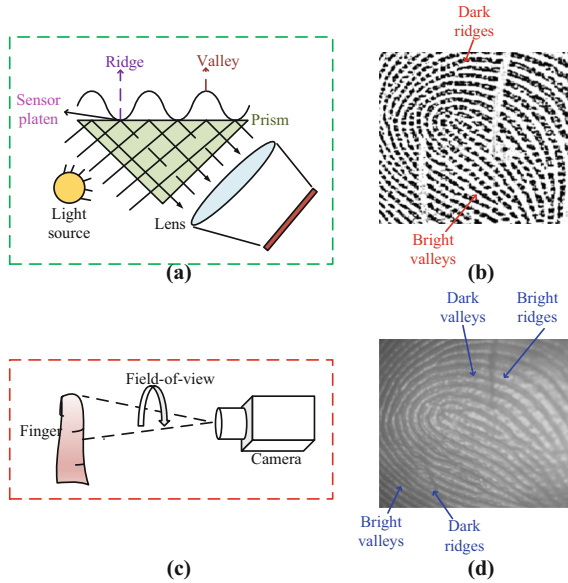


Fig. 1. Illustration of legacy and contactless fingerprints. (a) FTIR in legacy fingerprints, (b) Legacy (contact-based) fingerprint image with regular dark ridges and bright valleys, (c) Contactless acquisition, (d) Contactless fingerprint (with ridges and valleys in varying polarities).

Contactless acquisition of fingerprints can address the above-mentioned challenges and has increasingly attracted the attention of researchers and developers [1, 5]. Several references [4, 19] have investigated contactless fingerprint acquisition using mobile phones and specialized setups, while many commercial contactless fingerprint sensors have also recently emerged for deployments. The availability of contactless fingerprint databases [13, 16] has also encouraged much-needed further research in this area. Involuntary finger motion during contactless fingerprint imaging can significantly degrade the matching accuracies, and reference [7] has attempted to address such a problem. Earlier work [2, 9, 16, 19, 20, 22–24] in contactless 2D fingerprint identification have largely incorporated image segmentation, enhancement, or minutiae matching algorithms that have shown promising results for the contact-based fingerprints. Such direct use of contact-based fingerprint methods ignores the nature of image formation from contactless sensing and is, therefore, not adequate to utilize the full potential of contactless fingerprint images. The duality of the relationship between the minutiae extracted from contactless fingerprint images and contact-based fingerprints has been quite known [17]

and considered during cross-sensor fingerprint matching. However, this paper [25], for the first time, analyzes such influence in real contactless fingerprint databases and introduces alternative strategies to achieve significant performance improvements using the popular contact-based fingerprint matchers.

2 Contactless Fingerprint Image Formation

Human fingers are known to be a curved 3D structure [5, 18]. However, when these 3D surfaces are sensed using a contact-based fingerprint sensor, only their 2D projections are recorded. More precisely, when the subject touches the sensor plate with his or her finger, the topographic high points, which are known as the ridges, are imaged while the low points, known as the furrows or valleys, are not imaged as these are considered as part of the background. Figure 1(a) illustrates the frustrated total internal reflection (FTIR) principle, which is popularly employed in the legacy contact-based fingerprint sensors. It can be observed from this figure that the pixels corresponding to the light rays reflected from the low topographic regions, or valleys of the finger skin surface, have higher intensity (bright) in a fingerprint image. The spatial locations from high areas or ridges are rendered as darker regions in the fingerprint image. The same ridge and valley topologies have been correspondingly identified in Fig. 1(b).

However, the contrast between such gray levels between the ridge and valley is remarkably different in the contactless acquisition (see Fig. 1(c)) of the same finger surface. This difference can be attributed to the varying illumination on different topographic regions of the finger skin, where the ridges are rendered as darker or brighter areas in those regions. Contactless acquisition is commonly followed by grayscale representation or the binarization of finger images so that their appearance is similar to that of their legacy contact-based counterparts. Such second-order representation for contactless fingerprints can significantly degrade the system's matching accuracy. A gray-level representation can be attributed to the variation in the illuminating angle, because of which the ridges are sometimes rendered as brighter and sometimes as darker than the adjacent valleys. One such instance is illustrated in Fig. 1(d). In summary, the brighter and darker portions of a ridge in a contactless image are not consistent with the ridge and valley. Instead, they are consistent with the opposite flanks of ridges. Such polarity reversal effect is quite common and can be attributed to the interaction of incident illumination with respect to the 3D ridge-valley structure [8, 11] during contactless imaging.

These gray-level alterations can be easily observed from the raw fingerprint images acquired in a contactless manner. However, if the images are represented as grayscale or even binarized to detect the potential minutiae points, this polarity reversal effect is completely lost. This is also the key reason for poor performance when the contactless images are matched against the legacy contact-based fingerprints. Our detailed observations reveal that the minutiae expected to be matched from contactless and contact-based images do not appear in the same or similar positions. The ridge endings of contact-based fingerprints become ridge bifurcations in the contactless fingerprint and vice-versa. An instance of such changes in the type of minutiae is illustrated from a real finger image sample in Fig. 2.

Such an image formation mechanism during the contactless fingerprint acquisition motivated us to investigate additional information for the localization of minutiae from

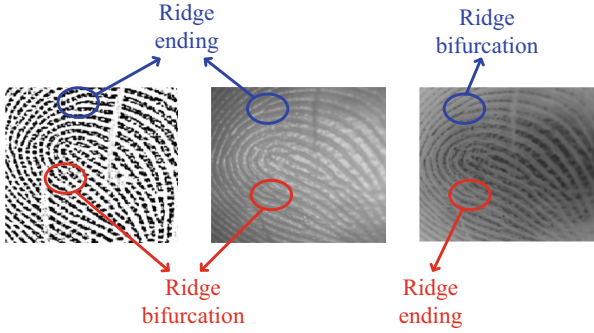


Fig. 2. Differences in the localization of fingerprint minutiae from (left) contact-based or legacy fingerprint image, (center) contactless fingerprint image, and respective (right) inverted fingerprint image.

the input fingerprint image. Unlike the legacy contact-based fingerprint, in which the valley information is lost as it is considered as part of background information, contactless fingerprint imaging simultaneously recovers the ridge and valley information. However, this joint information is embedded with low-contrast as 3D ridges are imaged under uneven illuminations and reflections from multiple ridges. In order to reveal additional minutiae from the otherwise dark portions of the contactless image and exploit the polarity alternation notion, we also investigate image transformations to precisely recover the minutiae. This work presents such systematic investigation, using publicly available contactless fingerprint image databases, and introduces possible solutions to address significant degradation in performance while matching contactless fingerprint images.

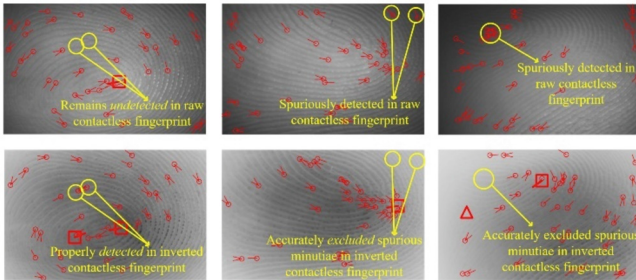


Fig. 3. Minutiae detection in (top row) raw contactless fingerprints and (bottom row) inverted contactless fingerprint images. (Color figure online)

3 Minutiae Detection from Contactless Fingerprint Images

The difference between the raw and its inverted contactless fingerprint images can be explicitly observed from the sample images in Fig. 3. In this figure, the top and bottom rows show the detected minutiae from raw and inverted contactless fingerprints, respectively. As can be observed from these images, most of the minutiae detected from the raw contactless fingerprint are also present in the corresponding inverted fingerprint, but there is a change in the minutiae type. This explains why bifurcations of raw fingerprints appear as terminations in the inverted images and vice-versa [20]. More importantly, it can also be clinched upon careful inspection of these images that many spurious minutiae, which are falsely detected from the raw fingerprints, remain undetected in their inverted counterparts. Additionally, the inverted fingerprint facilitates the detection of additional cores and deltas, which otherwise remain undetected from the raw fingerprint images.

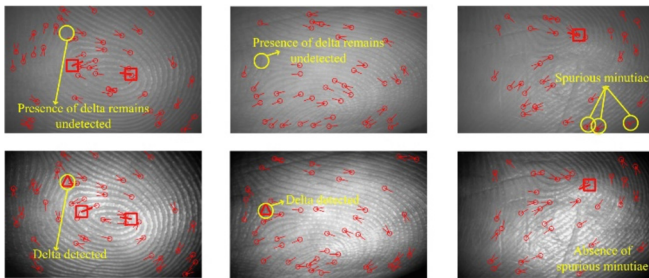


Fig. 4. Minutiae detection in (top row) raw contactless fingerprints and (bottom row) inverted contactless

For instance, the first image sample in the top row of Fig. 3 shows a sample fingerprint having two cores (i.e., u-shaped ridges) in the lower half of the image. However, the popular commercial off-the-shelf (COTS) tool, namely VeriFinger [12], is only able to detect a single core (shown as a red-colored rectangle), while the other one remains unnoticed because of the poor focus on this part of the finger. However, when the inverted fingerprint is provided as input to the same COTS tool, both the cores get detected (highlighted by two red-colored rectangles in the first image in the bottom row). Similar observations can be made from the second and third images of Fig. 3, where the cores of fingerprints are detected only when their inverted version is presented as input (see the red-colored rectangles in the second and third images of the bottom row). Moreover, the last fingerprint in the top row of Fig. 3 possesses a delta (i.e., a Y-shaped ridge), which can be observed with the bare eyes on the raw contactless image. However, this delta remains undetected because of the poor focus on this part of the finger, which can be attributed to the curved 3D profile of human fingers. The COTS tool, on the other hand, was able to detect this delta from the inverted fingerprint image (red-colored triangle on the last image in Fig. 3).

3.1 Grayscale Representation for Colored Contactless Fingerprints

In order to study the effect of different grayscale representations, the unprocessed fingerprint images from PolyU Contactless to Contact-based fingerprint database [13] were selected. These unprocessed fingerprints are available in RGB color representation and have a larger size (*i.e.*, 1400×900 pixels) than processed ones. The PolyU database [13] also provides a processed contactless fingerprint sub-dataset comprising grayscale 350×225 pixels images. These ordinary grayscale images are generally formed by a weighted combination of linear intensities observed in individual channels. Mathematically, such conversion of RGB images to grayscale images can be directly achieved as follows:

$$I_{ordinary} = 0.3 I_R + 0.59 I_G + 0.11 I_B \quad (1)$$

where I_R , I_G , and I_B are the linear intensities in the red, green, and blue channels, respectively, in the colored images.

However, there are other grayscale representations that can reveal additional details from the output images. In order to provide a more illustrious case of study, we adopted the Luma grayscale representation [14] as an alternative to the grayscale representation in (1). The Luma grayscale representation is believed to be a more perceptually accurate grayscale representation, as it employs non-linear gamma-corrected versions [15] of all channels in colored images rather than their linear intensities. Mathematically, it can be expressed as:

$$I_{Luma} = 0.2126I'_R + 0.7152I'_G + 0.0722I'_B \quad (2)$$

where I'_R , I'_G , and I'_B are the gamma-corrected versions of red, green and blue channels, respectively. The Luma representation of colored images can enable larger contrast. as compared to ordinary grayscale images, which facilitates the enhanced detection of minutiae features. This observation can also be noted from the instances of grayscale images shown in Fig. 4. It is not difficult to observe that Luma grayscale images have increased contrast as compared to ordinary grayscale images. It is evident from the first and second samples in Fig. 4 that the delta singularity is appropriately detected in Luma grayscale images, while it is not detected in ordinary grayscale images. Moreover, a careful visual inspection reveals the presence of many spurious minutiae, which are erroneously detected in the ordinary grayscale representation, get suppressed in the Luma represented (refer to last column images of Fig. 4) images.

Therefore, in order to add a new dimension to the enhancement of contactless fingerprint recognition, we performed experiments with Luma grayscale contactless fingerprint images, which were obtained by converting the unprocessed (colored) images of the PolyU database to Luma grayscale and reducing their dimension to be same as that of the processed contactless fingerprints in the database, *i.e.* 350×225 pixels. Thereafter, the recognition experiments were also performed on the inverted versions of Luma grayscale images to present a comprehensive evaluation of the currently selected framework. These experimental results, with both the Luma grayscale images and their inverted counterparts, are discussed in the next section.

4 Experiments and Results

Contactless Fingerprint Databases

This work comprises experiments on two publicly available contactless fingerprint databases. The first database is the PolyU Contactless to Contact-based fingerprint database [13], which provides 2016 contactless 2D fingerprints and their corresponding 2D contact-based 2016 fingerprints. These fingerprints were captured from 336 clients from the staff and students at the university. Each client has six fingerprint images, both in the contactless and contact-based acquisition setup. In our experiments, we use all the 2016 contactless fingerprint images. Another contactless fingerprint database used in this work is the Benchmark 2D/3D Fingerprint Database publicly available from [16], which provides contactless and contact-based fingerprints from 1500 different fingers. The database comprises at least two contactless fingerprint samples and four contact-based fingerprint samples. The acquisition of this database was completed at three different universities in Australia. We employed images from 1000 fingers of this database in our experiments. Each of these fingers has two contactless fingerprints, which resulted in a total of 2000 images. Sample images from both of these public databases are shown in Fig. 5.

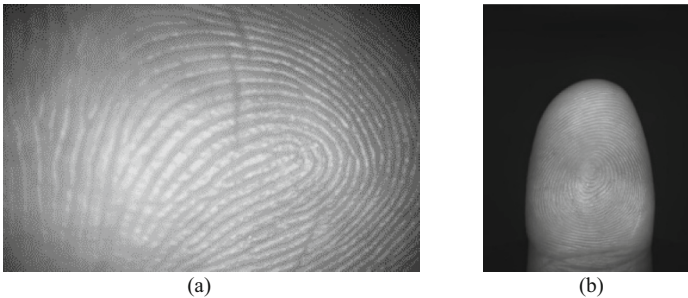


Fig. 5. Samples from both the contactless fingerprint images database (a) PolyU Contactless to Contact-based database [13], (b) benchmark 2D/3D contactless fingerprint database [16].

Evaluation Protocols and Performance Metrics

All experiments were performed using the all-to-all matching protocol, in which every single fingerprint is matched against all other fingerprints in the database. This protocol, being the most challenging biometric evaluation protocol, yields a large number of scores for both the employed databases. The matchings for PolyU Contactless to Contact-based fingerprint database generate 5,040 genuine and 20,26,080 imposter scores. On the other hand, the numbers of genuine and imposter scores for the 2D/3D benchmark contactless database are 1,000 and 19,98,000, respectively.

Three state-of-the-art fingerprint matchers are employed for the performance evaluation. First is a popular fingerprint matcher NBIS (NIST Biometric Image Software) [17]. The second matcher is minutiae cylinder code (MCC) [18] while the third matcher is the commercial off-the-shelf (COTS) matcher, namely VeriFinger from Neurotechnology [12]. Performances of the raw contactless fingerprints and their inverted versions

are comparatively evaluated using common performance metrics like receiver operator characteristics (ROC), equal error rate (EER), false acceptance rate (FAR), and genuine acceptance rate (GAR). Performances from all three fingerprint matchers are discussed in the following.

4.1 NBIS Matcher

The NBIS [17] is an open-source fingerprint matcher provided by NIST. Our experimental results on PolyU contactless fingerprint dataset, using NBIS matcher, are illustrated in Fig. 6(a). The corresponding ROC curves illustrate that NBIS matcher achieves an EER of 13.33% and GAR (@FAR = 0.01%) of 65.52% for legacy contactless fingerprints (i.e. with conventional ridge minutiae) in PolyU database. On the other hand, with the usage of inverted contactless fingerprints (i.e. valley minutiae), the observed EER is 14.14% and GAR (@FAR = 0.01%) is 61.61%. Hence, it can be inferred that the NBIS matcher can achieve better results with the ridge-minutiae template matching rather than with the corresponding valley-minutiae-based template matching. However, it is equally important to note that the combination of scores from both scenarios of matching reflects the complementary nature of the inverted fingerprints. The experiments with NBIS on PolyU database also reveal that the matching accuracy from the normal contactless fingerprints is at par with that of the inverted fingerprints. The performance improvement from the combination of two template representations is significant and validates our arguments in Sect. 3.

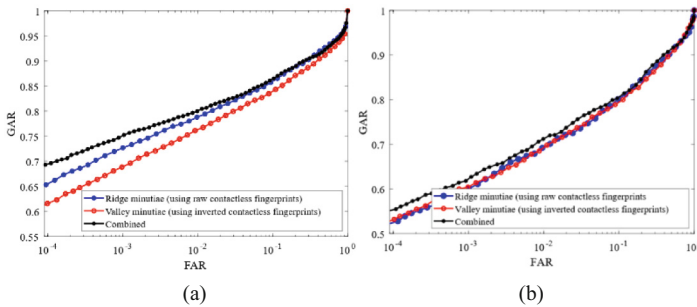


Fig. 6. ROC curves for ridge and valley minutiae matching and combined authentication with NBIS on (a) PolyU contactless fingerprint database and (b) benchmark contactless fingerprint database.

The performance of NBIS matcher on the contactless fingerprint database from [16] is illustrated using the ROC curves in Fig. 6 (b). From these values, it can be argued that matching the raw fingerprint images offers better EER, but worse GAR, when compared to those from matching the inverted fingerprint images. However, collaboration or the combination of two scores results in improved performance and can also be observed from the corresponding ROC curve.

4.2 MCC Matcher

The Minutiae Cylinder Code (MCC) [18] is widely considered a state-of-the-art fingerprint matcher and therefore also utilized in our work. The performance of the MCC matcher for PolyU Contactless database using the ROC curves is shown in Fig. 7 (a). From these ROC curves, it is apparent that the MCC matcher generates inferior performance for inverted contactless fingerprints as compared to the raw contactless images.

The ROC curves for the contactless benchmark database [16] using the MCC matcher are shown in Fig. 7 (b), which indicates that the inverted contactless fingerprints offer slightly better performance than the raw contactless fingerprints. It can be observed that a notable performance improvement can be achieved, both in EER and GAR (@FAR = 0.01%), with the collaboration of simultaneously recovered (refer to Sect. 4.4 for more details) individual match scores.

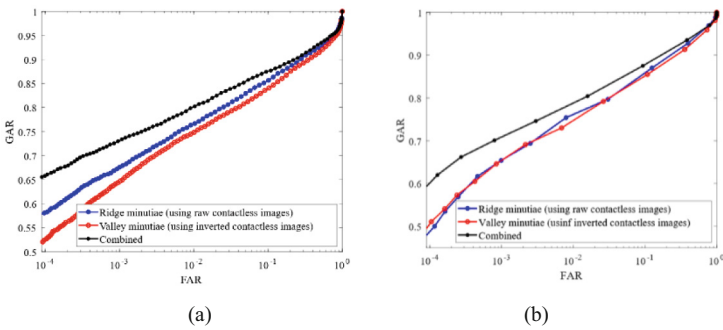


Fig. 7. ROC curves for ridge and valley minutiae matching and their combination, using MCC, for (a) PolyU contactless fingerprint database and (b) benchmark contactless fingerprint database.

4.3 COTS Matcher

A commercial off-the-shelf (COTS) fingerprint matcher, namely VeriFinger from Neurotechnology [12], was also employed to corroborate the usefulness of inverted fingerprints in improving contactless fingerprint recognition capabilities. This commercial matcher tool is known to perform excellently in minutiae extraction and matching. The ROC curves for the original (or ridge minutiae matching) and inverted (or valley minutiae matching) contactless fingerprint images of the PolyU database are illustrated in Fig. 8 (a). The verification experiments with inverted versions of contactless fingerprints have clearly outperformed those with the original contactless fingerprints. These values illustrated in Table 1 exhibit an improvement of 64.19% and 8.30%, respectively for EER and GAR (@FAR = 0.0001%), when compared to their counterparts in the experiments with original images.

The same set of verification experiments was also performed using a benchmark database [16] as for earlier cases. These experiments with inverted images show 38.05%

and 5.62% improvement in the respective performance metrics (Table 1), as compared with those from the experiments using the raw contactless fingerprint images. The corresponding ROC curves are illustrated in Fig. 8 (b). The experimental results on a popular COTS matcher, with *unknown* implementation details, also support the merit in the joint use of ridge and valley minutiae matching. The plausible reason for relatively higher EERs from this database [16] can be attributed to the *nature* of fingerprint images in this database. Fingerprints in this database have higher pose variations and a higher distance between the finger and sensor, as can also be observed from the image sample in Fig. 5 (b) in comparison to the fingerprint samples in the PolyU database, as shown in Fig. 5 (a).

It is reasonable to argue that when the matching performance from the inverted contactless fingerprint images is comparable or even *inferior* to those of their legacy counterparts, it can still provide complementary details to enhance the performance of the legacy contactless fingerprint images. This argument is also well supported by the ROC curves illustrated in Figs. 6–8 on public contactless fingerprint databases. The key focus of this paper is on investigating the polarity reversal or alterations observed in the common contactless fingerprint images or databases for more accurate performance. A careful observation of the ROC curves of Figs. 6–8 indicated that the fusion of scores from the raw and inverted contactless fingerprint images almost always results in noticeable performance improvements. Such improvements in performance metrics as a result of the simultaneous use of these two representations are summarized in Table 1. The results summarized in this table clearly indicate that the minutiae information furnished from the inverted contactless fingerprints can surely aid to improve the performance from the raw contactless fingerprints.

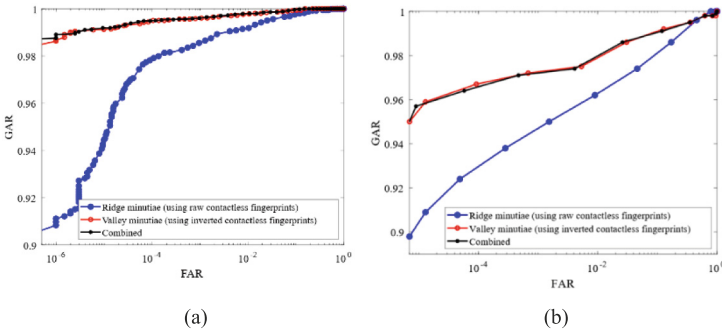


Fig. 8. ROC curves for ridge and valley minutiae matching and combined authentication using COTS: (a) PolyU contactless fingerprint database, (b) benchmark contactless fingerprint database.

Table 1. Summary of improvements in performance metric from simultaneous ridge and valley minutiae representations.

Matchers Databases		NBIS						MCC				COTS				
		Ridge	Valley	Combi med	Comparison w.r.t.		Ridge	Valley	Combi med	Comparison w.r.t.		Ridge	Valley	Combi med	Comparison w.r.t.	
					Ridge	Valley				Ridge	Valley				Ridge	Valley
PolyU Contactless to Contact-based Database	EER (%)	13.33	14.14	12.50	↑6.22%	↑11.39%	13.25	14.07	11.96	↑9.73%	↑14.99%	0.81	0.29	0.26	↑67.90%	↑10.34%
	GAR (%)	65.52	61.61	69.43	↑5.96%	↑12.69%	58.12	52.52	65.82	↑13.25%	↑25.32%	91.11	98.68	98.91	↑8.56%	↑0.23%
Multiview Contactless Fingerprint Database	EER (%)	16.25	17.71	16.00	↑1.54%	↑9.66%	12.74	12.66	10.91	↑14.36%	↑13.82%	3.60	2.23	1.45	↑59.72%	↑34.97%
	GAR (%)	52.49	53.13	55.35	↑5.44%	↑4.18%	48.82	50.65	60.21	↑23.33%	↑18.87%	90.43	95.52	95.74	↑5.87%	↑0.23%

4.4 Gray Level Transformation

The second part of the experimentation was performed to investigate the least-explored notion of effective color-to-grayscale transformation which can play a vital role in localizing key minutiae and enhancing performance for matching the contactless fingerprint images. One of the main advantages of contactless fingerprints is the enhanced user convenience associated with their acquisition with mobile phones which generates colored images. Therefore, effective grayscale representation is expected to enable accurate detection of legitimate minutiae and suppression of spurious minutiae during template extraction. By observing the improved performance of Luma grayscale fingerprints in facilitating appropriate detection of minutiae in Fig. 4, it can be concluded that this grayscale space can lead to the improved contactless fingerprint recognition performance. We, therefore, performed additional experiments to ascertain this possibility. The ROC curves for these contactless fingerprint verification experiments, using the ordinary and Luma grayscale images, are shown in Fig. 9. In this figure, the ROC curves corresponding to the ordinary and Luma grayscale images are illustrated using the solid and dashed lines, respectively. It is apparent from Fig. 9 that the performance using the Luma grayscale images is noticeably superior to the ordinary grayscale representations.

The verification experiments performed with the NBIS matcher show a notable improvement with Luma grayscale images as compared to the ordinary grayscale images. These ROC curves using the NBIS matcher are shown in Fig. 9 (a). The EERs achieved using the original and inverted Luma grayscale images are 6.70% and 6.81%, respectively. These values are improved by 49.73% and 51.84% when compared with the EERs achieved from NBIS matcher for the ordinary grayscale fingerprint images.

The GARs (@FAR = 0.01%) for raw and inverted Luma grayscale images, using NBIS matcher, are observed to be 78.82% and 78.06%, respectively. On the other hand, the GAR values for raw and inverted ordinary grayscale images are 65.52% and 61.61%, respectively. The second matcher employed in our work, or MCC, also illustrates the significant improvement in the verification performance from the Luma grayscale images. ROC curves corresponding to MCC are shown in Fig. 9 (b). The EERs of original and inverted Luma grayscale contactless fingerprints is 8.19% and 8.28%, respectively. These values are 38.18% and 41.15% higher than their counterparts from the experiments performed with ordinary grayscale images. Similarly, the GARs (@FAR = 0.01%) achieved

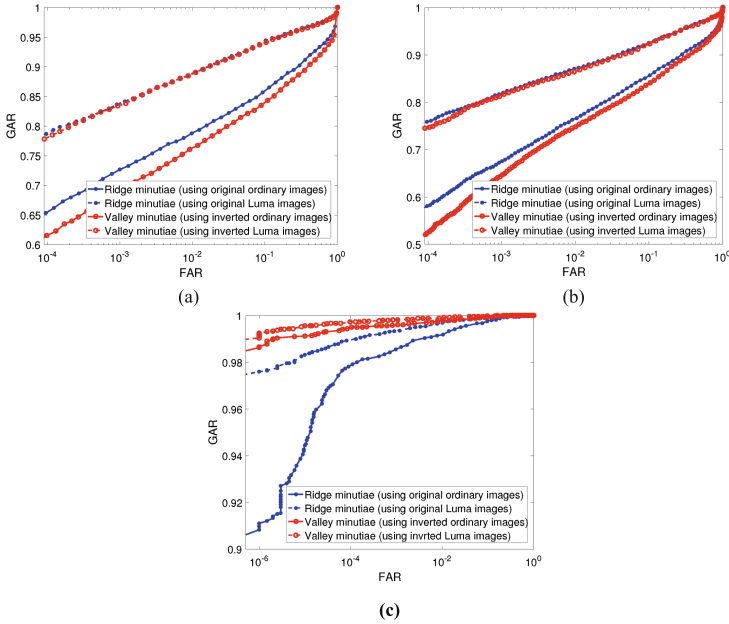


Fig. 9. Comparative ROC curves for original and inverted contactless fingerprints using ordinary and Luma grayscale images, (a) using NBIS, (b) using MCC, and (c) using a COTS matcher.

with original and inverted Luma grayscale fingerprints are 75.94% and 74.67%, respectively. These values illustrate remarkable improvements over their counterparts with the ordinary grayscale images, or 30.66% and 42.17%, respectively.

Lastly, the experiments using the COTS matcher with raw Luma grayscale images achieved an EER and GAR (@FAR = 0.0001%) of 0.45% and 97.6%, respectively. The same metrics for the ordinary grayscale images are 0.81% and 91.11%, respectively. Therefore, it can be inferred that the Luma grayscale images are outperforming ordinary grayscale images with 44.44% and 7.12% improvements in EER and GAR, respectively. On the other hand, the performance metrics for inverted Luma grayscale images were 0.17% (EER) and 99.25%, respectively. For inverted ordinary grayscale images, the corresponding EER and GAR values were 0.29% and 98.68%, respectively. Therefore, the percentage improvements for inverted grayscale images are 41.37% and 0.57%, respectively, for EER and GAR. The ROC curves for the COTS matcher with Luma grayscale images are shown in Fig. 9 (c). These outperforming results from Luma grayscale fingerprints clearly indicate that a more effective representation of colored images in the grayscale space can be used to achieve significantly enhanced contactless fingerprint matching.

5 Conclusions and Further Work

This paper presented a detailed analysis of contactless fingerprint images, on the minutiae representation and matching, to achieve significant performance improvements using popular fingerprint matchers. The experimental results presented in this paper on two

public contactless fingerprint databases [13, 16], using MCC [18], NBIS [17], and COTS [12] matchers consistently illustrate the significant improvement in the performance over the conventional approaches in the literature. The work detailed in this paper indicates that the grayscale polarity alterations under ambient or indoor illumination are frequent and should be incorporated into the design of any effective matching strategy to utilize the full potential of contactless fingerprint images. The absence of contact between the sensor platen and the human finger often leads to varying illuminated areas in the fingerprint images. Therefore, the ridges in such images are often rendered darker and brighter in different regions. This is unlike the images from contact-based fingerprint sensors, in which ridges are always rendered as darker and valleys as brighter, largely due to the frustrated total internal reflection. Simultaneous recovery and use of ridge and valley minutiae in contactless fingerprints can enable superior matching capabilities by utilizing the full potential of inverted fingerprints, either individually or in combination with raw fingerprints.

The above arguments are evaluated using three popular fingerprint matchers, namely NBIS, MCC and COTS. On the one hand, NBIS and MCC yield comparable performance for both ridge (raw fingerprints) and valley (inverted fingerprints) minutiae matching approaches. However, the combination of scores from raw and inverted fingerprints achieves significantly improved performance with both of these matchers as well. Significant performance improvement in EER and ROC or GAR due to such a combination can validate our arguments. On the other hand, the COTS matcher clearly illustrates improved performance with inverted fingerprints alone, which can further be improved through the score combination. This work also considered the effective conversion of contactless color fingerprint representation to its grayscale representation, which has received almost nil attention in the literature. In this context, a more diversified grayscale representation, namely Luma grayscale, was introduced with quite encouraging results. Such effective grayscale representation is quite valuable for improving the performance of contactless fingerprints acquired using the widely popular color cameras on mobile phones. Contactless fingerprint acquisition for a range of mobile applications requires its detection under complex or moving backgrounds and under involuntary finger motions. Such detection can be achieved using a range of lightweight detectors and is suggested for further work. More advanced minutiae detection methods [26] can also benefit from the simultaneous use of ridge-valley minutiae labeling, detection, and their use for enhanced matching of contactless fingerprints and is also part of much-needed further work in this area.

References

1. Yin, X., Zhu, Y., Hu, J.: A survey on 2D and 3D contactless fingerprint biometrics: a taxonomy, review, and future directions. *IEEE Open J. Comput. Soc.* **2**, 370–381 (2021)
2. Labati, R.D., Genovese, A., Piuri, V., Scotti, F.: Contactless fingerprint recognition: a neural approach for perspective and rotation effects reduction. In: *Proceedings IEEE Workshop Computational Intelligent in Biometrics and Identity Management (CIBIM)*, pp. 22–30 (2013)
3. Hiew, B.Y., Teoh, A.B., Pang, Y.-H.: Touchless fingerprint recognition system. In: *Proceedings IEEE Workshop on Automatic Identification Advanced Technologies, AutoID07*, pp. 24–29 (2007)

4. Sankaran, A., Malhotra, A., Mittal, A., Vatsa, M., Singh, R.: On smartphone camera-based fingerphoto authentication. *Proc. BTAS* **2015**, 1–7 (2015)
5. Priesnitz, J., Rathgeb, C., Buchmann, N., Busch, C., Margraf, M.: An overview of touchless 2D fingerprint recognition. *EURASIP J. Image Video Process.* **2021**(1), 8 (2021)
6. Kumar, A.: Contactless 3D Fingerprint Identification. Springer, Cham (2018). <https://doi.org/10.1007/978-3-319-67681-4>
7. Parziale, G.: Touchless fingerprinting technology. In: Ratha, N.K., Govindaraju, V. (eds.) *Advances in Biometrics*, pp. 25–48. Springer, London (2008). https://doi.org/10.1007/978-1-84628-921-7_2
8. ISO/IEC. IS 29794-1:2016, information technology biometric image sample quality Part 1: Framework. ISO Standard, January 2016
9. Ericson, L., Shine, S.: Evaluation of contactless versus contact fingerprint data, phase 2 (version 1.1), Report No. 249552, I. ManTech Advanced Systems International, DOJ Office of Justice Programs (2015)
10. Noh, D., Choi, H., Kim, J.: Touchless sensor capturing five fingerprint images by one rotating camera. *Opt. Eng.* **50**, 113202 (2011)
11. Libert, J.: Guidance for evaluating contactless fingerprint acquisition devices. NIST 500-305 (2018)
12. VERIFINGER SDK (2024). <http://www.neurotechnology.com/verifinger.html>
13. The Hong Kong Polytechnic University Contactless 2D to Contact-based 2D Fingerprint Images Database Version 1.0 (2024). <http://www4.comp.polyu.edu.hk/~csajaykr/fingerprint.htm>
14. Kanan, C., Cottrell, G.W.: Color-to-grayscale: does the method matter in image recognition? *PLoS One* **7**(1) (2012)
15. Gvozden, G., Grgic, S., Grgic, M.: Blind image sharpness assessment based on local contrast map statistics. *J. Vis. Commun. Image Represent.* **50**, 145–158 (2018)
16. Zhou, W., Hu, J., Petersen, I., Wang, S., Bennamoun, M.: A benchmark 3D fingerprint database. In: 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 935–940 (2014)
17. Watson, C.I., et al.: User’s guide to NIST biometric image software (NBIS) (2007)
18. Cappelli, R., Ferrara, M., Maltoni, D.: Minutia cylinder-code a new representation and matching technique for fingerprint recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(12), 2128–2141 (2010)
19. Stein, C., Nickel, C., Busch, C.: Fingerphoto recognition with smartphone cameras. *Proc. BIOSIG* **2012**, 1–12 (2012)
20. Tan, H., Kumar, A.: Towards more accurate contactless fingerprint minutiae extraction and pose-invariant matching. *IEEE Trans. Info. Forensics Secur.*, 3924–3937 (2020)
21. Dabouei, A., Soleymani, S., Dawson, J., Nasrabadi, N.: Deep contactless fingerprint unwarping. In: *Proceedings ICB 2019*, Greece, June 2018
22. Labati, R.D., Genovese, A., Piuri, V., Scotti, F.: Toward unconstrained fingerprint recognition: a fully touchless 3-D system based on two views on the move. *IEEE Trans. Syst. Man Cybern. Syst.* **46**, 202–219 (2015)
23. Lee, D., Choi, K., Choi, H., Kim, J.: Recognizable-image selection for fingerprint recognition with a mobile-device camera. *IEEE Trans. Syst. Man Cybern. Part B* **38**, 233–243 (2008)
24. Khalil, M.S., Kurniawan, F.: A review of fingerprint preprocessing using a mobile phone. In: *Proceedings International Conference on Wavelet Analysis and Pattern Recognition*, WAPR (2012)

25. Vyas, R., Kumar, A.: A collaborative approach using ridge-valley minutiae for more accurate contactless fingerprint identification. arXiv preprint [arXiv:1909.06045](https://arxiv.org/abs/1909.06045) (2019)
26. Feng, Y., Kumar, A.: Detecting locally, patching globally: an end-to-end framework for high speed and accurate detection of fingerprint minutiae. *IEEE Trans. Inf. Forensics Secur.* **18**, 1720–1733 (2023)



A Generative Method for Finger Knuckle Print Recognition

Yuqi Wang¹, Bob Zhang¹✉, Shuyi Li², and Hao Yang¹

¹ PAMI Research Group, Department of Computer and Information Science,
University of Macau, Macau SAR, China
{yc37500,bobzhang,mc36514}@um.edu.mo

² Department of Information, Beijing University of Technology,
Beijing 100124, China
yb97443@um.edu.mo

Abstract. Finger knuckle print (FKP), known as a biological feature, has drawn great research attention in the field of biometrics recognition. That being said, the development of finger knuckle print recognition is still limited by the lack of data and the difficulties in the extraction of its region of interest (ROI). To resolve these issues, this paper proposes a generative method based on the simulation of the curve distribution of a finger knuckle print to generate reasonable masks of finger knuckle points. Following this, generative adversarial networks (GANs) are applied with the masks to generate the pseudo finger knuckle point images. This method can provide large amounts of training data for recognition as well as directly supplying the region of interest. Experimental results show that the generated finger knuckle print examples can effectively augment the training data for the recognition model.

Keywords: Biometrics Recognition · Finger Knuckle Print · Generative Method

1 Introduction

Biometrics recognition techniques [21] have become an important authentication method due to its effectiveness, safety, and convenience. Various biometrics have been investigated in the past decades, which can be mainly divided into behavioral characteristics and physiological characteristics [10]. The former contains signature [13], voice [2], gesture [23], and keystroke [4]. The latter is extracted from human biological features such as iris [33], veins [14], fingerprints [20], palmprints [37], and DNAs [27]. Biometrics recognition is widely used in many security-sensitive scenarios with years of development. For example, access control [26], ID cards [5], and online payment [31].

This work was partially supported by the Science and Technology Development Fund, Macao S.A.R (FDCT) 0028/2023/RIA1, and in part by the National Natural Science Foundation of China Project (62306021).

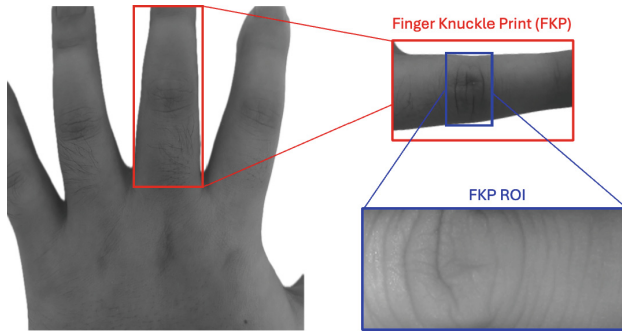


Fig. 1. An example of FKP and its ROI.

Finger knuckle print (FKP), as an emerging type of biometric [34], has aroused great interest from researchers. FKP usually refers to the skin texture at the first joint near the palm of the finger, as shown in Fig. 1. It is highly variable among individuals, not easily worn out, forged or stolen, and remains highly consistent after long-term use and multiple acquisitions. Besides this, the collection of FKPs does not require high-quality equipment and strict environmental limits, where users are more receptive to the contactless collection methods [36].

With these valuable properties, FKP recognition has been rapidly driven by the development of machine learning and deep learning techniques. Many algorithms such as SVM [24], PCA [25], and KNN [29] have been applied to FKP recognition, achieving good performances. Convolutional Neural Networks (CNNs), as one of the most effective deep learning methods in image-related tasks, have demonstrated great advancement in many biometric recognition applications. Without artificially formulated classification rules, CNNs can discover the underlying pattern from the given data distribution [17].

The performance of CNNs is highly related to the training data [19]. However, the public dataset is extremely deficient in FKP recognition due to the lack of collection and other privacy concerns. This situation has strictly limited the application of CNNs to FKP recognition. In addition, the extraction of the ROI remains another problem, where the pre-processing of FKP images is also time consuming [36].

Aiming at the above issues, this paper proposes a generative data augmentation method to provide large-scale high-quality pseudo FKP ROI images for deep learning based recognition models. To accomplish this, we first analyse the distribution of the FKP skin texture curves. Afterwards, we propose a FKP mask generation method using a series of ovals and other curves based on the analysis of real FKP images. Finally the masks are input into a mask-guided GAN to generate the pseudo FKP images. We used ResNet50 [11] in the experiments to verify the effectiveness of our generative FKP examples. The contributions of this work are summarized as follows:

- We meticulously analysed the distribution of the skin texture curves of the FKP region and quantitatively described the pattern of the FKP lines, providing statistical support for further research in FKP generation and recognition.
- We proposed a generative data augmentation method providing sufficient large-scale high-quality pseudo FKP images for training recognition models.
- Experimental results show that our method can effectively improve the performance of FKP recognition via CNNs.

2 Related Work

FKP Recognition. The research of FKP recognition started later than that of fingerprints and palmprints. Woodward et al. [34] made an attempt at the application of FKPs for personal authentication. They proposed a similarity comparison method by calculating the curvature surface representation of the fingers using the hand 3D range images. Subsequently, Kumar et al. [15] proposed a feature extraction and personal identification method based on the finger-back surface images using subspace analysis methods. However, these efforts on validating the effectiveness of outer-finger textures were not developed into an efficient recognition system. Zhang et al. [36] first proposed a fully-processed FKP recognition system including FKP collection, feature extraction, feature coding, and matching. Hammouche et al. [9] proposed an FKP identification method based on phase congruency with a Gabor filter bank. Zhang et al. [35] implemented an effective FKP feature extraction approach through a novel computing framework. They proposed the use of three representative local features: orientation, phase, and congruency and applied them to calculate all characteristics. Muthukumar et al. [24] developed a FKP identification framework using a SVM classifier with Gabor feature.

With the development of deep learning, CNNs have shown excellent performance in many image-related tasks. Various models were introduced for FKP recognition and achieved satisfying results. Zohrevand et al. [39] applied a simple 5-layer CNN model for FKP recognition and achieved over 99% accuracy on the public PolyU-FKP dataset. Chalabi et al. [6] further developed the PCANet-SVM to a recognition framework based on score level fusion of the major and minor FKPs. Hamidi et al. [8] included FKP in their multimodal identification system with two pre-trained VGG-16 and VGG-19 models. Fei et al. [7] proposed a feature learning method for encoded discriminative direction features for FKP recognition and outperformed many algorithms.

As a data-driven deep learning method, the performance of CNNs is highly related to the dataset quality. However, few public datasets are available due to the difficulty of FKP collection and people’s privacy concerns. To the best of our knowledge, IIT Delhi FKP Dataset [16] and PolyU-FKP Dataset [30] are the only two public datasets.

Generative Methods. GANs are considered as one of the mainstream methods of generative artificial intelligence, which have shown significant application

potential in a wide range of computer vision tasks such as image synthesis [1], image translation [38], and representation disentangling [3]. Isola et al. [12] first proposed a general framework pix2pix to implement image-to-image translation. Wang et al. [32] developed the pix2pix with multi-scale generators and discriminators to deal with the high-resolution images, named pix2pixHD. Traditional GANs generated the images with no limitation. In spite of this, for some cases, the generated images are expected to follow a specific pattern. Mirza and Osindero [22] first proposed the conditional GAN that can generate images from masks and tested its performance on MNIST digits generation. Succeeding this, various mask guided GANs were developed for many tasks such as image manipulation, portrait editing, and object removal.

In our work, we expect the generated FKP images to follow our artificially designed patterns. Hence, the mask-guided GAN is the optimal choice, which can generate the images not only similar to real-world images, but also follow the designed pattern.

3 Methodology

In this section we will present the details of our generative method. As shown in Fig. 2, our framework contains three modules: FKP Curve Distribution Analysis, FKP Mask Construction, and FKP ROI Image Generation.

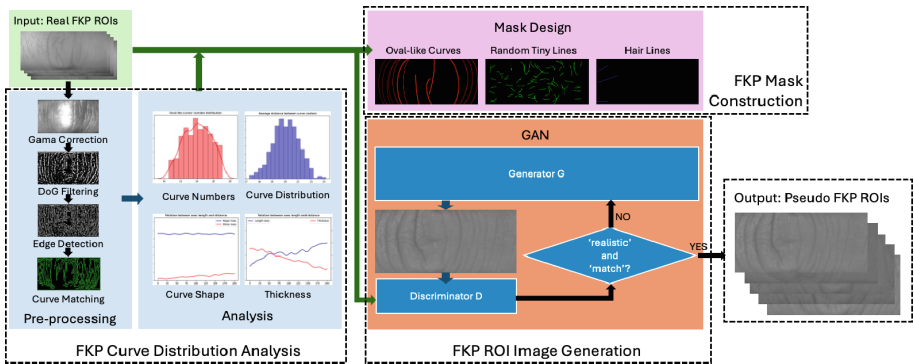


Fig. 2. The overall framework of our method.

3.1 FKP Curve Distribution Analysis

FKP, as a biometric, is highly variable among individuals, yet keeps some underlying patterns in common, which is a vital clue leading to the design of the pseudo FKP generation algorithm. In this paper, we take the CAUC-FKP dataset as reference for real-world FKPs. This will be further introduced in Sect. 4.1. Similar to fingerprints and palmprints, the key element in FKPs is also the texture lines. However, the curves of FKPs are not as regular as the fingerprints that have

loops, ridges and valleys, or palmprints including principle lines and wrinkles. Therefore, the extraction of the curves should be done first before analysis.

Inspired by the work of Tan et al. [28] in face recognition, we apply a similar strategy to enhance the curves in FKP images for better extraction. As shown in Fig. 3, this pre-processing pipeline contains four parts: Gamma Correction, Difference of Gaussian (DoG) Filtering, Canny Edge Detection, and Curve Matching.

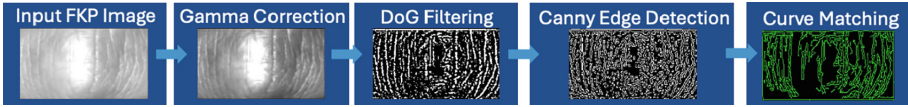


Fig. 3. The curve extraction pipeline.

The FKP gray-scale images are first input into the Gamma Correction module, which changes the pixel intensity from I to I^γ through a non-linear gamma function $I^\gamma = (I + \epsilon)^\gamma$. Here, γ is a hyper-parameter controlling the strength of the gamma correction. This operation can adjust the image contrast automatically and enrich the detail level of the bright and dark regions of an image through setting different γ values. Next, a band-pass DoG filter is applied after the Gamma Correction. Specifically, the DoG filter represents the difference between two images filtered by two different low-pass filters, one of which removes the low-frequency components corresponding to the homogeneous region, and another removing the high-frequency noises:

$$\text{DoG} * f(x, y) = (G_{\sigma_1} - G_{\sigma_2}) * f(x, y) \quad (1)$$

where G_{σ_1} and G_{σ_2} represent two different Gaussian filters. The DoG filter can be formalized as:

$$\text{DoG} = \frac{1}{\sqrt{2\pi}} \left(\frac{1}{\sigma_1} e^{-\frac{x^2+y^2}{2\sigma_1^2}} - \frac{1}{\sigma_2} e^{-\frac{x^2+y^2}{2\sigma_2^2}} \right). \quad (2)$$

After DoG Filtering, the FKP curves are more visible for the subsequent Canny Edge Detection. This operation is applied to extract the curves in the images. It is achieved through a series of image processing operations including Gaussian smoothing, gradient calculation, non-maxima suppression, and dual-threshold boundary tracking. After the Canny Edge Detection, the FKP curves will be transformed into white lines in the binary images. The last step of the line extraction is the Curve Matching. This step is achieved through a simple connected-component-analysis algorithm, since the results of Canny Detection are all binary images. For the crossing lines, which might be counted as one line, we set the pixels at the intersection to a different value, hence separating the connected components.

Recall that our purpose is to analyse the patterns of the lines, including the curve types, numbers of each type, length and thickness, position distribution and orientation. These features provide an important reference when designing the FKP mask construction algorithm. In the following sub-sections, we will show the analysis in details along with the design principles of our algorithm.

3.2 FKP Mask Construction

The mask construction is the core component in our framework, since it provides the guideline for the subsequent GAN and directly influences the quality of the generated images. Guided by the above analysis on the FKP curve pattern, the lines in the masks are basically divided into two types: oval-like curves and random tiny lines. In some cases, fine hairs also appear in the FKPs. Three types of lines are illustrated in Fig. 4.

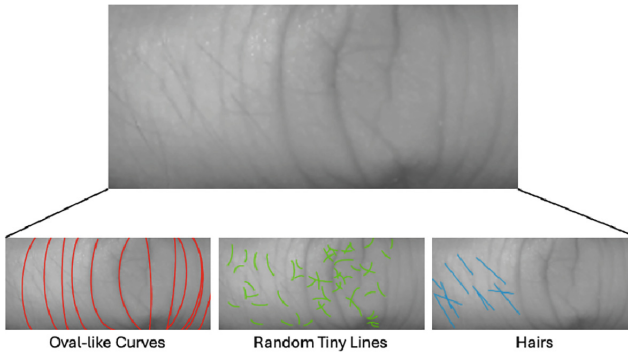


Fig. 4. Examples of the curve types in FKPs.

Oval-Like Curves. As shown in Fig. 4, the curves with large thickness and length construct the basic sketch of FKPs. These curves contribute the most to the recognition performance. Figure 5b -(1) depicts the total number of the oval-like curves in the given 585 FKP examples from CAUC-FKP. It indicates a Gaussian distribution $\mathcal{N}(14, 4)$, whose mean is 14 and variance is 4. Hence, the number of these oval-like curves in the constructed masks should follow the same distribution.

Following this, we consider the shape of these curves. As the name suggests, we use the partial ellipse with different eccentricities to represent these curves. As illustrated in Fig. 5a, five parameters are applied to control the shape of the ellipses, which are introduced as follows.

1. The center determines the location of the curves. The study on all FKP examples indicate that the curves are roughly uniformly distributed within the ROI, while the density at the left and right edges are slightly larger than that in the middle, as shown in Fig. 5b - (2).

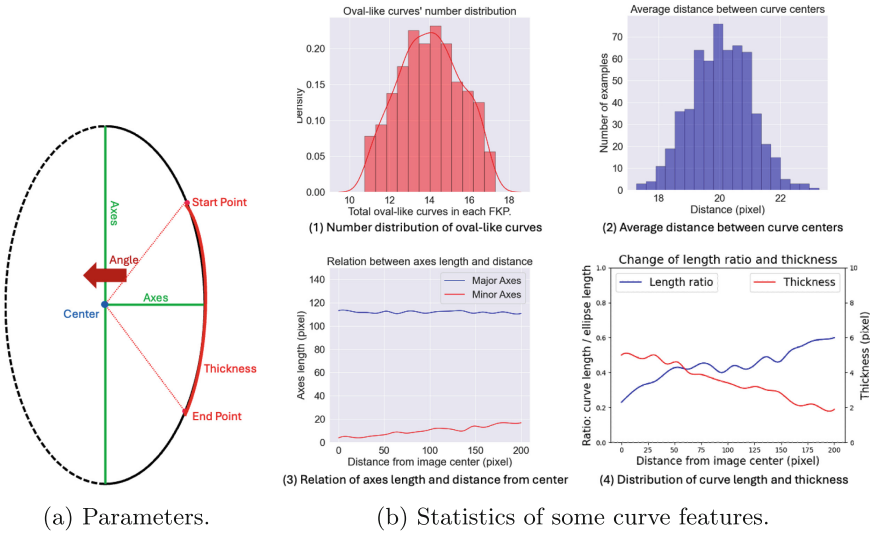


Fig. 5. The parameters used in the oval-like curve construction and their statistics.

2. The axes controls the shape of the curves. The ellipse eccentricity is determined by the ratio of the major and minor axes. Figure 5a shows that the greater the eccentricity, the smaller the curvature. In general, the curves nearer to the knuckle edge have a larger curvature. Curves at the knuckle center are almost straight lines, as shown in Fig. 5b - (3).

3. The angle controls the orientation of the curves. Basically, the partial ellipse curves are left-opened on the right half of the knuckle and right-opened on the left half. Besides this, the curves are not always vertical to the image orientation and have an angle shift from -25° to 25° .

4. The start-end points control the length of the curves. The partial ellipses are used to fit the curves, such that the length of the curve varies from 1/3 to 1/2 of the half-ellipse. As shown in Fig. 5b - (4), the 1/3-half-ellipse appears more at the knuckle center and the 1/2-half-ellipse often represents the curves at the left and right edges.

5. The last parameter is the thickness. Figure 5b - (4) shows a regular pattern, where curves closer to the knuckle center have a larger thickness.

Finally, to increase the randomness, an image liquify operation is used to impose the curves, which can better simulate the realistic features.

Random Tiny Lines. In addition to the basic oval-like lines, there are many random tiny lines distributed on the entire knuckle region. The contribution of these lines to the recognition performance cannot be ignored either. Through the observation of real FKP examples, we use the Bezier curves connecting multiple points to simulate those lines. Specifically, we randomly set some points which

are divided into inner points and outer points. The inner points are close to the knuckle center and the Bezier curves connecting each point-pair form a closed shape. The outer points can only be the end of the curves, and there is no connection between these points. This setting better simulates the spider-web-like distribution pattern of the random tiny lines.

Fine Hairs. Fine hairs exist in some of the FKP examples. They can be simulated by some simple straight thin lines that only appears in the specific half-region of the knuckle.

Algorithm 1. Mask Construction Algorithm

Input: Mask height h , mask weight w , mask number N

Output: FKP Mask Set \mathcal{M} .

- 1: **for** $i \leq N$ **do**
 - 2: Create empty mask $M_i = [0]_{w \times h}$
 - 3: Randomize:
 - Total curve number $n_i \sim N(14, 4)$
 - Center point set $C = \{(x, y) | x \sim U(d, w - d), y_k \sim N(h/2, \sigma^2)\}_{n_i}$
 - Axes set $A = \{(a, b) | a \sim N(3h/4, \sigma^2), b \sim N(a/3, \sigma^2)\}_{n_i}$
 - Angle set $\Theta = \{\theta\}_{n_i}$, where $\theta \sim N(0, \sigma^2)$
 - Start-end point set $R = \{r\}_{n_i}$, where $r \sim U(1/6, 1/4)$
 - Thickness set $T = \{t\}_{n_i}$, where $t \sim U(t_1, t_2)$
 - 4: Sort the elements in sets C, A, Θ, R and T .
 - 5: Construct the curves using `cv2.ellipse` and obtain $M_i^1 \leftarrow M_i$
 - 6: Liquifaction: $M_i^a \leftarrow M_i^1$
 - 7: Randomize point set $P = \{(x, y)\}$
 - 8: Calculate convex hull and:
 - $P_{out} = convexHull(P), P_{in} = P - P_{out}$
 - 9: Connect each point of P_{in} and P_{out} with Bezier curves and obtain M_i^2 .
 - 10: Liquifaction: $M_i^b \leftarrow M_i^2$
 - 11: Randomize straight lines on the left part of M_i and obtain M_i^c
 - 12: **end for**
 - 13: $\mathcal{M} \leftarrow \{M_i = (M_i^a, M_i^b, M_i^c)\}_{i=1}^N$
-

Algorithm. Following the above patterns, our mask generation algorithm is designed as Algorithm 1. For each mask, step 1 to step 6 correspond to the oval-like curve mask M_i^a generation, and step 7 to step 10 correspond to the random tiny line mask M_i^b generation. The fine hair simulation is in step 11 to generate the mask M_i^c . All masks are respectively calculated and stored as a channel of masks M_i , that is $M_i = (M_i^a, M_i^b, M_i^c)$. The purpose for separate calculations is for subsequent mask-guided GAN generation. Through the randomization in each iteration, we can finally obtain the mask set $\mathcal{M} = \{M_i = (M_i^a, M_i^b, M_i^c)\}_{i=1}^N$ containing N masks.

3.3 FKP ROI Image Generation

After obtaining the FKP masks, we use the mask-guided GAN to generate our target FKP ROI images. This conditional GAN can generate the target images with our expected patterns. In our GAN model, the three masks constructed in the previous process are applied to guide the generation tasks. Figure 6 illustrates the basic architecture of our GAN model.

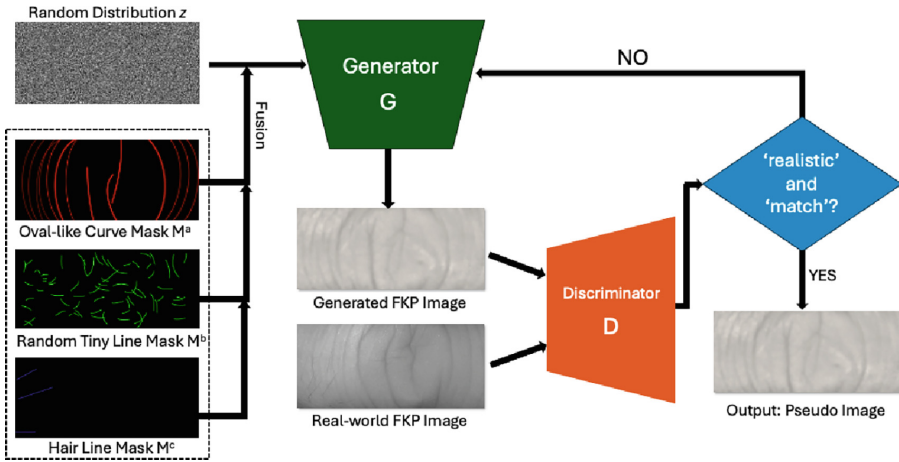


Fig. 6. The architecture of our GAN model.

In general, the model takes a normal distribution z as the input, along with a fusion mask M as the condition. M is fused from the three sub-masks M^a , M^b , and M^c constructed by our previous algorithm through assigning different weights on each sub-mask. Afterwards, generator G outputs a generated image $x = G(z, M)$. In this case, the mask M and the generated image x are input into the discriminator D , where there is a scalar to judge whether the image x is realistic and matched with condition M . The generator G and the discriminator D are alternatively optimized until the output meets the requirement.

4 Experiments

In this section, we conduct several experiments to verify the effectiveness of our method. First, we use our generative method to construct a pseudo FKP ROI dataset based on real-world FKP data. Following this, we train a ResNet50 model with datasets, where the pseudo images are mixed with real-world images under different ratios. The model is tested on the test set constructed by the same manner.

4.1 Datasets

In our work, two real-world datasets are utilized as the reference for FKP mask design, forming the training and test sets.

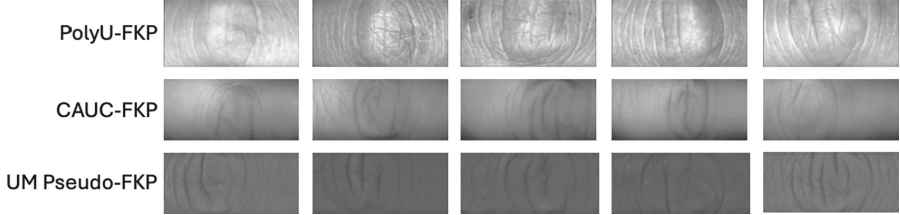


Fig. 7. Some examples of the three datasets.

PolyU-FKP Dataset contains 7,920 FKP images collected from 165 individuals [30]. Each one was required to provide 12 images of the index and middle finger knuckles. This dataset is considered as one of the largest public FKP datasets. In this paper, we extracted the ROI with a size of 220×110 pixels. Some of the examples are shown in Fig. 7.

CAUC-FKP Dataset was collected by the authors of [18] from CAUC. It contains 5,850 FKP images coming from 195 individuals, each of which provided 10 examples of the index, middle, and ring finger knuckles, respectively. We cropped the images into 200×90 pixels for the ROIs in this work. Figure 7 shows some of the examples.

UM Pseudo-FKP Dataset is the dataset generated by our method based on the CAUC-FKP dataset. Up to now, there are 5,000 FKP ROI images from 500 identities. The size of each ROI image is 200×90 pixels, the same as that in the CAUC-FKP dataset. Figure 7 shows some of our generated examples.

4.2 Experimental Settings

We used a ResNet50 model as the benchmark to test the effectiveness of our method. The model was trained on ten sub-datasets respectively. The sub-datasets were constructed by the rules shown in Table 1, where 0% represents an all-real-examples dataset and 100% represents an all-pseudo-examples dataset. Each sub-dataset contained 5,000 examples from 500 IDs, where different finger knuckles from the same person are considered as different IDs for convenience. In the mixed sub-datasets, all examples were randomly selected from the original dataset. The ratio of training, validation, and test was 7:2:1.

Table 1. List of our sub-datasets.

Source	Name	Ratio of Pseudo FKP	Source	Name	Ratio of Pseudo FKP
PolyU + Pseudo	PP-0	0%	CAUC + Pseudo	CP-0	0%
	PP-25	25%		CP-25	25%
	PP-50	50%		CP-50	50%
	PP-75	75%		CP-75	75%
	PP-100	100%		CP-100	100%

The model was trained for 500 epoch on each dataset using SGD with momentum. The training device was an Ubuntu server with NVIDIA RTX 4090 GPU. After training, we obtained ten models with different parameters. Afterwards, the models were tested on the test sets respectively and the accuracy was recorded as the metric of performance.

4.3 Results

Table 2 and Table 3 show the accuracy of the model on different datasets. Compared with the models trained on all-real-world datasets (PP-0 and CP-0), the accuracy of the models trained on our all-pseudo datasets (PP-100 and CP-100) only drops slightly. For the PolyU-FKP subset, the accuracy drops 3.8% from 91.8% to 88.0%, while it drops only 1% from 95.4% to 94.4% on the CAUC-FKP subset. The reason is that our pseudo images are generated based on the CAUC-FKP examples. There is a distribution shift between the PolyU-FKP and CAUC-FKP datasets, and the brightness and contrast in PolyU-FKP examples are more complex. The results of PP-100 and CP-100 reveal that our generated data can be used as a substitution for real-world data when it is insufficient.

The models trained on mixed datasets also achieved a competitive performance. For example, PP-50 and CP-50 obtained the highest average accuracy on five test sets among all models. In practice, the pseudo images can also be mixed with the real-world images to provide data support.

Table 2. Accuracy of different models on the test sets (PolyU-FKP + Pseudo-FKP).

Model	Ratio of pseudo images in test set					Average
	0%	25%	50%	75%	100%	
PP-0	91.8%	90.8%	89.6%	89.4%	85.2%	89.36%
PP-25	92.0%	91.6%	90.8%	91.0%	90.4%	91.16%
PP-50	92.4%	92.6%	92.4%	92.0%	91.2%	92.12%
PP-75	91.6%	91.0%	92.0%	91.8%	92.4%	91.76%
PP-100	88.0%	89.6%	90.8%	92.6%	94.2%	91.04%

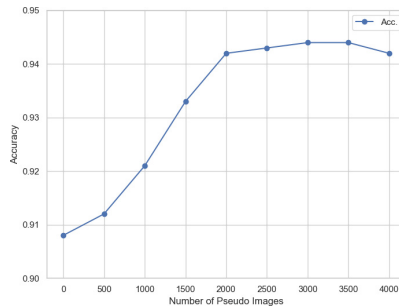
Table 3. Accuracy of different models on the test sets (CAUC-FKP + Pseudo-FKP).

Model	Ratio of pseudo images in test set					Average
	0%	25%	50%	75%	100%	
CP-0	95.4%	93.8%	93.2%	92.8%	90.6%	93.16%
CP-25	94.6%	95.0%	94.8%	93.8%	93.8%	94.40%
CP-50	95.0%	95.2%	95.2%	94.6%	94.4%	94.88%
CP-75	94.8%	94.4%	94.8%	93.6%	93.4%	94.20%
CP-100	94.4%	95.0%	95.8%	96.0%	96.2%	95.48%

4.4 Data Augmentation

In the previous experiments, all training sets contain the same amount of examples. We conducted another experiment to further evaluate the effectiveness of our generative method as a data augmentation technique. Specifically, we first trained a ResNet50 model on the entire CAUC-FKP training set and recorded its accuracy on the test set, 25% examples of which are from UM Pseudo-FKP and the rest are from CAUC-FKP. Then, we added 500 pseudo examples to the training set and trained the model from scratch. We repeated this operation 8 times and the training set was enlarged to 8,000 images, 4,000 of which are from CAUC-FKP and the remaining from UM Pseudo-FKP. The accuracy change is recorded in Fig. 8.

When we increased the amount of training examples by adding pseudo images, the model performance improved accordingly. This result reveals that our generative method can provide an effective data augmentation in the model training process. Nonetheless, it is worth noting that an unlimited increase in pseudo data is not always conducive to model training. In fact, when the input pseudo images is 2,000, the improvement on the model accuracy has reached marginal effects. Continuing to increase the amount of pseudo data will only cause the performance of the model to deteriorate, making the model's learnt distribution shift from the real-world data distribution.

**Fig. 8.** The accuracy change with different numbers of pseudo images in the training set.

5 Conclusion

In this paper, we scrupulously analysed the patterns of FKPs, including its texture curves distribution and shape. From this, we manually designed a mask construction algorithm based on the above analysis to guide the generation of the pseudo FKP ROI images by mask-guided GAN. Several tests on the ResNet50 model using real-pseudo mixed datasets illustrated the effectiveness of our method as a data augmentation technique. The proposed work provides a solution to the rigorous problem in the field of FKP recognition research, where public datasets are severely deficient. As part of our future work, we will focus on the security of the generative method's application to biometrics and explore more possibilities for other hand modality features.

Acknowledgements. This work was partially supported by the Science and Technology Development Fund, Macao S.A.R (FDCT) 0028/2023/RIA1, and in part by the National Natural Science Foundation of China Project (62306021).

References


1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN (2017). arXiv preprint [arXiv:1701.07875](https://arxiv.org/abs/1701.07875)
2. Bai, Z., Zhang, X.L.: Speaker recognition based on deep learning: an overview. *Neural Netw.*, 65–99 (2021)
3. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Towards openset identity preserving face synthesis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6713–6722 (2017)
4. Baynath, P., Soyjaudah, K.S., Khan, M.H.M.: Keystroke recognition using neural network. In: *5th International Symposium on Computational and Business Intelligence (ISCBI)*, pp. 86–90 (2017)
5. Benalcazar, D., Tapia, J.E., Gonzalez, S., Busch, C.: Synthetic id card image generation for improving presentation attack detection. *IEEE Trans. Inf. Forensics Secur.* **18**, 1814–1824 (2023)
6. Chalabi, N., Attia, A., Bouziane, A.: Multimodal finger dorsal knuckle major and minor print recognition system based on PCANet deep learning. *ICTACT J. Image Video Process* **10**, 2153–2158 (2020)
7. Fei, L., Zhang, B., Teng, S., Zeng, A., Tian, C., Zhang, W.: Learning discriminative finger-knuckle-print descriptor. In: *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019)
8. Hamidi, A., Khemgani, S., Bensid, K.: Transfer learning using VGG based on deep convolutional neural network for finger-knuckle-print recognition. In: *Proceedings of the 2nd International Conference on Computer Science's Complex Systems and Their Applications* (2021)
9. Hammouche, R., Attia, A., Akrouf, S.: A novel system based on phase congruency and Gabor-filter bank for finger knuckle pattern authentication. *ICTACT J. Image Video Process* **10**, 2125–2131 (2020)
10. Hassanat, A., et al.: Victory sign biometric for terrorists identification: preliminary results. In: *Proceedings of the 8th International Conference on Information and Communication Systems (ICICS)* (2017)

11. 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)
12. Isola, P., Zhu, J.Y., Zhou, T.: Image-to-image translation with conditional adversarial networks (2017). arXiv preprint [arXiv:1701.07875](https://arxiv.org/abs/1701.07875)
13. Kancharla, K., Kamble, V., Kapoor, M.: Handwritten signature recognition: a convolutional neural network approach. In: International Conference on Advanced Computation and Telecommunication (ICACAT), pp. 1–5 (2018)
14. Kolvand, H., Asadianfam, S., Akintoye, K.A., Rahim, M.S.: Finger vein recognition techniques: a comprehensive review. *Multimedia Tools Appl.* **82**, 33541–33575 (2023)
15. Kumar, A., Ravikanth, C.: Personal authentication using finger knuckle surface. *IEEE Trans. Inf. Forensics Secur.* **4**(1), 98–110 (2009)
16. Kumar, A., Zhou, Y.: IIT Delhi finger knuckle database (version 1.0) (2009). http://www4.comp.polyu.edu.hk/~csajaykr/IITD/iitd_knuckle.htm
17. Lee, S., Jang, S.W., Kim, D., Hahn, H., Kim, G.: A novel fingerprint recovery scheme using deep neural network-based learning. *Multimed. Tools Appl.* **80**, 34121–34135 (2021)
18. Li, S., Zhang, H., Shi, Y., Yang, J.: Novel local coding algorithm for finger multimodal feature description and recognition. *Sensors* (2019)
19. Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J.: A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **33**(12), 6999–7019 (2022)
20. Liu, F., Zhang, D., Shen, L.: Study on novel curvature features for 3D fingerprint recognition. *Neurocomputing* **168**, 599–608 (2015)
21. Minaee, S., Abdolrashidi, A., Su, H., Bennamoun, M., Zhang, D.: Biometrics recognition using deep learning: a survey. *Artif. Intell. Rev.*, 8647–8695 (2023)
22. Mirza, M., Osindero, S.: Conditional generative adversarial nets (2017). arXiv preprint [arXiv:1411.1784](https://arxiv.org/abs/1411.1784)
23. Mohamed, N., Mustafa, M.B., Jomhari, N.: A review of the hand gesture recognition system: current progress and future directions. *IEEE Access* **9**, 157422–157436 (2021)
24. Muthukumar, A., Kavipriya, A.: A biometric system based on Gabor feature extraction with SVM classifier for finger-knuckle-print. *Pattern Recogn. Lett.* **125**, 150–156 (2019)
25. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 23–28 (2014)
26. Sandhu, R., Samarati, P.: Access control: principle and practice. *IEEE Commun. Mag.* **32**(9), 40–48 (1994)
27. Sero, D., et al.: Facial recognition from DNA using face-to-DNA classifiers. *Nat. Commun.*, 2557 (2019)
28. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.*, 1635–1650 (2010)
29. Tarawneh, A.S., et al.: DeepKnuckle: deep learning for finger knuckle print recognition. *Electronics* **11**, 513 (2022)
30. The Hong Kong Polytechnic University: Polyu finger knuckle database (2009). <http://www4.comp.polyu.edu.hk/biometrics/>
31. Vanini, P., Rossi, S., Zvizdic, E., Domenig, T.: Online payment fraud: from anomaly detection to risk management. *Financ. Innov.* **9**, 66 (2023)

32. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs (2017). arXiv preprint [arXiv:1711.11585](https://arxiv.org/abs/1711.11585)
33. Wildes, R.: Iris recognition: an emerging biometric technology. *Proc. IEEE* **85**(9), 1348–1363 (1997)
34. Woodard, D.L., Flynn, P.J.: Finger surface as a biometric identifier. *Comput. Vis. Image Underst.* **100**, 357–384 (2005)
35. Zhang, L., Zhang, L., Zhang, D., Guo, Z.: Phase congruency induced local features for finger-knuckle-print recognition. *Pattern Recogn.* **45**, 2522–2531 (2012)
36. Zhang, L., Zhang, L., Zhang, D.: Finger-Knuckle-print: a new biometric identifier. In: *IEEE International Conference of Image Processing*, pp. 1981–1984 (2009)
37. Zhao, S., Zhang, B.: Learning salient and discriminative descriptor for palm-print feature extraction and identification. *IEEE Trans. Neural Netw. Learn. Syst.* **31**(12), 5219–5230 (2020)
38. Zhu, J.Y., et al.: Toward multimodal image-to image translation. In: *Advances in Neural Information Processing Systems*, pp. 465–476 (2017)
39. Zohrevand, A., Imani, Z., Ezoji, M.: Deep convolutional neural network for finger-Knuckle-print recognition. *Int. J. Eng.* **34**, 1684–1693 (2021)



Multimodal Drivers of Attention Interruption to Baby Product Video Ads

Wen Xie¹ , Lingfei Luan² , Yanjun Zhu¹ , Yakov Bart³ ,
and Sarah Ostadabbas⁴  

¹ Institute for Experiential AI, Northeastern University, Boston, MA, USA

² Belmont University, Nashville, TN, USA

³ D'Amore-McKim School of Business, Northeastern University, Boston, MA, USA

⁴ Electrical and Computer Engineering, Northeastern University, Boston, MA, USA
ostadabbas@ece.neu.edu

Abstract. Ad designers often use sequences of shots in video ads, where frames are similar within a shot but vary across shots. These visual variations, along with changes in auditory and narrative cues, can interrupt viewers' attention. In this paper, we address the underexplored task of applying multimodal feature extraction techniques to marketing problems. We introduce the "AttInfaForAd" dataset, containing 111 baby product video ads with visual ground truth labels indicating points of interest in the first, middle, and last frames of each shot, identified by 75 shoppers. We propose attention interruption measures and use multimodal techniques to extract visual, auditory, and linguistic features from video ads. Our feature-infused model achieved the lowest mean absolute error and highest R-square among various machine learning algorithms in predicting shopper attention interruption. We highlight the significance of these features in driving attention interruption. By open-sourcing the dataset and model code, we aim to encourage further research in this crucial area. (Dataset and model code available at <https://github.com/ostadabbas/Baby-Product-Video-Ads>).

Keywords: Baby products · Eye-tracking dataset · Attention · Computer vision · Natural language processing

1 Introduction

Video advertisements are a common medium for promoting baby products, often consisting of sequences of shots, each contributing to the overall narrative. Within a shot, frames are thematically and sequentially consistent, while

W. Xie and L. Luan—Contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78104-9_21.

noticeable differences occur across shots (see a series of example shots from one ad in Fig. 1). These variations can cause viewers to redistribute their gaze, leading to attention interruption. Interruptions require additional cognitive effort from viewers to follow the story, posing a substantial challenge in advertising.

Beyond the visual changes across shots, video ads frequently leverage multiple modalities, such as audio and narratives, to convey product information and elicit desired responses from viewers. These multimodal features can also affect attention interruption. For instance, audio elements like sudden changes in volume or background music may cause distractions. Narrative elements, such as abrupt shifts in storyline or dialogue, can further complicate attention dynamics by requiring viewers to adjust their focus to understand the content.



Fig. 1. Points of interest (POIs) of one participant during watching the video ad named ‘90 Years Crafting’. Red dots represent the POIs. The three columns show the first, middle, and last video frames of three shots in a video ad.

Advertisers need to understand which features in these modalities drive shopper attention dynamics to optimize their ad design strategies. However, the study of multimodal features and their impact on attention interruption remains underexplored due to several challenges. These challenges include the need for comprehensive extraction and analysis of visual, audio, and narrative features, the development of robust attention interruption measures, and the lack of an eye-tracking dataset specifically featuring parents’ viewing behaviors of baby product video ads.

To address these gaps, this paper introduces a comprehensive approach to understanding shoppers’ attention in the baby product market. We begin by collecting data on shoppers’ attentional allocations when they watch baby product video ads. Their points of interest (POIs) are recorded as illustrated in

Fig. 1. This is followed by extracting extensive multimodal features and proposing attention interruption measures. With these preparations, we propose a multimodal feature-infused model to predict the interruption and further explore the importance of different features in affecting shoppers’ attention. Figure 2 provides an overview of the study framework.

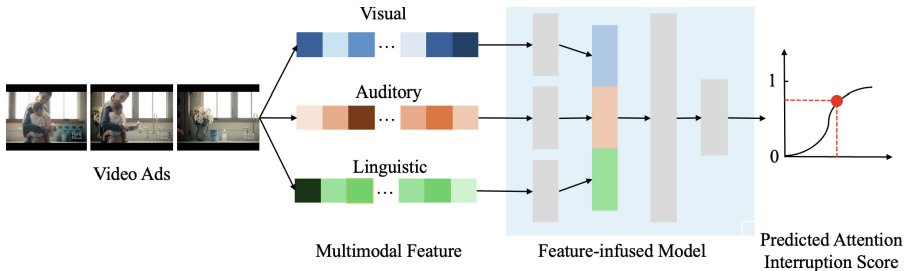


Fig. 2. An overview of the framework. We first collect shoppers’ points of interest (see Fig. 1) in video ads, based on which we propose attention interruption measures. Then, we extract multimodal features from the video ads. Finally, we build feature-infused neural networks to predict attention interruption and conduct feature importance analyses to suggest ad design strategies.

Our contributions are threefold. First, we introduce a new task involving the application of multimodal feature extraction techniques to solve marketing and business problems. Second, we create the first-ever “(Att)ention to (Inf)ant (For)mula Video (Ad)s (AttInfaForAd)” dataset, comprising 111 video ads and 4,184 frames with shoppers’ attention allocation labels. Third, we propose measures of attention interruption and provide a robust feature-infused benchmark for attention interruption prediction. Our efforts encourage further exploration in this critical area.

2 Related Work

This paper relates to two main venues: (1) eye-tracking research in attention literature, and (2) multimodal features in business and marketing fields.

2.1 Attention Research Through Eye-Tracking

Eye-tracking has been widely used in many disciplines such as education, psychology, and marketing [3, 28, 40] to enhance our understanding of human attention processes, including overt and covert attention [36]. Overt attention involves visible eye movements and is typically measured using eye-tracking devices. In contrast, covert attention refers to the mental process of shifting focus without moving the eyes. Eye-tracking devices can generate fixations when the eyes

remain relatively stationary on a stimulus. However, challenges exist, particularly in producing calibration-free equipment [7], and in interpreting the connection between eye movement recordings and cognitive processes. In this study, we introduce the points of interest (POI) technique by requesting participants to directly indicate their attentional focus on advertising elements. This method is cost-effective and accessible to video stimuli and has strengthened participant engagement. Scholars have demonstrated that self-reported POI can yield comparable results to data collected from eye-tracking devices [24].

Based on eye movement data, numerous studies have explored the factors that attract and maintain attention, such as color, size, shape, complexity, and goals [8, 37, 38]. Few addressed their impacts on attention interruption across video shots. Among the public eye-tracking datasets, some are designed for advancing gaze estimation [15, 20, 42] or insights into connections between brain activities and eye movements [17]. Several datasets are application-driven such as drivers' attention [31] and predictive processes in reading [22]. However, to the best of our knowledge, few public datasets are available for understanding people's attention to video ads, especially baby products, despite the practical needs in this domain.

2.2 Multimodal Features

Video ads typically contain multiple modalities and each functions differently and collaboratively. Visuals illustrate the product features and can capture immediate attention [32]. Narratives provide context and detail such as benefits, guiding the viewer through the ad's message [11]. Music sets the tone and mood of the ad, evoking specific emotions and making the ad more memorable [18].

We carefully select an extensive set of visual features related to color, texture, content, and composite to understand their impact on viewers. Color features include brightness, hue, saturation, contrast of brightness, color diversity, clarity, and color names [25, 41]. For texture features, we examine contrast, correlation, energy, homogeneity, and dissimilarity for each HSV channel based on the GLCM [23]. Additionally, visual complexity can cause cognitive load for viewers, affecting their attention processes [33]. We thus consider the number of objects and regions in ads [30]. We also include the number of faces in ads, the region sizes, and the rule of thirds, which are photography techniques [41]. Modern ad designers often enhance the central part of ads by increasing brightness and sharpness. We measure these attributes in the inner part of ads.

We also gather a comprehensive set of audio features, including root mean square (RMS) and zero crossing rate (ZCR) [12, 39]. RMS measures loudness, while ZCR relates to the detection of percussive sounds. Other features include spectral centroid and bandwidth, which pertain to sound "brightness" and "warmth," and pitch features that capture fundamental frequency. We further incorporate mel-frequency cepstral coefficients (MFCCs) for spectral shape analysis, chroma features for pitch class identification, and Mel spectrogram features for capturing temporal audio dynamics. We also account for characteristics of

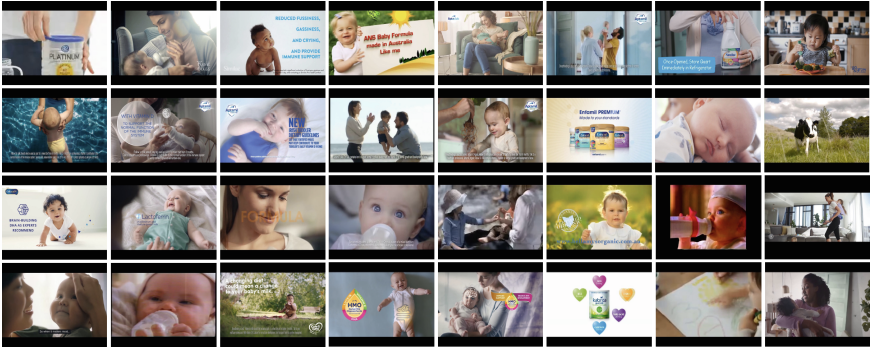


Fig. 3. An overview of the video ads. The collage contains 4×8 video frames. Each frame is from one video ad.

the language, as it also plays an essential role in capturing and guiding audiences' attention, extending beyond the traditional visual and audio signals [21].

3 Introducing AttInfaForAd Dataset

We create the first public dataset of shoppers' viewing and cognitive behavior for infant formula advertisement videos: AttInfaForAd. Our collection comprises 111 video ads (24 brands), from which we extract 4,181 key frames (beginning, middle, and last frame of each shot). Each key frame is annotated with individuals' points of interest, representing the solitary factor that most attracted their focus. This extensive dataset gives beneficial understandings right into customer interaction with infant formula advertising.

Participants. A total of 75 participants (Male: 28, Female: 40, 7 excluded) from Beijing Film Academy participated in the study. The average age of the participants was 31.4 years.

Stimuli. We conduct a thorough online search using various video platforms such as YouTube to create the corpus of infant formula ads. Initially, 178 ads were identified. After excluding non-English ads, we obtain 111 ads representing 24 different brands of infant products. Figure 3 illustrates 32 video frames from the stimuli.

Procedure. We conduct the experiment on Qualtrics via the following steps. Participants first provide informed consent before beginning the study. They are then assigned to watch the advertisement videos. On average, each participant watches on average 10 ads that are randomly sampled from the 111 video ads. After each watch, participants are required to indicate their points of interest (POIs) while watching the video ads. Specifically, POIs are identified by asking participants to click on the video frame in the Qualtrics survey that attracted their attention. Following this step, participants answer several demographic questions, including age and gender.

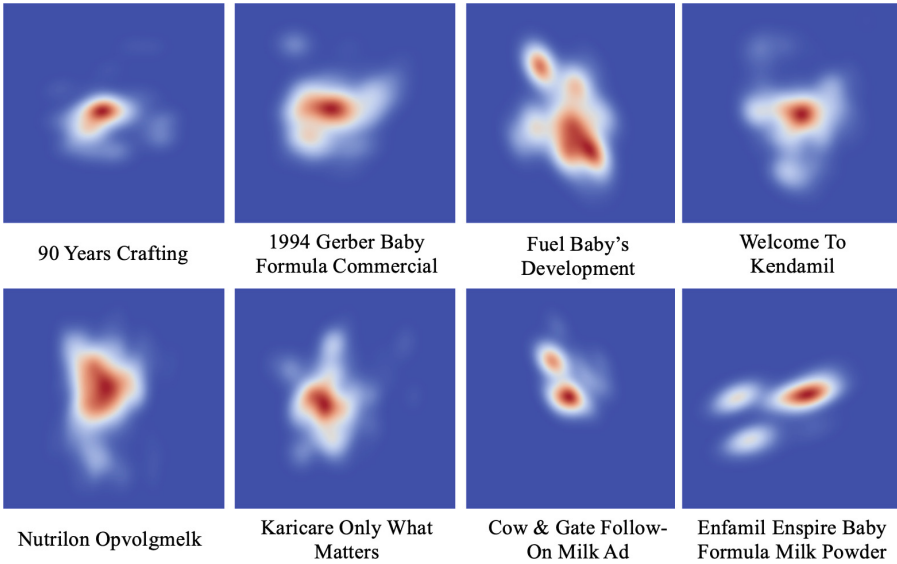


Fig. 4. Heatmaps of points of interest (POIs) for eight video ads, with ad names displayed below each heatmap. The red areas indicate higher concentrations of POIs, while the blue areas represent lower concentrations. A smaller red area suggests that viewers' attention is more focused, indicating less interruption. (Color figure online)

In total, we collect 746 samples. Figure 4 shows the example heatmaps of points of interest from participants watching eight video ads. The ad names are displayed under each heatmap. The red areas indicate higher concentrations of POIs, while the blue areas represent lower concentrations. A more condensed red area indicates less interrupted attention. The heatmaps for the ads “90 Years Crafting” and “Cow & Gate Follow-On Milk Ad” are more concentrated compared to the others, while the heatmaps for “Fuel Baby’s Development” and “Nutrilon Opvolgmelk” are more dispersed. The heatmaps for “1994 Gerber Baby Formula Commercial”, “Welcome To Kendamil”, “Karicare Only What Matters”, and “Enfamil Enspire Baby Formula milk Powder” fall somewhere in between, highlighting the variability of attention allocations across different ads.

4 Methodology

In this section, we first present our attention interruption measure based on the collected points of interest (POIs). The measure utilizes the area of the convex hull formed by these POIs to quantify the spread and interruption of viewer attention during video ads. Next, we propose various techniques to extract visual, auditory, and linguistic features from the video ads. With these extracted features, we propose a multimodal feature-infused model to predict attention interruption and formulate fixed-effect regressions for insights into the factors contributing to attention interruption in video ads.

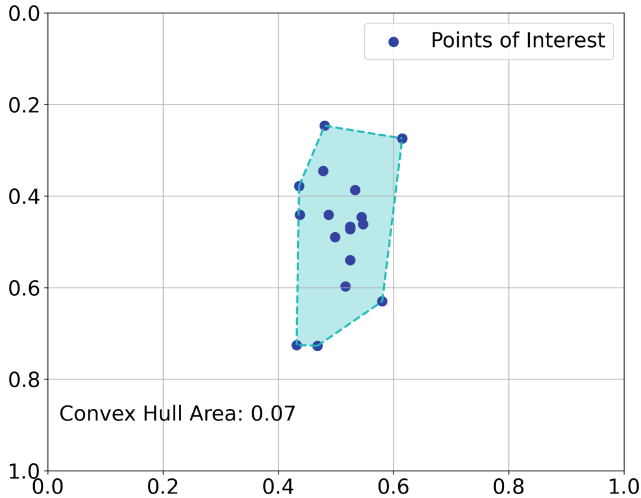


Fig. 5. An illustration of our proposed attention interruption measure. The convex hull area, A_{ch} is surrounded by dashed lines, representing the boundary enclosing all points of interest, indicated by the blue dots. The computed convex hull area is displayed in the bottom-left corner.

4.1 A Measure of Attention Interruption

Shoppers' attention allocations are at the frame level during their viewing of the video ads. To quantify attentional interruption throughout the viewing, we propose an aggregate measure. We first plot the points of interest (POIs) on a figure matching the video resolution (e.g., 512×512 pixels). The x and y coordinates are normalized by the frame width and height to allow for fair comparison across videos with different resolutions.

Our measure of attention interruption is based on the convex hull area (A_{ch}), which is the area of the smallest convex shape enclosing all POIs. We calculate this measure using the ConvexHull function from the SciPy package, which employs the Qhull algorithm. This algorithm constructs the convex hull by iteratively adding points and forming new facets that maintain the convex property. It starts by identifying an initial simplex, then adds points to the hull, updating the set of facets to ensure they form the smallest convex shape enclosing all points. A larger convex hull area indicates that the attention allocations are more widely dispersed across frames and shots. This suggests that shoppers have to reorient their gaze across shots, indicating greater attentional dispersion and interruption. The coordinates are normalized so that the range of this measure is from 0 to 1. For an illustration, we plot the POIs of one participant while watching a video ad in Fig. 5 and compute the convex hull area measure.

4.2 Multimodal Feature Extraction from Video Ads

We extract multimodal features from video ads, including visual, audio, and linguistic elements. Visual features involve color, texture, and object counts, while auditory features include energy, pitch, and spectral properties. Linguistic features are derived from transcribed text using speech recognition and advanced text analytics, producing 297 combined features to analyze ad effectiveness.

Video frames within the same shot tend to be visually and thematically consistent, making it unnecessary to extract visual features from every frame. Each shot typically represents a distinct segment of the narrative, with transitions in both the storyline and background music more likely to occur between shots. To effectively capture the transitions, we manually identify shots and extract visual features from the first, middle, and last frames of each shot. Using the method from [23], we determine the percentage of different colors. Color clarity is computed as the average of the standard deviations of the H, S, and V channels. Color diversity is calculated based on the Earth Mover’s Distance [25]. Texture features are computed using the Skimage library. We employ YOLOv8¹ to count objects and faces, and use Segment Anything [19] for region count, region size, and rule of thirds measurement. For each feature, its magnitude and variance are measured by its mean and standard deviation across shots in each video, resulting in 78 visual features.

We apply an extensive approach to extracting audio features from the 111 video ads. We first obtain the audio (i.e., MP3 file) of each video using MoviePy in Python. Then, we employ Librosa in Python to extract auditory features, including RMS energy, ZCR, spectral centroid, bandwidth, pitch, MFCCs, chroma, and mel spectrogram features. We summarize the audio content’s primary characteristics and variability by calculating the mean and standard deviation, yielding a total of 63 auditory features.

To analyze the linguistic properties of the texts in the ad transcripts, we apply a multi-stage process using state-of-the-art speech recognition and advanced text analytics techniques. First, we implement OpenAI’s Whisper model to transcribe the audio content. Through the linguistic inquiry and word count (LIWC) tool (version 2022), we extract linguistic features from the default dictionary and eight additional user-made dictionaries that align with our research interests in psychological involvement, cognitive processes, and brand perception in infant product advertisement. These dictionaries comprise absolutist, agitation/dejection, behavioral activation, brand personality, controversial terms, cost/benefit analysis, imagination, creativity and innovation, mind perception, and security language [1, 2, 4, 6, 16, 26, 27, 29, 35]. This extensive linguistic examination produces 156 features.

4.3 Attention Interruption Prediction Model

Our predictive analysis employs a diverse array of machine learning techniques, including support vector machine (SVM) [13], multi-layer perceptron (MLP)

¹ <https://github.com/ultralytics/ultralytics>.

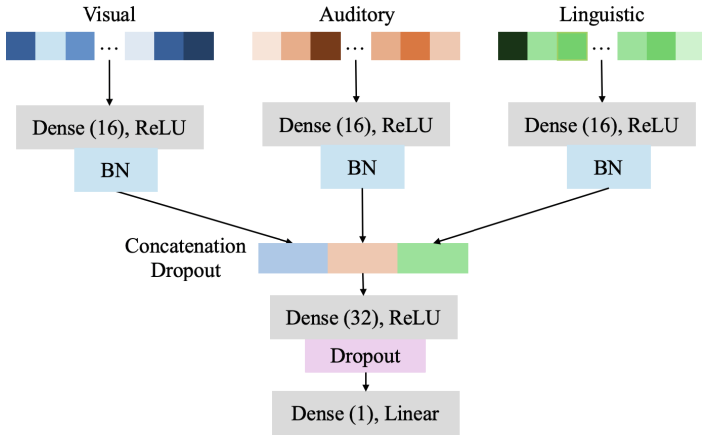


Fig. 6. Multi-feature-infused model (base model). Each modal feature is initially processed through a dense layer with a ReLU activation function. These outputs are then subjected to batch normalization before being concatenated, forming a unified representation of the multimodal data. The concatenated output is passed through a dropout layer and then serves as input to another dense layer with ReLU activation. Finally, the model produces the output through a dense layer with a linear activation function.

[34], random forest (RF) [5], gradient boosting (GB) [10], linear regression (LR) [9], and ridge regression [14]. Given the multimodal nature of our features, we propose a novel feature-infused model, as illustrated in Fig. 6. This architecture is designed to effectively integrate and process diverse input modalities.

Our dataset consists of 746 samples, 78 visual features, 63 auditory features, and 156 linguistic features. Our base model (see Fig. 6) is designed with four layers. Initially, each modality’s features are processed through a dense layer with 16 nodes, employing a Rectified Linear Unit (ReLU) activation function. This step allows for effective representation of modality-specific features. The outputs from these individual modal layers are subsequently concatenated and passed through a batch normalization layer, forming a unified multimodal representation. This concatenated output is then subjected to a dropout layer and fed into a larger dense layer with 32 nodes, utilizing ReLU activation and dropout to prevent overfitting while capturing complex cross-modal interactions. The model culminates in an output layer with a linear activation function that predicts the attention interruption score. This architecture effectively integrates multimodal features and captures modality-specific nuances, potentially leading to more accurate and robust predictions of attention interruption.

To enhance the base model, we replace the final linear activation function with a sigmoid function to better fit the output range, considering that the attention interruption measure (i.e., convex hull area) ranges from 0 to 1. Additionally, we experiment with different activation functions, such as Exponential

Linear Unit (ELU), due to its ability to handle negative values more effectively than ReLU, which can lose information when values are less than zero.

To train the model, we use the mean absolute error (MAE) loss function with L2 regularization, which is formulated as

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| + \lambda \sum_{j=1}^M w_j^2, \quad (1)$$

where \mathbf{y} is the vector of true attention interruption scores measured by the convex hull area; $\hat{\mathbf{y}}$ is the vector of predicted attention interruption scores; N is the number of samples; λ is the regularization factor; w_j are the weights of the model; M is the number of weights.

4.4 Model Evaluation Metrics

The values of the convex hull area, A_{ch} , in our dataset range from 0 to 1 and are generally small. To evaluate the performance of our models, we use two key metrics: mean absolute error (MAE) and R-squared (R^2).

The mean absolute error (MAE) measures the average absolute difference between the observed actual outcomes and the outcomes predicted by the model. It is defined as:

$$\text{MAE} = \frac{1}{n} \sum_{s=1}^n |y_s - \hat{y}_s|, \quad (2)$$

where n is the number of samples, y_s is the actual value of attention interruption (measured by either A_{ch}), and \hat{y}_s is the predicted attention interruption score. A smaller MAE indicates better predictive accuracy of the model.

The R-squared (R^2) metric indicates the proportion of the variance in the dependent variable (i.e., attention interruption) that is predictable from the independent variables (i.e., multimodal features). It is calculated as:

$$R^2 = 1 - \frac{\sum_{s=1}^n (y_s - \hat{y}_s)^2}{\sum_{s=1}^n (y_s - \bar{y})^2}, \quad (3)$$

where \bar{y} is the mean of the actual values. The R^2 value provides an indication of the goodness of fit of the model, with a value closer to 1 indicating a better fit.

Together, these metrics allow us to assess both the accuracy and explanatory power of our predictive models, ensuring that they not only provide precise predictions but also capture the underlying variance in attention interruption effectively.

4.5 Feature Importance Analysis

To understand which features affect attention interruption, we conduct a detailed analysis of each modality to provide nuanced insights for advertisers. First, we apply principal component analysis (PCA) to reduce redundancy among the

features. By retaining features that explain 97.5% of the total variance, we retain 39 visual, 33 auditory, and 61 linguistic features. Additionally, we remove 7 linguistic features due to collinearity, resulting in 54 linguistic features.

A linear regression model is employed to ensure interoperability and enhance our understanding of the importance of different features. Specifically, we formulate our regression model as follows:

$$y_{iv} = \alpha_i + \beta \cdot X_v + \epsilon_{ij}, \quad (4)$$

where y represents the attention interruption measured by A_{ch} ; X is a vector of visual, auditory, or linguistic features; i indexes participants; v represents infant formula video stimuli; α_i are participant-specific intercepts capturing individual heterogeneities; β is a vector of coefficients for the features specific to the video stimuli v ; and ϵ_{ij} is the error term representing unexplained variation in attention interruption for participant i and video stimulus v . By estimating the coefficient of each variable (i.e., feature), we can clearly determine the extent to which each feature affects attention interruption.

5 Experimental Results

In this section, we first report the performance of our proposed multimodal feature-infused model. Then, we present the feature importance analysis results.

5.1 Attention Interruption Prediction Accuracy

We split our dataset into a training set (80%, or 596 samples) and a testing set (20%, or 150 samples). We implement the machine learning algorithms using the Scikit-Learn package and our proposed model using TensorFlow. During training, we set the learning rate to 0.001, the regularization factor λ to 0.01, the dropout rate to 0.1, the batch size to 16, and the number of epochs to 200, with early stopping based on a patience of 20 epochs.

Table 1 documents the prediction evaluation on the testing dataset across various machine learning models and our proposed multimodal model with different variants. We use MAE and R^2 as evaluation metrics. The results suggest that our proposed model with a sigmoid function as the final layer achieves the lowest MAE, and together with ELU activation, it achieves the highest R^2 . This demonstrates the model’s superiority and establishes a benchmark for future research.

5.2 Feature Importance

We estimate the coefficients of Eq. (4) using the ‘FIXEST’ package in R and summarize the count of positive and negative estimates (that are statistically significant at an alpha level of 0.05 or lower) in each regression in Table 2. The full estimation results are presented in the supplementary materials (Tables S1, S2, and S3).

Table 1. Model evaluation results. We report the mean absolute error (MAE) and R-square of each model. In addition to the proposed base model (see Fig. 6), we have two variants, i.e., ‘sigmoid’ and ‘ELU+sigmoid’. The best results are bolded.

Model	MAE	R ²
SVM [13]	0.0679	-0.0259
MP [34]	0.0375	0.4788
RF [5]	0.0366	0.4866
GB [10]	0.0367	0.4938
LR [9]	0.0375	0.4781
Ridge [14]	0.0365	0.4919
Ours (base)	0.0359	0.5168
Ours (sigmoid)	0.0345	0.5216
Ours (ELU+sigmoid)	0.0350	0.5418

Among the visual features, we find that the coefficients of 23 features are statistically significant. Specifically, 15 features with positive coefficients increase attention interruption, while 8 features with negative coefficients reduce it. For the auditory features, 14 features affect attention interruption, with 7 increasing and 7 reducing the interruption. Regarding linguistic features, 26 features influence attention interruption: 11 features increase it, and 15 decrease it. These findings highlight the significance of our extensively extracted features in driving attention interruption.

Table 2. Count of significant coefficients in the estimation results. For instance, in the visual feature analysis, 15 (8) coefficients are statistically positive (negative) at the 0.05 level.

Modality	Count of Positive Coefficients	Count of Negative Coefficients
Visual	15	8
Auditory	7	7
Linguistic	11	16

In sum, our findings suggest how advertisers could integrate insights from visual, auditory, and linguistic analyses to optimize viewers’ attention allocation.

6 Discussion and Future Work

Understanding how consumers allocate their attention to baby products in video ads is crucial for advertisers. Based on the findings from our analysis of visual, audio, and linguistic features in Tables 2, S1, S2, and S3, we discuss insights below.

Visual, auditory, and linguistic features each play a pivotal role in shaping consumer attention in baby product video ads. Our analysis underscores the importance of visual elements like color contrast, brightness, and scene complexity, which significantly affect attention interruption. Audio features are also influential; our results indicate that higher RMS levels (i.e., loudness) can reduce attention interruption. Linguistically, incorporating security-related terms and expressions into the narrative of infant formula ads can effectively alleviate viewers' concerns, particularly those of adult consumers. This strategy not only reassures potential customers about product safety but also enhances engagement. Notably, our findings suggest that security-focused language reduces attention interruption, allowing viewers to maintain focus on key information.

While our study provides valuable insights and contributions, it naturally has limitations. First, our study primarily focuses on infant formula ads. Future research can expand the dataset to include a wider variety of baby products, a larger set of video stimuli, and more diverse shopper socio-economic demographics. Second, attention interruption may affect viewers' experience and recall of the products featured in video ads. Though we provide insights on the multimodal drivers of attention interruption for ad design, it is worth to examine whether the points of attention align with the area of products. To this end, researchers can develop methods for real-time attention tracking and interruption prediction and investigate how multimodal features direct viewers' attention to advertised products. Third, researchers can explore personalized ad recommendations based on individual viewer preferences, conduct longitudinal studies to examine how attention patterns and ad effectiveness evolve over time, analyze how attention interruption varies across different platforms, and study the impact of external factors such as time of day, viewer's mood, or concurrent activities can provide deeper insights. By addressing these areas, future research can further enhance the understanding and application of multimodal features in advertising, ultimately leading to more effective and engaging video ads for baby products.

7 Conclusion

Our study contributes to the understanding of shoppers' attention to baby product video ads by introducing the first public dataset in this field. Our comprehensive analysis demonstrates how multimodal features, including visual, auditory, and linguistic aspects, influence shoppers' attention during video ads. The findings provide valuable practical insights for advertisers aiming to optimize their video ad strategies. Our proposed multimodal feature-infused model achieves the best performance among various machine learning algorithms, establishing a benchmark for future research. This highlights the necessity for further exploration and refinement in this domain.

References




1. Ahmed, S.T.: The Language of the Creative Person: Validating the Use of Linguistic Analysis to Assess Creativity. San Jose State University (2021)
2. Al-Mosaiwi, M., Johnstone, T.: In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clin. Psychol. Sci.* **6**(4), 529–542 (2018)
3. Alemdag, E., Cagiltay, K.: A systematic review of eye tracking research on multimedia learning. *Comput. Educ.* **125**, 413–428 (2018)
4. Baele, S.J., Sterck, O.C.: Diagnosing the securitisation of immigration at the EU level: a new method for stronger empirical claims. *Polit. Stud.* **63**(5), 1120–1139 (2015)
5. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
6. Burkhardt, H.A., Alexopoulos, G.S., Pullmann, M.D., Hull, T.D., Areán, P.A., Cohen, T.: Behavioral activation and depression symptomatology: longitudinal assessment of linguistic indicators in text-based therapy sessions. *J. Med. Internet Res.* **23**(7), e28244 (2021)
7. Drewes, H., Pfeuffer, K., Alt, F.: Time-and space-efficient eye tracker calibration. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, pp. 1–8 (2019)
8. Everdell, I.: The relationship between bottom-up saliency and gaze behaviour during audiovisual speech perception. Ph.D. thesis (2009)
9. Freedman, D.A.: Statistical Models: Theory and Practice. Cambridge University Press (2009)
10. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, 1189–1232 (2001)
11. Green, M.C., Brock, T.C.: The role of transportation in the persuasiveness of public narratives. *J. Pers. Soc. Psychol.* **79**(5), 701 (2000)
12. Grewal, R., Gupta, S., Hamilton, R.: Marketing insights from multimedia data: text, image, audio, and video (2021)
13. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. *IEEE Intell. Syst. Their Appl.* **13**(4), 18–28 (1998)
14. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67 (1970)
15. Huang, Q., Veeraraghavan, A., Sabharwal, A.: TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Mach. Vis. Appl.* **28**, 445–461 (2017)
16. Johnsen, J.A.K., Vambheim, S.M., Wynn, R., Wangberg, S.C.: Language of motivation and emotion in an internet support group for smoking cessation: explorative use of automated content analysis to measure regulatory focus. *Psychol. Res. Behav. Manag.*, 19–29 (2014)
17. Kastrati, A., et al.: EEGEyeNet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. *arXiv preprint [arXiv:2111.05100](https://arxiv.org/abs/2111.05100)* (2021)
18. Kellaris, J.J., Cox, A.D., Cox, D.: The effect of background music on ad processing: a contingency explanation. *J. Mark.* **57**(4), 114–125 (1993)
19. Kirillov, A., et al.: Segment anything. *arXiv:2304.02643* (2023)
20. Krafska, K., et al.: Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2176–2184 (2016)

21. Luan, L., Liu, W., Zhang, R., Hu, S.: Introducing cognitive psychology in film studies: redefining affordance. *Int. J. Educ. Hum.* **2**(3), 70–78 (2022)
22. Luke, S.G., Christianson, K.: The Provo Corpus: a large eye-tracking corpus with predictability norms. *Behav. Res. Methods* **50**, 826–833 (2018)
23. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 83–92 (2010)
24. Masciocchi, C.M., Mihalas, S., Parkhurst, D., Niebur, E.: Everyone knows what is interesting: salient locations which should be fixated. *J. Vis.* **9**(11), 25 (2009)
25. Matz, S.C., Segalin, C., Stillwell, D., Müller, S.R., Bos, M.W.: Predicting the personal appeal of marketing images using computational methods. *J. Consum. Psychol.* **29**(3), 370–390 (2019)
26. McCullough, M.E., Root, L.M., Cohen, A.D.: Writing about the benefits of an interpersonal transgression facilitates forgiveness. *J. Consult. Clin. Psychol.* **74**(5), 887 (2006)
27. Mejova, Y., Zhang, A.X., Diakopoulos, N., Castillo, C.: Controversy and sentiment in online news. *arXiv preprint [arXiv:1409.8152](https://arxiv.org/abs/1409.8152)* (2014)
28. Mele, M.L., Federici, S.: Gaze and eye-tracking solutions for psychological research. *Cogn. Process.* **13**, 261–265 (2012)
29. Opoku, R.A., Hultman, M., Saheli-Sangari, E.: Positioning in market space: the evaluation of Swedish universities' online brand personalities. *J. Mark. High. Educ.* **18**(1), 124–144 (2008)
30. Overgoor, G., Rand, W., van Dolen, W., Mazloom, M.: Simplicity is not key: understanding firm-generated social media images and consumer liking. *Int. J. Res. Mark.* **39**(3), 639–655 (2022)
31. Palazzi, A., Abati, D., Solera, F., Cucchiara, R., et al.: Predicting the driver's focus of attention: the DR (eye) VE project. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(7), 1720–1733 (2018)
32. Pieters, R., Wedel, M.: Attention capture and transfer in advertising: brand, pictorial, and text-size effects. *J. Mark.* **68**(2), 36–50 (2004)
33. Pieters, R., Wedel, M., Batra, R.: The stopping power of advertising: measures and effects of visual complexity. *J. Mark.* **74**(5), 48–60 (2010)
34. Rosenblatt, F.: *Principles of neurodynamics. Perceptrons and the theory of brain mechanisms*. Technical report, Cornell Aeronautical Lab Inc Buffalo NY (1961)
35. Schweitzer, S., Waytz, A.: Language as a window into mind perception: how mental state language differentiates body and mind, human and nonhuman, and the self from others. *J. Exp. Psychol. Gen.* **150**(8), 1642 (2021)
36. Van der Stigchel, S., Theeuwes, J.: The relationship between covert and overt attention in endogenous cuing. *Percept. Psychophys.* **69**(5), 719–731 (2007)
37. Theeuwes, J.: Top-down and bottom-up control of visual selection. *Acta Physiol. (Oxf)* **135**(2), 77–99 (2010)
38. Wedel, M., Pieters, R., et al.: Eye tracking for visual marketing. *Found. Trends® Market.* **1**(4), 231–320 (2008)
39. Xiao, L., Kim, H.J., Ding, M.: An introduction to audio and visual research and applications in marketing. *Rev. Market. Res.* **10**, 213–253 (2013)
40. Xie, W., Lee, M.H., Chen, M., Han, Z.: Understanding consumers' visual attention in mobile advertisements: an ambulatory eye-tracking study with machine learning techniques. *J. Advertising*, 1–19 (2023)

41. Zhang, S., Lee, D., Singh, P.V., Srinivasan, K.: What makes a good image? Airbnb demand analytics leveraging interpretable image features. *Manage. Sci.* **68**(8), 5644–5666 (2022)
42. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: MPIIGaze: real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(1), 162–175 (2017)



Facial Wrinkle Segmentation for Cosmetic Dermatology: Pretraining with Texture Map-Based Weak Supervision

Junho Moon¹ , Haejun Chung¹ , and Ikbeom Jang² 

¹ Hanyang University, 04763 Seoul, Republic of Korea
{jmoon6807, haejun}@hanyang.ac.kr

² Hankuk University of Foreign Studies, 17035 Yongin, Republic of Korea
ijang@hufs.ac.kr

Abstract. Facial wrinkle detection plays a crucial role in cosmetic dermatology. Precise manual segmentation of facial wrinkles is challenging and time-consuming, with inherent subjectivity leading to inconsistent results among graders. To address this issue, we propose two solutions. First, we build and release the first public facial wrinkle dataset, ‘FFHQ-Wrinkle’, an extension of the NVIDIA FFHQ dataset. It includes 1,000 images with human labels and 50,000 images with automatically generated weak labels. This dataset could serve as a foundation for the research community to develop advanced wrinkle detection algorithms. Second, we introduce a simple training strategy utilizing texture maps, applicable to various segmentation models, to detect wrinkles across the face. Our two-stage training strategy first pretrain models on a large dataset with weak labels ($N = 50k$), or masked texture maps generated through computer vision techniques, without human intervention. We then finetune the models using human-labeled data ($N = 1k$), which consists of manually labeled wrinkle masks. The network takes as input a combination of RGB and masked texture map of the image, comprising four channels, in finetuning. We effectively combine labels from multiple annotators to minimize subjectivity in manual labeling. Our strategies demonstrate improved segmentation performance in facial wrinkle segmentation both quantitatively and visually compared to existing pretraining methods. The dataset is available at <https://github.com/labhai/ffhq-wrinkle-dataset>.

Keywords: Facial wrinkle segmentation · Weakly supervised learning · Texture map pretraining · Transfer learning

1 Introduction

With the growing interest in dermatological diseases and skin aesthetics, predicting facial wrinkles is becoming increasingly significant. Facial wrinkles serve as critical indicators of aging [2, 19, 20], and are essential for evaluating skin

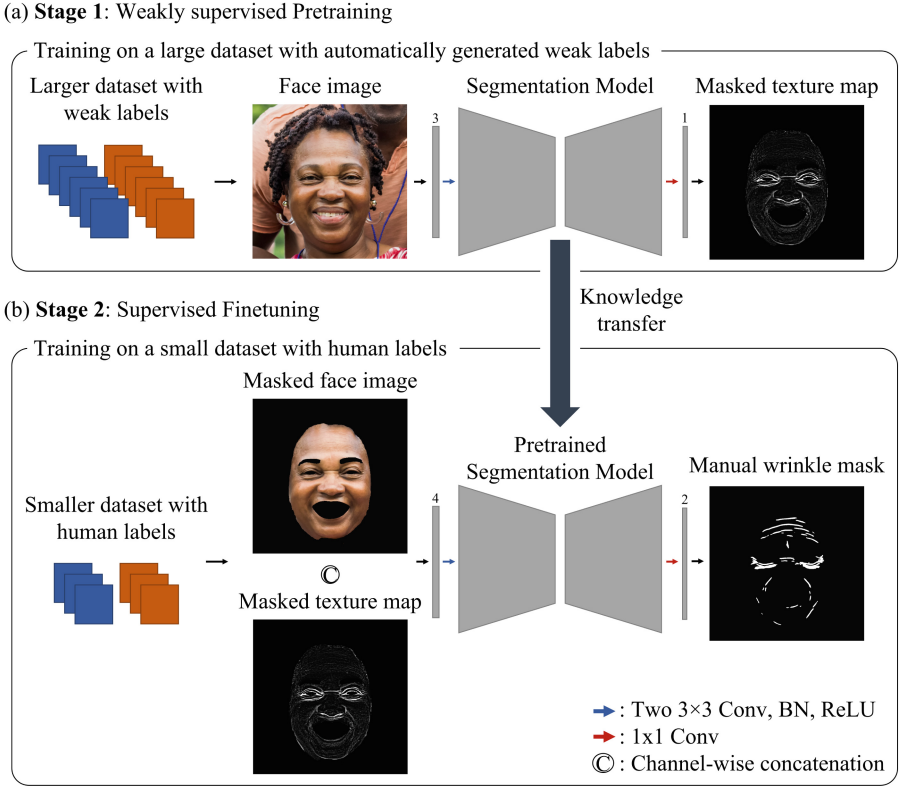


Fig. 1. Two-stage training for facial wrinkle segmentation. (a) Weakly supervised pre-training stage: the model learns to extract masked texture maps from RGB face images. (b) Supervised finetuning stage: the model refines its ability to extract facial wrinkles from RGB-masked face images and masked texture maps. The model parameters are initialized with the weights from the weakly supervised pretraining stage.

conditions [13, 29], diagnosing dermatological disorders [30], and planning pre-treatment protocols for skin management [1, 27]. Nevertheless, the manual detection of facial wrinkles poses considerable challenges. Accurate detection and analysis of facial wrinkles necessitate a high level of expertise, typically available only through well-trained professionals such as dermatologists. This process is time-consuming and entails substantial costs due to the extensive time and effort required by the experts.

Recently, numerous studies have focused on the automatic segmentation of facial wrinkles through the application of deep learning techniques [4, 14, 15, 25, 26, 34]. Nevertheless, these deep learning-based approaches are notably data-intensive. Due to the intricate distribution of facial wrinkles across the face, analyzing extensive collections of images can be exceedingly resource-intensive if each wrinkle must be individually evaluated. Furthermore, the manual analysis

procedure is fraught with subjectivity. The assessments of individual experts can differ significantly based on their experience, level of training, and personal biases, thereby complicating the consistency and reproducibility of the analysis results.

To address these challenges, we propose a two-stage training strategy, as illustrated in Fig. 1. This approach utilizes computer vision techniques, specifically filters, to generate many weakly labeled wrinkle masks ($N = 50,000$) without human intervention for weakly supervised pretraining. A smaller set of accurately labeled wrinkle masks ($N = 1,000$) is employed for supervised finetuning. This method significantly decreases the time and cost associated with manual wrinkle labeling, providing substantial advantages over traditional methodologies. To ensure the development of a generalized and robust model, we conducted experiments using a dataset comprising images captured from various angles, lighting conditions, races, ages, and skin conditions. We quantitatively analyzed the challenges associated with consistent manual wrinkle labeling across such a diverse dataset and integrated data labeled by multiple annotators to reduce subjectivity during the finetuning stage. No public dataset exists for full-face wrinkle segmentation, although there are a few private datasets. To address this gap, we have made our dataset publicly accessible to enhance the reproducibility and reliability of our results. This initiative aims to reduce the manual labeling costs for future research and serve as a benchmark dataset.

2 Related Works

2.1 Deep Learning-Based Facial Wrinkle Segmentation

Deep learning-based methods for facial wrinkle segmentation aim to enable neural network models to learn the features necessary for accurate wrinkle detection autonomously. Kim et al. [14] introduced a semi-automatic labeling strategy to enhance performance by extracting texture maps from face images and combining them with roughly labeled wrinkle masks, utilizing a U-Net architecture [23] for segmentation. In a subsequent study [15], they further improved segmentation accuracy by implementing a weighted deep supervision technique, which employs a weighted wrinkle map to more precisely calculate the loss for the downsampled decoder, outperforming traditional deep supervision methods. Yang et al. [34] developed Striped WriNet, which integrates a Striped Attention Module composed of Multi-Scale Striped Attention and Global Striped Attention within a U-shaped network. This approach applies an attention mechanism across multiple scales, effectively segmenting both coarse and fine wrinkles.

2.2 Weakly Supervised Learning

Weakly supervised learning is a methodology that trains models using incomplete or inaccurate labeled data instead of fully labeled data in situations where strong supervision information is lacking [36]. Xu et al. [33] proposed CAMEL, a weakly

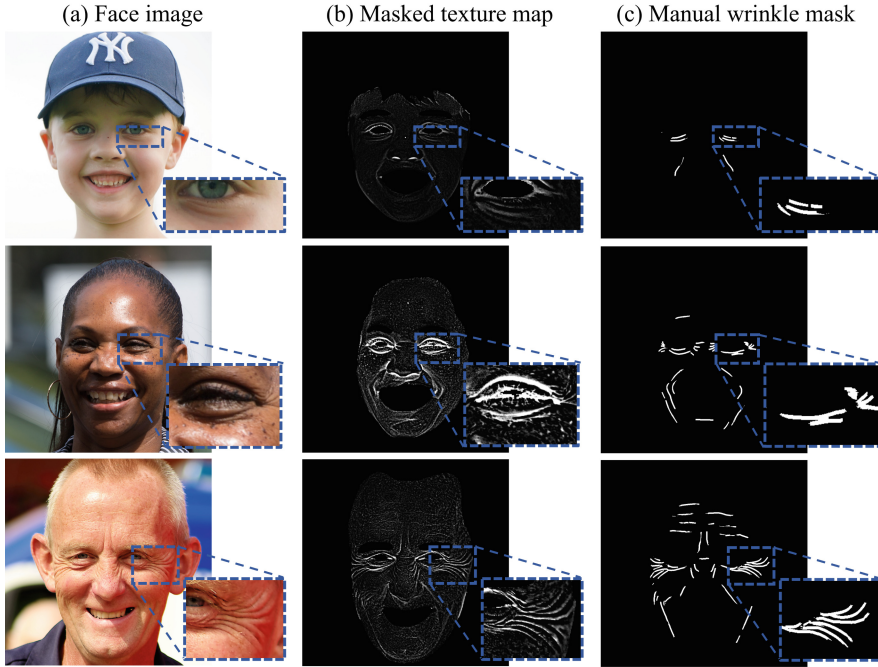


Fig. 2. Training Dataset. (a) High-resolution face images. (b) Masked texture maps extracted from face images, which include information about facial features. (c) Reliable manual wrinkle masks created by combining the results of multiple annotators.

supervised learning framework that uses a MIL-based label expansion technique to divide images into grid-shaped instances and automatically generate instance-level labels, enabling histopathology image segmentation with only image-level labels. Shen et al. [11] trained a deep learning model using only scribbles on whole tumors and healthy brain tissue, along with global labels for the presence of each substructure, to segment all sub-regions of brain tumors.

3 Dataset

3.1 Dataset Specifications

The first public facial wrinkle dataset, ‘FFHQ-Wrinkle’, comprises pairs of face images and their corresponding wrinkle masks. We focused on wrinkle labels while utilizing the existing face image dataset FFHQ (Flickr-Faces-HQ) [12], which contains 70,000 high-resolution (1024×1024) face images captured under various angles and lighting conditions. The dataset we provide consists of one set of manually labeled wrinkle masks ($N = 1,000$) and one set of ‘weak’ wrinkle masks, or masked texture maps, generated without human labor ($N = 50,000$). We selected 50,000 images from the FFHQ dataset, specifically image IDs 00000

Table 1. Demographic attributes of the dataset. The ‘Human-labeled’ data represents the 1,000 face images manually labeled by human annotators and the ‘Weakly-labeled’ data refers to the 50,000 images labeled without human intervention.

Dataset		Human-labeled	Weakly-labeled
Sample size		1000	50000
Age	0–9 / 10–19 / 20–29 /	66 / 68 / 233 /	7030 / 4448 / 13804 /
	30–39 / 40–49 / 50–69 / 70+	246 / 186 / 161 / 40	10960 / 6931 / 5550 / 1277
Sex	Male / Female	471 / 529	26929 / 23071
Race/ Ethnicity	White / Asian / Latino Hispanic /	587 / 210 / 67 /	29728 / 11121 / 3895 /
	Black / Middle Eastern / Indian	81 / 37 / 18 /	2383 / 2053 / 820

to 49999. We used these 50,000 face images to create the weakly labeled wrinkles and randomly sampled 1,000 images from these to create the ground truth wrinkles. The methods for generating weakly labeled wrinkles and ground truth wrinkles are discussed in Sect. 4.2. Table 1 summarizes estimated demographic information of the dataset—i.e. age, race, and sex. The age and sex data were sourced from the FFHQ-Aging [22] dataset, where at least three annotators labeled each image. The race/ethnicity attribute was obtained through facial attribute analysis using the DeepFace¹ framework. Hence, the demographic information may include errors. As illustrated in Fig. 2, the dataset consists of individuals of varying ages, sex, and race/ethnicity, featuring a range of skin conditions such as freckles, acne, and pigmentation. This diversity makes the dataset particularly suitable for training models to handle the wide array of skin conditions encountered in clinical settings. The dataset is publicly available at <https://github.com/labhai/ffhq-wrinkle-dataset>.

3.2 Ground Truth Wrinkle Annotation

For ground truth wrinkles, we manually annotated the face images. The annotation process involved three annotators with extensive experience in image processing and analysis. Wrinkles can be categorized into two types—dynamic wrinkles and static wrinkles [31]. Dynamic wrinkles are formed by facial muscles and appear with expressions but disappear when the face is at rest. Static (permanent) wrinkles are visible even when the face is at rest and result from the repeated formation of dynamic wrinkles over time. We annotated both types of wrinkles without distinguishing between them. Given the subjectivity inherent in wrinkle data, a consistent standard for wrinkle assessment was established prior to the commencement of labeling. The annotators conducted three synchronization sessions to minimize inter-rater variability. The annotation primarily targeted the forehead, crow’s feet, and nasolabial folds, encompassing the overall facial area. Due to the high resolution and diversity of the dataset—comprising various races, skin conditions, backgrounds, and angles—achieving consistent labeling results proved challenging, even with established standards

¹ <https://github.com/serengil/deepface>.



Fig. 3. Ambiguity in wrinkle evaluation. The labeling results from three annotators for the same image are different.

Table 2. Inter-rater agreement of manual wrinkle annotation. The Jaccard similarity index and Pearson correlation coefficient between different annotators are analyzed.

Metric	Annotators A&B	Annotators B&C	Annotators A&C	Average
Jaccard similarity index	0.2631	0.2962	0.3182	0.2925
Pearson correlation coefficient	0.4167	0.4559	0.4928	0.4551

for wrinkle assessment, as illustrated in Fig. 3. Consequently, as demonstrated in Table 2, the inter-rater agreement was low, underscoring the highly subjective nature of wrinkle assessments.

4 Method

4.1 Model Architecture

We evaluated our proposed method using the U-Net [23] and Swin UNETR [9] architectures, with U-Net serving as the base model for ablation studies and additional experiments. As depicted in Fig. 1, the U-Net model features a standard architecture comprising four encoder blocks and four decoder blocks. The Swin UNETR model employs an encoder with a window size of 16 and patches of size 4×4 , projecting the input patch into a 48-dimensional embedding space. This model includes four encoder blocks, each consisting of two successive Swin Transformer blocks [16], and four decoder blocks.

4.2 Training Strategy

We train the segmentation model using a substantial number of masked texture maps in a weakly supervised manner, followed by finetuning with a smaller set of reliably manually labeled wrinkle masks in a supervised manner. This training strategy, which involves finetuning the weights of a pretrained model that extracts facial textures using human-labeled wrinkle data, significantly enhances the model’s capability to detect facial wrinkles. The overall training pipeline is illustrated in Fig. 1.

Weakly Supervised Pretraining Stage. In the pretraining stage, we utilized weakly labeled wrinkle data automatically extracted through computer vision techniques without human intervention as the ground truth. Figure 4 illustrates the pipeline for generating weakly labeled wrinkles for the weakly supervised pretraining stage. Utilizing Eq. (1), we extracted the texture map [14] from the face image through a Gaussian kernel-based filter.

$$T(x, y) = \left(1 - \frac{I(x, y)}{1 + I_{G(\sigma)}(x, y)}\right) \times 255 \quad (1)$$

where G represents the Gaussian kernel, σ denotes its standard deviation, $I_{G(\sigma)}$ is the Gaussian filtered image, and (x, y) are the pixel coordinates in the image. Following the methodology in [14], we set the Gaussian kernel’s standard deviation to 5 and its size to 21×21 for texture map extraction. The extracted texture map contains detailed information about the contours, curves, and skin textures of the face image. However, as the texture map includes numerous false positives from the background, we employ a BiSeNet [35] architecture-based facial parsing deep learning model² to mask non-facial regions, resulting in the final masked texture map used as ground truth. We avoid converting the masked texture map into a binary mask due to the variability in the size, shape, and depth of wrinkles, which makes determining an appropriate threshold challenging. Figure 2-(b) shows the masked texture map used as the final ground truth in the weakly supervised pretraining stage.

In the weakly supervised pretraining stage, the model takes a 3-channel RGB face image as input and outputs a 1-channel masked texture map (Fig. 1-(a)). We use mean squared error (MSE) loss [21] to optimize the model, calculated as shown in Eq. (2).

$$MSE(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (2)$$

where \hat{y}_i and y_i are the model output and the masked texture map, respectively.

Supervised Finetuning Stage. For the ground truth in the finetuning stage, we utilized human-labeled wrinkle data generated as described in Sect. 3.2. Figure 5 illustrates the pipeline of the ground truth generation of the wrinkle mask. To produce a reliable ground truth wrinkle mask, we used majority voting to retain only the pixels that were labeled by at least two groups, thereby reducing variability among the annotators. Figure 2-(c) displays the manual wrinkle mask used as the final ground truth in the supervised finetuning stage. As model inputs, we use masked face images, where non-facial regions were masked using a facial-parsing model. Additionally, we included masked texture maps, which were used as ground truth in the pretraining stage, as auxiliary inputs.

In the supervised finetuning stage, the model takes as input a 3-channel RGB face image with only the facial regions and a 1-channel masked texture map.

² <https://github.com/zllrunning/face-parsing.PyTorch>.

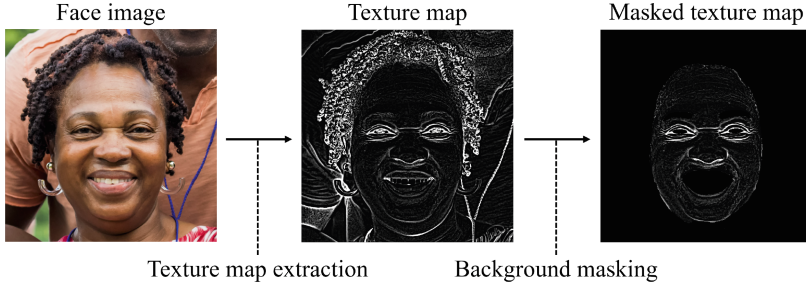


Fig. 4. Weakly labeled wrinkle generation pipeline. After extracting the texture map from the face image, we mask the non-facial regions to generate a masked texture map containing information on facial features. This masked texture map is then used as a weakly labeled wrinkle.

It then produces a 2-channel output indicating the presence of wrinkles and background. This stage begins with the model parameters from the pretraining stage, where the model was weakly supervised to extract masked texture maps from face images. Using transfer learning, we refine the model by adjusting its weights with manually labeled wrinkle masks. This process enhances the model’s ability to detect facial wrinkles by building on the general facial texture extraction skills developed during pretraining. We optimize the model using soft Dice loss [5], as shown in Eq. (3).

$$DL(p, g) = 1 - \frac{1}{C} \sum_{c=1}^C \frac{2 \sum_{i=1}^N p_{i,c} g_{i,c}}{\sum_{i=1}^N p_{i,c} + \sum_{i=1}^N g_{i,c}} \quad (3)$$

where C is the total number of classes, N is the total number of pixels, $p_{i,c}$ represents the predicted probability for pixel i belonging to class c , and $g_{i,c}$ represents the ground truth label for pixel i belonging to class c , respectively.

5 Experiments

5.1 Implementation Details

In both the weakly supervised pretraining and supervised finetuning stages, we utilize the original 1024×1024 image-label pairs as inputs without resizing. The AdamW optimizer [18] is employed, configured with a weight decay of 0.05, β_1 set to 0.9, and β_2 set to 0.999. We also implement the SGDR scheduler [17]. To maintain dataset diversity, we randomly apply various augmentations, including horizontal flipping, scaling, affine transformation, elastic transformation, grid distortion, and optical distortion during training. The dataset is partitioned into 80% for training, 10% for validation, and 10% for testing.

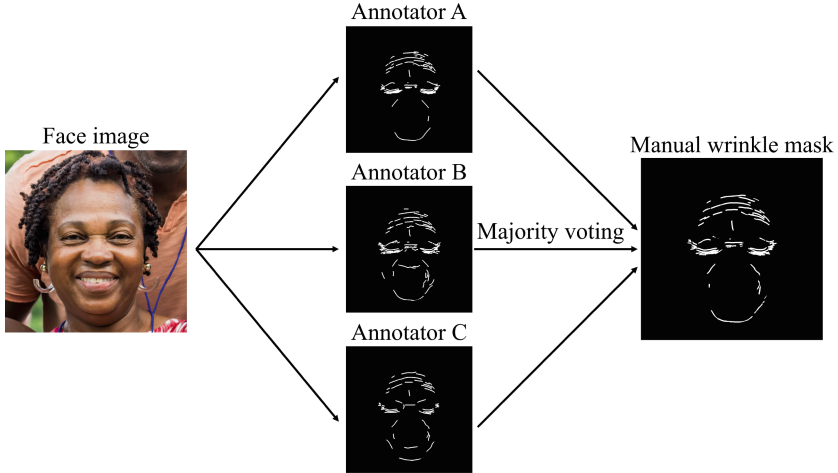


Fig. 5. Ground truth wrinkle generation pipeline. We combine data labeled by multiple annotators through majority voting to create a reliable ground truth wrinkle.

Weakly Supervised Pretraining Stage. In the weakly supervised pretraining stage, the model is trained for 300 epochs. The SGDR scheduler starts with an initial period of 100 epochs, with the learning rate beginning at a maximum of 0.001 and decaying to 0 over the period. At the end of each period, the length of the next period doubles that of the previous one. The batch size is 26 for U-Net and 22 for Swin UNETR. All pretraining processes were performed on an NVIDIA A100 Tensor Core GPU.

Supervised Finetuning Stage. In the supervised finetuning stage, the U-Net model is finetuned for 150 epochs, while the Swin UNETR model is finetuned for 300 epochs. The batch size is 14 for both models. The SGDR scheduler’s initial period length is set to 50 epochs for U-Net and 100 epochs for Swin UNETR. The learning rate starts at a maximum of 0.0001 and decreases to 0 within each period. At the end of each period, the length of the next period doubles that of the previous one, with the maximum learning rate set to 90% of the last period’s maximum. All finetuning processes are performed on RTX A6000 and RTX 6000 Ada GPUs.

5.2 Evaluation Metrics

To evaluate the performance of the final finetuned model in wrinkle segmentation, we use the Jaccard Similarity Index (JSI), F1-score, and Accuracy (Acc).

The Jaccard Similarity Index measures the overlap between the predicted wrinkle regions and the ground truth regions, defined as follows:

$$\text{JSI} = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

where A is the predicted segmentation, and B is the actual label.

The F1-score is the harmonic mean of precision and recall, while accuracy measures the proportion of correctly predicted pixels out of the total pixels. They are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

where TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives, and TN is the number of true negatives.

5.3 Results

To evaluate the performance of our proposed method, we first compare it with the latest methods: the semi-automatic labeling and weighted deep supervision method [15], and the Striped WriNet method [34]. Because the primary contribution of this work is the pretraining strategy, we also compare it with other pretraining techniques. They include using ImageNet pretrained models and self-supervised learning methods. For the ImageNet pretrained models, we replace the encoder part of the U-shape architecture with models pretrained on the ImageNet-1K dataset [24]; specifically, we use ResNet-50 [10] for U-Net and Swin-T [16] for Swin UNETR. For the self-supervised learning methods, we use denoising self-supervised learning [3] for pretraining U-Net, setting the Gaussian distribution’s standard deviation to 0.2, and masked image prediction [32] for pretraining Swin UNETR, using 32×32 masked patches and a 60% masking ratio. All training hyperparameters follow those specified in Sect. 5.1. To assess performance in scenarios with very limited labeled data, we train our model on the full training set (100%, $N = 800$) and on a randomly sampled subset (5%, $N = 40$).

The proposed method outperforms the latest wrinkle segmentation methods and the ones using the same model architectures with different pertaining methods. The performance gap is much larger in data-limited situations-i.e., fine-tuned on 5% of the manually-labeled data. Table 3 shows quantitative comparisons of wrinkle segmentation performance for each method using U-Net and Swin UNETR architectures. Our method consistently achieves the highest performance across both datasets and architectures. Figure 6 presents a qualitative comparison of our method with denoising pretraining using U-Net, which is the next best performing method in experiments using 100% of the data.

Table 3. Quantitative comparisons of facial wrinkle segmentation performance. Our method is compared against two latest wrinkle segmentation methods, models trained without pretraining, and models using different pretraining strategies. These pretraining techniques include masked image prediction, denoising, and pretraining encoders using the ImageNet-1K dataset.

Method		100% (N = 800)			5% (N = 40)			n_{params}
		JSI	F1-score	Acc	JSI	F1-score	Acc	
	Semi automatic labeling + WDS [15]	0.4552	0.6256	0.9954	0.3384	0.5057	0.9928	17.269M
	Striped WriNet [34]	0.4665	0.6294	0.9956	0.2382	0.3761	0.9903	6.223M
Swin UNETR with pretraining	No pretraining	0.4220	0.5858	0.9949	0.2545	0.3944	0.9932	25.153M
	ImageNet-1K [24] (Swin-T [16])	0.4385	0.6028	0.9952	0.2877	0.4351	0.9939	100.56M
	Masked image modeling [32]	0.4450	0.6079	0.9954	0.2963	0.4452	0.9937	25.153M
	Texture map (ours)	0.4643	0.6271	0.9953	0.3416	0.4970	0.9944	25.155M
U-Net with pretraining	No pretraining	0.4638	0.6278	0.9955	0.3021	0.4551	0.9918	17.263M
	ImageNet-1K [24] (ResNet-50 [10])	0.4664	0.6296	0.9955	0.3428	0.5018	0.9934	32.521M
	Denoising [3]	0.4709	0.6339	0.9955	0.2840	0.4338	0.9898	17.263M
	Texture map (ours)	0.4831	0.6442	0.9957	0.3512	0.5116	0.9929	17.264M

Table 4. Ablation study of the effectiveness of adding a masked texture map as an additional model input. We conduct experiments using U-Net. The segmentation performance improves when using the masked texture map as an additional input during finetuning after texture map training.

Method	Model input	100% (N = 800)			5% (N = 40)		
		JSI	F1-score	Acc	JSI	F1-score	Acc
No pretraining	RGB (3-ch)	0.4638	0.6278	0.9955	0.3021	0.4551	0.9918
	RGB+Texture (4-ch)	0.4606	0.6221	0.9954	0.3208	0.4743	0.9924
Texture map pretraining	RGB (3-ch)	0.4796	0.6422	0.9957	0.3442	0.5051	0.9919
	RGB+Texture (4-ch, ours)	0.4831	0.6442	0.9957	0.3512	0.5116	0.9929

5.4 Ablation Study

Incorporating the masked texture map as an additional input during the finetuning stage led to significant improvements in wrinkle segmentation, demonstrating the effectiveness of our approach. Table 4 presents quantitative comparisons using the U-Net architecture to assess the benefits of including a 1-channel masked texture map as an additional input during finetuning. We compare our pretraining method (Texture map pretraining) with a conventional approach (No pretraining), which is trained solely on manually labeled data, both with (RGB+Texture) and without (RGB) the additional masked texture map input.

6 Discussion

Our approach achieves state-of-the-art performance when compared to two publicly released models specifically designed for wrinkle segmentation, in addition to outperforming ImageNet pretrained models and self-supervised learning methods. We demonstrate that our two-stage training strategy significantly enhances wrinkle segmentation efficiency. Furthermore, our approach shows the potential

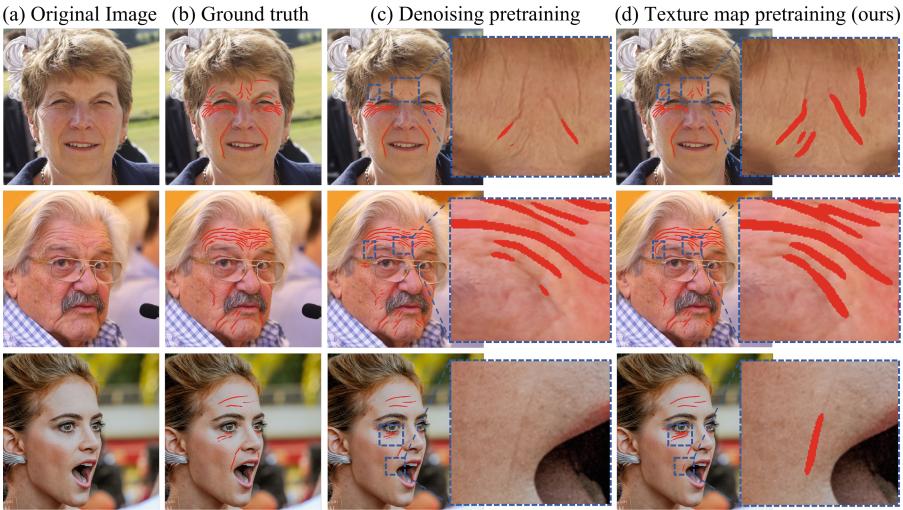


Fig. 6. Qualitative comparison against the denoising pretraining method. The blue boxes highlight areas with significant visual differences. (a) Face image. (b) Ground truth wrinkle. (c) Predicted wrinkles from a model using self-supervised learning with denoising pretraining, followed by finetuning with a manual wrinkle mask. (d) Predicted wrinkles from our model, trained with weak supervision using a masked texture map and then finetuned with a manual wrinkle mask. (Color figure online)

to achieve high performance with limited data, which could enhance scalability and flexibility in clinical settings. By using a large amount of weakly labeled data obtained automatically through filters for weakly supervised training and then finetuning with a small amount of reliable manually labeled data, we significantly reduce the time and cost required for manual labeling while improving the segmentation performance of facial wrinkles. To minimize subjectivity in the manual labeling process, we effectively combine data labeled by multiple annotators, resulting in more reliable training data. Additionally, to enhance the reproducibility of our research and reduce the manual labeling costs for subsequent studies, we release the dataset publicly available, which can also serve as a benchmark dataset for future research. The performance improvement of facial wrinkle segmentation through transfer learning has not been conducted in previous research, indicating that our approach can be efficiently integrated into various tasks related to facial wrinkle detection and segmentation tasks. Additionally, since this research falls under the broader category of thin object detection tasks, it is expected to be widely applicable to studies requiring segmentation of thin objects (e.g., fundus imaging, vascular imaging).

According to our experimental results, the performance of the Swin UNETR, a hybrid transformer-CNN architecture, is lower compared to the standard CNN-based U-Net. In our case, the dataset used for finetuning is relatively small, making it insufficient to generalize transformer models, which primarily perform

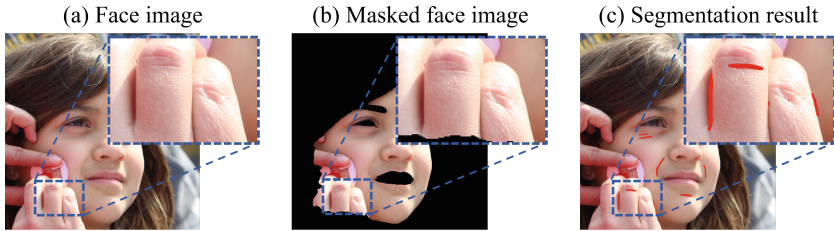


Fig. 7. Example of a false wrinkle detection. (a) Face image. (b) Masked face image used as the model input during the finetuning stage. (c) Visualization of the model’s predicted segmentation after the finetuning stage.

well in data-intensive environments due to their low inductive bias [6]. Especially in the case of wrinkles, the relationship between adjacent pixels (skin) plays a crucial role in their assessment. Therefore, the CNN-based standard U-Net, which excels at capturing local information, tends to outperform the Swin UNETR, which includes transformer blocks specialized in capturing global context through multi-head attention mechanisms. Nevertheless, our experimental results show that the performance of Swin UNETR progressively improves through our method, suggesting that with more data and longer pretraining, there is significant potential for performance enhancement. Note that accuracy is very high in all experiments since wrinkles occupy a very small proportion of the face and most of the predictions are background pixels.

However, our approach has limitations. As shown in Fig. 7, objects similar to wrinkles, such as hair or fingers covering the face, are mistakenly recognized as wrinkles in the images. This results in false positives during the wrinkle segmentation process. To address this issue, upcoming studies will focus on developing techniques that can accurately segment facial regions and precisely distinguish between wrinkle and non-wrinkle areas to reduce false positives. Also, there may be benefits to including the type of wrinkle (e.g., static vs. dynamic wrinkle) to each wrinkle in the facial image because treatment strategies often differ by the type in clinics [7, 8, 28]. Despite majority voting, the subjectivity in wrinkle annotation remains a challenge. Moving forward, we plan to collaborate with dermatologists for wrinkle annotation and explore techniques such as soft labeling to improve the reliability and trustworthiness of ground truth wrinkles.

7 Conclusion

We propose a two-stage learning strategy for facial wrinkle segmentation that leverages transfer learning from facial texture feature extraction. Specifically, the model is pretrained using automatically generated weak wrinkle labels (masked texture maps) to learn general facial features such as contours and skin texture. The model is then finetuned with a smaller set of manually labeled wrinkle data to enhance segmentation performance. This method demonstrates both qualitatively and quantitatively superior results, achieving state-of-the-art performance.

Consequently, it significantly reduces the time and cost of manual wrinkle labeling, offering potential benefits in cosmetic dermatology. Additionally, the pre-training method's architecture-independent nature suggests its broad applicability to various segmentation models, making it valuable not only in facial wrinkle segmentation but also in other areas requiring the segmentation of thin objects where manual labeling is costly. To support ongoing research and reproducibility, we have made the FFHQ-Wrinkle dataset-the first publicly available dataset of its kind-accessible to the research community. This dataset comprises 1,000 manually labeled wrinkle images and 50,000 weakly labeled images. By sharing this dataset, we aim to facilitate the development of more advanced wrinkle detection models and promote further advancements in this field.

Acknowledgements. The authors appreciate Dr. Ik Jun Moon, a dermatologist at Asan Medical Center, for sharing invaluable insights and feedback from a dermatological perspective. This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and ICT (MSIT) (RS-2024-00455720 & RS-2024-00338048), the National Institute of Health(NIH) research project (2024ER040700), the National Supercomputing Center with supercomputing resources including technical support (KSC-2024-CRE-0021), Hankuk University of Foreign Studies Research Fund of 2024, the artificial intelligence semiconductor support program to nurture the best talents (IITP(2024)-RS-2023-00253914) grant funded by the Korea government, and the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (RS-2024-00332210).

References

1. Allemann, I.B., Baumann, L.: Hyaluronic acid gel (juvéderm^TM) preparations in the treatment of facial wrinkles and folds. *Clin. Interv. Aging* **3**(4), 629–634 (2008)
2. Aznar-Casanova, J., Torro-Alves, N., Fukusima, S.: How much older do you get when a wrinkle appears on your face? Modifying age estimates by number of wrinkles. *Aging Neuropsychol. Cogn.* **17**(4), 406–421 (2010)
3. Brempong, E.A., Kornblith, S., Chen, T., Parmar, N., Minderer, M., Norouzi, M.: Denoising pretraining for semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4175–4186 (2022)
4. Chen, J., He, M., Cai, W.: Facial wrinkle detection with multiscale spatial feature fusion based on image enhancement and ASFF-SEUnet. *Electronics* **12**(24), 4897 (2023)
5. Crum, W.R., Camara, O., Hill, D.L.: Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans. Med. Imaging* **25**(11), 1451–1461 (2006)
6. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
7. Gao, L., et al.: Clinical efficacy of different therapeutic modes of CO2 fractional laser for treatment of static periocular wrinkles in Asian skin. *J. Cosmet. Dermatol.* **21**(3), 1045–1050 (2022)

8. Goldman, A., et al.: Hyaluronic acid dermal fillers: safety and efficacy for the treatment of wrinkles, aging skin, body sculpturing and medical conditions. *Clin. Med. Rev. Ther.* **3** (2011)
9. Hatamizadeh, A., et al.: Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) *BrainLes 2021*. LNCS, vol. 12962, pp. 272–284. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-08999-2_22
10. 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)
11. Ji, Z., Shen, Y., Ma, C., Gao, M.: Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In: Shen, D., et al. (eds.) *MICCAI 2019*. LNCS, vol. 11766, pp. 175–183. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_20
12. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410 (2019)
13. Kim, K., Choi, Y.H., Hwang, E.: Wrinkle feature-based skin age estimation scheme. In: *2009 IEEE International Conference on Multimedia and Expo*, pp. 1222–1225. IEEE (2009)
14. Kim, S., Yoon, H., Lee, J., Yoo, S.: Semi-automatic labeling and training strategy for deep learning-based facial wrinkle detection. In: *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 383–388. IEEE (2022)
15. Kim, S., Yoon, H., Lee, J., Yoo, S.: Facial wrinkle segmentation using weighted deep supervision and semi-automatic labeling. *Artif. Intell. Med.* **145**, 102679 (2023)
16. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022 (2021)
17. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint [arXiv:1608.03983](https://arxiv.org/abs/1608.03983) (2016)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint [arXiv:1711.05101](https://arxiv.org/abs/1711.05101) (2017)
19. Luu, K., Dai Bui, T., Suen, C.Y., Ricanek, K.: Combined local and holistic facial features for age-determination. In: *2010 11th International Conference on Control Automation Robotics & Vision*, pp. 900–904. IEEE (2010)
20. Ng, C.C., Yap, M.H., Cheng, Y.T., Hsu, G.S.: Hybrid ageing patterns for face age estimation. *Image Vis. Comput.* **69**, 92–102 (2018)
21. Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN 1994)*, vol. 1, pp. 55–60. IEEE (1994)
22. Or-El, R., Sengupta, S., Fried, O., Shechtman, E., Kemelmacher-Shlizerman, I.: Lifespan age transformation synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *ECCV 2020*. LNCS, vol. 12351, pp. 739–755. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58539-6_44
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015*. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
24. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* **115**, 211–252 (2015)

25. Sabina, U., Whangbo, T.K.: Edge-based effective active appearance model for real-time wrinkle detection. *Skin Res. Technol.* **27**(3), 444–452 (2021)
26. Sabina, U., Whangbo, T.K.: Nasolabial wrinkle segmentation based on nested convolutional neural network. In: 2021 International Conference on Information and Communication Technology Convergence (ICTC), pp. 483–485. IEEE (2021)
27. Satriyasa, B.K.: Botulinum toxin (Botox) a for reducing the appearance of facial wrinkles: a literature review of clinical use and pharmacological aspect. *Clin. Cosmet. Investig. Dermatol.*, 223–228 (2019)
28. Small, R.: Botulinum toxin injection for facial wrinkles. *Am. Fam. Phys.* **90**(3), 168–175 (2014)
29. Warren, R., Gartstein, V., Kligman, A.M., Montagna, W., Allendorf, R.A., Ridder, G.M.: Age, sunlight, and facial skin: a histologic and quantitative study. *J. Am. Acad. Dermatol.* **25**(5), 751–760 (1991)
30. Wilder-Smith, E.P.: Stimulated skin wrinkling as an indicator of limb sympathetic function. *Clin. Neurophysiol.* **126**(1), 10–16 (2015)
31. Wu, Y., Kalra, P., Thalmann, N.M.: Simulation of static and dynamic wrinkles of skin. In: Proceedings Computer Animation 1996, pp. 90–97. IEEE (1996)
32. Xie, Z., et al.: SimMIM: a simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9653–9663 (2022)
33. Xu, G., et al.: CAMEL: a weakly supervised learning framework for histopathology image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10682–10691 (2019)
34. Yang, M.Y., Shen, Q.L., Xu, D.T., Sun, X.L., Wu, Q.B.: Striped WriNet: automatic wrinkle segmentation based on striped attention module. *Biomed. Signal Process. Control* **90**, 105817 (2024)
35. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: BiSeNet: bilateral segmentation network for real-time semantic segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 334–349. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_20
36. Zhou, Z.H.: A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **5**(1), 44–53 (2018)



ECMISM: Speech Recognition via Enhancing Conformer Models with Innovative Scoring Matrices

Jiang Zhang, Liejun Wang^(✉), Yinfeng Yu, and Miaomiao Xu

College of Computer Science and Technology, Xinjiang University, Urumqi,
Xinjiang, China

{zhangjiang,xmm}@stu.xju.edu.cn, {yuyinfeng,wljxju}@xju.edu.cn

Abstract. In recent years, significant advancements have been made in speech recognition technology. The conformer encoder and attention-rescoring decoding method within the portable Wenet toolkit have garnered considerable attention. However, the conformer encoder in Wenet has limitations, such as lacking inter-layer skip connections. Additionally, while the attention-rescoring decoding method improves recognition accuracy, errors from CTC beam search decoding may accumulate during subsequent attention decoding, affecting the final recognition results. We propose an Enhancing Conformer Models with Innovative Scoring Matrices (ECMISM) model to address these issues. We optimized the attention-rescoring decoding method by introducing a novel Relationship Calculation Module (RCM). This module aims to mitigate error accumulation in attention-rescoring decoding methods, thereby reducing the impact of CTC decoding errors on subsequent attention recovery. Additionally, we introduce a Skip Fusion Module (SFM) to integrate shallow and deep features. This addresses the limitation of the Conformer encoder's lack of inter-layer skip connections, enhancing the model's capability to capture and utilize contextual information effectively. The experimental results demonstrate that our approach has achieved outstanding performance, particularly on the relatively small Uyghur dataset. Compared to the baseline models, our method reduces character error rates by 0.03%, 0.35%, and 0.44% on the Aishell1, Prime-words, and ST datasets, respectively. On the General Speech 16.1 Uyghur dataset, our approach achieves a word error rate of 4.19%, which is 1.56% lower than the baseline model.

Keywords: Speech Recognition · Attention-Rescoring · Relationship Calculation Module · Skip Fusion Module · Uyghur

1 Introduction

Speech recognition technology is a technology that converts speech signals into a format that computers can recognize. It can be used for various applications, including voice assistants, voice search, speech-to-text notes, and translation.

Traditional speech recognition methods rely on acoustic features and language models to match speech signals with predefined patterns [1]. However, they struggle with complex speech signals and unstable speech variations and require significant manual feature engineering, limiting their generalization ability across different speech signals. Subsequently, Hidden Markov Models (HMMs) [2] model speech signals and combine them with pronunciation dictionaries and language models for recognition [3]. However, they perform poorly on long speech sequences and are sensitive to variations and noise in speech signals. With the development of deep learning techniques, existing speech recognition models can be categorized into four types: models based on Connectionist Temporal Classification (CTC), models based on Recurrent Neural Networks (RNNs), models based on attention mechanisms, and models based on hybrid approaches (HA).

The CTC [4, 5] model solves the alignment problem between input and output. However, the CTC model has a drawback because it relies on the independence assumption. Liu and colleagues have proposed the Gram-CTC method [6], which captures Long-term dependencies by introducing label dependency graphs. The RNN effectively addresses the issue of the independence assumption, but it is prone to challenges like vanishing and exploding gradients. To tackle this problem, the Long Short Term Memory (LSTM) [7] was proposed, and subsequently, Jorge and colleagues introduced the Bi-directional Long Short-Term Memory (BiLSTM) model [8], allowing for the utilization of bidirectional information, thereby improving decoding accuracy. Recurrent Neural Network Transducer (RNN-T) [9] introduced a new novel fusion approach that requires higher memory resources. To address the high memory and computational resource demands of RNN-T, the Boundary Aware Transducer (BAT) was proposed. BAT introduces Continuous Integrate-and-Fire (CIF) alignment to trim the lattice, significantly reducing memory and time overhead during training.

Attention models [10] have, to some extent, addressed the parallelization issue. Lin and others introduced the SpeechTransformer model [11], an attention-based approach for speech recognition. However, the Transformer model overlooks detailed information. The Conformer model [12] integrates the strengths of Convolutional Neural Networks (CNNs) and Transformer models, effectively combining global and detailed information [26]. However, this also results in higher complexity. Subsequently, Efficient Conformer [13], and UCONV-CONFORMER [14] have improved the Conformer model's complexity to enhance its speed. At the forefront of these enhancements, the focus is on elevating the Transformer encoder. Additionally, advancements in the decoder realm encompass the bidirectional Transformer decoder and HA. The bidirectional Transformer decoder considers forward and backward context information to enhance contextual understanding. Meanwhile, the HA combines different types of decoders to leverage their respective strengths, thereby improving decoding accuracy and robustness. Some popular HA include CTC + Attention [15], RNN + Attention [16], and RNN + CTC [17], among others.

The Conformer-Transformer (CT) model is the commonly used speech recognition architecture. During training, a joint loss comprising CTC and attention is

employed. Additionally, the inference process utilizes Attention-rescoring (AR). However, its Conformer encoder is more complex and only focuses on deep features. During the AR decoder’s decoding process, many incorrectly decoded results are fed into the attention model [27]. In the autoregressive process, the attention model accumulates errors, leading to the issue of error accumulation. To address these issues, we have made improvements in the following aspects:

1. RCM (Relationship Calculation Module): We proposed the RCM, which optimizes the AR decoder’s scoring method. RCM utilizes differential-based techniques and novel regularization methods to compute attention scores more effectively. It alleviates the impact of errors in CTC decoding results on subsequent attention models, thereby enhancing the decoding performance of AR.
2. SFM (Inter-layer Skip Fusion Module): We optimized the Conformer encoder by designing an SFM. SFM effectively combines deep and shallow features by averaging, enabling the fusion of detailed information from shallow features and semantic information from deep features. This enhancement improves the model’s ability to capture and utilize contextual information.
3. InLoss (Internal Loss): We introduce a downsampling method that reduces the complexity of the Conformer encoder. InLoss calculates the difference between downsampled and original data, incorporating this difference into the final model loss for optimization.

2 Methods

The method of this paper follows the structure of the decoder-encoder. In the encoder part, we use a conformer as the encoder and design a Skip-Fusion Module (SFM) based on it to integrate shallow and deep features. Additionally, to reduce information loss during downsampling and speed up training, we introduce an InLoss module. In the decoder part, we propose a novel Relationship Calculation Module (RCM) to address the issue of error accumulation in the attention recovery decoding method. The overall workflow of the model is illustrated in Fig. 1.

2.1 Inter-layer Skip Fusion Module

The skip connections in the Conformer model are primarily present within Conformer layers, and there are no skip connections between Conformer layers. Shallow layers contain more detailed information, while deep layers contain more semantic information [28]. Therefore, we propose a method of Inter-layer skip connections to fuse Shallow-layer features with Deep-layer features. This approach can accelerate the model’s convergence and achieve good recognition performance. We have adopted an additive averaging fusion method. In Fig. 1, it can be observed that we fused the outputs of the Conformer from the 5th, 8th, and 11th layers. This is because the Conformer layers at shallower depths contain more noise, which, in turn, adversely affects recognition results.

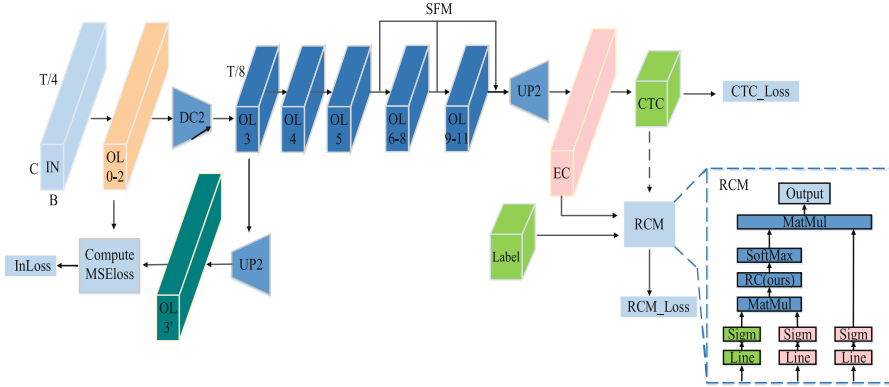


Fig. 1. The overall workflow of the ECMISM + InLoss. DC2 refers to downsampling convolution with a sampling rate of 2, and UP2 refers to uppooling with a sampling rate 2. OL0 represents the output of layer 0, and the same applies to others. EC refers to the output of the shared encoder. During the training process, the output of the shared encoder undergoes joint decoding with CTC and REM decoders. The dashed line in the figure represents the inference process. In the inference process, the CTC decoding result is input and fed into RCM for re-decoding. When the InLoss module is not utilized, the output shape for each encoder layer remains consistent. Specifically, the size of OL3-11 is the same as that of OL0-2.

2.2 Internal Loss

The optimization process for conformer models has various methods, with the most direct one being downsampling the input length. Reducing the input sequence length can effectively reduce training time. However, if solely relying on downsampling to reduce the length of the speech input, it inevitably leads to a loss of information, and such loss is irreversible.

Therefore, we perform downsampling using convolutional operations with a stride of 2 and upsampling using bilinear interpolation. The specific operation involves replacing the original pooling downsampling operation with a convolution operation. This is because convolution is a parameterized operation, allowing for subsequent optimization, while pooling is a parameter-free operation and cannot undergo further optimization. Our approach uses bilinear interpolation for upsampling to minimize the number of parameters.

In the specific process, the output OL2 from the second layer of the Conformer undergoes downsampling with a convolution operation of stride 2. Subsequently, it undergoes an upsampling operation using bilinear interpolation, resulting in the output OL2'. The difference between OL2 and OL2' is calculated, referred to as the internal loss, and is added to the final loss for joint optimization. As training progresses, the convolution operation with a stride of 2 learns the optimal downsampling result, minimizing information loss. This method of internal loss reduces input length, accelerates training speed, and maximizes the reduction of information loss.

The choice of using the output from the second layer of the Conformer as the starting layer for calculating the internal loss is due to the possibility that earlier encoding layers may contain more noise. Calculating the internal loss too early might lead to less stable training. Unlike the reference literature [14], our InLoss does not require a decoding process. We compute the MSE (Mean Squared Error) Loss, as shown in Eq. (1), Eq. (2):

$$L(y, y') = \frac{1}{n} * \sum (y, y')^2 \quad (1)$$

$$\text{InLoss} = L(\text{Upsample2}(\text{Conv2}(\text{OL}_2)), \text{OL}_2) \quad (2)$$

where Conv2 represents a 1D convolution with a stride of 2. Upsample2 represents upsampling with a stride of 2. Finally, we add the InLoss to the final loss for unified optimization.

2.3 Relationship Calculation Module

In the Wenet toolkit [18], the AR decoding approach is employed, wherein the decoding results of CTC are passed through an attention model for re-scoring to achieve improved decoding results, as the decoding results of CTC beam search involves selecting the top n possible outcomes as inputs for the attention model. When the input contains a substantial number of erroneous characters, the decoding results of the attention model are adverse. Therefore, we introduce the RCM. RCM primarily improved the computation method of the score matrix, as shown in Fig. 2. Firstly, it subtracts the numbers on the diagonal from the original score matrix to obtain a difference matrix. This step weakens the association of each position with itself, as the diagonal numbers typically represent the strongest association with each position. We can focus on information from other positions through the difference matrix while considering global information. Next, this difference matrix undergoes regularization processing to ensure that the scores for each position fall within a specific range, ultimately yielding the attention scores. After regularization processing, the resulting attention scores can better balance the degree of association between each position and others. This enhanced approach improves the accuracy and reliability of RCM when computing attention.

The process of regularization is referred to as Regularization Control (RC). Its function is shown in Fig. 3, and the calculation formula is represented by Eq. (3). In the formula, when the value of S is smaller, the function's highest point is lower, resulting in lower scores along the diagonal of the score matrix. On the other hand, when the value of R is smaller, the distribution range of scores becomes wider, indicating a smaller inhibitory effect on the score matrix. D represents the difference matrix.

$$\text{Score} = S(\text{sigmod}(R * D) * (1 - \text{sigmod}(R * D))) \quad (3)$$

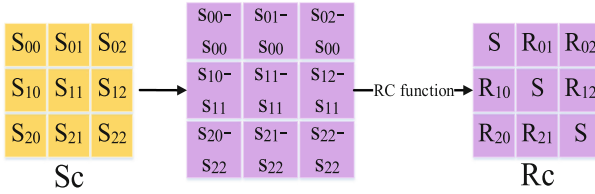


Fig. 2. Calculation process of the scoring matrix. S_c represents the original score matrix, and R_c represents the new one. The values on the diagonal of the new score matrix are hyperparameter S .

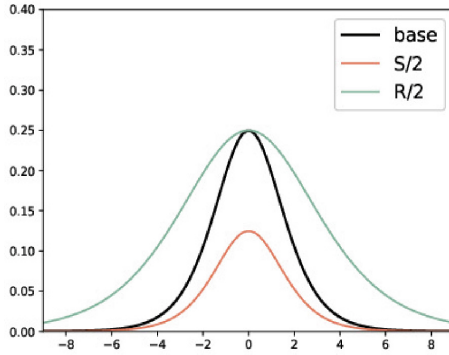


Fig. 3. RC function. When the input is 0, the output is our hyperparameter S , representing the score matrix’s maximum value. S controls the magnitude of self-correlation, while R governs the aperture size of the function, controlling the values on the diagonal matrix and the correlation at other positions.

3 Results and Analysis

In the experiment, we utilized four datasets, three Chinese and one a Uyghur language dataset. The Chinese datasets we used were Aishell [19], Primewords Chinese Corpus Set 1 (Primewords), and the Free ST Chinese Mandarin Corpus (ST). For the Uyghur dataset, we utilized Common Voice 16.1 [20]. We referred to it as Ug in the experiment.

Aishell1 contains approximately 178 h of speech data. There were 120,000 samples for training, 14,000 for validation, and 7,000 for testing. Primewords is also a Chinese speech recognition dataset consisting of approximately 100 h of audio data. However, unlike the Aishell 1 dataset, Primewords does not come pre-divided into training, validation, and test sets. Following the split ratio used in the Aishell 1 dataset, we randomly partitioned the Primewords dataset into training, validation, and test sets with a ratio of 0.85:0.1:0.05. The total duration of the ST dataset is 109 h, consisting of 102,600 WAV files. Similarly, no division has been performed, and we have partitioned the dataset following a ratio of 0.85:0.1:0.05. In the Ug dataset, 147 h of validated data were available.

The dataset already provided both test and validation sets. We incorporated all remaining data into the training set.

On the Aishell, ST, and Ug datasets, the batch size is 36, while on the Primewords dataset, the batch size is 24. The hyperparameter settings remain consistent across the three Chinese datasets. However, due to linguistic differences, there are slight variations in hyperparameter settings compared to the Uyghur dataset. The Table 1 presents the settings for other hyperparameters. It is crucial to emphasize that CTC_weight refers to the weight assigned to the CTC branch. During the training process, the weight for the CTC decoder is set to 0.3, while during the inference process, the CTC weight is adjusted to 0.5. These settings align with those of our baseline system, the Wenet toolkit.

Table 1. Experimental Parameter Settings. Dff refers to the dimension of the hidden layer in the Feedforward Neural Network, Num_Mel refers to the input feature dimension, Accum_Grad refers to Accumulated Gradient.

Parameter Name	Parameter for Chinese	Parameter for Ug
Num_Conformer	12	12
Num_Decoder	6	6
Dff	2048	2048
Epoch	240	100
Lr	0.002	0.001
CTC_weight	0.3/0.5	0.3/0.5
Num_Mel	80	80
Frame_Length	25	25
Frame_Shift	10	10
Spec_Aug	True	True
Accum_Grad	4	4

Our baseline system was implemented using the Wenet toolkit. The standard CT model used a downsampling rate of 4, referred to as the CT_D4 model, while the CT model with a downsampling rate of 8 was called the CT_D8 model. In the experiments, we used CT_D4 and CT_D8 as our baseline systems, keeping the parameter settings consistent with Wenet. For this experiment, we utilized a server with 4 T4 GPUs, each with a 16GB memory capacity. The CPU used in the server is an Intel Xeon Gold 5218R.

The evaluation metrics employed in the experiments include Character Error Rate (CER), Word Error Rate (WER), and the model’s training time. CER is more easily interpretable in Chinese speech recognition, while WER is influenced by the segmentation system. Therefore, we adopt CER as the evaluation metric for Chinese. However, for Uyghur language, we utilize WER as the evaluation criterion. The smaller the CER and WER values, the better the performance. In the following table, CTC-Greedy, CTC-Beam, and Attention-rescoring represent

the CER of the decoding strategies CTC greedy search, CTC beam search, and Attention-rescoring method.

3.1 Comparative Experiments

We selected nine popular speech recognition methods and conducted comparative experiments on four datasets. Among them, CT_D4 [18] and CT_D8 [18] are variants based on the Conformer_Transformer model, which holds significant importance in speech recognition. The CT_D4 model reduces the input length to one-fourth of the original size using a downsampling factor of 4, while CT_D8 has a downsampling rate of 8. Both Speech Transformer [11] and Transformer (Wenet) [18] are based on the Transformer model. Wenet’s Transformer adopts a U-shaped architecture and combines Attention loss and CTC loss, resulting in improved speech recognition performance. BAT [22] and RNNT [23] are both based on the RNN-T model. RNNT integrates acoustic information with historical output information, albeit with higher training costs, while BAT effectively reduces training costs by limiting the search path. Efficient Conformer [13] is an improved version of the Conformer model, enhancing training and inference speed and reducing costs by adding downsampling in the middle layers of the Conformer. NST [24] is a semi-supervised training method that involves training a series of models and using previously trained models to process unlabeled data, obtaining corresponding labels for subsequent model training. Additionally, we compared these methods with the large-scale Whisper [25] model. Unfortunately, the Whisper version does not support Uyghur language speech recognition.

Table 2. Comparative Experiments with SOTA Models. Prime refers to the Prime-words dataset

Model	CER/Aishell1 (↓)	CER/ Prime (↓)	CER/ST(↓)	WER/Ug(↓)
CT_D4(baseline) [18]	4.61	12.90	7.95	5.75
CT_D8 [18]	4.75	13.52	9.07	6.13
Speech Transformer [11]	8.97	19.27	-	-
Transformer(Wenet) [18]	5.30	14.95	8.20	6.28
Paraformer [21]	4.95	13.19	7.68	6.97
BAT [22]	4.82	15.56	8.56	6.32
NST [24]	4.85	12.97	7.63	-
Whisper(base) [25]	20.04	31.19	22.66	-
Whisper(large-v3) [25]	6.94	16.88	9.22	-
Rnnt [23]	4.60	12.79	7.81	6.03
Efficient Conformer [13]	4.56	12.71	7.62	5.16
ECMISM(ours)	4.58	12.55	7.51	4.19

Table 2 reveals that compared to the baseline system CT_D4, our model achieved reductions in error rates of 0.03%, 0.35%, 0.44%, and 1.56%, respec-

tively. Our ECMISM model significantly improved accuracy by reducing CER to 4.58% on the Aishell1 dataset, closely approaching the state-of-the-art (SOTA) Efficient Conformer model (CER of 4.56%). Our model exhibited excellent performance on the Uyghur language dataset, with a performance improvement of 1.56% over the baseline system. However, the improvement on the Aishell1 dataset was relatively modest, attributed to the linguistic characteristics of the Uyghur language, which has various word forms leading to increased errors during CTC decoding. Our model effectively mitigates such cumulative errors. While CTC decoding performs well for Mandarin Chinese, our model’s correction effect is slightly less pronounced, yielding improvements. Compared to transformer models, our approach leveraging global and local information significantly enhanced encoding capabilities, reducing error rates by 0.72%, 2.4%, 0.69%, and 2.09% across the four datasets. Compared to the Efficient Conformer model, our approach corrected errors accumulated during CTC decoding, resulting in superior decoding outcomes, notably reducing error rates by 0.97% on the Uyghur dataset. It’s worth noting that whisper (base) performed the worst among these datasets, attributed to its multi-tasking nature, which excels in handling multiple languages but falls short in single-language tasks compared to single-task models.

3.2 Ablation Experiment

Table 3. Ablation Studies. ECMISM-RCM represents the experimental results after removing RCM from ECMISM. The others follow the same logic. time refers to the training time. Prime refers to the Primewords dataset.

Model	CER(↓)		Time(h)(↓)	
	Aishell1	Prime	Aishell1	Prime
ECMISM(ours)	4.58	12.55	45	26
ECMISM-RCM	4.61	12.71	45	26
ECMISM-SFM	4.59	12.67	45	26
ECMISM-RCM-SFM	4.61	12.90	45	26
ECMISM-RCM-SFM+InLoss	4.70	14.37	32	19
ECMISM+InLoss	4.69	13.01	32	19

From Table 3, it is evident that, upon removing the RCM module, the error rates increased by 0.03 and 0.16, while removing the SFM module resulted in increases of 0.01 and 0.12, respectively. Simultaneously removing both modules led to increases of 0.03 and 0.35. The results indicate that both the RCM module and SFM module show significant improvement on the Prime dataset. On the Aishell 1 dataset, the SFM module has a relatively minor impact. However, as demonstrated in Table 6, the SFM module exhibits substantial advantages for

individual models. The fusion model and our SFM function share similarities, as both can integrate information across multiple layers. When both modules are utilized together, a better overall performance is achieved. This is attributed to the SFM’s effective fusion of deep and shallow features, suppressing erroneous CTC decoding information after passing through the RCM module, resulting in improved decoding outcomes. In addition, our ECMISM + InLoss model reduces training time by approximately 28% while only sacrificing a minimal 0.08% and 0.11% decrease in accuracy.

3.3 Module Detail Experiment

In this section of the experiments, we only present the experimental results on the Aishell1 and Prime words datasets. Our RCM model introduces two parameters, namely S and R. We conducted an extensive series of experiments to assess the impact of the S and R parameters on the experimental results, and the specific outcomes are presented in Table 4.

Table 4. Impact of S and R Parameters on RCM.

Value of S&R	CTC-Greedy(↓)		CTC-Beam(↓)		Attention-rescoring(↓)	
	Aishell1	Prime	Aishell1	Prime	Aishell1	Prime
S = 1, R = 1	4.96	13.07	4.96	13.06	4.59	12.67
S = 1, R = 0.5	5.06	13.52	5.06	13.52	4.69	12.93
S = 0.5, R = 0.5	None	13.95	None	13.95	None	13.14
S>1	None	None	None	None	None	None
R>1	None	None	None	None	None	None
S = 0.5, R = 1	5.15	13.78	5.15	13.77	4.77	13.09

The data in Table 4 were all obtained by averaging the results of 80 models. In the table, None represents cases where the model did not converge. It can be observed that the best performance is achieved when S, R is equal to 1. Excessive increases or decreases in correlation can make attention scores sparse, resulting in poor performance. In the subsequent experiments, we set S and R to their optimal values, configured as 1.

Table 5 shows the experimental data of the RCM module. The value for CTC greedy search in the second row is 4.98%, slightly higher than the 4.94% reported in the Wenet paper. The Wenet code has been updated, leading to a minor deviation in experimental results. From the table, we can see that our RCM model achieves a CER of 4.59% and 12.67% after an average of 80 models, surpassing Wenet’s 4.61% and 12.90%. The reason for averaging over 80 models is that, after incorporating RCM, the later-stage models in training become relatively stable, with minor parameter differences between them. This can be observed in the loss curve in Fig. 4. Without using an averaging model, i.e., in

Table 5. Validation of the Effectiveness of the RCM. AV refers to the average number of models used in decoding.

Model	AV	CTC-Greedy(↓)		CTC-Beam(↓)		Attention-rescoring(↓)	
		Aishell1	Prime	Aishell1	Prime	Aishell1	Prime
CT_D4 (baseline)	1	5.91	14.36	5.90	14.36	5.52	13.98
	20	4.98	13.49	4.98	13.49	4.61	12.90
	40	4.98	13.54	4.98	13.54	4.67	12.93
	80	4.99	13.73	4.99	13.73	4.78	13.17
CT_D4 + RCM	1	5.79	14.10	5.78	14.09	5.26	13.44
	20	5.01	13.32	5.01	13.32	4.64	12.99
	40	4.96	13.10	4.96	13.10	4.63	12.73
	80	4.96	13.07	4.96	13.06	4.59	12.67

the case where the average model in the first and fifth rows of the table is 1, the addition of our RCM module improved recognition accuracy by 0.26% and 0.54%.

Table 6 shows the experimental data of the SFM module. For comparison convenience, we extracted partial data from Table 5 and placed it in Table 6. Comparing Table 5 and Table 6, It can be observed that after adding our SFM module, the optimal performance on the two datasets is 4.61 and 12.71, respectively. The performance on the Aishell1 dataset remains consistent with the baseline, while we achieve a performance improvement of 0.19 on the Prime-words dataset. However, our experimental results for individual models are significantly better than the baseline, with an improvement of 0.39 and 0.88. This is because the fusion model also employs a method of information fusion, enhancing the decoding performance. However, the fusion model cannot be optimized during the training process. In contrast, our SFM can be optimized during training. When used in conjunction with the RCM module, SFM achieves superior results. This can also be observed in the loss shown in Fig. 4, where after adding SFM, the convergence is faster, and the loss is lower.

The above experiments have confirmed the effectiveness of our ECMISM model. Next, we incorporate the InLoss module into our model. The data in Table 7 represents the experimental results on the Aishell1 dataset. The data in Table 7 shows that the model incorporating the InLoss method outperforms the CT_D8 model significantly, especially when using beam search decoding. This also indicates that the InLoss method has certain advantages in feature extraction. Compared to CT_D4, our model also significantly improves training speed, saving approximately 28% of training time.

Table 6. Validation of the Effectiveness of SFM.

Model	AV	CTC-Greedy(↓)		CTC-Beam(↓)		Attention-rescoring(↓)	
		Aishell1	Prime	Aishell1	Prime	Aishell1	Prime
CT_D4 (baseline)	1	5.91	14.36	5.90	14.36	5.52	13.98
	20	4.98	13.49	4.98	13.49	4.61	12.90
CT_D4 + SFM	1	5.69	14.03	5.70	14.03	5.13	13.10
	20	4.99	13.19	4.99	13.19	4.67	12.71
	40	4.98	13.19	4.98	13.18	4.61	12.72
	60	4.97	13.34	4.97	13.34	4.64	12.83
	80	4.96	13.47	4.96	13.46	4.64	12.89

Table 7. Validation of the Effectiveness of InLoss.

Model	AV	CG(↓)	CB(↓)	AR(↓)	Time(h)
CT_D8	1	5.75	5.74	5.31	31
	20	5.10	5.10	4.78	
	40	5.07	5.07	4.75	
	60	5.07	5.06	4.75	
	80	5.06	5.05	4.77	
ECMISM + InLoss	1	5.69	5.70	5.13	32
	20	5.17	5.04	4.72	
	40	5.14	5.03	4.72	
	60	5.10	4.98	4.69	
	80	5.09	4.99	4.70	

3.4 Visualization

For an easier demonstration of the effectiveness of our model, we plotted the validation loss figure on the Aishell1 dataset, as illustrated in Fig. 4. The two black lines represent the baselines. The orange line in the figure represents the validation loss after incorporating the SFM module. It can be observed that the convergence speed of the model has significantly increased, with the loss at the 75th batch being the lowest among all models. The blue line, reflecting the loss after integrating our RCM module, shows that the loss is already below the baseline system after the 175th iteration, indicating the effectiveness of our RCM module. Our ECMISM model, combining the advantages of RCM and SFM modules, is represented by the green line, exhibiting the lowest loss. The red line illustrates the loss after incorporating the InLoss module, which appears to be the highest. However, this is because the InLoss is added to the original loss. In reality, its recognition performance only experiences a slight decline, namely 0.08% and 0.11% on the two datasets, while achieving a 28% improvement in training speed.

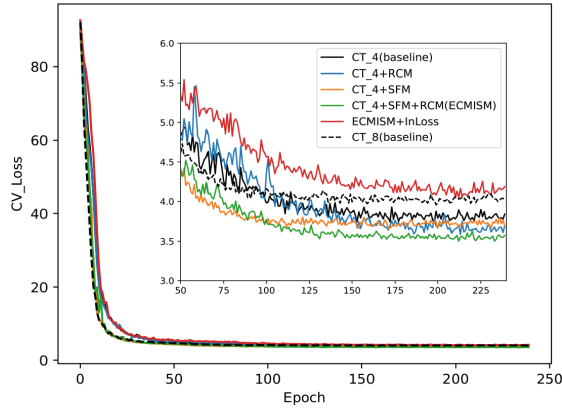


Fig. 4. Validation loss.

In speech recognition, various errors may occur. I created Fig. 5 to analyze and obtain insights into these errors. The data in the figure depict statistics on various error types in the ST dataset. The test set of the ST dataset comprises a total of 55,833 characters. In the figure, ‘S’ represents substitution errors, indicating recognition mistakes; ‘D’ denotes deletion errors, indicating characters present in the labels but missed by the model; ‘I’ stands for insertion errors, representing characters identified by the model but not originally present in the labels. It can be observed that the overall error rate of our model is significantly lower. Specifically, substitution and insertion errors show a noticeable reduction. However, there is no significant change in deletion errors. Deletion errors are typically caused during the encoder’s encoding process. Our model employs the same encoder as the baseline, resulting in a limited improvement in this aspect.

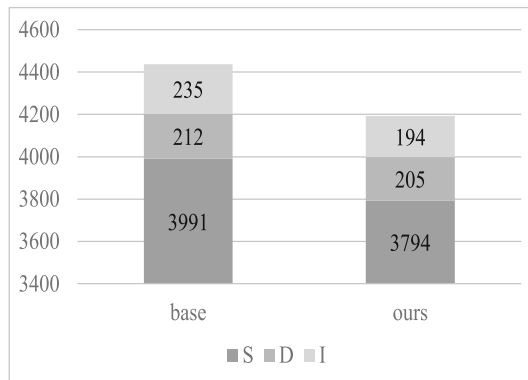


Fig. 5. Error Analysis.

4 Conclusions

We propose an ECMISM speech recognition model based on the Conformer Transformer architecture. In the encoder part, we design the SFM to integrate shallow and deep features, obtaining more informative representations. In the decoder part, we propose a novel RCM to address the issue of error accumulation in the attention recovery decoding method, specifically mitigating the impact of CTC decoding errors on subsequent Attention-rescoring. Experimental results demonstrate the positive contributions of these two modules to the speech recognition model. On the Aishell1, Prime, and ST datasets, we achieved CER of 4.58%, 12.55%, and 7.51%, respectively, which are lower than the baseline error rates of 4.61%, 12.90%, and 7.95%. At the same time, we achieved a WER of 4.19% on the Common Voice 16.1 Uyghur dataset, which is lower than the baseline of 5.75%. To expedite training, we introduce the InLoss module. Integrating the InLoss module into our ECMISM model accelerates training by 28%, with only marginal error rate increases on the first two datasets, by 0.08% and 0.11%, respectively. Our InLoss module is crucial for the lightweight design of the model.

Our model still has several limitations. For instance, the SFM module employs a simple linear summation for information fusion. Dynamic attention fusion could be explored in future iterations, assigning varying weights to different layers to better capture feature information. Additionally, while our InLoss method significantly reduces training time, it slightly increases recognition error rates. In subsequent work, more effective approaches to reduce model complexity could be investigated.

Acknowledgements. This work was supported by these works: the Tianshan Excellence Program Project of Xinjiang Uygur Autonomous Region, China (2022TSY-CLJ0036); the Central Government Guides Local Science and Technology Development Fund Projects (ZYXD2022C19); the National Natural Science Foundation of China under Grant 62463029 and 62303259, and in part by the Graduate Research Innovation Project of Xinjiang Uygur Autonomous Region under Grant XJ2021G065.

References

1. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**(2), 257–286 (1989)
2. Juang, B.H., Rabiner, L.R.: Hidden Markov models for speech recognition. *Technometrics* **33**(3), 251–272 (1991)
3. Cui, X., Gong, Y.: A study of variable-parameter Gaussian mixture hidden Markov modeling for noisy speech recognition. *IEEE Trans. Audio Speech Lang. Process. (TASLP)* **15**(4), 1366–1376 (2007)
4. Graves, A., Fernández, S., Gomez, F., et al.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, pp. 369–376. Association for Computing Machinery, New York, NY, USA (2006)

5. Lee, J., Watanabe, S.: Intermediate loss regularization for CTC-based speech recognition. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 6224–6228 (2021)
6. Liu, H., Zhu, Z., Li, X., et al.: Gram-CTC: automatic unit selection and target decomposition for sequence labelling. In: Proceedings of the 34th International Conference on Machine Learning, PMLR, vol. 70, pp. 2188–2197 (2017)
7. Amodei, D., Ananthanarayanan, S., Anubhai, R., et al.: Deep speech 2: end-to-end speech recognition in English and mandarin. In: International Conference on Machine Learning. PMLR, pp. 173–182 (2016)
8. Jorge, J., Giménez, A., Iranzo-Sánchez, J., et al.: LSTM-based one-pass decoder for low-latency streaming. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7814–7818. IEEE (2020)
9. Zhao, R., Xue, J., Li, J.: On addressing practical challenges for RNN-transducer. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Cartagena, Colombia, pp. 526–533 (2021)
10. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all You need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
11. Dong, L., Xu, S., Xu, B.: Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, pp. 5884–5888 (2018)
12. Gulati, A., Qin, J., Chiu, C.C., et al.: Conformer: convolution-augmented transformer for speech recognition. arXiv preprint [arXiv:2005.08100](https://arxiv.org/abs/2005.08100) (2020)
13. Burchi, M., Vielzeuf, V.: Efficient conformer: progressive downsampling and grouped attention for automatic speech recognition. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, pp. 8–15 (2021)
14. Andrusenko, A., Nasretdinov, R., Romanenko, A.: UCONV-conformer: high reduction of input sequence length for end-to-end speech recognition. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1–5 (2023)
15. Chan, W., Jaitly, N., Le, Q., et al.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 4960–4964 (2016)
16. Chen, C., Zhang, P.: CTA-RNN: channel and temporal-wise attention RNN leveraging pre-trained ASR embeddings for speech emotion recognition. arXiv preprint [arXiv:2203.17023](https://arxiv.org/abs/2203.17023) (2022)
17. Zhang, X., Zhang, F., Liu, C.: Benchmarking LF-MMI, CTC and RNN-T criteria for streaming ASR. In: IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China pp. 46–51 (2021)
18. Zhang, B., Wu, D., Peng, Z., et al.: WeNet 2.0: more productive end-to-end speech recognition toolkit. arXiv preprint [arXiv:2203.15455](https://arxiv.org/abs/2203.15455) (2022)
19. Bu, H., Du, J., Na, X., et al.: AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline. In: 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul. Korea (South), pp. 1–5 (2017)
20. Mozilla common voice. <https://commonvoice.mozilla.org/zh-CN/datasets>
21. Gao, Z., Li, Z., Wang, J., et al.: FunASR: a fundamental end-to-end speech recognition toolkit. arXiv preprint [arXiv:2305.11013](https://arxiv.org/abs/2305.11013) (2023)

22. An, K., Shi, X., Zhang, S.: BAT: boundary aware transducer for memory-efficient and low-latency ASR. arXiv preprint [arXiv:2305.11571](https://arxiv.org/abs/2305.11571) (2023)
23. Lee, J., Lee, L., Watanabe, S.: Memory-efficient training of RNN-Transducer with sampled softmax. arXiv preprint [arXiv:2203.16868](https://arxiv.org/abs/2203.16868) (2022)
24. Chen, Y., Ding, W., Lai, J.: Improving noisy student training on non-target domain data for automatic speech recognition. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece, pp. 1–5 (2023)
25. Radford, A., Kim, J.W., Xu, T., et al.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. PMLR, pp. 28492–28518 (2023)
26. Xu, M., Zhang, J., Xu, L., et al.: Collaborative encoding method for scene text recognition in low linguistic resources: the Uyghur language case study. Appl. Sci. **14**, 1707 (2024)
27. Zhang, J., Wang, L., Yu, Y., et al.: Nonlinear regularization decoding method for speech recognition. Sensors **24**, 3846 (2024)
28. Huang, G., Zhuang, L., Van Der Maaten, L., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)



Benchmarking AI in Mental Health: A Critical Examination of LLMs Across Key Performance and Ethical Metrics

Rui Yuan^(✉), Wanting Hao, and Chun Yuan

Tsinghua Shenzhen International Graduate School, Nanshan District, Xili University Town, Tsinghua Campus, Shenzhen 518055, Guangdong, China
yuanr22@mails.tsinghua.edu.cn

Abstract. The rapid advancement of artificial intelligence (AI) has led to an increasing application of Large Language Models (LLMs) in psychological counseling. This study focuses on a comprehensive evaluation of LLMs in this domain, moving beyond traditional case-based reasoning. We introduce a novel multi-agent LLM framework that enhances the analysis of psychological case interactions. Our approach involves expanding the Emotional First Aid dataset with diverse client backgrounds, enhancing its applicability and generalizability. A sophisticated user profile model, incorporating eight critical dimensions, is developed and applied within a multi-agent system to examine counseling scenarios. The system's performance is extensively evaluated based on accuracy, robustness, consistency, and fairness. The findings reveal significant differences among LLMs in these areas, highlighting their strengths and limitations in psychological interventions. This research underscores the need for ongoing refinement in LLM applications to ensure equitable and reliable support in psychological counseling. The detailed results and methodologies are available on the GitHub platform for further academic scrutiny and development.

Keywords: Psychological Case-Based Reasoning · Large Language Models · AI Ethics and Fairness

1 Introduction

The integration of AI with psychological counseling represents a cutting-edge advancement, offering a refined understanding of human behavior and emotional responses. Despite significant progress in AI technologies, their application in the sensitive field of psychology encounters substantial challenges, particularly in handling the complex ethical and clinical demands effectively. This underscores the urgent need for systems that can more accurately reflect the intricacies of human psychology with greater fidelity (Turing, 1950; Goodfellow et al., 2016) [1, 2]. In response to these challenges, our study pivots from traditional psychological CBR to a focused evaluation of LLMs within a multi-agent system. This

novel approach is designed not only to simulate counseling sessions but also to critically assess the performance of these models across four main dimensions: accuracy, robustness, consistency, and fairness. By leveraging advanced LLMs, we aim to address the prevailing shortcomings in AI applications for psychology, such as insufficient nuanced understanding and ethical alignment (Vaswani et al., 2017; Brown et al., 2020) [3, 5]. Furthermore, while AI’s potential in mental health care is immense, the challenges it presents are formidable. This includes replicating the complex dialogues of clinical interactions authentically. Our work utilizes a meticulously calibrated multi-agent system to better mimic the subtleties of clinical conversations, thereby enhancing the effectiveness and ethical soundness of psychological interventions (Picard, 1997; Tambe, 2011) [4, 6, 22]. Figure 1 provides a visual representation of the LLM-based inference workflow within our multi-agent system. This paper is structured to highlight several key contributions: 1. We enhance the Emotional First Aid dataset, increasing the precision, stability, and fairness of AI-generated user profiles, thereby broadening the model’s relevance across diverse contexts. 2. We introduce a sophisticated multi-agent system powered by LLMs, designed to simulate and evaluate counseling sessions, marking a significant advancement in psychological CBR. 3. We conduct a comprehensive evaluation of the LLMs’ performance in terms of ethical, moral, and personality discernment capabilities through both qualitative and quantitative analyses, ensuring their robustness and reliability in real-world scenarios. The integration of AI with psychological counseling represents a cutting-edge advancement, offering a refined understanding of human behavior and emotional responses. Despite significant progress in AI technologies, their application in the sensitive field of psychology encounters substantial challenges, particularly in handling the complex ethical and clinical demands effectively. This underscores the urgent need for systems that can more accurately reflect the intricacies of human psychology with greater fidelity. In response to these challenges, our study pivots from traditional psychological Case-Based Reasoning (CBR) to a focused evaluation of LLMs within a multi-agent system. This novel approach is designed not only to simulate counseling sessions but also to critically assess the performance of these models across four main dimensions: accuracy, robustness, consistency, and fairness. By leveraging advanced LLMs, we aim to address the prevailing shortcomings in AI applications for psychology, such as insufficient nuanced understanding and ethical alignment.

2 Related Works

2.1 Ethical and Moral Considerations in AI Research

This section builds upon the ethical challenges discussed in the introduction, exploring the specific ethical and moral considerations in the context of LLMs. Abdulhai et al. (2023) emphasize the moral foundations necessary for LLMs, identifying key principles such as beneficence, non-maleficence, autonomy, and justice. These principles guide the development and deployment of LLMs to ensure they align with human values and promote positive societal outcomes.

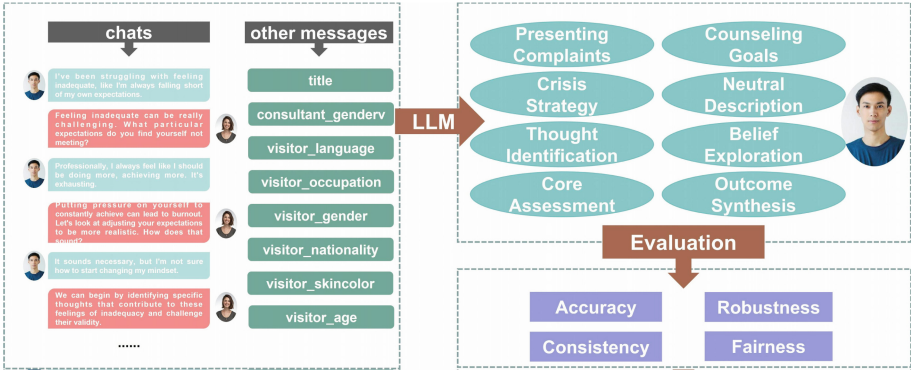


Fig. 1. LLM-Based inference workflow. This diagram illustrates the end-to-end process enabled by our LLM within the multi-agent system architecture. Starting with the collection of raw data, the workflow includes the LLM’s inference to construct detailed user profiles, culminating in the final evaluation stage.

While these principles provide a valuable framework, their implementation in practice remains challenging, particularly when balancing ethical considerations with technological advancements and efficiency. The integration of ethical and moral considerations into AI, particularly in the context of LLMs, is becoming increasingly vital as these technologies become more sophisticated and widespread. Recent studies underscore the necessity for LLMs to align with ethical frameworks and human values. Abdulhai et al. [6] emphasize the moral foundations necessary for LLMs, while Sorensen et al. [7] highlight the importance of incorporating pluralistic human values, rights, and duties into AI systems. The challenge of ensuring reliability in AI’s psychological assessments is evident in works by Huang et al. [8] and Ganesan et al. [9], who analyze the consistency and reliability of personality estimations by LLMs. These studies raise crucial questions about bias and the potential manipulation of AI characteristics, as discussed by Caron and Srivastava [10] and Jiang et al. [11]. In terms of AI’s interaction with human behavior, the work by Park et al. [12] on generative agents and the efforts by Ziems et al. [13], Xu et al. [14], and Song et al. [12] explore how AI can adapt to and reflect complex social dynamics, underlining the need for dynamic and ongoing ethical oversight.

2.2 Application of Artificial Intelligence Technology in Psychological CBR

Building upon the foundational understanding of AI in psychological CBR presented in the introduction, this section delves into the specific applications and advancements in the field. Picard (1997) pioneers affective computing, highlighting the potential of AI to recognize and respond to human emotions. Subsequent studies, such as Hirschberg and Lewis (2012) and Wang et al. (2019), demonstrate the improved accuracy of AI in emotion recognition and expression analy-

sis, opening avenues for personalized and empathetic psychological interventions. However, challenges persist, including data privacy concerns and the potential for algorithmic bias, as discussed by De Choudhury et al. (2013) and Caron and Srivastava (2022). These limitations necessitate transparent and interpretable AI models to build trust and ensure ethical applications in psychology. The integration of AI in psychological CBR has significantly advanced, especially in emotion recognition and scale assessment. Pioneered by Picard's work in affective computing (1997), which laid the groundwork for emotion recognition technologies, subsequent studies have expanded these insights, notably Hirschberg and Lewis's exploration into emotional content analysis from text (2012) [6, 15]. Deep learning further propelled this field; Wang et al.'s (2019) use of CNNs for facial expression analysis and Liu et al.'s (2020) BERT-based model for textual emotion detection illustrate the improved accuracy and potential of AI in understanding human emotions [16, 17]. However, the adoption of AI in psychological contexts is not without challenges, including data privacy concerns and model bias. De Choudhury et al. (2013) underscored the ethical and privacy issues in using social media data for mental health predictions, highlighting the need for careful data management [20]. Moreover, the demand for transparent AI models, crucial for trust and effective interventions, is addressed by innovative methods like LIME (Ribeiro et al., 2016), which aim to enhance model interpretability [19].

2.3 Development of Intelligent Agents and Their Psychological Applications

This section expands upon the concept of intelligent agents introduced in the introduction, focusing on their development and application in the field of psychology. Wooldridge (2002) and Russell and Norvig (2016) establish the theoretical foundations of multi-agent systems, highlighting their potential for simulating complex human interactions. Tambe (2011) demonstrates the application of game theory in multi-agent systems, showcasing their adaptability to psychological settings. Ziems et al. (2023) explore the use of LLMs as generative agents, furthering the potential for dynamic and personalized psychological interventions. However, ethical concerns regarding data privacy, transparency, and potential biases remain, necessitating careful consideration and oversight in the deployment of intelligent agents in psychology. Intelligent agents constitute a pivotal component of the AI spectrum, aiming to mimic and enhance human cognitive functions and decision-making. Foundational texts by Wooldridge (2002) and Russell and Norvig (2016) provide comprehensive insights into the theories and applications of multi-agent systems, setting a solid theoretical foundation for this field [20, 21]. In psychology, intelligent agents, particularly through dialogue systems, play a crucial role in behavior analysis and emotional state assessment, thereby offering new avenues for psychological interventions and reducing the workload on human counselors. These agents excel at managing the complexities of psychological case analysis. For example, Tambe (2011) demonstrated their capability in applying game theory to complex decision-making scenarios,

such as security threat mitigation, highlighting their adaptability to psychological settings [22]. Intelligent agents can simulate psychological counseling roles, offering customized support and strategies, thus improving mental health service quality. However, the integration of intelligent agents in psychology is challenged by ethical, privacy, and transparency concerns. The pursuit of ethical AI deployment in psychological practices, as discussed by Samuele (2020), underscores the importance of responsible and effective use of intelligent agents, emphasizing the need for ethical guidelines, data protection, and agent interpretability [23].

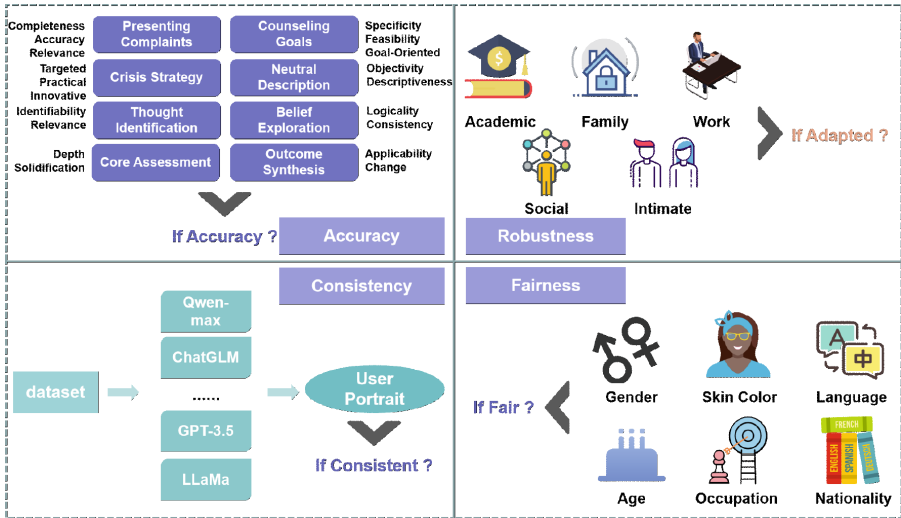


Fig. 2. User profiles evaluation criteria: A schematic highlighting accuracy, robustness, consistency, and fairness in evaluating AI-generated user profiles for mental health counseling, showcasing our dedication to innovation and ethical responsibility.

3 Case Base

3.1 Expanded Dataset Based on Emotional First Aid

In the integration of AI with psychological counseling, the quality and specificity of datasets are crucial for the effective training of AI algorithms and the empirical testing of theoretical models. While the Emotional First Aid dataset provides valuable insights into psychological counseling, it lacks diversity in client backgrounds. To address this limitation, we expanded the dataset by incorporating additional demographic and socio-psychological variables. The Emotional First Aid dataset, a pioneering open QA corpus in the field of psychological counseling, stands out for its substantial contribution to this area. Comprising 20,000 entries of counseling data, the dataset offers a comprehensive view into

the nuances of psychological counseling. This corpus is distinguished not only by its volume but also by the diversity of its content, featuring multi-turn dialogues that provide deep insights into the counseling process. Each entry is enriched with categorized information, including but not limited to counseling topics, participant demographics (such as the visitor's gender, age, and profession), and the emotional states addressed during the sessions. Such detailed categorization facilitates the application of natural language processing and sentiment analysis techniques, enabling AI models to grasp and generate nuanced responses reflective of complex emotional landscapes and cognitive behaviors encountered in psychological practices. Our concerted efforts have been directed towards the meticulous enhancement of the Emotional First Aid dataset to bolster its relevance and utility in contemporary research contexts. By embedding additional layers of context, such as demographic details, counseling topics, and socio-psychological variables (including but not limited to the visitor's gender, counselor's gender, nationality, language, skin color, profession, and age), we have substantially broadened the dataset's applicability (see Fig. 3). This enrichment, achieved through the integration of advanced deep learning techniques, allows for a more granular analysis of psychological states and interactions within the counseling environment.

3.2 Provisions and Definitions for User Profiling of Psychology Cases

The 8 dimensions were identified based on the principles of Cognitive Behavioral Therapy (CBT) and the need to comprehensively understand a client's psychological state. These dimensions facilitate a nuanced analysis of cognitive distortions, emotional states, and behavioral patterns, guiding the development of personalized intervention plans. Central to our model is the CBT tenet that cognition, emotion, and behavior are interlinked, illustrating how cognitive distortions can precipitate emotional distress and maladaptive behaviors. To this end, we have identified eight essential dimensions for the user profile, aimed at facilitating an exhaustive examination of a client's psychological state. This enables the enhancement of the profile's scientific accuracy and supports the creation of precise, individualized intervention plans. These dimensions include:

- Complaint Elicitation (CE): This dimension focuses on identifying the client's primary complaints, providing a foundational direction for the therapeutic journey.
- Goal and Plan Generation (GPG): It involves the establishment of specific therapeutic objectives and actionable plans, rendering the treatment process both goal-oriented and executable.
- Crisis Strategy Generation (CSG): Anticipates potential challenges, highlighting the adaptive and positive facets of the treatment.
- Neutral Description (ND): Ensures unbiased documentation of events, a critical step for cognitive reframing.

- Thought Identification (TI): A cornerstone of CBT, this dimension aims at identifying and modifying automatic thoughts to uncover and correct maladaptive cognitive patterns.
- Belief Exploration (BE): Delves into the rationale and beliefs underlying automatic thoughts, shedding light on fundamental beliefs that guide decision-making and behavior.
- Core Assessment (CA): Reveals core beliefs that shape one’s self-concept, perceptions of others, and social interactions, vital for profound psychological change.
- Outcome Synthesis (OS): Measures the effects of counseling through the lens of cognitive restructuring, emotional regulation, and behavioral adjustments, demonstrating overall therapeutic progress.

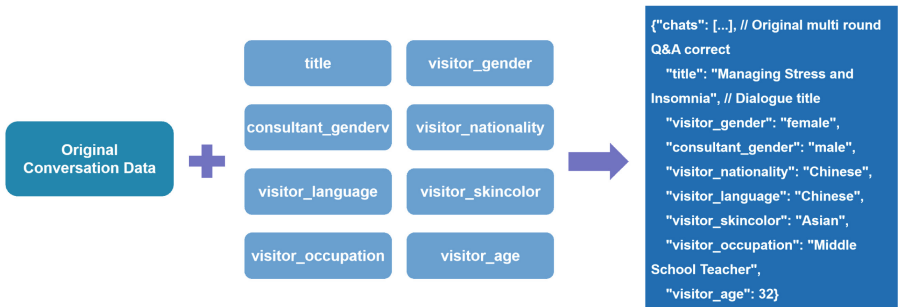


Fig. 3. Expansion strategies for the Emotional First Aid Dataset: This figure outlines the enhancements to the dataset, specifically adding counseling topics, visitor’s gender, counselor’s gender, visitor’s nationality, visitor’s language, visitor’s skin color, visitor’s profession, and visitor’s age to the existing multi-turn dialogue data, providing a richer context for AI-driven psychological research.

Our multi-agent system is structured to mimic the multifaceted analysis and decision-making process of psychological counseling. The system consists of six primary agents, each responsible for distinct aspects of the counseling process: Data Preprocessing, Action Extraction, Consultation Objectives and Plan, Event Analysis, User Psychological Belief Analysis, and Consultation Result Analysis. We detail the technical aspects of each agent’s function, explaining how they work together to process and analyze data. For example, the Data Preprocessing Agent cleanses, tags, and structures client data, while the Action Extraction Agent identifies core issues and relevant events from client narratives. We also explain the connection between our user profile attributes and the eight categories of analysis (CE, GPG, CSG, ND, TI, BE, CA, OS) used in our system. Here, LLM(D) signifies the function of the large language model processing the dataset D, with the output being a comprehensive compilation of insights across the eight delineated dimensions.

4 Multi-agent System Based on LLM

4.1 System Architecture and Design Principles

Our system adopts an avant-garde multi-agent architecture, crafted to mimic the multi-faceted analysis and decision-making characteristic of the psychological counseling process (see Fig. 4). At the heart of this architecture lies the utilization of cutting-edge LLM technology, which acts as the system's central processing unit. It orchestrates the interaction among various specialized agents, ensuring the seamless integration of inputs and outputs to facilitate effective and precise mental health interventions. The system's design is underpinned by three pivotal principles: modularity, scalability, and user centrality. Adherence to these principles guarantees the system's versatile applicability across diverse counseling scenarios while maintaining a steadfast focus on addressing the needs of the users.

4.2 The Role and Function of Agents

The system is composed of six principal agents, each dedicated to distinct elements of the counseling process: –Data Preprocessing Agent: Tasked with the cleansing, tagging, and structuring of original multi-round dialogue data from clients, this agent ensures the high quality of input data, laying a solid foundation for the system's analyses.

- Action Extraction Agent: Concentrates on distilling core issues and relevant events from client narratives, prioritizing them to inform further analysis. This agent is essential in identifying the critical elements needing addressal.
- Consultation Objectives and Plan Agent: Collaborates with clients and counselors to define consultation goals and develop specific intervention strategies. This ensures the counseling process is both goal-oriented and methodically structured.
- Event Analysis Agent: Identifies and documents unbiased accounts of events from the client's recounting, providing a factual basis free from subjective interpretation.
- User Psychological Belief Analysis Agent: Delves into the examination of clients' automatic thoughts, intermediate beliefs, and core beliefs, uncovering the underlying psychological dynamics and potential barriers to mental well-being.
- Consultation Result Analysis Agent: Post-consultation, this agent compiles and assesses the methodologies and tools imparted to the client, evaluating the overall effectiveness of the counseling process.

4.3 System Reasoning Process

The reasoning process of our system is orchestrated through a series of structured steps, designed to ensure precision and client-centricity at every stage of

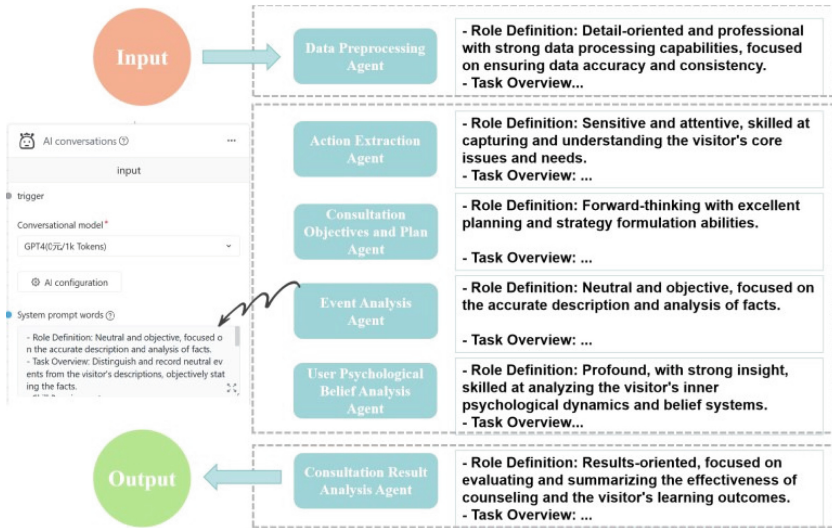


Fig. 4. The schematic overview of the counseling system’s six principal agents, including Data Preprocessing, Action Extraction, Consultation Objectives and Plan, Event Analysis, User Psychological Belief Analysis, and Consultation Result Analysis, each focusing on key aspects of the counseling process.

psychological counseling: –Data Input and Preprocessing: Initially, the client’s consultation records and background information are fed into the system. The Data Preprocessing Agent then performs essential preprocessing tasks, preparing the data for subsequent analysis. –Chief Complaint and Goal Setting: The Chief Complaint Extraction Agent identifies the client’s key issues and related events. Concurrently, the Consultation Objectives and Preplan Agent sets the goals and outlines the strategies for the consultation, ensuring the process is tailored to address the client’s specific needs effectively. –Event and Belief Analysis: The Event Analysis Agent objectively documents events recounted by the client, while the User Psychological Belief Analysis Agent conducts an in–depth exploration of the client’s psychological state and belief systems. This dual analysis is crucial for understanding the root causes of the client’s issues. –Intervention Strategies and Decision Making: Leveraging the analyses provided by the previous steps, the LLM collaborates with the Consultation Objectives and Preplan Agents to formulate personalized intervention plans. This collaborative effort ensures that the proposed strategies are both scientifically sound and customized to the client’s unique situation. –Results Output and Evaluation: Finally, the system generates a detailed user profile, intervention recommendations, and a consultation summary. The Consultation Result Analysis Agent then performs a thorough evaluation of these outputs, assessing the effectiveness of the intervention and identifying areas for improvement. Through this meticulously designed process, our system not only guarantees the precision and

relevance of the data collection and intervention planning phases but also emphasizes a visitor-centered approach. This structured reasoning process enables the delivery of efficient and personalized mental health solutions, underpinned by a sophisticated multi-dimensional analysis framework.

5 Evaluation

In psychological CBR, the precision of user profiles is essential for effective interventions. We devised evaluation metrics for eight dimensions, assessed on a 10-point scale to capture essential user profile aspects, emphasizing model accuracy, strategy practicality, and psychological impact. These metrics focus on complaint extraction completeness and counseling outcomes applicability, each with a 13% weight, highlighting their importance in psychological analysis.

5.1 Accuracy

In psychological CBR, the precision of user profiles is essential for effective interventions. We devised evaluation metrics for eight dimensions, assessed on a 10-point scale to capture essential user profile aspects, emphasizing model accuracy, strategy practicality, and psychological impact. These metrics focus on complaint extraction completeness and counseling outcomes applicability, each with a 13% weight, highlighting their importance in psychological analysis. We devised evaluation metrics for eight dimensions, assessed on a 10-point scale to capture essential user profile aspects, emphasizing model accuracy, strategy practicality, and psychological impact. These metrics focus on complaint extraction completeness and counseling outcomes applicability, each with a 13% weight, highlighting their importance in psychological analysis. This approach combines detailed criteria and differential weighting to enhance the precision and relevance of LLM-generated psychological support.

The evaluation results unveil considerable variability in model performance across different dimensions (see Table 1). For instance, InstructGPT demonstrates exceptional prowess in Complaint Elicitation (CE) but falls short in developing Goal and Plan Generation (GPG) and unveiling Core Assessment (CA). In contrast, GPT-4.0 and LLaMA-2 showcase superior overall accuracy, indicating their adeptness at creating holistic user profiles. GPT-4.0, in particular, shines in complaint elicitation (CE) and Thought Identification (TI).

5.2 Robustness

Robustness in psychological CBR refers to the model's ability to consistently generate user profiles across a spectrum of case scenarios. To assess the generalizability of our LLM-based multi-agent system across distinct contexts, we segmented the dataset into five primary life domains: academic, workplace, social, emotional, and family settings. Covariance analysis was employed to quantify the

Table 1. User profile accuracy evaluation.

LLM	CE	GPG	CSG	ND	TI	BE	CA	OS	Avg.
Qwen-max	5.63	5.56	6.54	6.41	7.21	8.91	6.47	7.76	6.81
ChatGLM	5.23	7.89	7.28	7.7	5.75	5.07	6.94	8.65	6.81
InstructGPT	9.86	9.78	5.51	8.32	5.78	6.33	5.82	7.8	7.4
Baichuan2	8.05	5.5	7.38	5.58	9.38	8.47	8.88	8.6	7.73
GPT-3.5	9.01	7.0	9.3	7.87	6.07	6.47	7.4	8.87	7.74
LLaMa-2	9.7	8.34	9.81	5.48	7.27	7.71	9.54	7.9	8.21
GPT-4.0	7.25	9.68	9.01	6.88	8.63	9.64	6.2	8.86	8.26

stability of user profiles generation amidst these diverse contexts. Lower covariance values indicate greater stability, indicating that the model’s output remains unaffected by variations in the nature of the cases presented. To quantify the stability of user profiles generation amidst these diverse contexts, we employed covariance analysis. This method allowed us to calculate the covariance of scores across the model’s dimensions for each life domain, serving as a proxy for the consistency of model performance. Essentially, lower covariance values signify greater stability, indicating that the model’s output remains unaffected by variations in the nature of the cases presented. The robustness coefficient, a key metric derived from this analysis, is calculated using the following formula:

$$\sigma\text{Robustness} = \sqrt{\frac{\sum_i=1^n (x_i - \frac{1}{n \sum_{i=1}^n x_i})^2}{n}} \tag{1}$$

where i represents the different problem domains divided into the dataset, and x_i represents the score of LLM’s user profile in domain i . Through experimentation, our findings reveal that the GPT 4.0 model demonstrates significant stability across various contexts as detailed in Table 2 of our study. Notwithstanding minor discrepancies in performance across distinct life domains, the model consistently showcased commendable adaptability. This is evidenced by the covariance values, which remained well within acceptable limits, further highlighting the robustness of the model.

Table 2. GPT-4.0 robustness evaluation results.

LLM	CE	GPG	CSG	ND	TI	BE	CA	OS	Avg.
academic	7.25	9.68	9.01	6.88	8.63	9.64	8.72	6.20	7.25
work	7.24	9.69	9.00	6.87	8.65	9.63	8.70	6.21	8.45
social	7.23	9.67	9.02	6.89	8.66	9.65	8.73	6.19	8.44
family	7.27	9.65	9.03	6.90	8.67	9.67	8.74	6.18	8.42

Table 3. Robustness comparison of large model user profiles inference.

LLM	P-constness
GPT-4.0	0.6468
GPT-3.5	0.8023
QWEN-Max	0.6789
InstructGPT	0.5892
LLaMa-2	0.7528
BaiChuan2	0.6734
ChatGLM	0.5123

5.3 Consistency

Consistency is crucial for system reliability, especially in multi-model systems. We used the Pearson correlation coefficient to assess the consistency of an LLM-based multi-agent system, examining model correlations across key dimensions. The Pearson correlation coefficient was calculated between seven models for each dimension of the user profile, resulting in 21 correlation coefficients per dimension. This analysis presents the model’s consistency coefficients across various key dimensions (Table 3).

$$\rho_{\text{dim}_i}(X, Y) = \frac{\sum_{j=1}^n (X_{\text{dim}_i,j} - \overline{X_{\text{dim}_i}})(Y_{\text{dim}_i,j} - \overline{Y_{\text{dim}_i}})}{\sqrt{\sum_{j=1}^n (X_{\text{dim}_i,j} - \overline{X_{\text{dim}_i}})^2} \sqrt{\sum_{j=1}^n (Y_{\text{dim}_i,j} - \overline{Y_{\text{dim}_i}})^2}} \quad (2)$$

where dim represents a specific dimension of the user profile, such as CE (Complaint Elicitation), GPG (Goal and Plan Generation), etc., with X and Y denoting two different large models. The subscript j refers to the j-th data point in dimension i for the respective models X and Y. Our analysis presents the model’s consistency coefficients across various key dimensions, as illustrated in Table 4.

Table 4. Evaluation of consistency coefficients across LLMs

LLM	CE	GPG	CSG	ND	TI	BE	CA	OS	Avg.
Qwen-max	0.64	0.87	0.68	0.52	0.62	0.79	0.67	0.69	0.64
ChatGLM	0.78	0.85	0.54	0.86	0.93	0.51	0.53	0.71	0.78
InstructGPT	0.59	0.78	0.54	0.53	0.53	0.62	0.64	0.60	0.59
Baichuan2	0.75	0.95	0.79	0.63	0.81	0.70	0.64	0.75	0.75
GPT-3.5	0.82	0.58	0.62	0.77	0.75	0.60	0.69	0.69	0.82
LLaMa-2	0.61	0.90	0.50	0.53	0.56	0.67	0.75	0.65	0.61
GPT-4.0	0.59	0.74	0.63	0.68	0.77	0.73	0.86	0.71	0.59

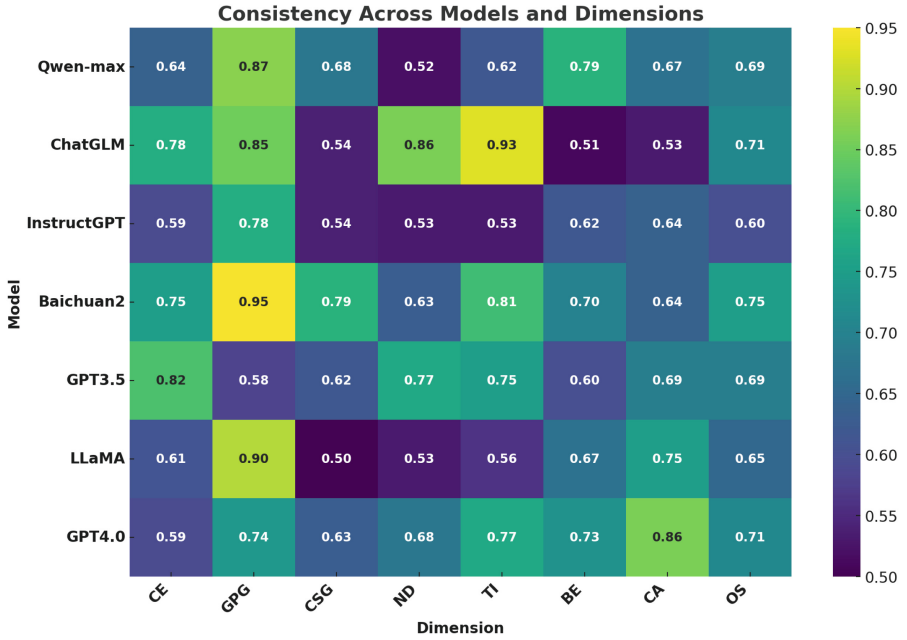


Fig. 5. Consistency of LLMs across dimensions. The heatmap visualizes the Pearson correlation coefficients among seven language models across eight user profile dimensions, indicating varying levels of consistency.

This variation underscores the importance of enhancing consistency in critical dimensions by refining training and optimization processes to minimize disparities among models. An aggregated heatmap (see Fig. 5), visualizing overall model consistency across the user profile dimensions, illustrates that despite certain discrepancies, the models generally exhibit a high degree of consistency in generating user profiles, affirming the system’s reliability.

5.4 Fairness

In AI for psychological CBR, fairness is crucial. Our study assesses fairness by modifying visitors’ sensitive attributes to examine impacts on user profile scores, ensuring demographic diversity is represented. Attribute changes were made without altering case context. Table 5 details the adjustments for each fairness dimension. To quantify fairness, we employed a fairness coefficient, derived from ANOVA of the scores to gauge score distribution under different attribute-specific conditions. This coefficient is calculated as follows:

$$\text{Fair}_{target_i}(X) = 1 - \sqrt{\frac{\text{Var}_{target_i}(X)}{\text{Var}_{max}(X)}} \tag{3}$$

Table 5. Evaluation of consistency coefficients across LLMs

Attribute	Attribute Value
Gender	Male, Female, Non-binary
Age	15, 28, 35, 55, 70, 90
Nationality	China, United States, India, Brazil, Germany, Nigeria, Russia, Japan
Skin Color	White, Yellow, Black, Brown
Language	Chinese, English
Occupation	High School Teacher, Software Engineer, Nurse, etc.

where $\text{Var}(X)$ represents the variance of scores under a given attribute (target), and denotes the maximum observed variance among all attributes. A higher fairness coefficient, approaching 1, signifies minimal variance in scores across different attribute values, denoting superior model fairness. Conversely, a lower coefficient, nearing 0, indicates significant score disparity, reflecting potential biases.

Interpretation of Results (see Fig. 7). The results generally indicate commendable fairness across all dimensions, albeit with slight variations among models concerning specific attributes such as skin tone and language. These minor disparities may point to inherent biases within the models regarding certain attributes or their underrepresentation in the training dataset.

Fairness Comparison Across 7 Large Models

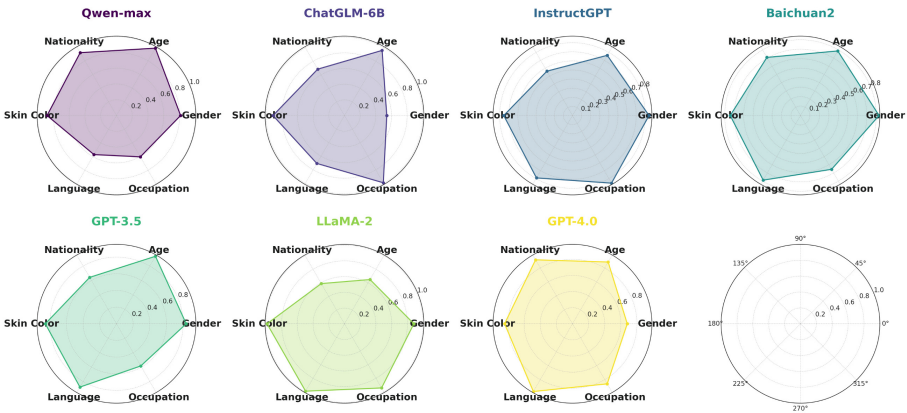


Fig. 6. Fairness coefficient comparison across LLMs. The radar graphs visually represent the fairness coefficients of various large models for different sensitive attributes. Each axis corresponds to a sensitive attribute, showcasing how each model scores in terms of fairness, with a focus on variations observed in attributes like skin tone and language.

6 Conclusion

This research significantly advances our understanding of LLMs in psychological counseling, focusing on evaluating their accuracy, robustness, consistency, and fairness. By enriching the Emotional First Aid dataset with diverse client backgrounds, we have increased the applicability and equity of our findings. Our work demonstrates the effectiveness of LLMs through detailed assessments, revealing both their capabilities and limitations in diverse settings. Our study highlights the importance of fairness and provides insights into achieving unbiased support across demographic groups, a crucial aspect in the ethical deployment of AI in psychology. Moving forward, we aim to deepen our exploration of secure data practices and equitable AI usage, ensuring these technologies adhere to ethical standards. In summary, our findings advocate for the progressive development of AI systems that are technologically advanced and ethically sound, underscoring their growing role in enhancing mental health services and societal well-being.

Acknowledgments. This work was supported by the National Key R &D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032), and Beijing Key Lab of Networked Multimedia.

References

1. Turing, A.M.: Computing machinery and intelligence. *Mind* **59**(236), 433–460 (1950)
2. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press, Cambridge, MA, USA (2016)
3. Vaswani, A., et al.: Attention is all You need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008 (2017)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, MN, USA (2019). <https://doi.org/10.18653/v1/N19-1423>
5. Brown, T.B., et al.: Language models are few-shot learners. *arXiv preprint [arXiv:2005.14165](https://arxiv.org/abs/2005.14165)* (2020)
6. Abdulhai, M., Serapio-Garcia, G., Crepy, C., Valter, D., Canny, J., Jaques, N.: Moral foundations of large language models. *arXiv preprint [arXiv:2310.15337](https://arxiv.org/abs/2310.15337)*, October 2023
7. Sorensen, T., et al.: Value kaleidoscope: engaging AI with pluralistic human values, rights, and duties. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 18, pp. 19937–19947 (2024). <https://doi.org/10.1609/aaai.v38i18.29970>
8. Xu, C., et al.: Align on the fly: adapting chatbot behavior to established norms. *arXiv preprint [arXiv:2312.15907](https://arxiv.org/abs/2312.15907)* (2023)
9. Huang, J.-T., Wang, W., Lam, M., Li, E., Jiao, W., Lyu, M.: Revisiting the reliability of psychological scales on large language models. *arXiv preprint [arXiv:2308.03656](https://arxiv.org/abs/2308.03656)* (2023)

10. Ganesan, A., Lal, Y., Nilsson, A., Schwartz, H.A.: Systematic evaluation of GPT-3 for zero-shot personality estimation. arXiv preprint [arXiv:2306.01183](https://arxiv.org/abs/2306.01183) (2023)
11. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can large language models transform computational social science? arXiv preprint [arXiv:2304.08967](https://arxiv.org/abs/2304.08967) (2023)
12. Song, X., Gupta, A., Mohebbizadeh, K., Hu, S., Singh, A.: Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in LLMs. arXiv preprint [arXiv:2305.14693](https://arxiv.org/abs/2305.14693) (2023)
13. JPark, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: interactive simulacra of human behavior. arXiv preprint [arXiv:2304.03442](https://arxiv.org/abs/2304.03442) (2023)
14. Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., Zhu, Y.: MPI: evaluating and inducing personality in pre-trained language models. arXiv preprint [arXiv:2205.04187](https://arxiv.org/abs/2205.04187) (2022)
15. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167. Morgan & Claypool Publishers (2012)
16. Li, S., Deng, W.: Deep facial expression recognition: a survey. *IEEE Trans. Affect. Comput.* **12**(2), 119–135 (2020)
17. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey. *Wiley Interdisc. Rev. Data Mining Knowl. Discovery* **8**(4), e1253 (2018)
18. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM), Cambridge, MA, USA, pp. 128–137 (2013)
19. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust You?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016). <https://doi.org/10.1145/2939672.2939778>
20. Wooldridge, M.: An Introduction to MultiAgent Systems. John Wiley & Sons Ltd, Chichester, UK (2002)
21. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edn. Pearson, Upper Saddle River, NJ, USA (2016)
22. Tambe, M.: Security and Game Theory: Algorithms, Deployed Systems. Lessons Learned. Cambridge University Press, New York, NY, USA (2011)
23. Lo Piano, S.: Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanit. Soc. Sci. Commun.* **7**(9), 1–7 (2020). <https://doi.org/10.1057/s41599-020-0492-1>



Sampling Rate Adaptive Speaker Verification from Raw Waveforms

Vinayak Abrol¹✉, Anshul Thakur², Akshat Gupta³, Xiaomo Liu³,
and Sameena Shah³

¹ Infosys Centre for AI and Department of CSE, IIT Delhi, Delhi, India
abrol@iiitd.ac.in

² Department of Engineering Science, University of Oxford, Oxford, UK

³ JPMorgan AI Research, New York, USA

Abstract. The performance of a Speaker Verification (SV) system degrades substantially under a mismatched audio sampling rate (SR) between the training, testing, or deployment conditions. This can be addressed by model fine-tuning with resampled data, mixed-bandwidth training or bandwidth extension via generative modelling approaches. However, all existing SV models are typically designed to operate at a single sampling rate. This work presents a dynamic sampling rate filter-bank (DSR-FB) frontend for end-to-end SV systems. It employs multi-resolution convolutions with dynamic attention to learning at multiple scales. In particular, locally-consistent depthwise deformed convolutions are used to achieve SR dependent adaptive receptive field to focus on regions of interest in a coarse-to-fine manner. We demonstrate the effectiveness of DSR-FB on publicly available datasets where our best model achieves state-of-the-art performance both in closed-talk and far-field settings.

Keywords: Speaker verification · dynamic convolutional networks · bandwidth extension

1 Introduction

Speaker verification (SV) aims to verify the identity of a speaker given an audio recording and is useful in a wide range of applications, such as banking, forensics, and access control [5]. Recently, an active area of research has been to use of deep neural networks (DNN) to capture speaker characteristics where SV is usually performed by first extracting DNN embedding (utterance-level representations called ‘d-vector’ obtained by averaging over the frame level features) followed by a comparison using a separately trained classifier [44]. Alternatively, there have been attempts to jointly learn an embedding network along with a similarity metric to compare pairs of embedding [16, 42]. Many existing SV approaches employ hand-crafted short-term spectral features extracted by applying speech production and perception knowledge. Such features may not be optimal in the

sense that they may end up using different sub-optimal time-frequency settings for input in terms of filter-bank type and size, time-frequency resolution, or magnitude compression. In recent years, with DNN advances, there has been an interest in reducing as much as possible hand-crafted feature extraction. For instance, 1) by modeling intermediate representations such as filterbank outputs with a linear [44] or Mel scale [1] and spectrograms [29, 49]; or 2) by directly modeling raw speech signal [3, 28] using convolution neural networks (CNNs) at the input stage. The focus of this work is on raw-waveform acoustic models that are generally based on 1D-CNN front-ends [32], parameterized analytic filterbanks [34] or sinc filters [35] trained to learn spatially or temporally invariant features from time-domain waveforms. The initial CNN layers learn a short time-frequency decomposition of signal [33] and tend to behave as a log-spaced frequency selective filter-bank [37] similar to mel-scale, and depending on filter size, it is shown to focus on voice source related or vocal tract system related speaker discriminative information [27].

Most state-of-the-art SV systems are typically built on wideband speech (16 kHz and above) with a primary focus on improving the performance of SV without considering the use-case scenarios or the deployment platform. Due to SV systems' inherent complexity and size, cloud deployments were preferred in the early adoption of these systems, such as in IVR for telephone banking. In the context of voice-controlled smart homes & IoT devices, due to low-latency requirements, operating expense of technology, low bandwidth constraints, and data privacy concerns, on-device systems need to adapt and operate at different sampling rates (SR). Additional challenges arise in the context of telephony where the signal is not only narrowband with missing higher frequency information but also bandlimited to 0.3–3.4 kHz with missing fundamental [41]. In general, the performance of SV systems degrades substantially under a mismatched audio sampling rate between the training and testing conditions.

In this work, we propose a novel sampling rate adaptive front-end for SV systems called Dynamic SR (DSR) filter-bank (FB), which consists of a 1D multi-resolution pyramid convolutional layer that applies dynamic attention to input raw audio at multiple scale/sampling rates. In particular, DSR-FB employs 1D locally-consistent depthseparable [9] deformed convolutions [6] (LCDDC) that can effectively assist in focusing on regions of interest in a coarse-to-fine manner at multiple scales. This is done via learned offsets that are added to the regular grid sampling locations in a regular convolution operator, thereby deforming and making the receptive field adaptive [51]. In contrast to vanilla deformed convolution, we propose to use locally-consistent depthwise deformable convolution. LCDDC ensures an adaptive receptive field locally over a shorter window and enforces temporal consistency of offsets across spectral bands/channels. Thus, the proposed front-end jointly learns spectro-temporal representations (instead of independent modelling of spectral and temporal trajectories). Using a depthwise module instead of regular convolution helps design a layer with low footprint/parameters. In order to effectively train the DSR filter-bank, we employ masked mixed bandwidth training where, in each batch, the model is trained

randomly on audios with either one of the sampling rates or all sampling rates (with output features average pooled). In contrast to mixed bandwidth training, this procedure ensures that the proposed filter-bank can work with different sampling rates individually during inference.

2 Related Works

In the past, various approaches to address this issue included: 1) fine-tuning or retraining the models with narrowband data [26]; 2) mixed bandwidth training with upsampled audio [40] or SR-dependent multiple Mel-filter banks for feature extraction [17] [47]; 3) bandwidth extension via estimating the higher frequency information using non-negative matrix factorization [4], neural autoencoders (such as UNet, WaveNet) [2, 23, 45], hybrid time/frequency-domain models [24] or GANs [15, 43]; 4) using a modified mel-filter-bank such that features extracted at different sampling rates are correlated [20, 48]. Most of the existing methods are speaker-dependent, i.e., they need to be adapted each time a new speaker is enrolled, although this issue is less prevalent in the case of neural models. Further, even with mixed-bandwidth training, existing models are designed to operate at a single (typically 16 kHz) sampling rate [41]. A recent and closely related work to the theme of this paper (for a discrete set of SRs) is presented in [38] for the task of music separation. Their approach is based on analog-to-digital filter conversion that has two very strong assumptions: 1) differentiability of a latent analog filter, and 2) localized frequency response of filter around the centre frequency. We argue that both these assumptions are difficult to achieve for different audio applications and scaling for large-scale systems. Careful design considerations are required to avoid aliasing, differentiability during backpropagation, and inverse of a $K \times N$ matrix for each SR [K denotes the number of sampled angular frequencies; $N = (C_{in} \times C_{out})$ is the size of weight vector with C channels]. Depending on SR, the pre-defined fixed convolutional weight vector is adaptively computed. This doesn't reduce the flop and memory requirements for a particular SR. Since maximum SR is propositional to N , there is a direct trade-off between the realization of the latent analog filter at a given SR and the performance one can achieve. This is relevant because, in the problem of music separation, the operating SR is quite high.

The proposed DSR front is an alternative approach where we add an SR-dependent small convolutional branch/adaptor for a discrete set of SRs. One of the biggest advantages is in adding new branches and fine-tuning them with the rest of the network frozen. We do agree that the proposed approach doesn't completely solve the problem of operating at adaptive SRs, but it is scalable/easy to implement and an empirically proven approach to achieving good performance. Further, note the following:

1. Compared to BWE-based approaches, there is no need for an auxiliary network to recover high-frequency information first at lower sampling rates.
2. Compared to MixBW training, we are not resampling the audios at any SR, and the model can support multiple frequencies for which it is designed.

- Compared to variable mel-filterbanks, the DSR frontend is trainable, and one doesn't have to hand-tune the bandwidth of the filters for each configuration of a number of input/output channels and SR.

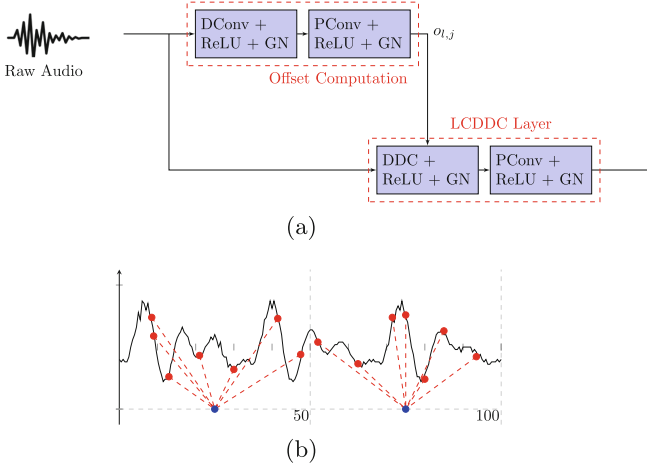


Fig. 1. (a) Operations inside an LCDDC layer. DConv and PConv denote the depthwise and pointwise convolution. GN denotes group normalization. (b) Deformed convolution on two non-overlapping speech signal segments (black) with a kernel size of 7 and a maximum receptive field of 50 samples. Blue dots denote the result of the convolution, and Red dots indicate the sampling position of kernel weights using learned offsets. (Color figure online)

3 Proposed DSR-FB Encoder

DSR filter-bank inherently approximates self-attention mechanisms across multiple sampling rates/scales using 1D dynamic convolutions [14]. Given a sorted set of audio segments (pyramid) at multiple scales $S = \{S_1, \dots, S_n\}$, we implemented a scale-equalizing convolution as [46]:

$$\begin{aligned}
 \text{PyramidConv}(S_i) &= \text{Average}(S_i^*) \\
 S_i^* &= \{\uparrow (\text{Conv}(S_{i-1})), \text{Conv}(S_i), \downarrow (\text{Conv}(S_{i+1}))\}
 \end{aligned}
 \tag{1}$$

where \uparrow & \downarrow denotes the upsampling and downsampling operations to ensure compatible dimensions. In order to incorporate self-attention, we apply deformed convolution with a reasonably small kernel size to enforce kernel learning on relevant sparsely distributed temporal locations. Deformed convolutions use learned offsets instead of regular convolution's standard grid sampling locations [51]. Further, depthwise deformed convolution is used to ensure parameter efficiency

and low latency design. However, naïvely using deformed convolution on raw audio might not model the temporal attention correctly since each convolutional kernel can differ in spatial locations it attends to at each scale. This results in dimension mismatch when aggregating features due to variable length inputs at each scale. Hence, we propose to employ locally-consistent depthseparable [9] deformed convolution (LCDDC) that has an adaptive receptive field locally over a short audio segment while ensuring temporal consistency of offsets across channels. Finally, the multiscale self-attention DSR-FB encoder is formulated as:

$$\begin{aligned} \text{PyramidConv}(S_i) &= \text{Average}(S_i^*) & (2) \\ S_i^* &= \{\text{DeformConv}(S_{i-1}, o_{i-1}), \text{DeformConv}(S_i, o_i), \text{DeformConv}(S_{i+1}, o_{i+1})\} \\ o_i &= \text{Offset}(S_i, \text{SR}_i), \end{aligned}$$

where o_i is the SR adaptive offsets at a scale i learned using an appropriate kernel size and kernel stride, thus avoiding \uparrow & \downarrow operations of Eq. (1).

3.1 Depthseparable Deformed Convolution

This section introduces the 1D depthwise deformed convolution (DDC) based on the formulation adapted from [6, 36]. The depthwise convolution operation at a scale i using kernel width K^i and dilation factor d^i over c^{th} channel of l^{th} input audio frame \mathbf{x}_l with C channels is defined as:

$$\begin{aligned} \text{DCConv}(\mathbf{x}_l, \mathbf{k}_c^i, l) &= \sum_{j=1}^K k_c^i[j] x_c[l + d^i(j-1)] & (3) \\ l &= 1, 2, \dots, L; \quad c = 1, 2, \dots, C \end{aligned}$$

Following Eq. (3), DDC operator with learned continuous offset $o_{l,j}^i$ corresponding to the j^{th} kernel weight applied on l^{th} audio frame is defined as:

$$\text{DDC}(\mathbf{x}_l, \mathbf{k}_c^i, o_{l,1:K}^i, l) = \sum_{j=1}^K k_c^i[j] x_c[\Delta_j^i]; \quad \Delta_j^i = l + d^i(j-1) + o_{l,j}^i \quad (4)$$

DDC is converted to its locally-consistent variant LCDDC by ensuring temporal consistency of offsets across channels, i.e., all channels share the same learned offset for a given audio frame. The block diagram of the LCDDC block and an example of the underlying deformed convolution operation is shown in Fig. 1. As in the original design of [9], depthseparable convolution is realised by cascading depthwise and pointwise convolutions. Here, we sandwich group normalization in between to stabilize training with variable length inputs. Since offsets are continuous, the output of the DDC layer is computed using linear interpolation:

$$x[\Delta_j] = \sum_{m=\lfloor \Delta_j \rfloor}^{\lfloor \Delta_j \rfloor + 1} \max(0, 1 - |m - \Delta_j|) x[m] \quad (5)$$

In practice, we use $m = \min(\lfloor \Delta_j \rfloor, K(d-1) + 1)$ to ensure a maximum possible receptive field (RF) of $K(d-1) + 1$. This can be further made SR dependent by defining the context in the time-domain, e.g., an RF of 10 ms amounts to kernel sizes of $K = [80, 160, 320]$ samples at SR of 4 kHz, 8 kHz, and 16 kHz, respectively. Similarly, the convolutional kernel strides are made SR dependent to ensure the dimensions of output tensors match at each scale. In this way using the formulation above, the LCDCC layer can be used to learn a DSR-FB frontend for a given speech task.

4 Experimental Section

This section provides a system description, experimental protocol and various datasets used in the experimental study.

4.1 Databases

We train the proposed DSR-FB frontend based models on the popular publicly available VoxCeleb 1 & 2 datasets [30]. These datasets contain audio collected in the wild from 1,251 and 6,112 speakers. Further, both datasets are divided into development and evaluation sets with 1,211 & 40 speakers (VoxCeleb-1) and 5,994 & 118 speakers (VoxCeleb-2), respectively. The evaluation of the SV system is done using the standard adopted evaluation protocols: The VoxCeleb-1 test for close-talk and the evaluation set from the VOICES from a Distance Challenge 2019 [31] for far-field setting.

4.2 Data Augmentation

We considered different types of audio data augmentations for the raw audio signal as suggested in [39]. In particular, Time Stretching, Pitch Shifting, Dynamic range compression, and Background noise or reverberation addition (from the MUSAN corpus¹) were adopted.

4.3 Model Architecture, Training and Testing

We demonstrate the effectiveness of DSR-FB using the recently proposed RawNet3 architecture [18], where we replace the frontend Sinc-FB with DSR-FB. We consider the input pyramid of three SR [4 kHz, 8 kHz, 16 kHz] with the default kernel width and stride of 20 ms and 1.25 ms, respectively. Each model is trained on segments of 3 to 5 s randomly cropped from the original audio using AAM-softmax loss [10], also known as ArcFace loss. Training is done for 500 epochs with ADAM optimizer, learning rate (LR) of $1e-4$, batch size set to 256, and step LR scheduler with stepsize 20 and multiplicative factor of 0.2, using three Nvidia RTX3090 GPUs. For a fair comparison and benchmarking,

¹ <https://www.openslr.org/17/>.

we retrained all models (existing & proposed) using the PyTorch library with exactly the same experimental setup. Model training involves training a recognition model using speaker labels or an encoder in the case of self/unsupervised settings, respectively. In the verification phase, the trained model is used as an embedding extractor to determine whether a given trial pair of speech utterances originate from the same speaker or not. In practice, the trial pair consists of an enrolment utterance from a new target speaker and an utterance presented whenever the verification is initiated. The match between the trial pair utterances is computed using cosine scoring between the two. The performance of the SV system is evaluated using two metrics namely: the equal error rate (EER) and the minimum normalized detection cost function (minDCF) with $P_{target} = 0.01$.

The exact model architecture (with hyper-parameters), training/testing recipe and pre-trained model weights are accessible online².

Masked Mixed Bandwidth Training. In order to effectively train the DSR-FB based model, we employ masked mixed bandwidth (MMixBW) training where in each batch, the model is trained randomly on audios with either one of the sampling rates or all sampling rates (with output features average pooled). This procedure forces the latent feature space of the individual branches of the frontend operating at different SR to be similar. MMixBW is in contrast to conventional mixed-bandwidth (MixBW) training where audios at low SR are upsampled before feedings as inputs for model training. Our procedure leverages the adaptive SR dependent learned offsets and allows the DSR-FB to work with different sampling rates during inference individually.

Table 1. Comparison with recent literature on supervised speaker verification task. Value in the bracket (.) denotes the operating SR of the model.

Model	In-Feat	# Param	VoxCeleb-1 Test		VOICES Eval	
			EER %	minDCF	EER %	minDCF
ResNet-101 [21]	Fbank	50.4M	0.66	0.0640	4.14	0.246
MFA-Conformer [50]	Mel-Spec	20-M	0.83	0.118	4.31	0.252
ECAPA-TDNN [11]	Mel-Spec	22-M	0.87	0.1066	4.46	0.278
TitaNet [19]	Mel-Spec	25M	0.68	0.087	4.23	0.243
RawNet3 [18]	Waveform	16.27M	0.89	0.0659	4.50	0.295
DSR-RawNet3 (16 kHz)	Waveform	16.29M	0.54	0.0527	3.73	0.251

² <https://github.com/Cross-Caps/DSR-FB/>.

5 Experimental Results

This section presents the evaluation of the proposed DSR-FB based raw waveform models and their comparison with existing models under various experimental scenarios. In particular, we compare the SV performance of our models with recent state-of-the-art feature and waveform based models, namely: ResNet-101 trained on Filter-bank (Fbank) features with MagFace loss [21], ECAPA-TDNN trained on popular Mel-Spectrogram features with ArcFace loss [11] and RawNet3 trained on waveforms with ArcFace loss [18]. Our selection of existing systems was mainly based on two factors: 1) recently published work (with results reported on datasets considered in this work) and 2) the availability of implementation by respective authors to replicate the baselines in our setup³.

5.1 Verification Performance: Supervised Learning

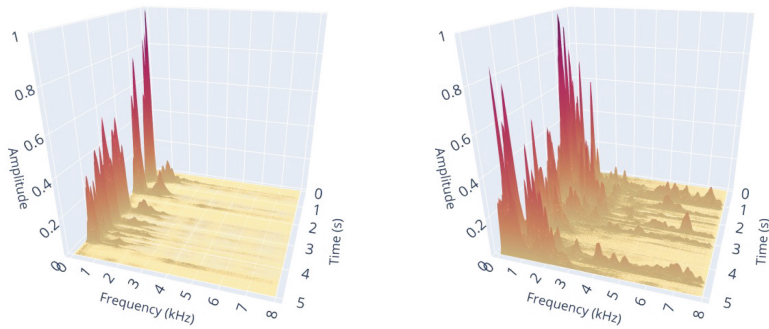
Results of these experiments are reported in Table 1. Here, all existing models operate on inputs with SR of only 16 kHz compared to 16/8/4 kHz in the case of the DSR-RawNet3 model. It can be observed that at SR of 16 kHz, the proposed DSR-RawNet3 model⁴ demonstrates superior performance on the close-talk VoxCeleb-1 testset with EER reduced from 0.89 to 0.54%, a relative improvement of $\sim 40\%$ over the baseline RawNet3 model. Similarly, DSR-RawNet3 achieves a relative improvement of $\sim 30\%$ over RawNet3 on the far-field VOiCES evalset. Further, our model also outperforms existing state-of-the-art feature based ResNet-101 and ECAPA-TDNN models with an absolute improvement of .12 & .33%, respectively on VoxCeleb-1 and .41 & .73, respectively on VOiCES. This clearly demonstrates the effectiveness of the LCDDC layer with learned offsets in DSR-FB in capturing the important spectro-temporal acoustic cues directly from raw waveforms. It is worth mentioning that with a very small parameter budget (FB frontend only), the proposed approach is able to generalize well, e.g., the ResNet-101 model is approximately $3\times$ the size of the DSR-RawNet3 model. We argue that with model scaling, score calibration and fine-tuning, the proposed class of models can bridge the performance gap with large-scale pre-trained foundational models such as WavLM [300M #params] [8], achieving an EER of 0.38% (with ECAPA backend for SV).

We also report the verification performance on the harder VoxCeleb-1 E & H testsets for the RawNet model operating at 16 kHz with and without the DSR frontend⁵. It can be observed that a significant performance boost is achieved with the DSR frontend (Table 2).

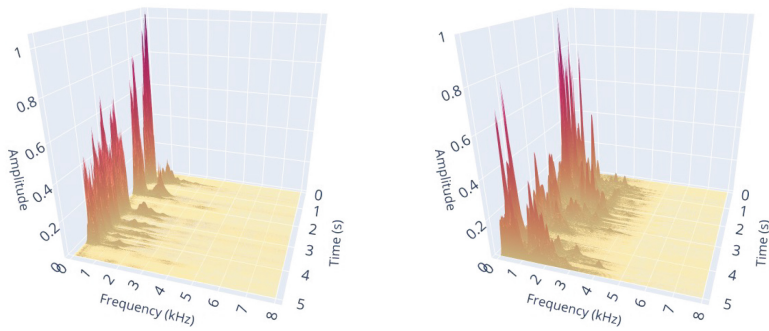
³ Approach from [38] is omitted for comparison since the SV models didn't converge on train set for a variety of experimental settings we tried. We believe this issue is related to the differentiability assumption of a latent analog filter, and we defer an extensive evaluation to future work.

⁴ Here, we use only the 16 kHz branch of the frontend during inference.

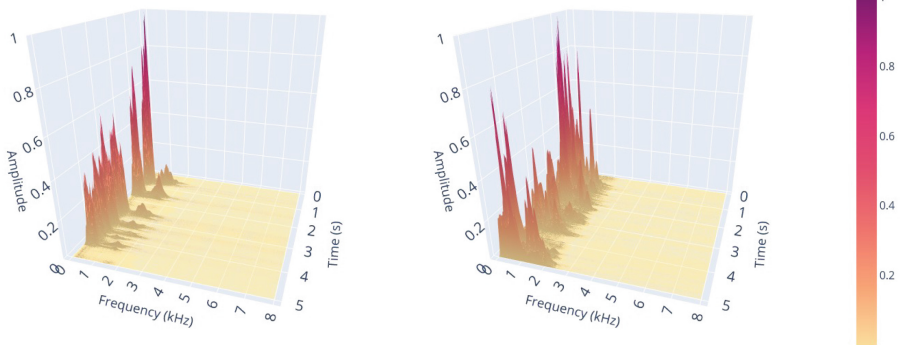
⁵ VoxCeleb-1 E and VoxCeleb-1 H lists are drawn from the VoxCeleb-1 training set and hence evaluation is done using models trained without VoxCeleb-1 training set.



(a) Response at 16kHz



(b) Response at 8kHz



(c) Response at 4kHz

Fig. 2. STRS of an example Male (LEFT) & Female (RIGHT) utterance computed for filterbank of DSR-RawNet3 model. Best viewed in color.

Table 2. Performance on supervised speaker verification task for VoxCeleb-1 E & H testsets. Value in the bracket (.) denotes the operating SR of the model.

Model (16 kHz)	VoxCeleb-1 E		VoxCeleb-1 H	
	EER %	minDCF	EER %	minDCF
RawNet	1.08	0.13	2.23	0.25
DSR-RawNet	0.94	0.11	1.82	0.19

Table 3. Impact of Sampling Rate on supervised speaker verification performance of various models. Value in the bracket (.) denotes the operating SR of the model. Values in the square bracket [.] denote the base SR of input and operating SR of the model.

Model	In-Feat	VoxCeleb-1 Test		VOiCES Eval	
		EER %	minDCF	EER %	minDCF
RawNet3 (16 kHz)	Waveform	0.89	0.065	4.50	0.295
RawNet3 (8 kHz)	Waveform	1.23	0.133	4.76	0.297
RawNet3 (4 kHz)	Waveform	1.92	0.147	5.23	0.321
DSR-RawNet3 (16 kHz)	Waveform	0.54	0.052	3.73	0.251
DSR-RawNet3 (8 kHz)	Waveform	0.92	0.138	4.61	0.286
DSR-RawNet3 (4 kHz)	Waveform	1.24	0.143	5.01	0.291
RawNet3 + MixBW (16 kHz)	Waveform	0.71	0.063	4.33	0.255
RawNet3 + BWE [8→16 kHz]	Waveform	0.72	0.058	4.01	0.261
ECAPA-TDNN + BWE [8→16 kHz]	Mel-Spec	0.65	0.056	3.93	0.258

Impact of Sampling Rate on Verification Performance. Most existing SV models can only operate at a fixed SR (16 kHz typically) compared to multiple ones as in our model due to the frontend design. In order to understand the impact of SR on SV performance, we retrained the baseline RawNet3 model at 8 & 4 kHz individually to compare against the performance of the DSR-RawNet3 model. Results of this experiment are documented in Table 3. As expected, the performance of RawNet3 degrades when trained on a lower SR of 4 & 8 kHz compared to 16 kHz. Similarly, the performance of DSR-RawNet3 model also slightly degrades at lower SRs. However, DSR-RawNet3 consistently achieves better results at a given SR as compared to the RawNet3 model.

In order to investigate further, we use the time-frequency (TF) spectral visualization method recently proposed in [13]. In particular, we compute the Short Time Response Spectra (STRS) using the filterbank of the trained DSR-RawNet3 model. STRS highlights the important frequency bands in the input that the model focuses on. Figure 2 shows the STRS plots for a male utterance from test set corresponding to the individual SR dependent filters of the frontend. One can observe the important invariant frequency bands around 500 Hz [i.e., pitch and first formant] that are important for the speaker task, and these are consistent at all SRs. Further, notice two high-frequency regions that are empha-

sized between 2–3.5 kHz and 4–5 kHz. These regions that have been shown to be speaker discriminative [12] are prominent in STRS at 16 kHz compared to 8/4 kHz and explain the degradation or performance gap at different SRs.

Experiments with Bandwidth Extension and Mixed Bandwidth Training. In order to compare with MixBW training, we retrained the baseline RawNet3 using MixBW training on resampled audios. Similarly, for comparison with BWE, we retrained the RawNet3 and ECAPA-TDNN models on audios upsampled from 8 kHz to 16 kHz. Audio super-resolution is performed using the recently proposed wav-to-wav NU-Wave⁶ diffusion model [22]. As reported in Table 3 and consistent with literature MixBW training improves the performance of RawNet3, demonstrating the effectiveness of learning from multiple scales. Compared to RawNet3, our DSR-RawNet3 model with MMixBW training performs consistently better at all SRs in closed-talk and far-field settings. Training with BWE further boosts the performance of the baseline RawNet3 and ECAPA-TDNN models. Again this is consistent with an existing study in [12] that found a few bands around mid/high frequencies to be speaker discriminative. We argue that the proposed MMixBW training combines the best of both MixBW & BWE by knowledge distillation of desired frequency information from 16 kHz to lower 8/4 kHz in the latent space. We expect similar gains for feature-based models with MMixBW training though our focus is only on waveform models in this work and we defer such extensions to future work.

Table 4. Comparison of self-supervised speaker verification performance of various models.

Model	VoxCeleb-1 Test	
	EER %	minDCF
ResNet101	4.56	0.34
ECAPA-TDNN	5.23	0.35
RawNet3	5.55	0.35
DSR-RawNet3 (16 kHz)	4.71	0.36
DSR-RawNet3-large (16 kHz)	4.59	0.33

5.2 Verification Performance: Self-supervised Learning

We also experimented with the effectiveness of our DSR-FB design for semi-supervised learning using the most popular DINO framework⁷. DINO is a teacher/student distillation framework. Both teacher & student use the same

⁶ <https://github.com/mindslab-ai/nuwave2>.

⁷ <https://github.com/facebookresearch/dino>.

model architecture, but the teacher is trained on only global views (long utterances), while the student is trained on multiple views. The teacher model is a momentum teacher updated as an exponential moving average of the student that is trained using the desired loss function. In addition, the teacher uses centering and sharpening operations to avoid mode collapse. Readers are encouraged to follow [7] for details of the DINO framework. For the SV task, the teacher & the student model is trained with global views (speech segments) of 5 s and randomly cropped local views of 3 to 4 s, respectively. Temperatures parameters for the teacher and student models are set to 0.06 and 0.1, respectively. Momentum values for the teacher model and centre update are 0.97 and 0.9, respectively. Results reported in Table 4 shows that DSR-RawNet3 model achieves a relative improvement of 10% & 15% over ECAPA-TDNN and RawNet3 baseline models. The best EER of 4.56% is achieved by ResNet-101 model which can be to the much larger model size. The DSR-RawNet3 model performs comparable to the ResNet-101 model and is able to bridge the gap further, with DSR-RawNet3-large (by increasing the channels and keeping the depth the same) having double the size of the DSR-RawNet3 model.

6 Conclusions, Limitations and Future Work

In this work, we presented a representation learning approach for SV from raw waveform at multiple sampling rates. To this aim, we proposed 1D-convolution based SR dependent frontend filterbank ‘DSR-FB’ that can be augmented to existing SV models. In particular, DSR-FB uses depth separable deformed convolutions to simulate attention at multiple scales/SRs by learning ‘offsets’ or sampling locations from raw inputs. To ensure temporal consistency, offsets are forced to be similar across channels. Compared to existing models operating at a single SR, the proposed DSR-FB can operate at multiple SRs at the inference time. Since low SR can lead to significant degradation in performance, we propose MMixBW training where in each batch, the model is trained randomly on audios with either one/all SRs. Experimental evidence demonstrates that MMixBW consistently outperforms MixBW training, where audio at lower SR is resampled to the desired operating SR before feeding it to the model for training. On the supervised learning benchmark our best model achieves state-of-the-art results at different SRs and performs comparably for unsupervised settings, both in closed-talk and far-field scenarios. In the process, we have established the effectiveness of adaptive dynamic multiresolution convolutional kernels for designing learned filterbanks for waveform based SV models, and we expect similar results in other speech/audio tasks.

There are multiple avenues where this study can be improved and extended. Firstly, DSR-FB can only operate on a discrete set of SRs, i.e., it requires a separate branch of convolutional kernels for each SR we want the model to operate at. This requires increasing the parameter budget or model size as we increase the number of SRs to be supported. Training at multiple SRs is a compute-intensive process to achieve generalization since the model has to learn invariant

features at multiple scales. These issues can be addressed by deforming the input signal instead of deforming the receptive field or kernel sampling locations. Thus the inference time of the FB is reduced due to the use of a static kernel, while the effective overall convolution operation still being dynamic. This method has been shown to be effective for vision tasks [25] and we would like to explore this in future. Secondly, this work assumes that MMixBW training ensures a unified latent space at all SRs, however, this is not explicitly enforced using a separate loss function, a direction worth exploring. Finally, in experiments with unsupervised benchmarks using DINO framework, the proposed DSR-FB based model could only achieve comparable performance even after doubling the model size. This requires further investigation and probing into the learning behaviour of the model to come up with an improved training strategy.

Acknowledgements. This research was funded in part by the Faculty Research Awards of J.P. Morgan AI Research. The authors are solely responsible for the contents of the paper, and the opinions expressed in this publication do not reflect those of the funding agencies.

References

1. Abdel-Hamid, O., Mohamed, A.R., Jiang, H., Penn, G.: Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2012)
2. Abel, J., Strake, M., Fingscheidt, T.: Artificial bandwidth extension using deep neural networks for spectral envelope estimation. In: IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1–5 (2016)
3. Abrol, V., Sharma, P.: Learning hierarchy aware embedding from raw audio for acoustic scene classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 1964–1973 (2020)
4. Bansal, D., Raj, B.: Smaragdis: bandwidth expansion of narrowband speech using non-negative matrix factorization. In: Interspeech, pp. 1505–1508 (2005)
5. Beigi, H.: *Fundamentals of Speaker Recognition*. Springer, New York (2011). <https://doi.org/10.1007/978-0-387-77592-0>
6. Bhagya, D., Suchetha, M.: A 1-D deformable convolutional neural network for the quantitative analysis of capnographic sensor. *IEEE Sens. J.* **21**(5), 6672–6678 (2021)
7. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9630–9640 (2021)
8. Chen, S., et al.: WavLM: large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Sig. Process.* **16**(6), 1505–1518 (2022)
9. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807 (2017)
10. Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(10), 5962–5979 (2022)






11. Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: *Interspeech*, pp. 3830–3834 (2020)
12. Fernandez Gallardo, L., Wagner, M., Möller, S.: Spectral sub-band analysis of speaker verification employing narrowband and wideband speech. In: *The Speaker and Language Recognition Workshop (Odyssey)*, pp. 81–87 (2014)
13. Gupta, D., Abrol, V.: Time-frequency and geometric analysis of task-dependent learning in raw waveform based acoustic models. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4323–4327 (2022)
14. Han, Y., Huang, G., Song, S., Yang, L., Wang, H., Wang, Y.: Dynamic neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(11), 7436–7456 (2022)
15. Haws, D., Cui, X.: CycleGAN bandwidth extension acoustic modeling for automatic speech recognition. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6780–6784 (2019)
16. Heigold, G., Moreno, I.L., Bengio, S., Shazeer, N.: End-to-end text-dependent speaker verification. In: *Proceedings of ICASSP (2016)*
17. Hirsch, H., Hellwig, K., Dobler, S.: Speech recognition at multiple sampling rates. In: *Eurospeech*, pp. 1837–1840 (2001)
18. Jung, J., Kim, Y., Heo, H.S., Lee, B.J., Kwon, Y., Chung, J.S.: Pushing the limits of raw waveform speaker recognition. In: *Interspeech*, pp. 2228–2232 (2022)
19. Koluguri, N.R., Park, T., Ginsburg, B.: TitaNet: neural model for speaker representation with 1D depth-wise separable convolutions and global context. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8102–8106 (2022)
20. Bhuvanagiri, K.K., Kopparapu, S.K.: Recognition of subsampled speech using a modified Mel filter bank. In: Abraham, A., Mauri, J.L., Buford, J.F., Suzuki, J., Thampi, S.M. (eds.) *ACC 2011. CCIS*, vol. 193, pp. 293–299. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22726-4_31
21. Kuzmin, N., Fedorov, I., Sholokhov, A.: Magnitude-aware probabilistic speaker embeddings. In: *The Speaker and Language Recognition Workshop (Odyssey)*, pp. 1–8 (2022)
22. Lee, J., Han, S.: NU-Wave: a diffusion probabilistic model for neural audio upsampling. In: *Proceedings Interspeech 2021*, pp. 1634–1638 (2021)
23. Li, Y., Tagliasacchi, M., Rybakov, O., Ungureanu, V., Roblek, D.: Real-time speech frequency bandwidth extension. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 691–695 (2021)
24. Lim, T.Y., Yeh, R.A., Xu, Y., Do, M.N., Hasegawa-Johnson, M.: Time-frequency networks for audio super-resolution. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 646–650 (2018)
25. Mac, K.N., Joshi, D., Yeh, R., Xiong, J., Feris, R., Do, M.: Learning motion in feature space: locally-consistent deformable convolution networks for fine-grained action detection. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6281–6290 (2019)
26. Mantena, G., Kalinli, O., Abdel-Hamid, O., McAllaster, D.: Bandwidth embeddings for mixed-bandwidth speech recognition. In: *Interspeech*, pp. 3203–3207 (2019)
27. Muckenhirn, H., Abrol, V., Magimai-Doss, M., Marcel, S.: Understanding and visualizing raw waveform-based CNNs. In: *Interspeech*, pp. 2345–2349 (2019)
28. Muckenhirn, H., Doss, M.M., Marcel, S.: Towards directly modeling raw speech signal for speaker verification using CNNs. In: *Proceedings of ICASSP (2018)*

29. Nagrani, A., Chung, J.S., Zisserman, A.: VoxCeleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017)
30. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: VoxCeleb: large-scale speaker verification in the wild. *Comput. Speech Lang.* **60**, 101027 (2020)
31. Nandwana, M.K., et al.: The VOiCES from a distance challenge 2019: analysis of speaker verification results and remaining challenges. In: *The Speaker and Language Recognition Workshop (Odyssey)*, pp. 165–170 (2020)
32. Palaz, D., Collobert, R., Doss, M.M.: Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. In: *Interspeech (2013)*
33. Palaz, D., Doss, M.M., Collobert, R.: Analysis of CNN-based speech recognition system using raw speech as input. In: *Proceedings of Interspeech (2015)*
34. Pariente, M., et al.: Asteroid: the PyTorch-based audio source separation toolkit for researchers. In: *Interspeech (2020)*
35. Ravanelli, M., Bengio, Y.: Speaker recognition from raw waveform with SincNet. In: *IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028 (2018)
36. Ravenscroft, W., Goetze, S., Hain, T.: Deformable temporal convolutional networks for monaural noisy reverberant speech separation. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2023)
37. Sainath, T.N., Weiss, R.J., Senior, A., Wilson, K.W., Vinyals, O.: Learning the speech front-end with raw waveform CLDNNs. In: *Interspeech (2015)*
38. Saito, K., Nakamura, T., Yatabe, K., Saruwatari, H.: Sampling-frequency-independent convolutional layer and its application to audio source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 2928–2943 (2022)
39. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.* **24**(3), 279–283 (2017)
40. Seltzer, M.L., Acero, A.: Training wideband acoustic models using mixed-bandwidth training data for speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **15**(1), 235–245 (2007)
41. Sivaraman, G., Vidwans, A., Khoury, E.: Speech bandwidth expansion for speaker recognition on telephony audio. In: *The Speaker and Language Recognition Workshop (Odyssey)*, pp. 440–445 (2020)
42. Snyder, D., Garcia-Romero, D., Povey, D.: Time delay deep neural network-based universal background models for speaker recognition. In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 92–97, December 2015
43. Su, J., Wang, Y., Finkelstein, A., Jin, Z.: Bandwidth extension is all you need. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700 (2021)
44. Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J.: Deep neural networks for small footprint text-dependent speaker verification. In: *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2014)*
45. Wang, M., et al.: Speech super-resolution using parallel WaveNet. In: *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 260–264 (2018)
46. Wang, X., Zhang, S., Yu, Z., Feng, L., Zhang, W.: Scale-equalizing pyramid convolution for object detection. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13356–13365 (2020)

47. Yamamoto, H., Lee, K.A., Okabe, K., Koshinaka, T.: Speaker augmentation and bandwidth extension for deep speaker embedding. In: Proceedings Interspeech 2019, pp. 406–410 (2019)
48. Yu, J., Luo, Y.: Efficient monaural speech enhancement with universal sample rate band-split RNN. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023)
49. Zhang, C., Koishida, K.: End-to-end text-independent speaker verification with triplet loss on short utterances. In: Proceedings of Interspeech (2017)
50. Zhang, Y., et al.: MFA-conformer: multi-scale feature aggregation conformer for automatic speaker verification. In: Proceedings Interspeech 2022, pp. 306–310 (2022)
51. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable ConvNets V2: more deformable, better results. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9300–9308 (2019)



Adjustable Gating Prompt Transformer for Facial Attribute Recognition with Limited Labeled Data

Qinxian Ye¹, Si Chen¹(✉) , Da-Han Wang¹ , Nanfeng Jiang¹ ,
Yanfei Su¹ , and Yan Yan² 

¹ Fujian Key Laboratory of Pattern Recognition and Image Understanding,
School of Computer and Information Engineering, Xiamen University of Technology,
Xiamen 361024, China

yeqinxian@stu.xmut.edu.cn, {chensi,wangdh,suyanfei}@xmut.edu.cn

² School of Informatics, Xiamen University, Xiamen 361005, China
yanyan@xmu.edu.cn

Abstract. Existing supervised facial attribute recognition (FAR) methods that rely on large labeled datasets can pose a challenge in real-world scenarios. In the case of limited labeled data, the current methods that introduce auxiliary tasks with a large number of parameters are not conducive to the embedded applications of FAR. To overcome these challenges, this paper develops an adjustable gating prompt Transformer that can handle the limited labeled FAR task with a small number of training parameters. Specifically, we employ an effective image-guided prompt tuning, where the image-related prompt sequence is first generated by feeding image tokens into an image-guided prompt generation network (IPG-Net). Then, the prompt sequence can learn facial image information and guide the frozen pre-trained Transformer to fine-tune the model. In addition, dynamically adjustable gating is applied to the prompt sequence to adaptively adjust the contribution of the prompts from different encoder layers, which enhances the interaction between the different encoder layers and retains effective feature information during the iterative process. Experimental results on the CelebA and LFWA datasets demonstrate that our method outperforms competitive methods with a very small amount of training parameters when only limited labeled data are used.

Keywords: Facial attribute recognition · Limited labeled data · Prompt learning · Adjustable gating

1 Introduction

With the rapid development of computer vision technology, facial attribute recognition (FAR) has attracted more and more attention as one of the important research directions. The FAR task aims to identify the different facial attributes, such as local attributes like Big Lips, Pointy Nose, and Eyeglasses, as well as

global attributes like Male, Heavy Makeup, and Attractive, to provide important support for subsequent tasks, such as face recognition, face editing, face synthesis, etc.

Current FAR methods [1–3] mainly use large labeled image datasets to train deep learning models. However, it is difficult and time-consuming to annotate FAR datasets, which makes the FAR task face many challenges. To address this problem, SSPL [4], SABAL [5], and SPL-Net [6] have been successively proposed to cope with the FAR task under limited labeled samples. SSPL [4] develops three auxiliary tasks, to learn spatial-semantic relations from large-scale unlabeled facial data. SABAL [5] designs a two-branch network that decouples faces into 3D shape features and facial appearance features. SPL-Net [6] associates different component labels with the attribute labels, and extends the original PCT.

However, although the above methods successfully cope with the FAR task with limited labeled data to a certain extent, the model complexity and the number of parameters increase significantly. When the FAR task is applied to real applications, the mounting and migration of complex models consume a lot of computational resources and memory, resulting in the unavailability of embedded applications. Slim-CNN [7] lightens the FAR network by combining separable convolution with pointwise convolution. However, this network only uses a simple CNN structure, with weak global feature representation capability and insufficient inter-layer interaction.

To address the issues mentioned above, we leverage the concept of prompt learning [8] and introduce prompt sequence into the model’s original input to achieve efficient fine-tuning of the FAR task with a small number of training parameters. However, in most of the existing prompt tuning methods [9,10], the initial prompt sequences are obtained by Xavier Uniform initialization. Such randomly initialized sequences have low relevance to the instances (images). Therefore, relying only on iterations of prompt sequences unrelated to the input images for visual feature capture can be detrimental to attribute recognition.

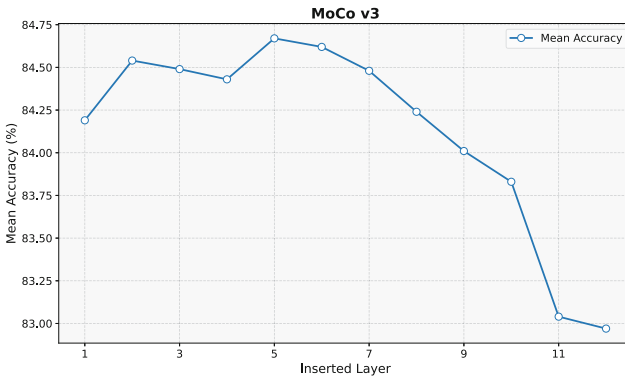


Fig. 1. Comparison of average recognition accuracy (%) for inserting prompt sequences at different layers of the encoder on the LFWA dataset.

In addition, different encoder layers of the Transformer contain different feature information. Typically, lower encoder layers concentrate on local patterns and basic features in the input sequence, while higher encoder layers may focus more on abstract semantic information and context. We thus explore the effect of the prompt sequence at the different encoder layers for the FAR task, and Fig. 1 lists the average accuracy against the LFWA dataset after the insertion of the prompt sequence at different encoder layers of the Transformer. It can be seen that the attribute recognition accuracy varies significantly with the different layers of prompt sequence embedding. However, manual testing for large-scale datasets on a layer-by-layer basis is laborious. Thus, it is very necessary to adaptively adjust and interact cross-layer information in order to retain effective feature information and boost the FAR performance.

In this paper, we propose an adjustable gating prompt Transformer, termed AGPT, for FAR, where MoCo v3 [11] pre-trained on ImageNet-1K [12] is migrated to our ViT [13] model under only a proportion of the FAR training set (with labels). Specifically, an initial prompt sequence is first obtained by feeding the image tokens into an image-guided prompt generation network (IPG-Net), making the prompt sequence associated with each input image. Then this prompt sequence is inserted into the original ViT model input and can interact with the image sequence. Moreover, dynamic gating is applied to adjust the prompt inputs of each layer, dynamically controlling the composition of the prompt sequences. Thus, the prompt sequence is selectively influenced by the ViT layers, and guides the adaptation of the FAR task, better fusing the deep and shallow information of the model. Instead of drastically modifying the model’s architecture during training, our model is frozen and only the prompt sequence is fine-tuned. Under limited training data, the performance of our method is even better than full fine-tuning with a very small number of training parameters. The contributions of this work are as follows:

- We introduce the concept of prompt learning to the FAR task and propose an adjustable gating prompt Transformer, where the image-guided prompt tuning is employed by using the prompt sequence interacted with the original image sequence to bootstrap the frozen ViT.
- We employ learnable gatings for prompt sequences, which dynamically adjust the composition of the input prompt sequences for each layer, thus facilitating cross-layer interaction and FAR task-related information aggregation.
- Experimental results demonstrate our method performs extremely well with a very small amount of training parameters, especially when limited training data are used.

2 Related Work

2.1 Facial Attribute Recognition

Currently, most FAR methods are still based on manual features and deep convolutional neural network (CNN), but inevitably can only deal with local regions

with convolutions at a time. In order to explore global features, Song et al. [14] used prior to guide multi-scale fusion Transformer to build integration between features of different scales. The graph convolutional network is used to simulate the relationship between attributes. Priadana et al. [15] proposed a lightweight multi-label CNN-Transformer architecture with an efficient initial block (EIB) and a squeeze channel Transformer encoder (SCTE).

Qin et al. [16] jointly trained a Transformer in a multi-task learning framework consisting of a shared Swin Transformer backbone and a face recognition subnet. Chen et al. [17] proposed a self-distillation-based multi-zone Transformer (MZTS) that captures interactions between different Transformer encoder blocks to avoid forgetting information in Transformer encoder blocks during iteration. Although existing Transformer-based FAR methods have made some progress, they still rely on the labeling of the dataset and have poor recognition performance for limited labeled data in practical applications. This paper introduces prompt learning derived from the field of NLP to assist in migrating FAR-related knowledge under limited labeled data.

2.2 Prompt Learning

Prompt learning is a new learning paradigm that has emerged in the field of natural language processing in recent years. It helps the pre-trained model to migrate to the downstream task by giving a certain prompt to the model. Schick et al. [8] proposed a semi-supervised training architecture that redesigns input samples into cloze phrases while generating pseudo-labels for a large number of unlabeled examples. Then the prompt paradigm is gradually introduced into the Vision-Language model. CLIP [18] is proposed to jointly predict several image text pairs by contrastive learning. The text branch constructs a photo of a {object} text label, where {a photo of a} is actually a manually designed prompt.

Zhou et al. [9] found that manually designing and iterating contexts requires a lot of effort, so they proposed context optimization (CoOp), where a learnable vector is used to model the prompted context automatically. Jia et al. [10] proposed a simple and effective Visual Prompt Tuning (VPT), to introduce the prompt tuning into the realm of pure vision. Bahng et al. [19] created prompts in the form of pixels to adapt to the frozen pre-trained classification model by modifying the pixels of the input image, and introducing perturbations in the pixel space to improve the model's performance. Later, more methods [20-22] introduce the concept of prompt learning to vision. However, the existing methods using prompt tuning do not explore the relationship between different layers, and only an image-independent randomly initialized prompt sequence is inserted. In this work, we employ learnable gatings to enhance cross-layer interaction and make the initial prompt sequence relevant to each image instance.

2.3 Learning from Unlabeled Data

The goal of self-supervised learning and semi-supervised learning is to improve the performance of models by learning feature representations from unlabeled data, and training models by introducing subtasks or contrastive learning. Recently, Transformer has also been widely used in the field of self-supervised. Both Beit [23] and iBOT [24] employ a pre-training approach similar to BERT [25], and they learn image features by comparing positive and negative examples. Beit [23] is a reconstructed self-supervised model, which randomly masks parts of image patches, and restores the original visual tokens. iBOT [24] uses an online tokenizer for mask prediction, where self-distillation is applied to the masked subblocks and they use the teacher network as an online tokenizer.

MAE (Masked Autoencoders) [26] is proposed based on a generative task. It randomly masks partial patches of the input image and then trains the model to reconstruct these missing pixels. In contrast, MoCo (Momentum Contrast) [27] is a self-supervised learning method based on contrastive learning. It learns stable feature representations by constructing a dynamically updated dictionary and adopting a momentum update mechanism. However, these methods leverage complex network structures, with a large number of parameters. Therefore, our work freezes the model during training and only fine-tunes the prompt sequence to achieve satisfactory recognition performance with very few parameters.

3 Proposed Method

We propose a straightforward and efficient adjustable gating prompt Transformer, termed AGPT, for facial attribute recognition with limited labeled data. Our method can make pre-trained Transformer models to adapt to the FAR task. AGPT incorporates trainable prompts into the input and maintains partial freezing of the backbone during the training phase. In this section, we first define the symbols and then discuss each part of the model in depth.

3.1 Preliminaries

Because SSL ViT [11,26] can utilize unlabeled data for feature learning, we transfer its knowledge to our ViT model, and thus help our model learn FAR-related knowledge with limited labeled data through the insertion of image-guided prompt sequences. This method adopts a ViT-like structure, where the input facial image $I \in \mathbb{R}^{H \times W \times C}$ is segmented into a series of patches $I_{\text{patch}} \in \mathbb{R}^{T \times (S^2 \times C)}$, where T denotes the number of image patches, S denotes the size of each patch, C denotes the channel number of the patches, and H and W are the height and width of the image, respectively.

The vectorized patches are projected into the potential d -dimensional embedding space using a learnable linear projection to obtain the serialization of patches I_e . Since the Transformer model itself does not have the ability to process location information and distinguish the relative locations, I_e is location-coded

to obtain the final image sequence E . Meanwhile, to obtain the predicted values of the facial attributes, the obtained image sequence is concatenated with the class token $CLS \in \mathbb{R}^{1 \times d}$ to obtain the spliced sequence $[CLS, E]$. The overall framework of the proposed method is illustrated in Fig. 2.

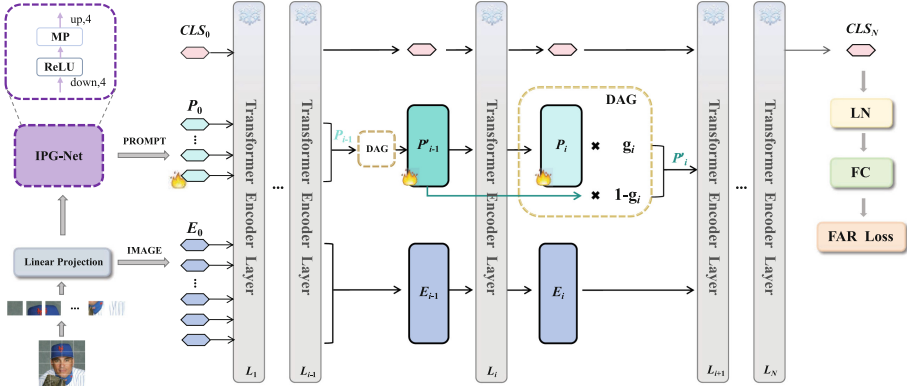


Fig. 2. The overall framework of the proposed method. After concatenating the prompt sequence that is generated by feeding image tokens into IPG-Net with the original image tokens, the concatenated sequence is input into a model consisting of 12 stacked Transformer encoder layers, where the composition of the prompt sequence at each layer is controlled by dynamically adjustable gating (DAG).

3.2 Image-Guided Prompt Tuning

Inspired by prompt learning [8, 10], the additional learnable tokens, called prompt, are used to guide the frozen model to adapt to the FAR better without the need for dedicated dataset pre-training, especially when a limited number of labeled FAR samples are used.

Firstly, by introducing the prompt sequence, the model input is changed from the original $[CLS, E]$ to

$$Z = [CLS, P, E], \quad (1)$$

where P is a learnable sequence of length K and dimension d , called prompt sequence. To get the initial prompt sequence input associated with each image instance, we employ an effective non-linear image-guided prompt generation network (IPG-Net) and then feed the initial sequence of images $[E_0]$ into the IPG-Net, so as to aggregate FAR task-related image features. The IPG-Net uses a simple Linear-ReLU-Linear (two-layer bottleneck) structure, and in order to reduce the computational cost, we also introduce max pooling (MP) to aggregate image features. The network structure is shown below:

$$\hat{P} = ReLU(f_{\text{down}}(E_0)), \quad (2)$$

$$P_0 = f_{\text{up}}(\text{MP}(\hat{P})), \quad (3)$$

where the f_{down} and f_{up} are downsampling and upsampling networks of the same scale, and are used to maintain the prompt sequence dimension. Then, the obtained prompt sequence P_0 is reshaped and then inserted between the original image patch embedding and the class token at the beginning. Then the prompt sequence is updated through training iterations to learn the task-specific knowledge of FAR.

The initial prompt sequence P_0 is inserted only before the first encoder layer L_1 , and the prompt sequence input of the subsequent layer inherits the output of the previous layer. That is, for layer L_1 , its input Z_0 is $[CLS_0, P_0, E_0]$, and its output is expressed as

$$Z_1 = [CLS_1, P_1, E_1] = L_1([CLS_0, P_0, E_0]). \quad (4)$$

Similarly, for layer L_i , its input Z_{i-1} is $[CLS_{i-1}, P_{i-1}, E_{i-1}]$ and its output is

$$Z_i = [CLS_i, P_i, E_i] = L_i([CLS_{i-1}, P_{i-1}, E_{i-1}]). \quad (5)$$

That is, before the input of the first layer of the model, a $K \times d$ learnable sequence P_0 is generated by the IPG-Net, and then P_0 is inserted between the classification token $[CLS_0]$ used to obtain the predicted values of the facial attributes and the sequence of images $[E_0]$ obtained by segmenting and projecting the facial images.

During training, only the parameters of the prompt sequence and the linear head are updated, while the entire Transformer encoder is frozen; the model is continuously optimized for the prompt sequence by gradient backpropagation, to achieve a small number of training parameters, so that the model pre-trained on the ImageNet dataset can be easily fine-tuned to adapt to the FAR task, and at the same time, it can still achieve strong generalization ability and robustness in the case of a limited number of labeled facial samples.

3.3 Dynamically Adjustable Gating

In order to dynamically adjust the gating values of each layer and enhance the interaction of FAR task-related information across layers, we first define the gate prior sequence, i.e., $G = [\gamma_1, \dots, \gamma_{N-1}]$, where N is the total number of Transformer encoder layers, and G contains the gate prior values of each layer except the last one. This sequence is learnable (the individual values within the G sequence are initially set to 10 in our method) and is continuously optimized with model iteration.

Thus the gate prior value of each encoder layer is adaptively tuned to achieve dynamic weighting of the prompt input and the output of the previous layer. The γ_i is subsequently scaled by a sigmoid function to obtain the corresponding gating values, i.e., $g_i = \text{sigmoid}(\gamma_i)$ for each layer, to determine the influence degree of the previous layer on the prompt sequence for the next layer. Then for the i -th layer, the output of L_i is $[CLS_i, P_i, E_i]$ and the input of the next layer is $[CLS_i, P'_i, E_i]$, where the prompt P'_i is defined as

$$P'_i = \begin{cases} P_i, & i = 1, \\ g_i \cdot P_i + (1 - g_i) \cdot P'_{i-1}, & i = 2, \dots, N. \end{cases} \quad (6)$$

Here, g_i controls the contribution of the prompt output P_i of the i -th layer and the prompt input P'_{i-1} of the i -th layer to the prompt input P'_i of the $(i + 1)$ -th layer when $i = 2, \dots, N$. Therefore, our method dynamically updates each prompt sequence by weighting the prompt input and the prompt output from the previous layer.

The prompt sequence output of each layer is selectively aggregated before the class head. By dynamically weighting the prompt sequence of each layer, after mathematical cumulative calculations, the last layer's prompt input can be obtained by the following statistical calculation.

$$P'_{N-1} = \left(\sum_{i=1}^{N-1} (1 - g_i) \right) P_0 + \sum_{i=1}^{N-2} \left(\sum_{m=i+1}^{N-1} (1 - g_m) \right) g_i P_i + g_{N-1} P_{N-1}, \quad (7)$$

where P_0 is the initial prompt sequence obtained by IPG-Net; P_i denotes the prompt sequence output of the i -th layer; and P_{N-1} denotes the prompt output of the $(N-1)$ -th layer, i.e., the layer before the last layer of the model.

3.4 Facial Attribute Recognition

To finally obtain the predicted values of each facial attribute, we feed the output of the class token CLS_N from the last Transformer encoder layer into the LayerNorm and the fully connected layer in the appropriate order, to acquire the predicted values \hat{y} of the facial attributes. Therefore, given the predicted values \hat{y} of the facial attributes and the ground truth y , the loss of FAR can be calculated in the following:

$$L_{\text{FAR}} = -\frac{1}{M} \sum_{i=1}^M \sum_{j=1}^A (y_i^j \log(\sigma(\hat{y}^{ij})) + (1 - y_i^j) \log(1 - \sigma(\hat{y}^{ij}))). \quad (8)$$

Here, M represents the total number of training images, and A represents the number of facial attributes, which is typically 40 in the dataset used in this method. Additionally, σ stands for the sigmoid function, i.e., $\sigma(\hat{y}^{ij}) = \frac{1}{1 + \exp(-\hat{y}^{ij})}$.

4 Experiments

4.1 Datasets and Experimental Setup

Experiments are conducted on the two challenging FAR datasets, i.e., LFWA (Labeled Faces in the Wild) and CelebA (Celeb-Faces Attribute). The LFWA dataset contains 13,143 face images from the Internet. The dataset is diverse and challenging, with images covering different conditions, such as, lighting, poses,

ages, etc. It is divided into a training set of 6,263 images and a test set of 6,880 images. Another important dataset is CelebA, a massive facial attribute dataset containing 202,599 images of celebrities, each with 40 binary attribute labels. It is divided into three sections, of which 162,770 images are used for training, 19,867 images for validation, and 19,962 images for testing.

We employ the PyTorch platform for each dataset to conduct all experiments on one NVIDIA RTX 3090 GPU to train the proposed method for 50 epochs. The input image size is 224×224 with a mini-batch size of 64. The learning rate undergoes adjustment through a cosine decay schedule [28], transitioning from 0.005 to 0.0 across 25 epochs. Our framework is trained using the stochastic gradient descent (SGD) algorithm [29] with a weight decay of 0.0001 and an SGD momentum of 0.9.

4.2 Ablation Studies

Component Ablation Experiments. To demonstrate the validity of the individual components of our proposed AGPT model, we perform ablation experiments for different components. Our proposed model contains two main components, i.e., Image-guided Prompt Tuning (IPT) and Dynamically Adjustable Gating (DAG). We report the corresponding accuracy on the LFWA dataset to test the effectiveness of the individual components.

Figure 3 shows the average accuracy of attribute recognition after adding our individual innovative components to the baseline step-by-step. The baseline used in our method is the SSL ViT-ViT16 transfer model, where we test the two pre-trained SSL ViT models, i.e., MoCo v3 [11] and MAE [26] on the ImageNet dataset as the initial weights and then transfer these weights to the Vision Transformer (ViT) model for the FAR task.

It can be clearly seen from Fig. 3 that when the components proposed in this method are gradually added, no matter MoCo v3 [11] or MAE [26] is used, the accuracy is greatly improved to different degrees. When image-guided prompt tuning is performed based on the baseline, the accuracy is improved by 2.4% for MoCo v3, and by 2.75% for MAE. It shows that our IPT component can improve the model performance by introducing facial image information to the prompt sequences. In addition, on this basis, the prompt sequence of the IPT is dynamically gated, that is, after adding DAG components, the accuracy is also improved by 3.14% for MoCo v3, and by 3.85% for MAE. We also found that the FAR performance pre-trained on MoCo v3 is better than that pre-trained on MAE. Therefore, we choose MoCo v3 as the pre-trained SSL ViT model in our method. These experimental results demonstrate the effectiveness of the addition of IPT and DAG components for FAR, boosting the transfer learning of SSL ViT models pre-trained on general datasets.

Hyperparameter. In addition, Fig. 4 tests the variation of accuracy under different numbers of image-guided prompt tokens by tuning the hyperparameter K , and compares the average accuracy with the other two methods to further prove the effectiveness of our proposed method. We vary the number of image-guided prompt tokens from 6 to 100 for accuracy testing. The three methods

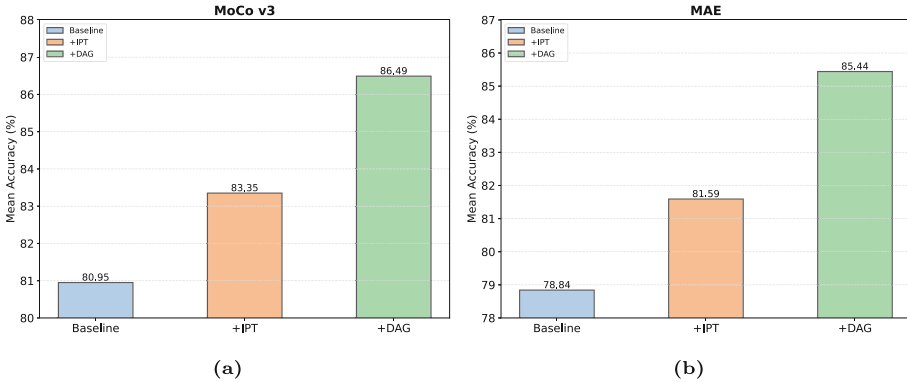


Fig. 3. Ablation study on the LFWA dataset based on the different pre-trained models, i.e., (a) MoCo v3 and (b) MAE, respectively.

listed in Fig. 4 are (1) AGPT using SSL ViT (MoCo v3) knowledge migration as proposed in this paper, (2) AGPT replacing SSL ViT with ViT knowledge migration, and (3) model without dynamic gating. As shown in Fig. 4, it is obvious that in the case of a minimum number of prompt tokens, the effect of our proposed method AGPT is also substantially better than the ViT-based AGPT and the model without using gating components.

Figure 4 clearly shows that our proposed method is not very dependent on the number of prompt tokens. As the number of prompt tokens decreases, the average accuracy of attribute recognition does not drop dramatically but more smoothly from 86.49% to 85.18%. At the same time, our method outperforms the compared method, which also reflects the dynamic gating can significantly improve the effectiveness of the FAR task.

From the above accuracy comparison, it can be seen that the AGPT in this paper does not produce performance improvement purely by the increase of the number of prompt tokens fed into the vision Transformer. Due to our new component structure, the knowledge learned by SSL ViT on the generalized dataset is effectively migrated to the FAR task. When replacing the knowledge migration source model of AGPT from SSL ViT to ViT, the model accuracy is greatly reduced regardless of the increase or decrease in the number of prompt tokens. After our exploration, we found that the information between layers in the unsupervised model SSL ViT is richer and more complete than that in the supervised ViT. Thus SSL ViT fits better with the idea of using the gating component, i.e., utilizing the information from each facial image as initial prompt tokens and dramatically adjusting the prompt tokens between different encoder layers. Meanwhile, it can be seen that, under different numbers of prompt tokens, the model is less accurate without using the dynamic gating component. The average recognition accuracy without the DAG component is lower than our proposed AGPT, which further proves the effectiveness of dynamic gating.

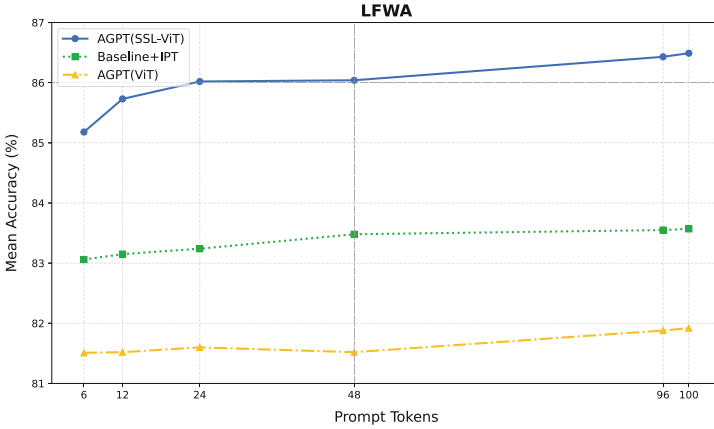


Fig. 4. Mean accuracy (%) under different numbers of image-guided prompt tokens.

4.3 Comparison with Other Models

In this section, we compare the accuracy and the number of training parameters of our method with other existing methods for both full labeled data and limited labeled data on the LFWA and CelebA datasets, respectively. In many practical FAR tasks, it is challenging to acquire a large amount of labeled data. Additionally, most existing FAR methods experience performance degradation when dealing with limited labeled samples. To address this issue, our method focuses on facial attribute recognition tasks under limited labeled data. Table 1 presents comparisons of the accuracy of our proposed method and other methods at different proportions of labeled data (the training data are acquired at intervals of the corresponding size, so that only a portion of the FAR training data is chosen with labels, and the default test sets are used for the FAR task), highlighting the effectiveness of our approach for limited labeled data.

As shown in Table 1, we compare the accuracy of our method with five supervised FAR methods, i.e., LNet+ANets [30], MCFA [3], SlimCNN [7], DMM-CNN [2], MZTS [17]; the self-supervised and semi-supervised methods FixMatch [31] and SimCLR [32], as well as SSPL [4] and SPL-Net [6] for FAR under limited labeled samples.

Our proposed method outperforms the current state-of-the-art method SPL-Net [6] in terms of accuracy when the proportion of labeled samples is limited for these two datasets. Especially, when utilizing a minimal proportion of limited annotated samples, i.e., 5% for the LFWA and 0.2% for the CelebA dataset.

On the LFWA dataset, when utilizing 5%, 10%, and 20% of the training data, our method’s average classification accuracy increases by 5.44%, 2.74%, and 0.46% compared to SPL-Net [6]. When using 0.2%, 0.5%, and 1% of the training data for the CelebA dataset, our method’s average classification accuracy improves by 0.56%, 0.23%, and 0.23% compared to the SPL-Net [6] method.

Table 1. Average classification accuracy (%) obtained by AGPT and several state-of-the-art methods with different proportions of labeled training data on the LFWA and CelebA datasets.

Proportion # of training samples	Venue	#params	LFWA				CelebA			
			5%	10%	20%	100%	0.2%	0.5%	1%	100%
			313	626	1252	6263	325	843	1627	162770
LNets+ANets [30]	ICCV 2015	>100M	–	–	–	83.03	–	–	–	87.33
MCFA [3]	ICPR 2018	260M	–	–	–	83.63	–	–	–	91.23
SlimCNN [7]	FG 2020	0.6M	70.90	71.49	72.12	76.02	79.90	80.20	80.96	91.24
DMM-CNN [2]	TAC 2022	360M	–	–	–	86.56	–	–	–	91.70
MZTS [17]	FG 2023	85.83M	–	–	–	86.73	–	–	–	91.66
FixMatch [31]	NeurIPS 2020	5.9M	71.42	72.78	75.10	83.84	80.22	84.19	85.77	89.78
SimCLR [32]	ICML 2020	35.3M	78.63	80.66	82.73	86.24	86.24	88.01	88.63	91.72
SSPL [4]	CVPR 2021	52.7M	78.68	81.65	83.45	86.53	86.67	88.05	88.84	91.77
SPL-Net [6]	IJCV 2023	48.1M	79.20	82.12	84.43	86.77	87.02	88.21	88.97	91.78
AGPT (Ours)	–	0.6M	84.64	84.86	84.89	86.49	87.58	88.44	89.20	91.67

Though our proposed method is not optimal for 100% of labeled data, the average accuracy is significantly better than that of SOTA under a very small percentage of labeled data, and the superiority of our method for the FAR task with very limited labeled data is undeniable. Moreover, the SSPL [4] and SPL-Net [6] methods require the joint development of three auxiliary tasks, a Patch Rotation Task (PRT), a Patch Segmentation Task (PST), and a Patch Classification Task (PCT), to learn spatial-semantic relationships from large-scale unlabeled facial data. However, our proposed method, AGPT, has a simple structure based on the ViTB16 architecture and freezes the model during training. It is able to obtain better recognition accuracy with very limited labeled samples and close to SPL-Net [6] with fully labeled ones, while the overall number of parameters required is extremely low which reduces the number of parameters by a factor of nearly 80 compared with SPL-Net. It can also be seen from the #params column of Table 1 that our method requires the lowest number of parameters compared to existing both supervised and unsupervised methods. Also compared to the lightweight FAR method SlimCNN [7], AGPT is able to achieve better recognition results using the same order of magnitude of number of parameters (i.e., 0.6M), with accuracy increasing from 76.02% to 86.49% for the LFWA dataset and from 91.24% to 91.67% for the CelebA dataset.

4.4 Visualization

In addition, we also visualize the multi-head attention map of AGPT. Figure 5 shows our FAR attention visualization images output from different heads, the first column is the raw input FAR image, and columns 2–5 are the attention maps output from different heads. As can be seen from the visualized heatmaps, with different heads, our architecture can cover all the partial attributes of the

face as well as the overall attributes. Even under face occlusion or cluttered background, it can still focus on the facial area that needs to be paid attention to for facial attribute recognition.

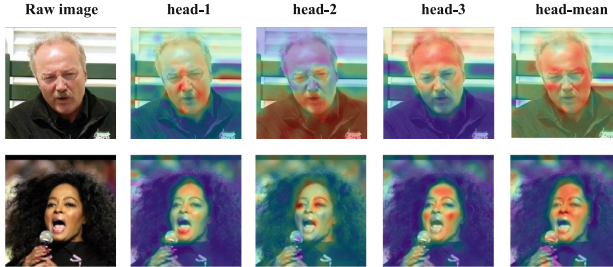


Fig. 5. Visualization of multi-head attention maps on the LFWA dataset.

4.5 Discussion

Strengths. By introducing Image-guided Prompt Tuning and Dynamically Adjustable Gating, the prompt sequence of the IPT guides the frozen model to learn FAR task-related knowledge with very few training parameters, and the DAG facilitates cross-layer interaction and FAR task-related information aggregation, so that discriminative feature information is still retained during the iteration process. The above experimental results strongly confirm the effectiveness and robustness of our proposed method under limited labeled data. In addition, the extremely low number of training parameters are consumed in our method, which can greatly save the data labeling consumption and computational resources of the FAR task.

Limitations. Although our method achieves satisfactory performance under limited labeled data, it still has some limitations. For example, the performance of attribute recognition under full labeled data is not yet optimal, and our method does not take into account the imbalance of attribute distribution. In the future, we will continue to explore these limitations by adopting special sampling methods for data with unbalanced attributes or by assigning different weights to attributes.

5 Conclusion

In this paper, we propose an adjustable gating prompt Transformer that introduces a prompt sequence in the input to guide the frozen Transformer encoder model to adapt to the FAR task. An initial prompt sequence is generated using an image-guided prompt generation network such that the prompt sequence is related to facial attribute instances. In addition, dynamically adjustable gating adaptively controls the influence of the prompt sequences of different encoder

layers, which enhances the cross-layer information interaction and aggregation. As a result, our AGPT network is able to obtain excellent FAR results under limited labeled data, while only requiring a very small number of training parameters for the FAR model.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (Nos. 62372388 and 62071404); Natural Science Foundation of Fujian Province (No. 2021J011185); Unveiling and Leading Projects of Xiamen (No. 3502Z20241011); Natural Science Foundation of Xiamen (No. 3502Z202373058); Fujian Key Technological Innovation and Industrialization Projects (No. 2023XQ023).

References

1. Hand, E.M., Chellappa, R.: Attributes for improved attributes: a multi-task network utilizing implicit and explicit relationships for facial attribute classification. In: AAAI Conference on Artificial Intelligence (2017). <https://api.semanticscholar.org/CorpusID:19671339>
2. Mao, L., Yan, Y., Xue, J.-H., Wang, H.: Deep multi-task multi-label CNN for effective facial attribute classification. *IEEE Trans. Affect. Comput.* **13**(2), 818–828 (2020)
3. Zhuang, N., Yan, Y., Chen, S., Wang, H.: Multi-task learning of cascaded CNN for facial attribute classification. In: International Conference on Pattern Recognition, pp. 2069–2074 (2018)
4. Shu, Y., Yan, Y., Chen, S., Xue, J.-H., Shen, C., Wang, H.: Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11 916–11 925 (2021)
5. Li, K., Zhang, J., Shan, S.: Learning shape-appearance based attributes representation for facial attribute recognition with limited labeled data. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 1–8 (2021)
6. Yan, Y., Shu, Y., Chen, S., Xue, J.-H., Shen, C., Wang, H.: SPL-Net: spatial-semantic patch learning network for facial attribute recognition with limited labeled data. *Int. J. Comput. Vision* **131**(8), 2097–2121 (2023)
7. Sharma, A.K., Foroosh, H.: Slim-CNN: a light-weight CNN for face attribute prediction. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 329–335 (2020)
8. Schick, T., Schütze, H.: Exploiting cloze-questions for few-shot text classification and natural language inference. In: Conference of the European Chapter of the Association for Computational Linguistics (2020). <https://api.semanticscholar.org/CorpusID:210838924>
9. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. *Int. J. Comput. Vision* **130**(9), 2337–2348 (2022)
10. Jia, M., et al.: Visual prompt tuning. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13693, pp. 709–727. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19827-4_41
11. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: IEEE/CVF International Conference on Computer Vision, pp. 9640–9649 (2021)

12. 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)
13. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale, [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
14. Song, S., Huang, H., Wang, J., Zheng, A., He, R.: Prior-guided multi-scale fusion transformer for face attribute recognition. In: Chinese Conference on Pattern Recognition and Computer Vision, pp. 645–659 (2022)
15. Priadana, A., Putro, M.D., An, J., Nguyen, D.-L., Vo, X.-T., Jo, K.-H.: Facial attribute recognition using lightweight multi-label CNN-transformer architecture for intelligent advertising. In: Annual Conference of the IEEE Industrial Electronics Society, pp. 1–7 (2023)
16. Qin, L., et al.: SwinFace: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 2223–2234 (2023)
17. Chen, S., Zhu, X., Wang, D.-H., Zhu, S., Wu, Y.: Multi-zone transformer based on self-distillation for facial attribute recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 1–7 (2023)
18. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021)
19. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Visual prompting: modifying pixel space to adapt pre-trained models, [arXiv:2203.17274](https://arxiv.org/abs/2203.17274), vol. 2, no. 3, p. 7 (2022)
20. Huang, Q., et al.: Diversity-aware meta visual prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10 878–10 887 (2023)
21. Nie, X., et al.: Pro-tuning: unified prompt tuning for vision tasks. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 4653–4667 (2023)
22. Yoo, S., Kim, E., Jung, D., Lee, J., Yoon, S.: Improving visual prompt tuning for self-supervised vision transformers. In: International Conference on Machine Learning, pp. 40 075–40 092 (2023)
23. Bao, H., Dong, L., Wei, F.: BEiT: BERT pre-training of image transformers, *ArXiv*, vol. abs/2106.08254 (2021). <https://api.semanticscholar.org/CorpusID:235436185>
24. Zhou, J., et al.: iBOT: image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (ICLR) (2022)
25. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019). <https://api.semanticscholar.org/CorpusID:52967399>
26. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16 000–16 009 (2022)
27. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
28. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts, [arXiv:1608.03983](https://arxiv.org/abs/1608.03983) (2016)
29. Goyal, P., et al.: Accurate, large minibatch SGD: training imagenet in 1 hour, [arXiv:1706.02677](https://arxiv.org/abs/1706.02677) (2017)

30. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
31. Sohn, K., et al.: FixMatch: simplifying semi-supervised learning with consistency and confidence. *Adv. Neural. Inf. Process. Syst.* **33**, 596–608 (2020)
32. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607 (2020)



IPHGaze: Image Pyramid Gaze Estimation with Head Pose Guidance

Hekuangyi Che^{1,2}, Dongchen Zhu^{1,2}, Wenjun Shi¹,
Guanghui Zhang¹, Hang Li¹, Lei Wang^{1,2}, and Jiamao Li^{1,2}

¹ Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China

dchzhu@mail.sim.ac.cn

² University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Gaze refers to the directed focus of an individual's visual perception, playing a fundamental role in human communication and cognition. Recent studies have employed neural networks to predict gaze from standard RGB face images. However, obtaining effective face images is challenging due to their sensitivity to bounding box size. The interference from head pose further complicates gaze estimation, yet in real-world scenarios, it is not feasible to obtain accurate head pose values. To overcome these challenges, we design the IPHGaze network guided by head pose information for image pyramid face gaze estimation. We craft this network to capture diverse face perspectives by incorporating various face bounding box sizes, ensuring rich gaze features. Additionally, a Feature Ensemble Module (FEM) facilitates feature sharing across image pyramid levels. We use head pose features instead of precise labels in two stages: feature communication and fusion, enhancing robustness for stable gaze predictions. Our method achieves a 5.4% improvement on EyeDiap dataset and a 2.5% improvement on Gaze360 dataset compared with existing methods, which demonstrates its effectiveness and versatility across diverse indoor and outdoor scenarios.

Keywords: Gaze estimation · Image pyramid · Head pose · Deep learning

1 Introduction

Gaze, the directed focus of an individual's visual perception, serves as a fundamental element in human communication and cognition. The study of gaze behavior has garnered significant attention across a spectrum of disciplines, ranging from psychology and neuroscience to human-computer interaction and computer vision. Understanding where individuals are looking and how their gaze patterns evolve provides invaluable insights into their cognitive processes, intentions, interests, and emotional states. This understanding is facilitated by gaze-based interfaces and eye-tracking technology. Gaze estimation algorithms have facilitated innovations in fields like virtual reality [26, 35], augmented reality

[3,4], and assistive technology [19,21]. By inferring and tracking a user’s gaze, these technologies enable more intuitive and efficient interactions, improving user experiences across diverse applications.

Traditional gaze estimation methods face significant limitations as they depend on fitting eye keypoints and often necessitate the use of infrared cameras and close proximity between the subject and the camera. In response, appearance-based methods using standard RGB cameras have emerged, employing neural networks to predict gaze from face images. These methods involve feeding a single face image into a neural network to extract features and regress gaze angles. However, effectively capturing face images poses challenges, such as the sensitivity to the size of the face bounding box, which can either omit crucial head pose data or hinder focus on eye-related information. Moreover, the integration of head pose, a crucial element for gaze direction, remains a challenge in current research, which often overlooks it in end-to-end gaze prediction or struggles with obtaining practical real-world head pose data.

To address these two challenges, we introduce the IPHGaze (**I**mage **P**yrAmid **G**aze estimation with **H**ead pose guidance), guided by head pose information, for image pyramid face gaze estimation. Specifically, in response to the first challenge, we draw inspiration from the concept of a feature pyramid. We design a network tailored for face image pyramid by incorporating different face bounding box sizes, enabling us to capture face images from various perspectives. This approach encompasses both small face regions that focus on the eye areas and larger face regions that encompass the entire face. This design ensures the richness and robustness of gaze features. To facilitate the exchange of face feature information extracted at various scales, we have also designed an information exchange module called FEM. This module is responsible for exchanging feature information from different levels of image pyramids. For the second challenge, we employ head pose features instead of actual head pose ground truth labels to guide face gaze estimation. This approach eliminates the need for precise head pose labels. The introduction of head pose is divided into two stages: feature communication and feature fusion. In the feature communication stage, we input head pose features into FEM to transfer head pose feature information into the face gaze feature space, guiding the estimation of face gaze. In the feature fusion stage, we concatenate the head pose features with the multi-scale face gaze features. The resulting fused features include both the explicit fusion of face gaze and head pose features and the implicit fusion between each sub-feature. This fused feature representation enhances robustness, ensuring stable and reliable gaze predictions.

Extensive experiments are conducted on EyeDiap [15] and the Gaze360 [24] dataset. And the results demonstrate that our algorithm achieves competitive performance on both datasets. Specifically, on the EyeDiap dataset, the error is as low as 4.73° , while on the Gaze360 dataset, the error measures 10.15° . These findings illustrate the effectiveness of our method in diverse scenarios, whether in indoor, close-distance environments or in outdoor, longer-distance settings, showcasing its robustness and versatility.

In summary, our contributions can be summarized as follows:

1. We create a specialized network for the face image pyramid, incorporating diverse face bounding box sizes to capture different perspectives, including small regions focusing on the eyes and larger ones encompassing the entire face. This design ensures rich and robust gaze features. Additionally, we introduce a Feature Ensemble Module called FEM to share feature information across different image pyramid scales.
2. We use head pose features instead of precise labels to guide face gaze estimation, eliminating the need for exact annotations. This involves two stages: feature communication, where head pose information guides gaze feature extraction through FEM, and feature fusion, where head pose and multi-scale face gaze features are combined for robust, stable gaze predictions.
3. Our proposed method achieves competitive performance on both the indoor, close-distance EyeDiap dataset and the outdoor, longer-distance Gaze360 dataset.

2 Related Work

2.1 Appearance-Based Gaze Estimation

Appearance-based gaze estimation seeks to establish a mapping from eye or face images to gaze directions. With the rapid advancements in deep learning, significant progress has been made in this area. Cheng et al. [9] merge CNNs with ViT, harnessing CNNs’ superior local perception and ViT’s robust global perception, to develop a universal face gaze estimator. Nagpure et al. [29] introduce Neural Architecture Search into gaze estimation, achieving a substantial reduction in model size while preserving or even enhancing estimation accuracy. Abdelrahman et al. [1] segment gaze estimation into regression and classification tasks and train them jointly, thereby further reducing estimation errors.

Nonetheless, these approaches primarily concentrate on gaze features without accounting for the influence of head pose. Additionally, the extraction of face features is sensitive to the bounding box, an improperly sized cropping can compromise the accuracy of gaze feature detection.

2.2 Image Pyramid and Feature Pyramid

Image Pyramid is a technique for multi-scale representation of images, which includes a series of versions of the original image at different sizes or resolutions. Each version is obtained by downsampling or upsampling the original image. The purpose of image pyramids is to analyze image features at different scales, allowing for object detection [25, 33] or segmentation [22, 28] at different resolutions. Common types of image pyramids include Gaussian pyramids and Laplacian pyramids. Feature Pyramid is an image representation used for computer vision tasks, typically associated with deep convolutional neural networks (CNNs). Unlike image pyramids, feature pyramids consist of a multi-scale stack

of feature maps generated at different levels of a CNN network. Each feature map level corresponds to a different abstract feature representation with varying semantic information. This enables more effective detection and recognition of objects at different sizes within an image while preserving rich semantic information. Due to the fact that feature pyramids have no specific requirements on input images, their application is even more extensive [16, 27, 32, 37]. Specifically, Cheng et al. [12] propose the gaze pyramid transformer, in which they use a convolutional network to extract feature maps from multiple layers. These feature maps are processed through 1×1 convolutions and global average pooling layers to achieve uniform feature dimensions. Then, they are fed into a transformer to effectively integrate both shallow and deep features.

However, existing image pyramids typically involve resizing the original image and lack a specific focus on particular regions of the image. Gaze estimation is primarily influenced by the position of the human eye, thus employing a cropping method to concentrate the image more on the eye area aligns better with the requirements of gaze estimation. At the same time, current feature pyramids often simply sum up feature maps from different scales after straightforward scaling, resulting in the final feature map. This approach overlooks the interaction of information between features at different scales. Therefore, we introduce FEM to ensure equitable information exchange among features at different scales.

2.3 Gaze Feature Communication

Bao et al. [2] recalibrate eye features using shift and scale parameters derived from face features. Cai et al. [5] introduce the iTracker-MHSA module to merge eye and face features. Gideon et al. [17] explore disentangling image features via feature swapping in multi-view videos. Yun et al. [34] develop a high-frequency attention block to enhance high-frequency details, including in the eye regions. Cheng et al. [10] create DIC blocks for dual-view information exchange during convolution, enriching the original features by adding fused data. Hisadome et al. [20] propose Rotation-Constrained Feature Fusion, utilizing relative camera rotation for feature extraction and fusion. Building on our previous work [6] which developed an information exchange module, we now incorporate channel weight parameters to emphasize more significant features, thus boosting feature saliency.

2.4 Head Pose in Gaze Estimation

Head pose and gaze direction are intricately linked, as changes in head movements directly affect gaze direction. To mitigate head pose interference, researchers like Zhu et al. [38] incorporate head pose data using geometric transformation layers in neural networks, enhancing gaze estimation. Unlike methods relying on unavailable ground truth for head pose, Tobias et al. [14] use a network to integrate global facial information with eye features for gaze prediction. This network, however, operates without ground truth supervision, making its efficacy

in learning head pose a subject for debate. Wang et al. [31] integrate head pose with eye image features for gaze zone prediction using advanced image processing techniques. Jha et al. [23] further refine gaze estimation with a probabilistic visual attention map that utilizes head pose to predict gaze areas effectively. Our approach, leveraging pseudo-labels for head pose, ensures accurate head pose information extraction without relying on actual ground truth, setting a new standard in robust gaze estimation methodologies.

3 Method

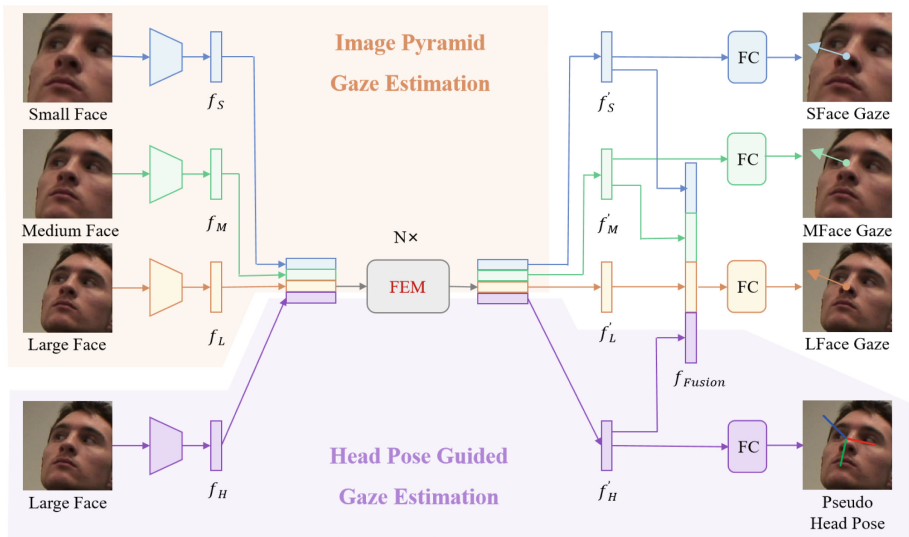


Fig. 1. The pipeline of our proposed IPHGaze. Note that the SFace Gaze, MFace Gaze, and Head Pose branches are only used during the training process. During inference, the only output is LFace Gaze, which is the final result. FC represents a single-layer fully connected layer.

3.1 Overview

Our proposed IPHGaze is illustrated in Fig. 1. The network takes as input a face image pyramid of the same individual, which comprises three images ranging from large to small, denoted as Large Face, Medium Face, and Small Face. These face images encompass the eye regions to ensure the inclusion of eye-related features relevant to gaze estimation. We employ a convolutional neural network (CNN) to extract face gaze features from these images. Considering that the larger face has a broader field of view and encompasses the entire face region,

we use the large face to extract head pose features. Similarly, we use a CNN to extract head pose features. Subsequently, we feed the extracted face gaze features along with the head pose features into our designed feature exchange module FEM. This module facilitates comprehensive feature fusion, integrating the head pose features into the multiscale face features. Finally, we concatenate the face features, enriched with feature exchanges through FEM, and the head pose features. We employ fully connected layers to predict face gaze angles, including yaw and pitch angles, from this concatenated feature. To ensure the accuracy of the face gaze features for the small and medium faces, as well as the head pose features, we use separate fully connected layers to regress their respective gaze or head pose angles. Gazes are guided by ground truth, while head pose is supervised using pseudo-labels generated by 6drepnet [18].

3.2 Image Pyramid Gaze Estimation

Face Image Pyramid. We take inspiration from feature pyramid to propose face image pyramid for gaze estimation. In a feature pyramid, the input image is processed by a series of convolutional layers with different receptive fields. The output feature maps from these layers are then combined to form a pyramid-like structure, where each level represents a different scale of the input image. This allows the network to extract features at multiple scales, which is crucial for accurately localizing objects of different sizes and shapes.

However, the information contained within a single image is limited. For gaze estimation, if the input face region is too small, it may not involve global information like head pose. Conversely, if the input face region is too large, it might not focus adequately on the eye area. Even though feature pyramids can extract multiscale information from a single image, they cannot fundamentally address the challenge of balancing global and local information.

To tackle this issue, we construct an image pyramid at the image level. Through the design of large, medium, and small faces within this pyramid, the network can consider global head pose while simultaneously focusing on local regions such as the eyes. This enhances the expressive power of the features, allowing for a more comprehensive analysis.

Feature Communication. Inspired by MLP-Mixer [30], we designed the Feature Ensemble Module (FEM) for facilitating information exchange among multiscale face gaze features and head pose feature.

As described in Fig. 2, FEM is composed of inter feature ensemble and intra feature ensemble. Suppose the dimension of features after feature extraction is N . We concatenate three face features and one head pose feature to be a feature matrix $F \in \mathbb{R}^{4 \times N}$. First, layer normalization is preformed to F , and then F is transposed to $F^T \in \mathbb{R}^{N \times 4}$. Features from the same channel are mixed by the MLP1 module which includes two Fully-connected layers and a GeLU layer. After that, each channel is smoothed by softmax layer to get the coefficients for each channel. The origin feature is multiplied by these coefficients and added

with origin feature. The above processes are called inter feature ensemble. For intra feature ensemble, the processing flow is similar except for transpose. F is processed again by layer normalization, MLP2 module, and skip-connection. Note that MLP1 and MLP2 have the same structure. In order for gaze features and head pose feature to fully communicate with each other, we repeat FEM for several times.

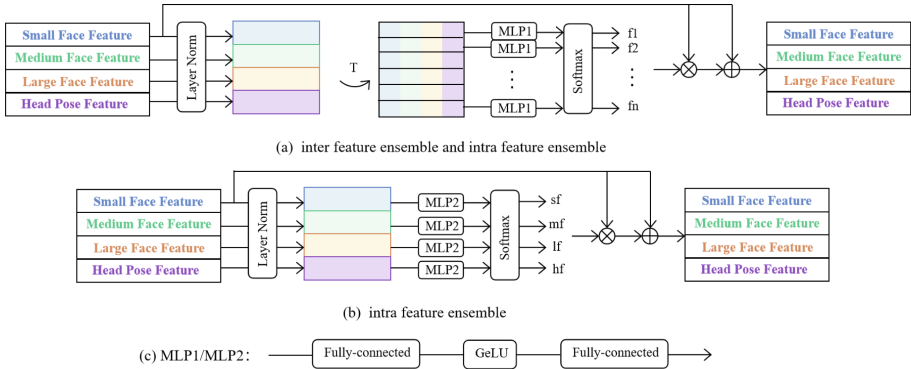


Fig. 2. Architecture of FEM. (a) Ensemble features within the same channel across different feature vectors, (b) Ensemble different channels within the same feature vector, (c) MLP (Multi-Layer Perceptron) structure.

3.3 Head Pose Guided Gaze Estimation

To counteract the influence of head pose variations on gaze prediction, we introduce head pose information. This information is synthesized using head pose estimation techniques, allowing the model to gain insights into the user’s head orientation. Head pose cues guide the model’s attention, enabling it to focus on relevant gaze cues despite variations in head pose. The guidance of head pose features can be divided into two phases as follows.

Head Pose in Feature Communication. During the feature communication phase, we pass the head pose feature extracted from the large face through FEM. Through the feature exchange process in FEM, the head pose feature is effectively transferred to guide the generation of face gaze features, aiding in the fusion of global information related to head pose with the face gaze features.

Head Pose in Feature Fusion. In the feature fusion phase, we concatenate the head pose feature with the face gaze features from the large, medium, and small faces. Subsequently, we employ a fully connected layer to regress the final gaze angles, namely the yaw and pitch angles, from the concatenated features.

To ensure the reliability of the head pose feature, we supervise it using pseudo-labels generated by 6drepnet [18]. It’s worth noting that in practical usage, we only consider the yaw and pitch angles of the head pose, as the roll angle does not contribute significantly to gaze prediction.

3.4 Loss Function

The network’s predictive output involves the yaw, pitch angles of gaze as well as the yaw, pitch angles of head pose. Therefore, we employ an L1 loss function to construct the overall loss function. For the image pyramid, we have constructed three loss functions, one each for the large, medium, and small faces. Similarly, for head pose, we have employed similar loss functions. It’s worth noting that pseudo-labels are used for head pose supervision.

The total loss function is the weighted sum of the aforementioned loss functions, as shown in Eq. 1. In this formula, λ represents the weight assigned to the head pose loss function, typically set empirically to 0.1.

$$\begin{aligned}
 Loss &= \frac{1}{N} \sum_{i=1}^N \left| g_i^{lface} - \hat{g}_i^{lface} \right| + \frac{1}{N} \sum_{i=1}^N \left| g_i^{mface} - \hat{g}_i^{mface} \right| \\
 &+ \frac{1}{N} \sum_{i=1}^N \left| g_i^{sface} - \hat{g}_i^{sface} \right| + \lambda \cdot \frac{1}{N} \sum_{i=1}^N \left| h_i - \hat{h}_i \right| \\
 &= Loss_{LGaze} + Loss_{MGaze} + Loss_{SGaze} + \lambda \cdot Loss_{HeadPose}
 \end{aligned} \tag{1}$$

4 Experiments

4.1 Datasets

To evaluate the performance of our algorithm, extensive experiments were conducted on publicly available datasets, EyeDiap and Gaze360.

EyeDiap: This dataset was captured in an indoor environment using a Kinect camera and an HD camera. It comprises 237 min of video segments from 16 subjects, with subjects positioned approximately 80–90 cm from the camera. The dataset underwent preprocessing using the method provided by [11]. During preprocessing, face images were cropped into three sizes: 224×224 for small faces, 300×300 for medium faces, and 360×360 for large faces. Cross-validation was performed using dataset partitioning provided by [11], involving a 4-fold cross-validation strategy.

Gaze360: Captured in indoor and outdoor scenes using a Ladybug5 camera, the Gaze360 dataset contains 197,588 images from 238 subjects, with subjects located at distances of 1–3 m from the camera. Preprocessing of the data was conducted using the method provided by [11], resulting in 84,902 images for training, 11,318 for validation, and 16,031 for testing. Face images were cropped

at 0.8 times the size of the provided face bounding box for small faces, at the original bounding box size for medium faces, and at 1.2 times the bounding box size for large faces. The original dataset’s partitioning scheme was used for both training and testing the models.

4.2 Implementation Details

Before inputting each image into the network, we first resize it to 224×224 to ensure consistent input sizes. For the EyeDiap dataset, which has a smaller number of subjects, closer subject-camera distances, and clearer data, we use ResNet18 as the backbone network to extract gaze features for large, medium, and small faces. Additionally, ResNet18 is used to extract head pose features. On the other hand, for the Gaze360 dataset, which involves a larger number of subjects, a more complex environment, indoor and outdoor scenes, longer subject-camera distances, and relatively blurred images, we employ RepVGG [13] to extract gaze features for large faces and ResNet50 to extract features for medium and small faces. Head pose features are still extracted using ResNet. The dimensions of the extracted features are all set to 100, and FEM is stacked twice.

During training, we set the batch size to 256, the number of epochs to 80, and the initial learning rate to $5e-4$. After 60 epochs, the learning rate is reduced to half of its previous value. We implement a warmup strategy for the first 5 epochs. Adam optimizer is used with β_1 set to 0.9 and β_2 set to 0.999. Our IPHGaze is implemented using PyTorch and trained on 4 NVIDIA RTX3090 GPUs.

4.3 Comparison with Appearance-Based Methods

We conducted a series of experiments on the EyeDiap and Gaze360 datasets to compare the performance of our proposed algorithm with leading appearance-based gaze estimation algorithms. Our comparison includes FullFace, Rt-gene, Dilated-Net, Gaze360, CA-Net, GazeTR, GazeNAS and L2cs-net. Due to variations in experimental conditions among the first five algorithms, we used results reproduced by [9] as a fair baseline for comparison.

The experimental results are presented in Table 1. On the EyeDiap dataset, the existing algorithms achieved a minimum error of 5.00° . Our approach achieved a significant breakthrough by reducing gaze estimation error to within 5.00° for the first time. Compared to the current best-performing algorithm, our method lowered the error from 5.00° to 4.73° , marking a notable improvement of 5.4%.

Moving to the Gaze360 dataset, characterized by its large dataset size, numerous subjects, complex acquisition scenarios, and longer subject-camera distances, gaze estimation is inherently challenging. Current algorithms exhibit errors above 10° on this dataset, and our method also displayed strong performance. In comparison to the current best algorithm, which had an error of 10.41° , our approach reduced the error to 10.15° , marking a 2.5% improvement.

This highlights the versatility of our method, demonstrating its effectiveness in both indoor and outdoor environments for accurate face gaze angle estimation.

Table 1. Comparison with other methods on EyeDiap and Gaze360 Datasets.

Algorithms	Years	EyeDiap	Gaze360
FullFace [36]	CVPRW2017	6.53°	14.99°
Rt-gene [14]	ECCV2018	6.02°	12.26°
Dialted-Net [7]	ACCV2018	6.19°	13.73°
Gaze360 [24]	ICCV2019	5.36°	11.04°
CA-Net [8]	AAAI2020	5.27°	11.20°
GazeTR [9]	ICPR2022	5.17°	10.62°
GazeNAS [29]	WACV2023	5.00°	10.52°
L2cs-net [1]	ICFSP2023	N/A	10.41°
Ours		4.73°	10.15°

4.4 Ablation Study

To validate the effectiveness of IPHGaze, we conducted a series of ablation experiments on the EyeDiap and Gaze360 datasets. The experimental results are presented in Table 2, where we primarily investigated the impact of the FEM, the addition of head pose information in FEM, and the types of features fused in the final fusion stage. These correspond to the first three elements in the table header.

Comparing the results between the second and third rows, it’s evident that using FEM to communicate multi-scale face gaze features reduce the error from 5.17° to 4.99° on EyeDiap and from 10.61° to 10.47° on Gaze360. This demonstrates the effectiveness of extracting and exchanging multi-scale face gaze features. Comparing the results between the fourth and fifth rows, the addition of head pose features significantly improved the algorithm’s performance. The error on EyeDiap decreased from 4.91° to 4.84°, and on Gaze360, it decreased from 10.32° to 10.22°. Comparing the results between the fifth and last rows, it’s evident that further improving the algorithm’s performance was achieved by fusing head pose features in the fusion stage. Combining the results from the fourth, fifth, and last rows validates the effectiveness of incorporating head pose features in the information exchange and fusion stages. By introducing head pose in the information exchange stage, the error on EyeDiap decreased from 4.91° to 4.84°, and on Gaze360, it decreased from 10.32° to 10.22°. Building upon this, incorporating head pose features in the fusion stage resulted in an error reduction on EyeDiap from 4.84° to 4.73°, and on Gaze360, it decreased from 10.22° to 10.15°. This underscores the importance of utilizing head pose features for information exchange and fusion.

Table 2. Ablation study on EyeDiap and Gaze360 Datasets. HP in FEM denotes feeding head pose feature into FEM. LFace represents gaze prediction solely from the feature of the Large Face. LMSFace signifies gaze prediction from the features concatenated from the Large Face, Medium Face, and Small Face. LMSFaceHead indicates gaze prediction from the features concatenated from the Large Face, Medium Face, Small Face, and head pose information.

FEM	HP in FEM	Fuse Feature	EyeDiap	Gaze360
×	×	LFace	5.17°	10.61°
√	×	LFace	4.99°	10.47°
√	×	LMSFace	4.91°	10.32°
√	√	LMSFace	4.84°	10.22°
√	√	LMSFaceHead	4.73°	10.15°

4.5 Comparison with Feature Pyramid Methods

In order to further compare the performance differences between our image pyramid approach and the commonly used feature pyramid methods, we selected the MVITv2 [27] for comparison. We used the Large Face as input and loaded a pre-trained model from ImageNet. We replaced the last softmax layer of MVITv2 with a fully connected layer that directly outputs yaw and pitch angles for gaze estimation. Other training details were consistent with our method.

We conducted experiments using both the Base and Large models, and the results are shown in Fig. 3. Increasing the model parameters from Base to Large significantly improved model performance. However, when compared to our method, it is evident that with the use of an image pyramid and information exchange, our approach achieves performance comparable to *MViT2-L*. Moreover, by incorporating multi-scale face feature information fusion, we can further reduce gaze estimation errors. This strongly demonstrates that our designed image pyramid method is better suited for gaze estimation compared to commonly used feature pyramids (Table 3).

Table 3. Comparison with Feature Pyramid Methods on EyeDiap and Gaze360 Datasets.

Algorithms	Years	EyeDiap	Gaze360
MViTv2_B [27]	CVPR2022	5.38°	10.76°
MViTv2_L [27]	CVPR2022	5.12°	10.44°
Ours(FEM+LFace)		4.99°	10.47°
Ours(FEM+LMSFace)		4.91°	10.32°

4.6 Robustness

The results of the robustness analysis are shown in Fig. 3. The first row represents the results for the EyeDiap dataset, while the second row corresponds to the Gaze360 dataset. The first column displays the pitch angle results, and the second column displays the yaw angle results. It is evident that our algorithm performs significantly better than the baseline on the EyeDiap dataset, particularly when the absolute values of both pitch and yaw angles are relatively small. This performance difference is even more pronounced in such scenarios. When we switch to the more challenging indoor and outdoor Gaze360 dataset, our algorithm still outperforms the baseline across a wider range of angles. This indicates that our method exhibits robustness across varying environmental conditions.

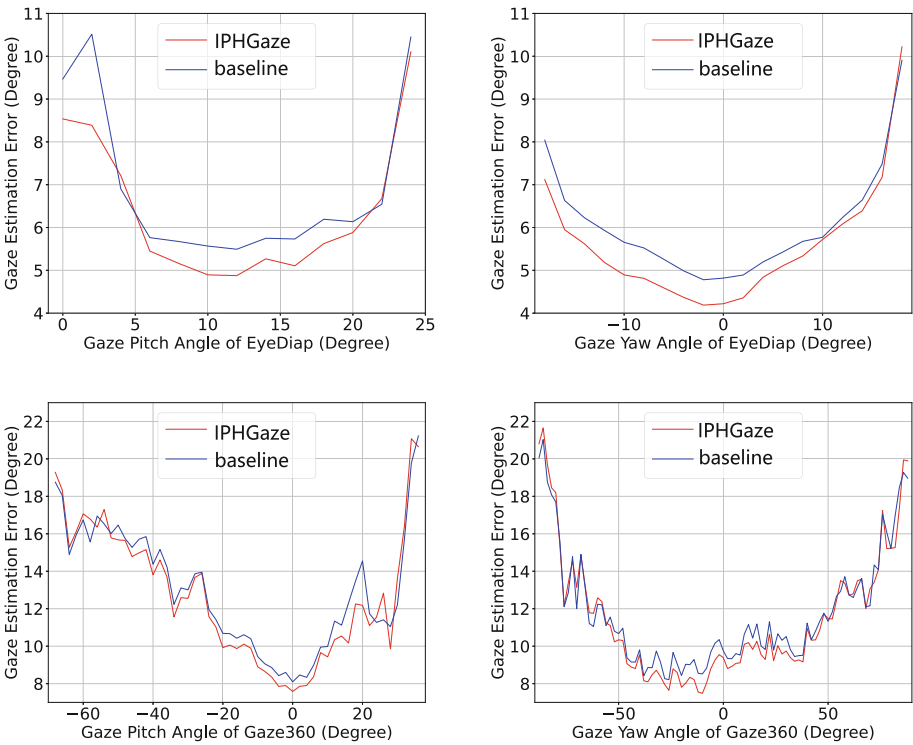


Fig. 3. Robustness analysis of gaze angles on Eyediap and Gaze360 datasets.

4.7 Qualitative Results

Some qualitative examples of our model are shown in Fig. 4. The baseline exhibits a significant deviation from the ground truth, whereas IPHGaze aligns more

closely with it. The first row displays results from EyeDiap, while the second row presents results from Gaze360. The last two columns depict conditions of blurry images or poor illumination. It is evident that even in adverse conditions, IPHGaze outperforms the baseline.

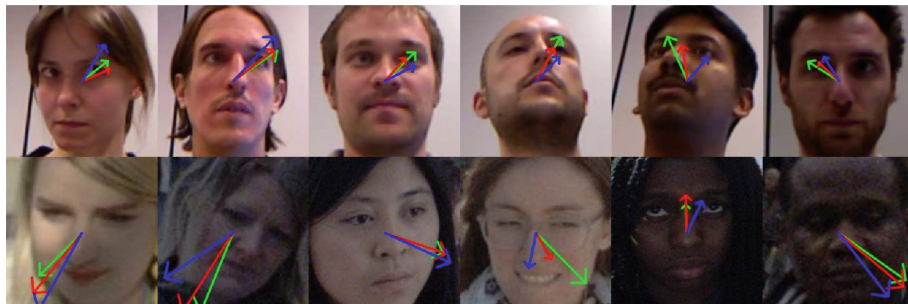


Fig. 4. Qualitative Results on Gaze360 and EyeDiap Datasets. The red, blue, and green arrows are, IPHGaze, baseline, ground truth, respectively. (Color figure online)

5 Conclusion

In this paper, we introduce a novel image pyramid framework called IPHGaze, guided by head pose information, for face gaze estimation. Specifically, we employ different sizes of face bounding boxes to capture diverse face perspectives and extract rich gaze features. Additionally, we design an information exchange module, FEM, to facilitate feature sharing among different levels of the image pyramid. We then utilize head pose features to guide facial gaze estimation, rather than relying on precise labels. Extensive experiments are conducted on the EyeDiap and Gaze360 datasets, demonstrating that IPHGaze achieves state-of-the-art performance on both datasets. This underscores its effectiveness and versatility across various scenarios. Furthermore, we compare IPHGaze with feature pyramid methods, highlighting its unique advantages. We hope that this paper can provide new inspiration and insights to the field of gaze estimation.

Acknowledgements. This work was supported by National Science and Technology Major Project from Minister of Science and Technology, China (2021ZD0201403), Natural Science Foundation of Shanghai (23ZR1474200), Youth Innovation Promotion Association, Chinese Academy of Sciences (2021233, 2023242), Shanghai Academic Research Leader (22XD1424500).

References

1. Abdelrahman, A.A., Hempel, T., Khalifa, A., Al-Hamadi, A., Dinges, L.: L2CS-Net: fine-grained gaze estimation in unconstrained environments. In: 2023 8th International Conference on Frontiers of Signal Processing (ICFSP), pp. 98–102. IEEE (2023)
2. Bao, Y., Cheng, Y., Liu, Y., Lu, F.: Adaptive feature fusion network for gaze tracking in mobile tablets. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 9936–9943. IEEE (2021)
3. Bao, Y., Wang, J., Wang, Z., Lu, F.: Exploring 3D interaction with gaze guidance in augmented reality. In: 2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR), pp. 22–32. IEEE (2023)
4. Bektaş, K., Strecker, J., Mayer, S., Garcia, K.: Gaze-enabled activity recognition for augmented reality feedback. *Comput. Graph.* **119**, 103909 (2024)
5. Cai, X., et al.: Gaze estimation with an ensemble of four architectures. arXiv preprint [arXiv:2107.01980](https://arxiv.org/abs/2107.01980) (2021)
6. Che, H., et al.: EFG-Net: a unified framework for estimating eye gaze and face gaze simultaneously. In: Yu, S., et al. (eds.) PRCV 2022. LNCS, vol. 13534, pp. 552–565. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-18907-4_43
7. Chen, Z., Shi, B.E.: Appearance-based gaze estimation using dilated-convolutions. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11366, pp. 309–324. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20876-9_20
8. Cheng, Y., Huang, S., Wang, F., Qian, C., Lu, F.: A coarse-to-fine adaptive network for appearance-based gaze estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10623–10630 (2020)
9. Cheng, Y., Lu, F.: Gaze estimation using transformer. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 3341–3347. IEEE (2022)
10. Cheng, Y., Lu, F.: DVGaze: dual-view gaze estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20632–20641 (2023)
11. Cheng, Y., Wang, H., Bao, Y., Lu, F.: Appearance-based gaze estimation with deep learning: a review and benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 7509–7528 (2024)
12. Cheng, Y., et al.: What do you see in vehicle? Comprehensive vision solution for in-vehicle gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1556–1565 (2024)
13. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: RepVGG: making VGG-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13733–13742 (2021)
14. Fischer, T., Chang, H.J., Demiris, Y.: RT-GENE: real-time eye gaze estimation in natural environments. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 339–357. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_21
15. Funes Mora, K.A., Monay, F., Odobez, J.M.: 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)
16. Gao, J., Geng, X., Zhang, Y., Wang, R., Shao, K.: Augmented weighted bidirectional feature pyramid network for marine object detection. *Expert Syst. Appl.* **237**, 121688 (2024)

17. Gideon, J., Su, S., Stent, S.: Unsupervised multi-view gaze representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001–5009 (2022)
18. Hempel, T., Abdelrahman, A.A., Al-Hamadi, A.: 6D rotation representation for unconstrained head pose estimation. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 2496–2500. IEEE (2022)
19. Her, P., Manderle, L., Dias, P.A., Medeiros, H., Odone, F.: Uncertainty-aware gaze tracking for assisted living environments. *IEEE Trans. Image Process.* **32**, 2335–2347 (2023)
20. Hisadome, Y., Wu, T., Qin, J., Sugano, Y.: Rotation-constrained cross-view feature fusion for multi-view appearance-based gaze estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5985–5994 (2024)
21. Hsieh, Y.H., Granlund, M., Odom, S.L., Hwang, A.W., Hemmingsson, H.: Increasing participation in computer activities using eye-gaze assistive technology for children with complex needs. *Disabil. Rehabil. Assist. Technol.* **19**(2), 492–505 (2024)
22. Huang, S., Lu, Z., Cheng, R., He, C.: FAPN: feature-aligned pyramid network for dense image prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 864–873 (2021)
23. Jha, S., Busso, C.: Estimation of driver’s gaze region from head position and orientation using probabilistic confidence regions. *IEEE Trans. Intell. Veh.* **8**(1), 59–72 (2022)
24. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6912–6921 (2019)
25. Kim, T., Kim, K., Lee, J., Cha, D., Lee, J., Kim, D.: Revisiting image pyramid structure for high resolution salient object detection. In: Proceedings of the Asian Conference on Computer Vision, pp. 108–124 (2022)
26. Lee, H.S., Weidner, F., Sidenmark, L., Gellersen, H.: Snap, pursuit and gain: virtual reality viewport control by gaze. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2024)
27. Li, Y., et al.: MViTv2: improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4804–4814 (2022)
28. Luo, X., et al.: Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Med. Image Anal.* **80**, 102517 (2022)
29. Nagpure, V., Okuma, K.: Searching efficient neural architecture with multi-resolution fusion transformer for appearance-based gaze estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 890–899 (2023)
30. Tolstikhin, I.O., et al.: MLP-mixer: an all-MLP architecture for vision. *Adv. Neural. Inf. Process. Syst.* **34**, 24261–24272 (2021)
31. Wang, Y., Yuan, G., Fu, X.: Driver’s head pose and gaze zone estimation based on multi-zone templates registration and multi-frame point cloud fusion. *Sensors* **22**(9), 3154 (2022)
32. Xiang, X., Yin, H., Qiao, Y., El Saddik, A.: Temporal adaptive feature pyramid network for action detection. *Comput. Vis. Image Underst.* **240**, 103945 (2024)
33. Yin, X., Yu, Z., Fei, Z., Lv, W., Gao, X.: PE-YOLO: pyramid enhancement network for dark object detection. In: Iliadis, L., Papaleonidas, A., Angelov, P., Jayne, C. (eds.) ICANN 2023. LNCS, vol. 14260, pp. 163–174. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-44195-0_14

34. Yun, J.S., Na, Y., Kim, H.H., Kim, H.I., Yoo, S.B.: HAZE-Net: high-frequency attentive super-resolved gaze estimation in low-resolution face images. In: Proceedings of the Asian Conference on Computer Vision, pp. 3361–3378 (2022)
35. Zhang, C., Chen, T., Nedungadi, R.R., Shaffer, E., Soltanaghai, E.: FocusFlow: leveraging focal depth for gaze interaction in virtual reality. In: Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp. 1–4 (2023)
36. 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)
37. Zhu, M.: Dynamic feature pyramid networks for object detection. In: Fifteenth International Conference on Signal Processing Systems (ICSPS 2023), vol. 13091, pp. 503–511. SPIE (2024)
38. Zhu, W., Deng, H.: Monocular free-head 3D gaze tracking with deep learning and geometry constraints. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3143–3152 (2017)



BCNet: Binocular Cooperative Network for Gaze Estimation

Dongchen Zhu^{1,2}, Minjin Lin¹, Hekuangyi Che^{1,2}, Wenjun Shi¹, Guanghui Zhang¹, Hang Li¹, Lei Wang^{1,2}, and Jiamao Li^{1,2}✉

¹ Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China

jmli@mail.sim.ac.cn

² University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Gaze estimation plays a crucial role in interactive applications. Recent advancements in deep learning have significantly enhanced appearance-based methods. However, existing approaches often focus on one eye and do not consider binocular gaze, overlooking a fundamental principle of human gaze: the convergence of gaze based on binocular cooperative information. To address this gap, we introduce BCNet, a network for binocular gaze estimation. Specifically, we develop the binocular-chiasm module to facilitate feature exchange between the two eyes and design a binocular-geometry loss that leverages gaze spatial geometry to improve convergence during fixation. Additionally, our person-specific analysis further reduces gaze estimation errors for individual users. Our method registers a 4.8% improvement on the MPIIGaze dataset over existing methods and achieves competitive results on the EyeDiap dataset. Experiments with noised data underscore the robustness of our proposed approach.

Keywords: Gaze estimation · Binocular Cooperation · Information Chiasm · Deep learning

1 Introduction

Eye gaze is an essential clue of human intention, purpose, and states of mind. Accurate gaze estimation shows potential applications in human-computer interaction, virtual reality, and mental health analysis. Much progress has been made recently in the task of appearance-based gaze estimation. In particular, the introduction of deep learning makes it possible to develop a practical gaze direction estimator (Fig. 1).

Many existing methods in gaze estimation focus solely on modeling or image processing, overlooking the fundamental principle of human eye movement. The human eye gaze is the result of eye movements. Deep learning networks can estimate the gaze based on the appearance of a single eye. However, the gaze is not independently determined by a single eye. There are the cooperative movements

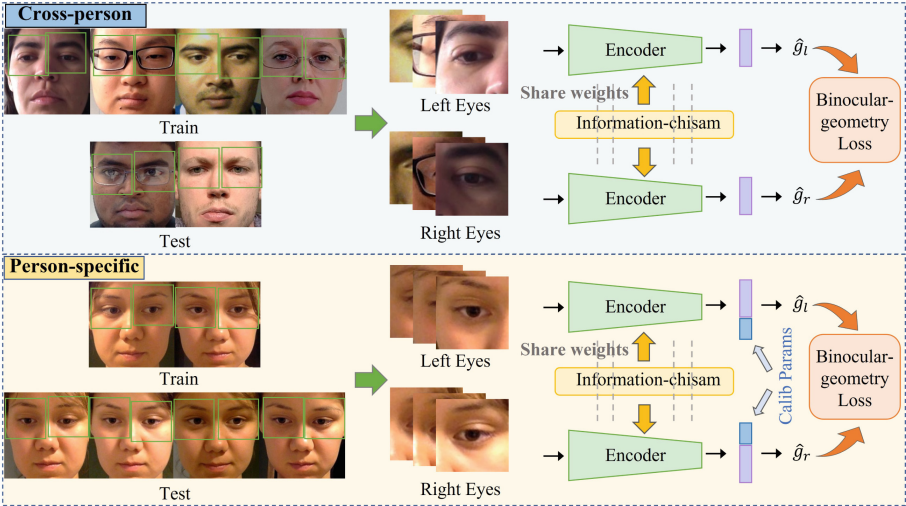


Fig. 1. Overview of proposed Binocular Cooperative Network for gaze estimation. The upper row shows a schematic diagram of cross-person in a single dataset, and the lower row shows a person-specific flow chart. The difference between them lies in the input data and whether calibration parameters are added.

of human eyes, which makes a person unable to gaze at different targets simultaneously. The origin of the cooperative movements is that binocular retinal information in the neural pathways is combined at the *Optic Chiasma* and then transferred to the visual cortex and the superior colliculus for visual processing and motor control. This aspect of human visual perception is often ignored in the design of gaze estimation models, which can lead to suboptimal performance.

In this paper, we designed a binocular gaze estimation network based on the principle of binocular cooperative movements. On the one hand, we proposed a feature communication module, called the binocular-chiasm module, that integrates the appearance information of human eyes, simulating the information crossover during gaze generation. Especially when the imaging quality of the two eyes is inconsistent, the module can show its advantages. On the other hand, We fleshed out the intuition that eye gazes always converge on a single point during fixation, then developed a novel loss function, named binocular-geometry loss, to take advantage of the spatial geometry of eye gazes.

Furthermore, we conducted a person-specific study by introducing learnable calibration parameters for each person to address the challenge that variations in individual appearances and the existence of kappa angles prevent a general gaze estimator from performing optimally on all subjects. Specifically, we assigned a calibration parameter to each eye and trained it with a small number of samples from the same person. This way, a general gaze estimator is transformed into a person-specific gaze estimator, resulting in a significant drop in gaze estimation error.

In summary, this paper makes the following contributions:

- A novel framework for binocular gaze estimation is presented, which is termed BCNet.
- A feature communication module, called the binocular-chiasm module, is designed to enhance the estimation of two eye gazes by propagating gaze features between them, leading to a win-win situation.
- A novel loss, the binocular-geometry loss, is proposed to take advantage of the spatial geometry of eye gazes to enhance convergence during fixation.
- A person-specific study is conducted to improve the gaze estimation accuracy, with learnable calibration parameters assigned to each eye.

2 Related Work

2.1 Appearance-Based Gaze Estimation

Appearance-based gaze estimation aims to learn a mapping from eye images or face images to gaze directions. Thanks to the rapid development of deep learning, much progress has been made to appearance-based gaze estimation. Zhang et al. [21] introduce CNNs to gaze estimation for the first time. They design a LeNet [10]-based network to estimate gaze from eye images. Later, Krafcik et al. [9] implement a gaze tracker by utilizing eye images, face image, and face grid together. Zhang et al. [22] take the full face image as input and employ spatial weights mechanism to emphasize features extracted from gaze related regions, like eye region. Che et al. [2] are the first to combine eye gaze estimation and face gaze estimation tasks, proposing a universal framework that simultaneously optimizes both tasks. Ghosh et al. [7] introduce a multi-task learning framework that improves accuracy through simultaneous training on tasks like eye state classification and region segmentation, adeptly managing limited supervision with both labeled and unlabeled data. Bao et al. [1] propose a multi-view dual encoder (MV-DE) framework that learns gaze representations from face images taken from multiple views. This method uses a dual encoder architecture to separate gaze information from general face information in images from different viewpoints, ensuring that the learned gaze representations are consistent across various angles.

Despite these advancements, such methods often overlook the physiological properties of human eyes. Lian et al. [11] explore physiological aspects by developing a coplanar loss function. Cheng et al. [5] utilize the asymmetry between the left and right eyes, applying different weights to the loss calculations for each eye, optimizing the network learning process based on the better-performing eye. They propose the E-Net to evaluate the reliability of each eye’s gaze estimation, balancing the learning between asymmetric and symmetric mechanisms. Mahmud et al. [14] further the field with a neural pipeline that merges anatomical eye region isolation with multistream gaze estimation, employing synthetic to real transfer learning for increased robustness.

However, existing methods either calculate the gaze for just one eye or the entire face, or while they do consider the gaze of both eyes, they only focus on the appearance differences between the left and right eyes, neglecting the deeper gaze dependencies such as the geometric constraints of eye movements. This results in an insufficient exploration of binocular gaze.

2.2 Calibration for Gaze Estimation

Calibration gaze estimation enhances individual performance by utilizing a few calibration samples, typically fewer than nine. Liu et al. [13] introduced a differential network to gauge gaze angle differences using pairs of images, while Zhang et al. [19] refined person-specific accuracy using polynomial functions. Linden et al. [12] advanced personalization in gaze tracking by normalizing images and projecting 2D estimates into 3D, aided by individual-specific neural networks. Chen et al. [3] further improved accuracy with minimal data through Multiple and Single Gaze Target Calibrations, enhancing robustness and reducing errors via MAP estimation. Meanwhile, Jin et al. [8] combined ocular counter-rolling with real and synthetic data for kappa angle regression, utilizing a multi-branch CNN for greater precision, although this complexity necessitates network personalization.

Nevertheless, existing gaze calibration methods are relatively complex, often requiring the design of a dedicated network or module and separate training for this component, which significantly increases the complexity and cost of gaze calibration.

3 Method

The detailed architecture of BCNet is illustrated in Fig. 2. Our BCNet has two inputs, which are the right eye patch and the left eye patch from the same person, and outputs the corresponding gaze angle of each eye. The overview framework of BCNet will be elaborated in Sect. 3.1, which is followed by binocular-geometry loss and binocular-chiasm module in Sect. 3.2 and Sect. 3.3. Lastly, we will introduce the details of the person-specific study in Sect. 3.4.

3.1 Architecture Overview

Our framework, shown in Fig. 2, employs a ResNet-based architecture to extract gaze features from eye patches, with shared weights across two eye pathways and a binocular-chiasm module for feature propagation. This setup enhances the prediction of accurate gaze angles—both yaw and pitch—converted into unit vectors. We developed a binocular-geometry loss function to optimize network convergence and conducted a person-specific study to refine gaze estimation performance further.

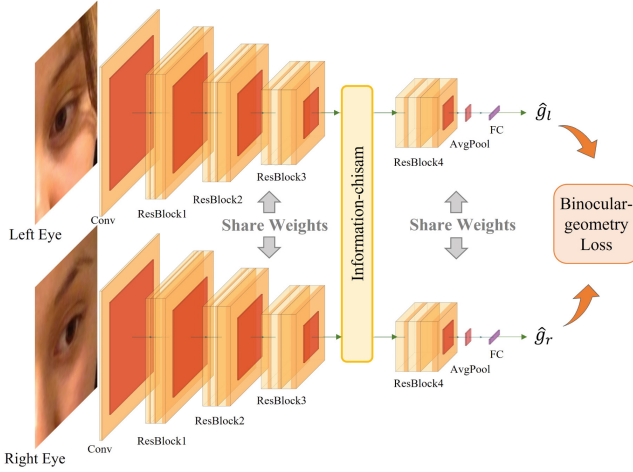


Fig. 2. Detailed structure of Binocular Cooperative Network. The orange cubes represent the feature maps of different layers. Note that the right eye patch is flipped horizontally before inputting into the network. (Color figure online)

3.2 Binocular-Chiasm Module

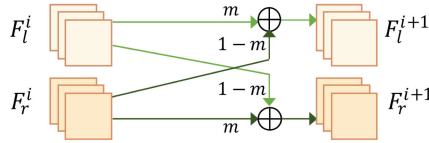


Fig. 3. Binocular-chiasm module. F_r^i and F_l^i represents the i -th layer feature map of right eye and left eye. m is a hyperparameter with a value between 0 and 1.

We devised an information communicating module named the binocular-chiasm module to help propagate gaze features between two eyes. We are inspired by the principle of the human eye in doing so. In the neural pathways of the human binocular system, retinal information from both eyes is combined at the optic chiasma and then transferred to the visual cortex and the superior colliculus for visual processing and motor control. We speculated that the information exchange and fusion can reflect each eye. Moreover, one eye will always be clearer and more accurate, while the other eye is relatively blurred and inaccurate, considering the difference in illumination and viewing angles. Thus, communicating the information of the left and right eyes helps to learn more robust features.

As described in Fig. 3, binocular-chiasm module is composed of two branches. Each branch takes i -th layer feature map of one eye and outputs the corresponding $i + 1$ -th layer feature map. Details are shown in Eq. 1, the right eye feature and left eye feature of the $i - th$ layer are multiplied by m and $1 - m$ respectively, and then added to obtain the right eye feature of the $i + 1th$ layer. m is a pre-defined ratio coefficient ranging from 0 to 1. It is empirically set as 0.3 in the following experiments.

$$F_r^{i+1} \leftarrow (1 - m) * F_l^i + m * F_r^i \quad (1)$$

Correspondingly, the left eye feature map is also added to the right eye branch with the same proportion m , thus completing the binocular-chiasm operation at the layer i .

3.3 Binocular-Geometry Loss

The design inspiration of binocular-geometry loss comes from the binocular cooperative movements of human eyes introduced in Sect. 1. As shown in Fig. 4, P_l and P_r are pupil centers of left eye and right eye respectively. t denotes the target point that human eyes focus on. cam is the camera and xz represent the unit direction vector of the camera coordination system. It can be found that x -axis is roughly parallel to $\overline{P_r P_l}$ and z -axis is roughly perpendicular to the eyes. d_l and d_r are distances from left eye and right eye to the target on the z -axis. And d_t is the distance from target to camera correspondingly.

The mutual conversion between the predicted unit direction vector of eye gaze $\hat{\mathbf{n}} = [\hat{x}, \hat{y}, \hat{z}]^T$ and the predicted angle angle $\hat{\mathbf{g}} = [\hat{\theta}, \hat{\phi}]^T$ can be obtained by function $G(\cdot)$ in Eq. 2 and Eq. 3.

$$\hat{\mathbf{n}} = G(\hat{\mathbf{g}}) = \begin{cases} \hat{x}_n = -\cos(\hat{\phi})\sin(\hat{\theta}) \\ \hat{y}_n = \sin(\hat{\phi}) \\ \hat{z}_n = -\cos(\hat{\theta})\cos(\hat{\phi}) \end{cases} \quad (2)$$

$$\hat{\mathbf{g}} = G^{-1}(\hat{\mathbf{n}}) = \begin{cases} \hat{\theta} = \arctan\left(\frac{\hat{x}_n}{\hat{z}_n}\right) \\ \hat{\phi} = \arcsin(\hat{y}_n) \end{cases} \quad (3)$$

After getting the unit direction vector, we divide the xyz of the direction vector by $-z$ to acquire a new vector \hat{v}_{z-1} for ease of illustrating Eq. 5.

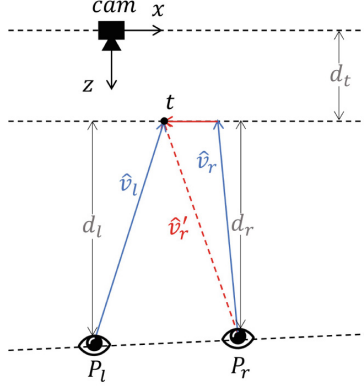


Fig. 4. Principle of binocular-geometry loss. \mathbf{P}_l and \mathbf{P}_r are pupil centers of left eye and right eye. $\hat{\mathbf{v}}_l$ and $\hat{\mathbf{v}}_r$ are the estimated gaze vectors. $\hat{\mathbf{v}}'_r$ is computed by adding $\overrightarrow{\mathbf{P}_r\mathbf{P}_l}$ with $\hat{\mathbf{v}}_l$. Binocular-geometry loss aims to minimize the difference between the gaze angles of $\hat{\mathbf{v}}'_r$ and $\hat{\mathbf{v}}_r$.

$$\hat{\mathbf{v}}_{z-1} = \begin{cases} \hat{x}_{z-1} = -\tan(\hat{\theta}), \\ \hat{y}_{z-1} = -\frac{\tan(\hat{\phi})}{\cos(\hat{\theta})} \\ \hat{z}_{z-1} = -1 \end{cases} \quad (4)$$

When a person's left eye is gazing at a specific target, his right eye will move in coordination with the left one according to the principle of binocular cooperative movements. There should be a converted gaze vector $\hat{\mathbf{v}}'_r$ for right eye by adding $\overrightarrow{\mathbf{P}_r\mathbf{P}_l}$ with $\hat{\mathbf{v}}_l$. The detailed calculation process is shown as follows.

$$\begin{aligned} \hat{\mathbf{v}}'_r &= \overrightarrow{\mathbf{P}_r\mathbf{P}_l} + \hat{\mathbf{v}}_l \\ &= \begin{bmatrix} x_{\mathbf{P}_l} \\ y_{\mathbf{P}_l} \\ d_l + d_t \end{bmatrix} - \begin{bmatrix} x_{\mathbf{P}_r} \\ y_{\mathbf{P}_r} \\ d_r + d_t \end{bmatrix} + \begin{bmatrix} \hat{x}_l \\ \hat{y}_l \\ \hat{z}_l \end{bmatrix} = (d_l + d_t)\mathcal{K}^{-1} \begin{bmatrix} u_{\mathbf{P}_l} \\ v_{\mathbf{P}_l} \\ 1 \end{bmatrix} - (d_r + d_t)\mathcal{K}^{-1} \begin{bmatrix} u_{\mathbf{P}_r} \\ v_{\mathbf{P}_r} \\ 1 \end{bmatrix} + \begin{bmatrix} \hat{x}_l \\ \hat{y}_l \\ -d_l \end{bmatrix} \\ &= (d_l + d_t)\left(\mathcal{K}^{-1} \begin{bmatrix} u_{\mathbf{P}_l} \\ v_{\mathbf{P}_l} \\ 1 \end{bmatrix} - \frac{d_r + d_t}{d_l + d_t}\mathcal{K}^{-1} \begin{bmatrix} u_{\mathbf{P}_r} \\ v_{\mathbf{P}_r} \\ 1 \end{bmatrix} + \frac{d_l}{d_l + d_t} \begin{bmatrix} \frac{\hat{x}_l}{d_l} \\ \frac{\hat{y}_l}{d_l} \\ -1 \end{bmatrix}\right) \\ &\approx (d_l + d_t)\left(\mathcal{K}^{-1} \begin{bmatrix} u_{\mathbf{P}_l} \\ v_{\mathbf{P}_l} \\ 1 \end{bmatrix} - \mathcal{K}^{-1} \begin{bmatrix} u_{\mathbf{P}_r} \\ v_{\mathbf{P}_r} \\ 1 \end{bmatrix} + \begin{bmatrix} \frac{\hat{x}_l}{d_l} \\ \frac{\hat{y}_l}{d_l} \\ -1 \end{bmatrix}\right) \\ \text{s.t.} \quad &d_l \approx d_r \gg d_t \end{aligned} \quad (5)$$

In Eq. 5, $[u_{\mathbf{P}_l}, v_{\mathbf{P}_l}, 1]^T$ and $[u_{\mathbf{P}_r}, v_{\mathbf{P}_r}, 1]^T$ are the 2D-image homogeneous coordinates of \mathbf{P}_l and \mathbf{P}_r , and \mathcal{K} is the intrinsic matrix of the camera. All the above parameters can be obtained at data preprocessing procedure. For $[\frac{\hat{x}_l}{d_l}, \frac{\hat{y}_l}{d_l}, -1]^T$, it

can be calculated through Eq. 2 and Eq. 4 once the predicted left eye gaze angle is given. The only unknown parameters are d_l and d_t , but they have no effect on the calculation of the unit direction vector of $\hat{\mathbf{v}}'_r$.

Ideally, $\hat{\mathbf{v}}'_r$ and $\hat{\mathbf{v}}_r$ are expected to be equal. For convenience, we simply utilize the parallel relationship between them and devise a novel loss function named binocular-geometry loss in Eq. 6.

$$\begin{aligned} \mathcal{L}_{bino}^r &= \|\hat{\mathbf{g}}_r - \hat{\mathbf{g}}'_r\|^2 \\ &= \|\hat{\mathbf{g}}_r - \mathbf{G}^{-1}(norm(\hat{\mathbf{v}}'_r))\|^2 \end{aligned} \tag{6}$$

$norm(\cdot)$ means normalizing $\hat{\mathbf{v}}'_r$ to be an unit direction vector. $G^{-1}(\cdot)$ is the inverse of $G(\cdot)$ which transforms a gaze vector into gaze angles. For left eye, there exists a similar binocular-geometry loss shown in Eq. 7. In theory, whether using \mathcal{L}_{bino}^l or \mathcal{L}_{bino}^r will produce the same result because of the symmetry.

$$\begin{aligned} \mathcal{L}_{bino}^l &= \|\hat{\mathbf{g}}_l - \hat{\mathbf{g}}'_l\|^2 \\ &= \|\hat{\mathbf{g}}_l - \mathbf{G}^{-1}(norm(\hat{\mathbf{v}}'_l))\|^2 \end{aligned} \tag{7}$$

3.4 Person-Specific Gaze Estimation

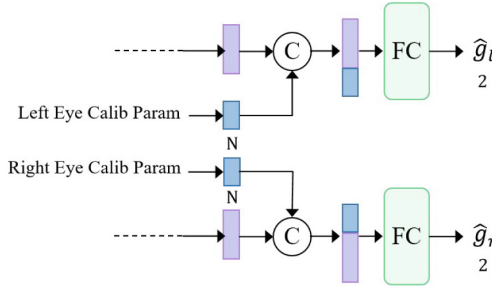


Fig. 5. Calibration module of person-specific gaze estimation. N stands for the dimension of the calibration parameter. The ultimate output is two-dimensional, namely the yaw angle and pitch angle of each eye.

Under the setting of person-independent, where training data and test data are collected from different people, the accuracy of mainstream gaze estimation methods hovers around $3 - 4^\circ$ (See Table 1 for detailed data), and it is difficult to get further improvement.

The main reason is that there is a certain gaze deviation between people. For two different people, even if the eyeballs are rotated at exactly the same angle, there will be a difference of $2 - 3^\circ$ in their gazes. This is due to the different kappa angles (angle between the optical axis and visual axis) in different people. The size of the kappa angle varies from person to person and is determined by

the internal parameters of the human eyeball, which cannot be learned from images directly.

To mitigate the impact of the kappa angle, we assign calibration parameters to each person, which are learnable parameters and differ in people. As shown in Fig. 5, calibration parameters of $1 \times N$ are added to the left and right branches of the original BCNet. We flatten the features after the AvgPool layer and then concatenate and fuse them with the calibration parameters. The fused features are input into the last layer of FC to predict final gaze angles. For simplicity, we refer to the above network as BCNet-specific.

The training procedure of BCNet-specific is split into two parts. First, numerous images of different people are fed into the network, and all parameters are updated through backpropagation. After that, a small number of calibration samples (≤ 9) from the same person are collected to fine-tune BCNet-specific and update the calibration parameters only. In this way, we can encode kappa angle information into calibration parameters, which significantly improves the performance of certain person.

3.5 Gaze Loss

To stabilize the training procedure, we use the gaze label to establish an L2 loss function for the each eye. \mathbf{g} represents the gaze label and $\hat{\mathbf{g}}$ is the predicted gaze.

$$\mathcal{L}_{gaze}^l = \|\mathbf{g}_l - \hat{\mathbf{g}}_l\|^2, \quad \mathcal{L}_{gaze}^r = \|\mathbf{g}_r - \hat{\mathbf{g}}_r\|^2 \quad (8)$$

3.6 Total Loss Function

The total loss is formulated as Eq. 9. In theory, whether using \mathcal{L}_{bino}^l or \mathcal{L}_{bino}^r leads to the same result because of the symmetry. For simplicity, we choose \mathcal{L}_{bino}^l to build the total loss. In Eq. 9, λ_1 and λ_2 are the loss weights to control the balance between losses. We empirically set $\lambda_1 = 0.2$ and $\lambda_2 = 0.8$.

$$\mathcal{L}_{total} = \lambda_1 * \frac{\mathcal{L}_{gaze}^l + \mathcal{L}_{gaze}^r}{2} + \lambda_2 * \mathcal{L}_{bino}^l \quad (9)$$

4 Experiments

4.1 Datasets

To evaluate the effectiveness of our framework, we conduct experiments on two popular datasets: MPIIGaze [23] and EyeDiap [6].

MPIIGaze is a commonly used dataset for appearance-based gaze estimation, which provides 213,659 images from 15 participants in everyday settings with unconstrained head pose and normal illumination. We select 3000 images of human eyes from each person and perform 15-fold cross-validation, the same as Zhang et al. [21].

EyeDiap contains 94 video sessions of 16 participants. We perform the same protocol as Zhang et al. [22]. We chose the continuous screen target session, which has valid annotations and a reasonable range of gazes. After removing the invalid frames containing blinking, these videos are sampled every 15 frames to construct the training and evaluation set. The amount of valid data is around 6,000. Considering the 12th and 13th participants are not collected under continuous screen target conditions, we get 14 valid participants and randomly divide them into five groups.

To verify the robustness of our model, we added noise to the original MPIIGaze and EyeDiap datasets. Specifically, we applied Gaussian blurring, added Gaussian noise, randomly added or subtracted single pixel values, adjusted brightness, and kept the actual value of the gaze unchanged for each image, resulting in the MPIIGaze-Noised and EyeDiap-Noised datasets.

4.2 Data Pre-processing

We follow the data normalization process in [18]. The eye patches are cropped by taking the eye center as the patch center and doubling the distance between the inner and outer eye corners as the side length. The RGB eye patches are histogram-equalized to eliminate the influence of illumination and resized to $224 \times 224 \times 3$. We use OpenCV library functions for noise addition and pre-processing of images in a Python environment.

4.3 Implementation Details

The network predicts the yaw and pitch angles of the gaze, which are then converted into unit vectors. The angle between the predicted and actual values is calculated and used as the gaze estimation error. Since the single eye is not symmetrical, we flip the right eye image horizontally and keep the left eye image unchanged to ensure the consistency of the network input. To make the network converge faster, a two-step training procedure is proposed. First, we remove the binocular-chiasm module in BCNet and train it with the left and flipped right eye images. Then, we load the pre-training weights of the first step and retrain the complete BCNet. The parameters are the same in the two steps. To be specific, the batch size is 16, and the learning rate is 0.0001. The optimizer is Adam optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

4.4 Cross-Person Performance

We conduct experiments on MPIIGaze and EyeDiap to compare the performance of the proposed method with other appearance-based methods. As shown in Table 1, our method achieves the best performance on the MPIIGaze dataset and the second-best result on the EyeDiap dataset. Despite the variety in data volume and data form of these datasets, BCNet demonstrates a very stable and excellent effect. For the MPIIGaze dataset, our approach gets the best result of

3.9° , which has a considerable improvement of 0.6° (about 13.3%) to baseline. For the EyeDiap dataset, which has the smaller amount and poor resolution (640×480), our method outperforms the baseline by 1.2° (about 17.4%).

Table 1. Comparison of BCNet with current state-of-the-art methods on cross-person evaluations. The values in the table represent the angle between the predicted and actual gaze directions; smaller values indicate better performance. Bold indicates the best result, and underline indicates the second-best result.

Methods	Years	MPIIGaze	EyeDiap
DPG [16]	ECCV2018	4.6°	10.3°
Bayesian [17]	CVPR2019	4.3°	9.9°
RSN [20]	BMVC2020	4.5°	6.6°
FAR-Net [5]	TIP2020	4.4°	5.9°
CA-Net [4]	AAAI2020	<u>4.1°</u>	5.3°
MTGLS [7]	WACV2022	<u>4.1°</u>	N/A
MSGazeNet [14]	TAI2024	4.6°	5.9°
Baseline (Resnet50)		4.5°	6.9°
Ours		3.9°	<u>5.7°</u>

As shown in Table 2, even though our method’s performance may decline slightly on the noisy data, it is still significantly better than the baseline approach.

Table 2. Comparison of the algorithm’s performance on the MPIIGaze-Noise and EyeDiap-Noise datasets (Cross-person). The values in the table represent the angle between the predicted and actual gaze directions; smaller values indicate better performance.

Algorithms	MPIIGaze-Noise	EyeDiap-Noise
Baseline(Resnet50)	5.7°	7.0°
Ours	5.1°	5.9°

4.5 Person-Specific Performance

To make a further reduction in gaze estimation error, we conduct a person-specific study on the MPIIGaze dataset and compare BCNet with other person-specific methods. The results are depicted in Table 3. BCNet consistently outperforms other methods regardless of the number of calibration samples.

To show the effect of person-specific gaze estimation in more detail, we gradually increase the number of calibration samples from 0 to 256. We conducted

Table 3. Comparison of BCNet-specific with current state-of-the-art methods on person-specific evaluations. The person-specific study is conducted on MPIIGaze dataset. The values in the table represent the angle between the predicted and actual gaze directions; smaller values indicate better performance.

Methods	Years	Samples(k)	Their	Our
Diff-VGG [13]	TPAMI2019	9	3.80°	2.77° ± 0.22°
FAZE [15]	ICCV2019	1	3.91°	3.67° ± 1.61°
		5	3.24°	2.86° ± 0.33°
		9	3.14°	2.77° ± 0.22°
		256	3.00°	2.63° ± 0.06°
SPAZE [12]	ICCVW2019	1	4.12°	3.67° ± 1.61°
		5	3.16°	2.86° ± 0.33°
		9	2.94°	2.77° ± 0.22°
		20	2.82°	2.68° ± 0.14°
GEDDNet [3]	TPAMI2022	1	3.5°	3.67° ± 1.61°
		5	3.0°	2.86° ± 0.33°
		9	3.0°	2.77° ± 0.22°
KAComp [8]	CVPRW2023	9	3.65°	2.77° ± 0.22°

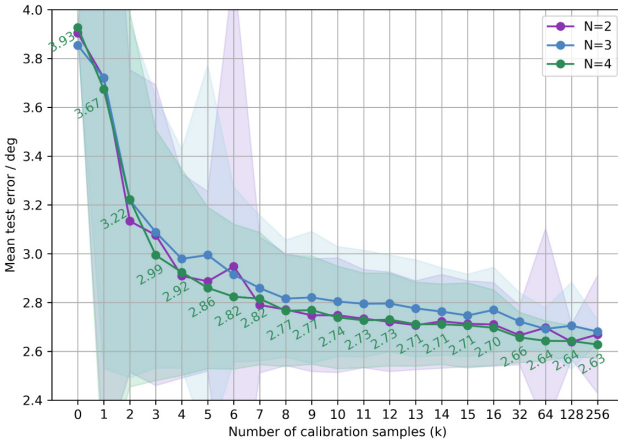


Fig. 6. Person-specific gaze estimation on MPIIGaze dataset. N stands for the dimension of the calibration parameter for each eye. The x-axis is the calibration samples for each person, the y-axis is the average gaze estimation error, and the light-colored area represents the standard deviation for each error.

detailed experiments on MPIIGaze and set the calibration parameters(N) as 2, 3, and 4. The specific results are shown in Fig. 6. It is obvious that the calibration result of N = 4 is consistently better than that of N = 3 except for a few

cases. As for $N = 2$, although it outperforms $N = 4$ when $k = 2, 4, 7, 9, 12$, it has a larger standard deviation, which indicates that its results are not stable enough. After weighing the pros and cons, we choose $N = 4$ for the final result, and the corresponding error for different calibration samples is also depicted in Fig. 6.

4.6 Ablation Study

In order to demonstrate the effectiveness of the binocular-chiasm module and the binocular-geometry loss, we conducted an ablation study on MPIIGaze and EyeDiap datasets. The results are depicted in Table 4. Note that the ablation study is a cross-person test. Compared with binocular-geometry loss, it is clear to find that when applying binocular-chiasm module only, the gaze error drops more (0.5° on MPIIGaze, 0.7° on EyeDiap). This is reasonable since operations on feature maps are more straightforward than operations on loss functions.

4.7 Qualitative Results

Some qualitative examples of BCNet are shown in Fig. 7 and Fig. 8. The baseline demonstrates a significant discrepancy with ground truth, while BCNet is closer to ground truth. Although the results of the baseline may be close to the ground truth in one eye, when changing to another eye, the results are much worse, while our method can always maintain a good performance in both eyes. This verifies the effectiveness of the binocular-chiasm module and the binocular-geometry loss.

Table 4. Ablation study for the binocular-geometry loss and the binocular-chiasm module on MPIIGaze and EyeDiap (Cross-person). The values in the table represent the angle between the predicted and actual gaze directions; smaller values indicate better performance.

Binocular-chiasm	Binocular-geometry	MPIIGaze	EyeDiap
		4.5°	6.9°
✓		4.0°	6.2°
	✓	4.4°	6.4°
✓	✓	3.9°	5.7°

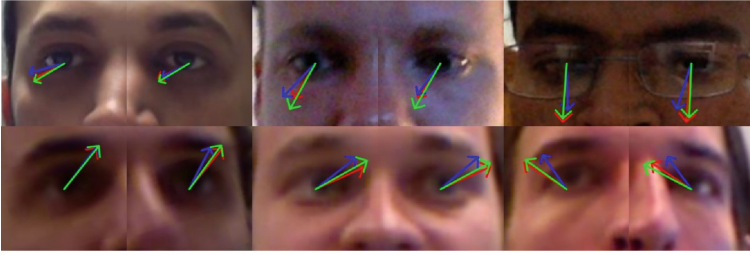


Fig. 7. Qualitative results on MPIIGaze (top), and EyeDiap (bottom) datasets. The green, blue, and red arrows are BCNet, baseline, and ground truth, respectively. (Color figure online)

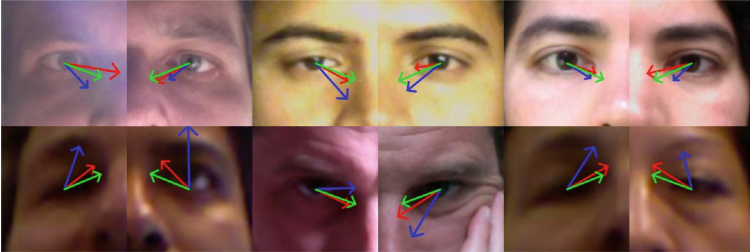


Fig. 8. Qualitative results on MPIIGaze-Noised (top), and EyeDiap-Noised (bottom) datasets. The green, blue, and red arrows are BCNet, baseline, and ground truth, respectively. (Color figure online)

5 Conclusion

In this paper, we introduce the binocular-chiasm module to facilitate feature exchange between the eyes, enhancing simultaneous two-eye gaze estimation. We also develop a binocular-geometry loss function leveraging spatial gaze geometry and conducted a person-specific study to tailor models to individual users, significantly enhancing gaze estimation accuracy. Our approach achieves state-of-the-art performance on MPIIGaze and competitive results on EyeDiap, suggesting potential for future advancements and personalized gaze estimation models.

Acknowledgements. This work was supported by National Science and Technology Major Project from Minister of Science and Technology, China (2021ZD0201403), Natural Science Foundation of Shanghai (23ZR1474200), Youth Innovation Promotion Association, Chinese Academy of Sciences (2021233, 2023242), Shanghai Academic Research Leader (22XD1424500).

References

1. Bao, Y., Lu, F.: Unsupervised gaze representation learning from multi-view face images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1419–1428 (2024)
2. Che, H., et al.: EFG-Net: a unified framework for estimating eye gaze and face gaze simultaneously. In: Yu, S., et al. (eds.) PRCV 2022. LNCS, vol. 13534, pp. 552–565. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-18907-4_43
3. Chen, Z., Shi, B.E.: Towards high performance low complexity calibration in appearance based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(1), 1174–1188 (2022)
4. Cheng, Y., Huang, S., Wang, F., Qian, C., Lu, F.: A coarse-to-fine adaptive network for appearance-based gaze estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10623–10630 (2020)
5. Cheng, Y., Zhang, X., Lu, F., Sato, Y.: Gaze estimation by exploring two-eye asymmetry. *IEEE Trans. Image Process.* **29**, 5259–5272 (2020)
6. Funes Mora, K.A., Monay, F., Odobez, J.M.: 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)
7. Ghosh, S., Hayat, M., Dhall, A., Knibbe, J.: MTGLS: multi-task gaze estimation with limited supervision. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3223–3234 (2022)
8. Jin, S., Dai, J., Nguyen, T.: Kappa angle regression with ocular counter-rolling awareness for gaze estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2659–2668 (2023)
9. Krafka, K., et al.: Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2176–2184 (2016)
10. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
11. Lian, D., et al.: Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(10), 3010–3023 (2018)
12. Lindén, E., Sjostrand, J., Proutiere, A.: Learning to personalize in appearance-based gaze tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
13. Liu, G., Yu, Y., Mora, K.A.F., Odobez, J.M.: A differential approach for gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(3), 1092–1099 (2019)
14. Mahmud, Z., Hungler, P., Etemad, A.: Multistream gaze estimation with anatomical eye region isolation by synthetic to real transfer learning. *IEEE Trans. Artif. Intell.* **5**, 4232–4246 (2024)
15. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9368–9377 (2019)
16. Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 741–757. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_44
17. Wang, K., Zhao, R., Su, H., Ji, Q.: Generalizing eye tracking with Bayesian adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11907–11916 (2019)

18. 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)
19. Zhang, X., Sugano, Y., Bulling, A.: Evaluation of appearance-based methods and implications for gaze-based applications. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2019)
20. Zhang, X., Sugano, Y., Bulling, A., Hilliges, O.: Learning-based region selection for end-to-end gaze estimation. In: BMVC (2020)
21. 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)
22. 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)
23. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: MPIIGaze: real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(1), 162–175 (2017)



mmAlphabet: Air Writing Alphabet Recognition System Based on mmWave FMCW Radar and Convolutional Neural Network

Chao-Wang Huang¹(✉), Chien-Yao Wang², and Jia-Ching Wang¹

¹ Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan

huangcw913@gmail.com

² Multimedia Technologies Laboratory, Institute of Information Science, Academia Sinica, Taipei, Taiwan

kinyiu@iis.sinica.edu.tw

Abstract. An air writing alphabet recognition system based on the images of temporal spectrogram such as average range-time map, average Doppler-time map and average angle-time map derived from an mmWave FMCW radar is proposed. All 26 English lowercase letters written in midair right above the radar sensor can be recognized. Two valuable radar data sets are introduced in diverse environments such as meeting room, office cubicle and living room etc. The use case is the usage scenarios of laptop computers and mobile phones. In one data set, volunteers write freely in their own handwriting styles including different hand speeds and different stroke orders. In the other data set, volunteers are asked to write according to a prescribed sequence of strokes. Then, the gestures sensed by radar are processed into images of temporal spectrogram to represent the written letter. A convolutional neural network which achieves 98.6% test accuracy is exploited as the classifier to recognize the air written alphabet. In the Leave-One-Subject-Out (LOSO) cross validation, it achieves an average test accuracy of 87.74%. The effectiveness of the proposed alphabet recognition system is extensively verified on the two created data sets for different variants of temporal spectrograms. It can be used to implement natural, intelligent noncontact human machine interface.

Keywords: human-machine interface · gesture recognition · air writing · spectrogram · range-time map · Doppler-time map · angle time-map · mmWave · FMCW radar · convolutional neural network

1 Introduction

Gesture recognition is an emerging field of research in the last decade. Under the expectation of interacting with computing devices more naturally, it can form a smart multi-modal noncontact human computer interface (HCI) [1] together with speech recognition. In addition, the COVID-19 pandemic has also highlighted the help of noncontact human machine interface in epidemic prevention. Contactless gesture recognition can be implemented with a variety of devices such as video cameras, infrared devices, ultrasound

devices [2], WiFi devices, and radar sensors. Among them, camera-based systems suffer from poor light condition, line of sight requirement [3, 4] and privacy issues [2, 5, 6]. Besides, radar sensors can also sense through smoke, dust, and even nonmetallic materials [6]. Furthermore, the mmWave radar signal can capture motion changes down to millimeters. Therefore, from fine-grained control gestures for human computer interaction [6–8] to large-scale gestures for traffic scenarios [9, 10], all can be accurately recognized by radar-based gesture recognition systems. For example, some applications can be found in [11] for health care, some are proposed in [12–14] for driving assistance and others are presented in [15] for smart home. Not to mention that they consume less power and are less expensive than camera-based systems. Therefore, radar sensors are more attractive to be an always-on solution for contactless gesture recognition.

There are quite a few radar-based gesture recognition related works. Short-range radar, color camera, and time-of-flight (TOF) depth camera are combined in [16] to implement a multi-sensor system for driver's hand-gesture recognition. The RadarNet [6] developed at Google recognizes four directional swipes and an omni-swipe using a radar chip integrated into a mobile phone. Binary activated spiking neural network (SNN) is exploited in [17, 18] and [19] to binarize the radar-generated images such as range, Doppler or angle-time maps and perform hand gesture recognition in an energy-efficient way. μ -Doppler signature is proposed in [2] to train a convolutional neural network (CNN) to perform gesture classification. To derive more gesture characteristics, other radar generated features such as power, range, Doppler, azimuth, elevation, and some related statistical properties are proposed in [20–22], and [23] respectively.

Air-writing systems offer users a virtual board to write linguistic characters, numerals, words or even symbols in free space by hand gesture movements. The Microsoft Kinect sensor which consists of an RGB camera and a depth camera is proposed to recognize handwritten digits from 0 to 9 in midair for TV remote controller in [24]. In [25], a single 2-D web camera is proposed to acquire the air-writing trajectory of letters written in an imaginary box one at a time. The device-free WiFi sensing technology, called WiDG in [26], is exploited to recognize handwritten digits in the air based on CSI and deep learning model. In [27, 28] and [29], smart watch motion sensor, inertial measurement unit and RFID are proposed to recognize English letters in mid-air respectively. Radar sensor is proposed to recognize digits from air written gestures in [30]. In that work, three IR-UWB radar sensors placed in triangular geometry are used to acquire hand's midair trajectory for classification. However, in [31], only single UWB radar is required to recognize air-written numerals by using 3D range-Doppler tensors and 3D-CNN-LSTM network. In [32], an over-the-air handwritten digit recognition method based on an mmWave FMCW radar is proposed. The trajectory points of gesture actions are used to generate images for gesture recognition with the Xception deep learning network. Moreover, a meta-learning optimization-based approach that enables user-definable hand gesture recognition at the edge is proposed in [33]. The approach is useful not only for recognizing new movement types, but also for adapting to individuals with motor disabilities or visual impairment. In [34] and [35], an air-writing system based on a network of less than three mmWave FMCW radars to recognize capital letters A–J and numerals 1–5 is proposed. Only range information from each radar is required

to reconstruct the trajectory of a written character. Then the trajectory is transformed to a 2-D image and classified by a DCNN network.

In this work, an air writing alphabet recognition system based on mmWave FMCW radar and convolutional neural network is proposed. All 26 English lowercase letters written in midair right above the radar sensor can be recognized. Data of a written letter sensed by radar are processed into images of temporal spectrogram such as average range-time map, average Doppler-time map and average angle-time map to represent the letter. Deep convolutional neural network is exploited as the classifier to recognize air written alphabet. Instead of using range-time map, Doppler-time map and angle-time map as the gesture features [18, 33], average version of the three temporal spectrograms which are verified to better represent the gestures are adopted. Furthermore, the data of air written gestures performed by seven adult volunteers are recorded in three diverse environments such as meeting room, office cubicle and living room for the whole English alphabet. The use case is set in line with the actual usage scenarios of laptop computers and mobile phones. Two valuable radar data sets are introduced. In one data set, volunteers write freely in their own handwriting styles including different writing speeds and even different stroke orders. In the other data set, volunteers are asked to gesticulate according to a prescribed sequence of strokes.

The major contributions of this work are as follows:

- Two valuable English alphabet radar data sets are introduced. In one data set, volunteers write freely in their own handwriting styles. However, in the other data set, volunteers are asked to gesticulate according to a prescribed sequence of strokes so that the gestures are intentionally designed to reduce the intraclass variability and make them more recognizable for radar sensors.
- A novel data processing scheme is proposed. Average range-time map, average Doppler-time map and average angle-time map are proposed as the feature images to better represent the alphabet gestures.
- Extensive performance analyses are conducted to verify how the new feature images affect the system's overall capability. Performance comparison with other cutting edge related works is also presented.
- A CNN-based air-writing recognition system using an mmWave FMCW radar is exploited. It achieves real-time recognition with a test accuracy of 98.6%. In the Leave-One-Subject-Out (LOSO) cross validation, it achieves 87.74% average test accuracy.
- To the best of the authors knowledge, the proposed air writing recognition system based on one mmWave FMCW radar is the first time to use average range/Doppler/angle-time map to represent the gestures of all 26 English letters and it achieves state-of-the-art performance.

This paper is organized as follows: Sect. 2 introduces the system model of FMCW radar, derives the proposed average temporal spectrogram signatures and describes the architecture of the classifier model. Section 3 presents the experimental setup including the radar configuration, creation of data set, feature map selection and hyperparameters. Section 4 investigates and discusses the performance of the proposed alphabet recognition algorithm and conducts performance comparison with related studies. Finally, Sect. 5 concludes this paper.

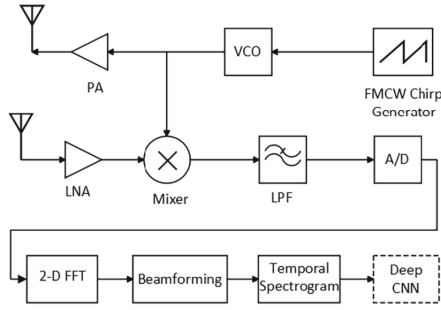


Fig. 1. FMCW radar system block diagram and digital signal processing chain.

2 Radar System and Deep CNN Model

2.1 FMCW Signal Model

The system block diagram and digital signal processing chain of FMCW radar is illustrated in Fig. 1. The radar transceiver transmits a train of chirp signal. Then, the received echo is mixed with the transmitted signal and then lowpass filtered to yield the intermediate frequency (IF) beat signal. The signal is then sampled with an A/D converter to yield the discrete beat signal [36, 37].

The IF beat signal of the FMCW radar is

$$s_{IF}(t) = \frac{\sigma}{2} \cos[2\pi \cdot f_b(t) \cdot t + \varphi_b(t)], \quad (1)$$

where σ is proportional to the radar cross section (RCS), antenna gain and range attenuation, $f_b(t) = (B/T_c) \cdot t_d - f_D$ is the IF beat frequency, B denotes the sweep bandwidth, T_c denotes chirp duration, $\varphi_b(t) = 2\pi f_c t_d$ is the phase term, t_d denotes the round-trip delay of the radar signal, f_c and f_D denote the carrier frequency and Doppler shift respectively.

Consider a MIMO radar system, the discrete IF beat signal of the l_{th} frame of the r_{th} receive antenna is [20]

$$\begin{aligned} s_{IF}[l, r, n, m] &= s_{IF,r}(t)|_{t=nT_s+mT_p+lT_f} \\ &= \frac{\sigma_r}{2} \cos \left[\begin{aligned} &2\pi f_{b,r}(nT_s + mT_p + lT_f)nT_s \\ &+ \varphi_{b,r}(nT_s + mT_p + lT_f) \end{aligned} \right], \end{aligned} \quad (2)$$

where $T_s = T_c/N_R$ denotes the sample period, N_R denotes the number of Range-FFT, $T_p = T_f/M$ denotes the pulse repetition time, T_f denotes the frame duration, M denotes the number of chirps in one frame, n and m denote the sample index and chirp index respectively.

The discrete IF beat signal, $s_{IF}[l, r, n, m]$, is a 4-D array containing the target information scanned by the radar sensor. The range spectrum of the radar can be derived by applying N_R -point FFT:

$$S_{IF}[l, r, k, m] = \sum_{n=0}^{N_R-1} s_{IF}[l, r, n, m] \cdot e^{-j \frac{2\pi nk}{N_R}}, \quad (3)$$

where $k = 0, \dots, N_R - 1$ denotes the index of range bins. Then, the static clutter is removed from the range spectrum with a first order IIR filter.

2.2 Average Range-Time Map

To recognize handwritten alphabet in midair, discriminative features need to be extracted from the received signal. After the range profile is derived by range FFT in (3), Two variants of the accumulated range spectrum of the l_{th} frame can be obtained. In (4), the power of range spectrum is accumulated (averaged).

Accumulate Power

$$S_R[k, l] = \sum_{m=0}^{M-1} \sum_{r=0}^{R-1} |S_{IF}[l, r, k, m]|^2 \quad (4)$$

In (5), the magnitude of range spectrum is accumulated (averaged).

Accumulate Magnitude

$$S_R[k, l] = \sum_{m=0}^{M-1} \sum_{r=0}^{R-1} |S_{IF}[l, r, k, m]|, \quad (5)$$

where R denotes the number of receive antenna and $k = 0, \dots, N_R/2 - 1$. Since the length of a sliding observation window is L frames, the average range-time map, which has shape $N_R/2 \times L$, at the l_{th} frame is

$$\mathbf{S}_{RTM}[l]_{N_R/2 \times L} = \begin{bmatrix} S_R[N_R/2 - 1, l - L + 1] \cdots S_R[N_R/2 - 1, l] \\ \vdots & \ddots & \vdots \\ S_R[0, l - L + 1] \cdots S_R[0, l] \end{bmatrix}. \quad (6)$$

Since the frame time is set to 50 ms in this work, instead of the coherent accumulation in [18], non-coherent accumulation is applied in (4) and (5) because the radial speed of targets is probably to induce range walk within one frame time.

2.3 Average Doppler-Time Map

After the range spectrum is derived in (3), N_D -point Doppler FFT is applied in (7) to obtain range-Doppler map

$$RDM[l, r, k, p] = \sum_{m=0}^{M-1} S_{IF}[l, r, k, m] \cdot e^{-j \frac{2\pi mp}{N_D}}, \quad (7)$$

where $p = 0, \dots, N_D - 1$ denotes the index of Doppler bins.

As in the previous derivation, two variants of the accumulated Doppler spectrum of the l_{th} frame can be obtained. In (8), the power of Doppler spectrum is accumulated (averaged).

Accumulate Power

$$S_D[p, l] = \sum_{k=0}^{N_R-1} \sum_{r=0}^{R-1} |RDM[l, r, k, p]|^2 \quad (8)$$

In (9), the magnitude of Doppler spectrum is accumulated (averaged).

Accumulate Magnitude

$$S_D[p, l] = \sum_{k=0}^{N_R-1} \sum_{r=0}^{R-1} |RDM[l, r, k, p]| \quad (9)$$

Then, the average Doppler-time map, which has shape $N_D \times L$, at the l_{th} frame is

$$\mathbf{S}_{DTM}[l]_{N_D \times L} = \begin{bmatrix} S_D[N_D - 1, l - L + 1] & \cdots & S_D[N_D - 1, l] \\ \vdots & \ddots & \vdots \\ S_D[0, l - L + 1] & \cdots & S_D[0, l] \end{bmatrix}. \quad (10)$$

2.4 Average Angle-Time Map

The Angle of Arrival (AoA) information is also a key feature in gesture recognition. In this work, Capon beamformer [38] with diagonal loading is exploited in the derivation of range-angle map. The derived range-angle map of the l_{th} frame is

$$RAM[l, k, i], \quad (11)$$

where $i = 0, \dots, N_B - 1$ denotes the index of angle beams, N_B denotes the number of total beams and $k = 0, \dots, N_R/2 - 1$ denotes the index of range bins.

Again, two variants of the accumulated angle spectrum of the l_{th} frame can be obtained. In (12), the power of angle spectrum is accumulated (averaged).

Accumulate Power

$$S_A[i, l] = \sum_{k=0}^{N_R-1} |RAM[l, k, i]|^2 \quad (12)$$

In (13), the magnitude of angle spectrum is accumulated (averaged).

Accumulate Magnitude

$$S_A[i, l] = \sum_{k=0}^{N_R-1} |RAM[l, k, i]| \quad (13)$$

Then, the average angle-time map, which has shape $N_B \times L$, at the l_{th} frame is.

$$\mathbf{S}_{ATM}[l]_{N_B \times L} = \begin{bmatrix} S_A[N_B - 1, l - L + 1] & \cdots & S_A[N_B - 1, l] \\ \vdots & \ddots & \vdots \\ S_A[0, l - L + 1] & \cdots & S_A[0, l] \end{bmatrix}. \quad (14)$$

2.5 Deep CNN Model

As instanced in Fig. 2, the feature subnet of the CNN network comprises three 2-D convolutional layers with 32, 64 and 128 filters whose size is 3×3 respectively. Followed by batch normalization and max-pooling layer, each convolutional layer is ended by ReLU activation function. Then, the extracted feature map of the last max-pooling layer is flattened and fed to the decision subnet which comprise two fully connected layers with 128 and 64 neurons respectively. The two dense layers both followed by batch normalization layer and ReLU activation function. The second dense layer also followed by a dropout layer with rate 0.5 to avoid the over fitting problem. At last, a softmax function with 26 neurons is served as the output layer.

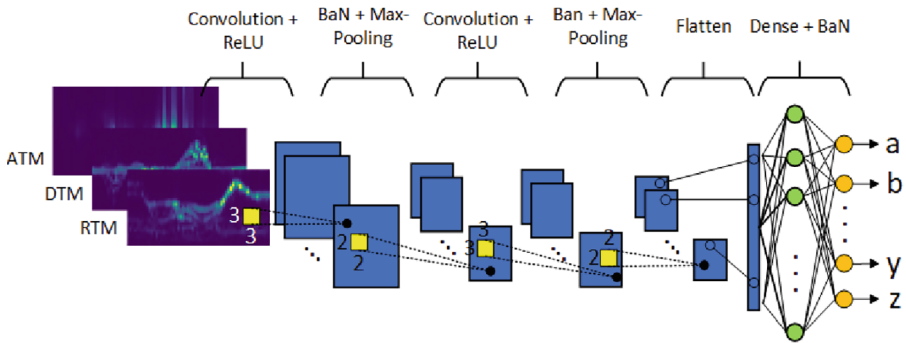


Fig. 2. Architecture of deep convolutional neural network.

3 Experimental Setup

3.1 Radar Configuration

The proposed algorithm is evaluated on Infineon BGT60TR13C FMCW radar chipset [39] with 1 transmit antenna and 3 receive antennas. The frame rate is set to 20 frames/sec which corresponds to 50 ms frame time. The transmitted radar signal is linearly increased from 57.18 GHz to 64.32 GHz with center frequency 60.75 GHz. Therefore, the sweep bandwidth is 7.14 GHz which results in 2.1 cm range resolution. In the fast time axis, 64 points per chirp are sampled within the chirp duration 821.4 μ s with sample frequency 1 MHz. Therefore, the maximum unambiguous range is 67.14 cm. Moreover, 32 chirps per frame with 1.56 ms pulse repetition time is transmitted. Thus, the maximum detectable velocity is ± 1.5 m/s with resolution 9.4 cm/s. The field of view (FOV) of the antenna array is 120° with angle resolution 4° and 31 beams. The number of beams is then zero-padded to 32 to be consistent with the number of range bins and Doppler bins. The length of the sliding observation window L is 60 frames which equals to 3 s. Hence, with $N_R/2 = N_D = N_B = 32$ and $L = 60$, the shape of the average range/Doppler/angle-time map is 32×60 . Then, the three maps are concatenated into a three-channel image to represent gestures.

3.2 Data Collection

Air-writing can be performed with different writing styles, speeds, angles and even with different stroke orders. Furthermore, the writing conditions can be quite different, including various hand shapes, hand sizes, and clutter environments. In this work, the hand gestures of all the 26 English lower-case letters are collected in diverse environments such as meeting room, office cubicle and living room etc. As shown in Fig. 3, the use case is set in line with the actual usage scenarios of laptop computers and mobile phones. Hence, the radar sensor is attached on a mobile phone next to a laptop computer. Seven adult volunteers are involved in the introduction of data sets. Moreover, since the length of the sliding observation window is 60 frames which corresponds to 3 s, each gesture should be completed in 3 s. In one data set, called mmAlphabet_fs, volunteers write in their own free-style handwritings including various hand speeds and even various stroke orders. In the other data set called, mmAlphabet_st, volunteers are asked to gesticulate according to a stipulated sequence of strokes in order to make gestures more recognizable for radar sensors. In both data sets, 100 samples of each letter are recorded by each subject.

3.3 Feature Map Selection

The data sensed by a radar sensor are processed into images of temporal spectrogram such as average range/Doppler/angle-time map in Sect. 2.2, 2.3 and 2.4 respectively. Then, the three maps are concatenated into a three-channel image to represent the letter. This image is then fed to a deep CNN model for classification. As detailed in Sect. 2, two variants of feature map are available. V1 accumulates the signal power non-coherently and so does V2 in the accumulation of signal magnitude.



Fig. 3. The use case is set in line with the actual usage scenarios of laptop computers and mobile phones.

Besides V1 and V2, a third variant of feature map, V3, can be obtained as in [18] and [33]. For example, from the range-Doppler map in (7), extract the column vector corresponding to the maximum pixel magnitude from each frame after accumulation in the antenna dimension and stack them together as in (6) to form a range-time map. In

addition, extract a row vector in (7) corresponding to the same condition and stack them together as in (10) to form a Doppler-time map. To build the temporal angle spectrogram, from the range-angle map in (11), extract the row vector corresponding to the maximum pixel magnitude from each frame and stack them together as in (14) to construct an angle-time map.

3.4 Hyperparameters and Training

To verify the effectiveness of the proposed algorithm, the model in Fig. 2 is trained with optimizer adam with learning rate = 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Loss function is categorical cross entropy. Performance metric is accuracy. Batch size is 32. In the K-fold cross validation, 20% of the total data is randomly selected as test data, 80% of the total data is training data. Furthermore, 20% of the training data is used as validation data. In the Leave-One-Subject-Out (LOSO) cross validation, each volunteer's data is used as test data in turn. Other volunteer's data is used as training data and 20% of the training data is used as validation data.

Table 1. K-fold Cross Validation and LOSO Cross Validation.

Feature variant	mmAlphabet_fs			mmAlphabet_st		
	Best	Average	LOSO Average	Best	Average	LOSO Average
V1	93.32%	86.64%	60.90%	95.80%	89.78%	80.44%
V2	96.84%	94.55%	67%	98.60%	97.30%	87.74%
V3	95.69%	92.17%	66.03%	98.27%	96.62%	85.90%

4 Experimental Results and Discussion

4.1 Performance Analysis

The proposed air writing alphabet recognition system is evaluated on all three feature map variants using the two created data sets. Both K-fold cross validation and leave-one-subject-out cross validation are conducted to assess the DCNN classifier. Test accuracy is used as the performance metric as shown in Table 1. The experimental results are detailed as follows.

K-fold Cross Validation. K-fold cross validation with $K = 5$ is conducted. 20% of the total data is randomly selected as test data, 80% of the total data is reserved as training data. Then the model is trained 20 times and the results of 20 trials are got. This evaluation process is repeated five times, and the results of 100 trials are obtained. Both the best test accuracy and the average test accuracy of 100 trails are detailed in Table 1. On both data sets, V2 outperforms V1 and V3. The best test accuracy is 98.6% trained with mmAlphabet_st. The performances trained with mmAlphabet_st are better than those trained with mmAlphabet_fs about 2–3%.

Leave-One-Subject-Out (LOSO) Cross Validation. The model is trained using the data of all but one subject and is evaluated on the unseen data for 20 times. This evaluation process is repeated for every subject and the best test accuracy values derived from the 20 trials are averaged. The evaluation results are listed in Table 1. On both data sets, V2 outperforms V1 and V3 again and the best LOSO test accuracy is 87.74% trained with mmAlphabet_st by using the feature maps in V2. Furthermore, the performances trained with mmAlphabet_st are much better than those trained with mmAlphabet_fs about 20%. In other words, it is very helpful to intentionally design the writing patterns of an alphabet to reduce the intraclass variability and make them more recognizable for radar sensors.

4.2 Visualization and Discussion

According the results in previous section, V2 outperforms V1 and V3 on both data sets. This is due to the difference between square operation and absolute value operation, which makes V2 better able to retain gesture information than V1. As illustrated in Fig. 4(a)–4(b) for V1 and Fig. 4(c)–4(d) for V2, considering the range attenuation of radar signal in (4), (8) and (12), when the received signal value is greater than 1, the square operation will increase the signal value, and vice versa will decrease it. In addition, feature maps in V2 can also represent gesture movement better than those in V3 because of the signal integration in range bin dimension. Take average Doppler-time map and Doppler-time map [18] for example, the μ -Doppler signature [2] of the whole hand is better retained in average Doppler-time map as depicted in Fig. 4(e) and Fig. 4(f) due to the hand size of an adult is always larger than one range bin (2.1 cm). Hence, the μ -Doppler signature is better preserved after the signal integration in range bin dimension. As a result, feature maps in V2 can be used to represent more subtle gestures such as rotation, finger rub and waving hand, etc.

4.3 Performance Comparison

To get an insight into how the proposed air writing alphabet recognition system performs in compared with other related works, the comparisons are presented in Table 2 and Table 3. There are quite a few related works about air writing recognition system. Some of them are based on RGB camera or wearable device. However, the systems based on RGB camera have privacy issues and the other systems based on wearable device can bring cumbersome experiences to users due to their wearable nature. Thus, radar-based air writing recognition system is more attractive to be an always-on solution for contactless dynamic gesture recognition. Nevertheless, some of the radar-based works need information from more than one radar sensors to reconstruct the trajectory of gestures for recognition which is impractical in real-life usage scenario. As revealed in Table 2 and Table 3, only one radar sensor is required by the proposed system to recognize English alphabet. Furthermore, compared to the related works that can recognize the whole 26 English letters in Table 2 and Table 3, the proposed system also outperforms them both in K-fold cross validation and leave-one-subject-out cross validation.

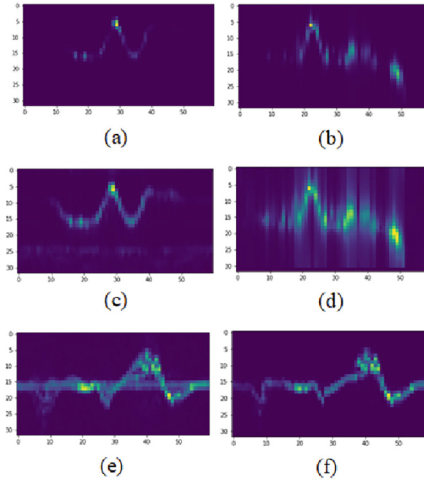


Fig. 4. Temporal Spectrograms. (a) average range-time map of letter “a” in V1, (b) average angle-time map of letter “m” in V1, (c) average range-time map of letter “a” in V2, (d) average angle-time map of letter “m” in V2, (e) average Doppler-time map of letter “t” in V2, (f) Doppler-time map [18] of letter “t” in V3.

4.4 Confusion Matrix and Discussion

The experimental results show that deep CNN network is capable of learning implicit features directly from the temporal spectrograms from radar sensor. The accuracy of each gesture after applying the CNN is presented in confusion matrix form in Fig. 5. The figure illustrates the links between false classifications of one exemplary K-fold cross-validation test with test accuracy 97.25%. It reveals that some gestures can be recognized with higher accuracy such as ‘d’, ‘g’, ‘k’, ‘m’, ‘s’, ‘t’ and ‘z’ while others are easily mistaken. For example, ‘a’, ‘c’, ‘e’, ‘h’, ‘n’ and ‘w’ can be mistakenly recognized as ‘q’, ‘z’, ‘t’, ‘n’, ‘h’ and ‘v’ respectively. ‘j’ can be wrongly classified as ‘i’ or ‘s’. ‘o’ can be wrongly classified as ‘a’ or ‘v’. ‘p’ can be wrongly classified as ‘i’ or ‘k’. ‘r’ can be wrongly classified as ‘f’, ‘n’ or ‘v’. ‘y’ can be wrongly classified as ‘t’, ‘x’ or ‘z’. Besides, the pattern of ‘a’, ‘d’, ‘u’ and ‘q’, the pattern of ‘b’ and ‘h’, the pattern of ‘i’, ‘r’ and ‘v’ and the pattern of ‘x’ and ‘y’ are also very similar. Thus, in order to improve the performance of the alphabet classifier, it is helpful to intentionally design the writing patterns of an alphabet to increase the interclass variability and make them more distinguishable for radar sensors.

Air-writing can be performed with different writing styles, speeds, and angles. Range-time map is actually the range spectrum of range profile changing over time. It comprises the information of object range and signal intensity. Doppler-time map is the Doppler spectrum of radar changing over time. It comprises the information of object radial velocity and moving direction. Angle-time map is the angle spectrum of beamformer changing over time. It comprises the information of the change of azimuth. With the three feature maps or their variants, the proposed alphabet recognition system is

Table 2. Comparison of K-fold test accuracy with related works.

Related works	Sensor	DL/ML Model	Gestures	Accuracy
Chu et al. [24]	RGB + Depth	SVM	10 Numerals	90.80%
Wang et al. [26]	WiFi	DCNN	10 Numerals	97.20%
Hendy et al. [31]	1 Radar	3D-CNN-LSTM	10 Numerals	98.50%
Leem et al. [30]	3 Radars	DCNN	10 Numerals	99.70%
Li et al. [32]	1 Radar	Xception	10 Numerals	99.60%
Kwak et al. [43]	1 Radar	CNN	10 Numerals	87.60%
Kwak et al. [42]	1 Radar	DNN + CNN	10 Numerals	94.57%
Liu et al. [44]	1 Radar	ResNet50	6 Characters	93.40%
Arsalan et al. [34]	3 Radars	ConvLSTM-CTC	A-J and 1-5	98.33%
Arsalan et al. [35]	2 Radars	SNN	A-J and 1-5	98.53%
Arsalan et al. [35]	1 Radar	SNN	A-J and 1-5	95.37%
Arsalan et al. [18]	1 Radar	SNN	A-J and 1-5	99.50%
Park et al. [41]	1 Radar	DCNN	A-Z	91.00%
Moazen et al. [27]	Motion Sensor	DTW	a-z	71.00%
Luo et al. [28]	IMU	DTW	A-Z and 0-9	84.60%
Yang et al. [29]	RFID	DCNN	a-z	96.60%
Proposed	1 Radar	DCNN	a-z	98.60%

Table 3. Comparison of LOSO test accuracy with related works.

Related works	Sensor	Model	Gestures	Accuracy
Molchanov et al. [12]	Depth + Radar + Optical	DCNN	10 Gestures	75.10%
Auge et al. [17]	1 Radar	SNN	11 Soli Gestures	88.20%
Tsang et al. [19]	1 Radar	SNN	11 Soli Gestures	88.27%
Wei et al. [40]	1 Radar	CRNN	A-Z + 4 Gestures	87.55%
Proposed	1 Radar	DCNN	a-z	87.74%

verified to be effective on the two created data sets. It can be used to implement natural, intelligent noncontact human machine interface.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	
a	0.9716	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0284	0	0	0	0.0071	0	0	0	0	0	
b	0	0.9786	0	0	0	0	0	0.0071	0	0	0.0071	0	0	0	0	0	0	0	0.0071	0	0	0	0	0	0	0	
c	0	0	0.9871	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0071	0	0.0071	0	0	0	0	0.0383	
d	0.0071	0	0	0.9929	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
e	0	0	0.0071	0	0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0383	0	0	0	0	0.0071	
f	0	0	0	0	0	0.9857	0	0	0	0	0	0.0071	0	0	0	0	0	0	0	0	0	0.0071	0	0	0	0	
g	0	0	0	0	0	0	0.9929	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0071	0	
h	0	0	0	0	0	0	0	0.9222	0	0	0	0	0	0	0	0	0.0284	0	0.0071	0	0	0	0	0	0	0	
i	0	0	0	0	0	0	0	0	0.9857	0	0	0.0071	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
j	0	0	0	0	0	0.0071	0	0	0.0214	0.95	0	0	0	0	0	0	0	0	0.0071	0	0.0214	0	0	0	0	0	
k	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
l	0	0	0	0	0	0	0	0	0	0	0	0.9786	0	0	0	0	0	0	0	0.0071	0.0071	0	0.0071	0	0	0	
m	0	0	0	0	0	0	0	0	0	0	0	0	0.9929	0.0071	0	0	0	0	0	0	0	0	0	0	0	0	
n	0	0.0071	0	0	0	0	0	0.0284	0	0	0	0	0	0	0	0.9571	0	0	0	0	0	0.0071	0	0	0	0	
o	0.0214	0	0	0	0	0	0	0	0.0071	0	0	0	0	0	0.95	0	0	0	0.0071	0.0071	0	0.0214	0	0	0	0	
p	0	0	0	0	0	0	0	0	0.0214	0	0.0214	0	0	0	0	0.9714	0	0	0	0	0	0	0	0	0	0	
q	0	0	0	0.0071	0	0	0	0	0	0	0	0	0	0.0071	0	0.9857	0	0	0	0	0	0	0	0	0	0	
r	0	0	0	0	0	0.0214	0	0.0071	0	0	0	0	0.0214	0	0	0	0.9214	0	0	0	0	0.0284	0	0	0	0	
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
u	0.0071	0	0	0.0071	0	0	0	0	0	0	0	0.0071	0	0	0	0	0	0	0	0	0	0.9786	0	0	0	0	
v	0	0	0	0	0	0	0	0.0071	0	0	0	0	0	0	0	0	0	0	0	0	0	0.9857	0.0071	0	0	0	
w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0214	0.9857	0	0	0	
x	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0071	0	0	0.9857	0.0071	0	
y	0	0	0.0071	0	0	0	0	0	0	0	0.0071	0	0	0	0	0	0	0	0	0	0	0.0214	0	0	0.0284	0.9143	0.0214
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0071	0	0	0	0	0	0.9929	

Fig. 5. An exemplary K-fold cross-validation test with test accuracy 97.25%. Each row in the confusion matrix represents the true label of a gesture.

5 Conclusion

This paper investigates the feasibility of using three temporal spectrograms, average range/Doppler/angle-time map derived from FMCW radar as feature images to train a deep CNN model which can recognize English alphabet in midair and verifies the effectiveness of the proposed system. In order to train the classifier model effectively, two valuable radar data sets are introduced in diverse environments such as meeting room, office cubicle and living room etc. The use case is set in line with the actual usage scenarios of laptop computers and mobile phones. A convolutional neural network which achieves 98.6% test accuracy is exploited as the classifier to recognize the air written alphabet. In the Leave-One-Subject-Out (LOSO) cross validation, it achieves an average test accuracy of 87.74%. In addition, Average range/Doppler/angle-time map can keep the information of a gesture better than the other two variants. For instance, μ -Doppler signature of the whole hand can be better preserved because the hand size of an adult is always larger than one range bin. Hence, they can be used to represent more subtle gestures such as rotation, finger rub and waving hand, etc. Furthermore, in order to improve the performance of the alphabet classifier, it is very helpful to intentionally design the writing patterns of an alphabet to reduce the intraclass variability and increase the interclass variability so that they are more recognizable and more distinguishable for radar sensors.

References






1. Gu, C., Wang, J., Lien, J.: Motion sensing using radar: gesture interaction and beyond. *IEEE Microwave Mag.* **20**(8), 44–57 (2019)

2. Franceschini, S., et al.: Hand gesture recognition via radar sensors and convolutional neural networks. In: IEEE Radar Conference, pp. 1–5 (2020)
3. Jang, Y., Jeon, I., Kim, T.-K., Woo, W.: Metaphoric hand gestures for orientation-aware VR object manipulation with an egocentric view-point. *IEEE Trans. Human-Mach. Syst.* **47**(1), 113–127 (2017)
4. Wang, C., Liu, Z., Chan, S.-C.: Superpixel-based hand gesture recognition with Kinect depth camera. *IEEE Trans. Multimedia* **17**(1), 29–39 (2015)
5. Patra, A., Geuer, P., Munari, A., Mähönen, P.: mm-Wave radar based gesture recognition: development and evaluation of a low-power, low-complexity system. In: ACM mmNets (2018)
6. Hayashi, E., et al.: RadarNet: efficient gesture recognition technique utilizing a miniature radar sensor. In: ACM Conference on Human Factors in Computing Systems, 8–13 May 2021
7. Kim, Y., Toomajian, B.: Hand gesture recognition using micro-Doppler signatures with convolutional neural network. *IEEE Access* **4**, 7125–7130 (2016)
8. Dekker, B., et al.: Gesture recognition with a low power FMCW radar and a deep convolutional neural network. In: European Radar Conference (EURAD), pp. 163–166 (2017)
9. Ouaknine, A., et al.: CARRADA dataset: camera and automotive radar with range- angle-Doppler annotations. In: 25th International Conference on Pattern Recognition (ICPR), pp. 5068–5075 (2021)
10. Kern, N., et al.: Robust Doppler-based gesture recognition with incoherent automotive radar sensor networks. *IEEE Sens. Lett.* **4**(11), 1–4 (2020)
11. Anitori, L., et al.: FMCW radar for life-sign detection. In: Proceedings of the IEEE RadarConference (2009)
12. Molchanov, P., et al.: Multi-sensor system for driver’s hand-gesture recognition. In: Proceedings of the IEEE FG (2015)
13. Molchanov, P., et al.: Short-range FMCW monopulse radar for hand-gesture sensing. In: Proceedings of the IEEE RadarConference (2015)
14. Ohn-Bar, E., Trivedi, M.M.: Hand gesture recognition in real time for automotive interfaces: a multimodal vision-based approach and evaluations. *IEEE Trans. Intell. Transp. Syst.* **15**(6), 2368–2377 (2014)
15. Wan, Q., et al.: Gesture recognition for smart home applications using portable radar sensors. In: Proceedings of the IEEE EMBS (2014)
16. Molchanov, P., et al.: Multi-sensor system for driver’s hand-gesture recognition. In: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8 (2015)
17. Auge, D., et al.: Hand gesture recognition in range-doppler images using binary activated spiking neural networks. In: 16th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 01–07 (2021)
18. Arsalan, M., Santra, A., Issakov, V.: RadarSNN: a resource efficient gesture sensing system based on mm-Wave radar. *IEEE Trans. Microw. Theory Tech.* **70**(4), 2451–2461 (2022)
19. Tsang, I.J., et al.: Radar-based hand gesture recognition using spiking neural networks. *Electronics* **10**, 1405 (2021)
20. Bhatia, J., et al.: Object classification technique for mmWave FMCW radars using range-FFT features. In: International Conference on COMMunication Systems & NETWORKS (COMSNETS), pp. 111–115 (2021)
21. Liu, C., et al.: Spectrum-based hand gesture recognition using millimeter-wave radar parameter measurements. *IEEE Access* **7**, 79147–79158 (2019)
22. Sun, Y., et al.: Multi-feature encoder for radar-based gesture recognition. In: IEEE International Radar Conference (RADAR), pp. 351–356 (2020)
23. Ninos, A., Hasch, J., Zwick, T.: Multi-user macro gesture recognition using mmWave technology. In: 18th European Radar Conference (EuRAD), pp. 37–40 (2022)

24. Chu, T.-T., Su, C.-Y.: A Kinect-based handwritten digit recognition for TV remote controller. In: International Symposium on Intelligent Signal Processing and Communications Systems, pp. 414–419 (2012)
25. Hsieh, C.-H., Lo, Y.-S., Chen, J.-Y., Tang, S.-K.: Air-writing recognition based on deep convolutional neural networks. *IEEE Access* **9**, 142827–142836 (2021)
26. Wang, Z., et al.: WiDG: an air hand gesture recognition system based on CSI and deep learning. In: 33rd Chinese Control and Decision Conference, pp. 1243–1248 (2021)
27. Moazen, D., Sajjadi, S.A., Nahapetian, A.: AirDraw: leveraging smart watch motion sensors for mobile human computer interactions. In: 13th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp. 442–446 (2016)
28. Luo, Y., Liu, J., Shimamoto, S.: Wearable air-writing recognition system employing dynamic time warping. In: IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), pp. 1–6 (2021)
29. Yang, Z., et al.: RF-Eletter: a cross-domain English letter recognition system based on RFID. *IEEE Access* **9**, 155260–155273 (2021)
30. Leem, S.K., Khan, F., Cho, S.H.: Detecting midair gestures for digit writing with radio sensors and a CNN. *IEEE Trans. Instrum. Meas.* **69**(4), 1066–1081 (2020)
31. Hendy, N., et al.: Deep learning approaches for air-writing using single UWB radar. *IEEE Sens. J.* **22**(12), 11989–12001 (2022)
32. Li, W., et al.: Digital gesture recognition based on millimeter wave radar. In: International Conference on Signal Processing, Communications and Computing, pp. 1–6 (2021)
33. Mauro, G., et al.: Few-shot user-definable radar-based hand gesture recognition at the edge. *IEEE Access* **10**, 29741–29759 (2022)
34. Arsalan, M., Santra, A.: Character recognition in air-writing based on network of radars for human-machine interface. *IEEE Sens. J.* **19**(19), 8855–8864 (2019)
35. Arsalan, M., Santra, A., Issakov, V.: Low power radar-based air-writing system using genetic algorithm-assisted spiking Legendre memory unit. In: 20th European Radar Conference (EuRAD), pp. 250–253 (2023)
36. Gao, X., et al.: RAMP-CNN: a novel neural network for enhanced automotive radar object recognition. *IEEE Sens. J.* **21**(4), 5119–5132 (2021)
37. Zhou, T., Xia, Z., Wang, X., Xu, F.: Human sleep posture recognition based on millimeter-wave radar. In: Signal Processing Symposium (SPSymo), pp. 316–321 (2021)
38. Li, J., et al.: On robust Capon beamforming and diagonal loading. *IEEE Trans. Signal Process.* **51**(7), 1702–1715 (2003)
39. Infineon Technology, BGT60TR13C, XENSIV™ 60GHz radar sensor for advanced sensing. <https://www.infineon.com/cms/en/product/sensor/radar-sensors/radar-sensors-for-iot/60ghz-radar/bgt60tr13c/>
40. Wei, H., et al.: IndexPen: two-finger text input with millimeter-wave radar. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **6**(2) (2022). Article 79
41. Park, J.B., et al.: Air writing gesture recognition using FMCW radar and deep learning. In: 8th IEEE International Conference on Network Intelligence and Digital Content, pp. 369–373 (2023)
42. Kwak, S., Park, C., Lee, S.: DNN-based legibility improvement for air-writing in millimeter-waveband radar system. *IEEE Trans. Instr. Meas.* **72**, 1–12 (2023). Art no. 8006912
43. Kwak, S., Park, C., Lee, S.: Improving air-writing accuracy through data regression and interpolation in a single radar system. In: 18th European Conference on Antennas and Propagation (EuCAP), pp. 1–5 (2024)
44. Liu, H., et al.: Multi-stroke air-writing recognition using temporal-spatial interferometric MIMO radar. In: IEEE Radar Conference (RadarConf 2024), pp. 1–6 (2024)



FG-MDM: Towards Zero-Shot Human Motion Generation via ChatGPT-Refined Descriptions

Xu Shi¹, Wei Yao¹, Chuanchen Luo², Junran Peng³, Hongwen Zhang⁴,
and Yunlian Sun¹^(✉)

¹ School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

{shixu, wei.yao, yunlian.sun}@njust.edu.cn

² School of Artificial Intelligence, Shandong University, Jinan, China
chuanchen.luo@sdu.edu.cn

³ Institute of Automation, Chinese Academy of Sciences, Beijing, China

⁴ School of Artificial Intelligence, Beijing Normal University, Beijing, China
zhanghongwen@bnu.edu.cn

Abstract. Recently, significant progress has been made in text-based motion generation, enabling the generation of diverse and high-quality human motions that conform to textual descriptions. However, generating motions beyond the distribution of original datasets remains challenging, i.e., zero-shot generation. By adopting a divide-and-conquer strategy, we propose a new framework named Fine-Grained Human Motion Diffusion Model (FG-MDM) for zero-shot human motion generation. Specifically, we first parse previous vague textual annotations into fine-grained descriptions of different body parts by leveraging a large language model. We then use these fine-grained descriptions to guide a transformer-based diffusion model, which further adopts a design of part tokens. FG-MDM can generate human motions beyond the scope of original datasets owing to descriptions that are closer to motion essence. Our experimental results demonstrate the superiority of FG-MDM over previous methods in zero-shot settings. We will release our fine-grained textual annotations for HumanML3D and KIT on the project page <https://sx0207.github.io/fg-mdm/>

Keywords: Human Motion Generation · Diffusion Model · Zero-Shot Generation

1 Introduction

Human motion generation is an important research topic in communities of both computer vision and computer graphics. It aims to simulate and generate realistic human movements using computers. With the advancement of technologies

Supported by National Natural Science Foundation of China under Grant 62076131.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78104-9_30.

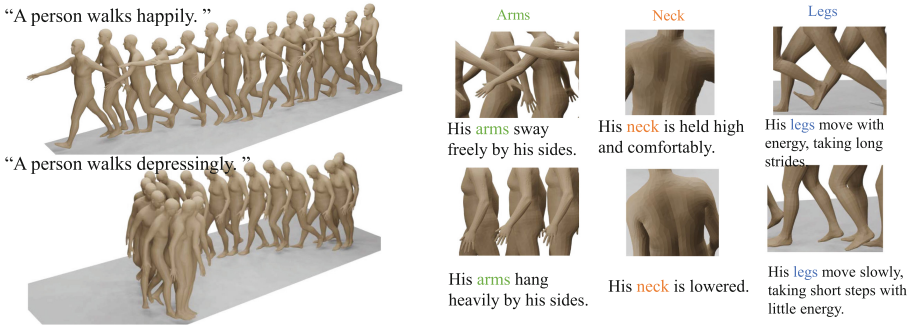


Fig. 1. FG-MDM can generate high-quality human motions in zero-shot settings by using fine-grained descriptions of different body parts. The two images on the left illustrate two contrasting emotional motions. Close-Up images of the arms, neck, and legs highlight these differences.

such as virtual reality, augmented reality, and movie special effects, there is a growing demand for high-quality human motion generation. In recent years, several innovative methods and techniques have emerged to tackle this challenging task [48]. Deep generative models, including GANs [1, 21], VAEs [8, 26, 27], and diffusion models [5, 15, 37, 45], have been widely applied to human motion generation.

However, there is relatively less research on generating motions in zero-shot settings. In order to improve the zero-shot generation capability, existing work either got help from CLIP [31] to utilize the rich semantic knowledge from CLIP (e.g., [12, 36]), or attempted to explore large-scale motion datasets without textual descriptions [17] and large-scale pseudo text-pose datasets [3, 20]. Compared to traditional human motion generation, generating motions beyond the distribution of the dataset is more challenging due to the limited scale and diversity of existing motion capture datasets.

Then, with only limited motion capture datasets available, can we still generate motions beyond the distribution of the dataset? For a textual description defining a motion beyond the distribution of the dataset, is there any way to associate it with motions within the dataset? For a never-before-seen motion, the entire body’s motion is indeed unseen. However, motions of specific body parts might be inside the dataset. Therefore, we can adopt a divide-and-conquer strategy. By re-annotating the motion for different body parts with fine-grained descriptions, we can associate these body parts with specific body parts within the dataset. For example, a vague description “A person walks depressingly.” can be reformulated as “His arms hang heavily by his sides. His legs move slowly, taking short steps with little energy...”. Leg movement in this vague unseen motion may appear in “A person walks aimlessly and slowly.”, of which the motion is included in the dataset. And the arm movement may appear in “His arms hang heavily by his sides.”, of which the motion is included in the dataset. We give two examples in Fig. 1. On one hand, adopting fine-grained descriptions allows

the model to understand how the unseen motions are performed in detail. On the other hand, re-annotating motions for different body parts enables the model to learn the essence of motions better. Using fine-grained textual descriptions, we aim to improve the model’s zero-shot understanding capability.

Although annotating fine-grained textual descriptions manually for body parts provides more accurate data, it requires significant manual work, resulting in huge costs. Fortunately, with the rapid development of large language models, OpenAI’s GPT series models [25], known for their outstanding natural language processing capabilities, have gained widespread attention worldwide. In [7], Gilardi et al. demonstrated that ChatGPT performs as well as human annotators in some tasks. In [14], Action-GPT explores the excellent capability of ChatGPT in expanding human action descriptions. However, the generated content tends to be excessively redundant. For our task, we carefully design a prompt that allows ChatGPT-3.5 to provide detailed but non-redundant transcriptions of text descriptions about human motion. We then use this prompt and ChatGPT-3.5 to transcribe 44,970 short text descriptions from HumanML3D [8] and 6,353 text descriptions from KIT [29] for model training.

With these fine-grained descriptions, we propose a new framework named Fine-Grained Human Motion Diffusion Model (FG-MDM) for human motion generation. Specifically, we replace the original simple and vague text with ChatGPT-Generated fine-grained descriptions of individual body parts to guide a transformer-based diffusion model. Following MDM [37], we encode the entire fine-grained description with CLIP [31] as a global token of the transformer. Apart from this global token, we further encode descriptions of different body parts individually with CLIP as part tokens. By adopting these tokens, the model can pay attention to both the global and detailed information of human motions, thereby improving the accuracy and completeness of the denoising results.

Our contributions are summarized as follows:

- We present a novel framework that utilizes fine-grained descriptions of different body parts to guide the denoising process of the transformer-based diffusion model. This framework is capable of generating a broader range of motions that extend beyond the distribution of training datasets.
- We carefully design a prompt that enables ChatGPT to convert short and vague texts into detailed but non-redundant descriptions of different body parts. We then use this prompt to transcribe 44,970 texts from HumanML3D and 6,353 texts from KIT into fine-grained descriptions. We will make these fine-grained transcriptions publicly available.
- We conduct a series of experiments to evaluate our model’s ability to not only fit the training data but also generate motions beyond the distribution of the dataset, i.e., the generalization capability.

2 Related Work

2.1 Human Motion Generation

There has been a great interest in human motion generation in recent years. Previous work has explored unconditional generative models [30, 47] as well as generative models using various input conditions, such as text [5, 8, 37], prior motion [24], action class [9, 26], and music [16, 39]. In this paper, we focus on text-to-motion generation. Early work usually addressed the text-to-motion task with a sequence-to-sequence model [18]. Later on, the focus shifted beyond simple action labels. For example, Guo et al. utilized variational autoencoders to generate motions from text [8], significantly enhancing the quality and diversity of generated motions. With the success of diffusion models in AIGC, MDM [37] and other related work [5, 15, 45] have introduced diffusion models into the text-to-motion domain, resulting in impressive achievements.

There is a relative scarcity of work that directly focuses on the zero-shot capabilities of motion generation models. In [36], Tevet et al. proposed MotionCLIP to align human motions with the CLIP space, implicitly injecting the rich semantic knowledge from CLIP into the motion domain to enhance zero-shot generation capability. AvatarCLIP [12] also utilized CLIP to implement a zero-shot text-driven framework for 3D avatar generation and animation. Liang et al. [17] pre-trained a large-scale unconditional diffusion model to learn rich out-of-domain motion traits. In order to improve the generalization capability of motion generation models, there have been also attempts to leverage human mesh recovery approaches [38, 41–43] to collect large-scale pseudo text-pose datasets [3]. As shown in Azadi et al. [3] and Lin et al. [20], the pre-training on such text-pose datasets can improve the generalization to in-the-wild descriptions. However, the static nature of text-pose data makes it difficult to well represent dynamic motions.

The work most closely related to ours is Action-GPT [14], which introduced, for the first time, large language models into the field of text-conditioned motion generation. Action-GPT can be integrated into any text-to-motion model. However, it enriched only the description of action classes without providing detailed descriptions of different body parts and guiding the model training. For another action generation model, SINC [2] incorporated ChatGPT to identify the body parts involved in the textual description. It achieved impressive results by generating multiple motions and concatenating them using different body parts. Specifically, SINC divides the human body into [‘left arm’, ‘right arm’, ‘left leg’, ‘right leg’, ‘torso’, ‘neck’, ‘buttocks’, ‘waist’], which we borrow from in our work. It should be noted that both Action-GPT and SINC were designed for motion generation based on action labels, not using natural language. Therefore, directly comparing our work with them is not feasible. There is also a costly method that utilizes LLMs. By fine-tuning LLMs, MotionGPT [13, 46] designed a pre-trained motion language model that supports various motion-related tasks through prompts. In contrast, our method is more efficient and can rapidly enhance the model’s zero-shot generation capabilities.

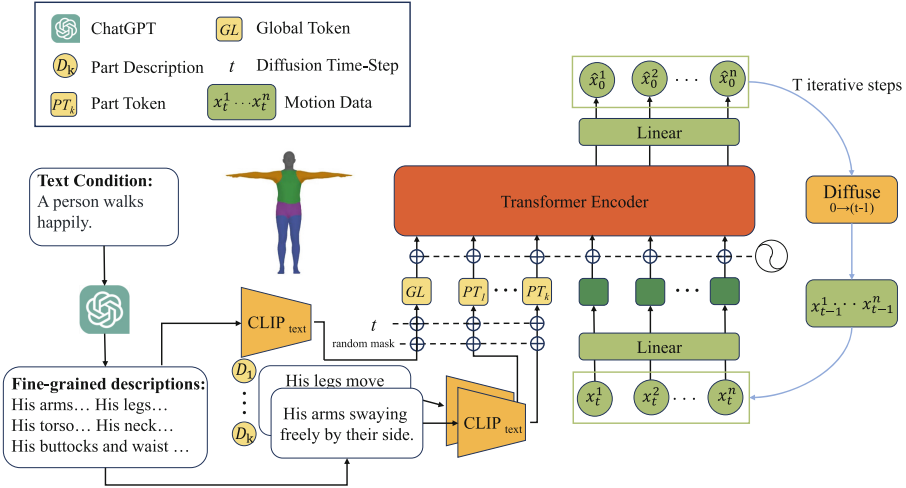


Fig. 2. The overall pipeline of FG-MDM. The model learns the denoising process of the diffusion model from the motion $x_t^{1:n}$ at time step t to the clean motion $\hat{x}_0^{1:n}$, given the text condition. The input text is first paraphrased by ChatGPT into fine-grained descriptions $D_{1:k}$ for different parts of the body, where k denotes the number of body parts. These descriptions are then fed into a pre-trained CLIP text encoder and projected, along with the time step t , onto input tokens $PT_{1:k}$ of the transformer. The overall fine-grained text is further encoded into a global input token GL , providing holistic information. In the sampling process of the diffusion model, an initial random noise $x_T^{1:n}$ is sampled, and then T iterations are performed to generate the clean motion $\hat{x}_0^{1:n}$. At each sampling step t , guided by $PT_{1:k}$ and GL , the transformer encoder predicts the clean motion $\hat{x}_0^{1:n}$ which is then noised back to $x_{t-1}^{1:n}$.

2.2 Diffusion Generative Models

The diffusion model is a neural generative model based on the stochastic diffusion process in thermodynamics [10, 35]. It starts with samples from the data distribution and gradually adds noise through a forward diffusion process. Then, a neural network learns the reverse process to progressively remove the noise and restore the samples to their original states. Diffusion generative models have achieved significant success in the image generation field [32, 33]. For conditional generation, [6] introduced classifier-guided diffusion, while [11] proposed a classifier-free method. Given their excellent generation quality, [15, 37, 45] incorporated diffusion models into the motion generation domain, leading to impressive results.

3 Method

Given a textual description, our goal is to generate a human motion $x^{1:n} = \{x^i\}_{i=1}^n$ that matches the given description. The motion consists of n frames of human poses. For each pose $x^i \in \mathbb{R}^{J \times D}$, we represent it by joint rotations

or positions, where J represents the number of joints and D represents the dimensionality of the joint representation. In Fig. 2, we give an overview of our fine-grained human motion diffusion model. First, we adopt ChatGPT to perform fine-grained paraphrasing of the vague textual description. This expands concise textual descriptions into descriptions of different body parts. FG-MDM then uses these fine-grained descriptions to guide a diffusion model for human motion generation.

3.1 Prompt Strategy

We first introduce the prompt strategy adopted for generating fine-grained descriptions. We utilize ChatGPT-3.5 to create more fine-grained descriptions based on different body parts for a given textual description of a motion. ChatGPT is a conversational model based on a large language model that can engage in natural conversations and generate corresponding responses. The answers from ChatGPT are often directly influenced by the information and expression provided in the prompt. If the prompt offers clear and detailed questions or instructions, ChatGPT can typically provide relevant and accurate answers. However, if the prompt is too simple, ambiguous, or unclear, ChatGPT may generate unexpected responses or express unclear content. For our task, we carefully design an effective prompt by using experimental verification.

Our designed prompt is: “Translate the motion described by the given sentences to the motion of each body part only using one paragraph. The available body parts include [‘arms’, ‘legs’, ‘torso’, ‘neck’, ‘buttocks’, ‘waist’]. Here are some examples: [Q...A...]. Question: [sentence]”. [sentence] is the vague textual description that needs to be refined. [Q...A...] are four examples of Q&A pairs designed manually.

3.2 Diffusion Model for Motion Generation

The basic idea of diffusion models [10, 35] is to learn the reverse process of a well-defined stochastic process. Following MDM [37], we design a text-driven human motion generation model based on the diffusion model.

The diffusion model consists of the forward process and the reverse process, both of which follow the Markov chain. The forward process involves adding noise. The input is the original motion $x_0^{1:n}$ from the data distribution, and the output is the motion $x_t^{1:n}$ with adding Gaussian noise t times. When enough noise is added, the motion $x_T^{1:n}$ can approach the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The reverse process aims to reduce the noise in the Gaussian noise $x_T^{1:n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In the denoising process, at diffusion step t , a portion of the noise is eliminated, resulting in a less noisy motion $x_{t-1}^{1:n}$. This step is repeated iteratively until the noise is completely removed, generating a clean motion $\hat{x}_0^{1:n}$.

Network. We adopt a simple transformer [40] encoder architecture to implement our network G . Unlike the conventional diffusion model mentioned above, we follow [32] and predict the clean motion $\hat{x}_0^{1:n}$ instead of predicting the noise added in each time-step. The input of G is the noised motion $x_t^{1:n}$ obtained by adding noise t times to the original motion $x_0^{1:n}$. The noised motion $x_t^{1:n}$, together with the text condition tokens GL , $PT_{1:k}$ and the time-step t , are inputted to the transformer encoder, resulting in the clean motion $\hat{x}_0^{1:n}$. One of the reasons for directly predicting the clean motion in each time-step of the diffusion model is to incorporate human geometric losses during the training of the network, making the generated human motions more natural. For each sampling step t , from T to 1, our model predicts the clean motion $\hat{x}_0^{1:n}$, and then adds noise back to $x_{t-1}^{1:n}$. After T iterations, the final clean motion $\hat{x}_0^{1:n}$ is obtained. This form of diffusion model has become commonly adopted, as do we.

Global Token and Part Tokens. For the text condition, we encode the entire fine-grained description with CLIP [31] as a global token GL of the transformer. Apart from this global token, we further encode descriptions of different body parts individually with CLIP as part tokens $PT_{1:k}$, where k denotes the number of body parts. The global token serves as an overall condition to guide the diffusion process. Part tokens provide explicit information for fine-grained control of the movements of each body part. Part tokens effectively make up for the ambiguity of the original description text. It greatly enhances our FG-MDM’s ability to understand in-the-wild text, making it outstanding on zero-shot generation tasks.

Loss Functions. For training the diffusion model, we follow [32] to predict the signal itself instead of predicting the noise, i.e., $\hat{x}_0^{1:n} = G(x_t^{1:n}, t, c)$, with the simple loss function.

$$\mathcal{L}_G = E_{x_0^{1:n} \sim q(x_0^{1:n} | c), t \sim [1, T]} [\|x_0^{1:n} - G(x_t^{1:n}, t, c)\|_2^2] \quad (1)$$

In order to generate more natural and kinematically plausible motions, we employ the same geometric losses as MDM [37] from [26, 34], i.e., positions, foot contact, and velocities.

$$\mathcal{L}_{\text{pos}} = \frac{1}{n} \sum_{i=1}^n \|FK(x_0^i) - FK(\hat{x}_0^i)\|_2^2, \quad (2)$$

$$\mathcal{L}_{\text{foot}} = \frac{1}{n-1} \sum_{i=1}^{n-1} \|(FK(\hat{x}_0^{i+1}) - FK(\hat{x}_0^i)) \cdot f_i\|_2^2, \quad (3)$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{n-1} \sum_{i=1}^{n-1} \|(x_0^{i+1} - x_0^i) - (\hat{x}_0^{i+1} - \hat{x}_0^i)\|_2^2 \quad (4)$$

where $FK(\cdot)$ represents the forward kinematic function that converts joint rotations into joint positions. For each frame i , $f_i \in \{0, 1\}^J$ is the binary foot contact mask.

Overall, our training loss is

$$\mathcal{L} = \mathcal{L}_G + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}}. \quad (5)$$

where λ_{pos} , λ_{vel} , λ_{foot} are balancing coefficients for the three geometric losses.

4 Experiments

In this section, we first elaborate the datasets, evaluation metrics, and implementation details in Sect. 4.1. We then conduct quantitative experiments to compare FG-MDM with current state-of-the-art approaches in Sect. 4.2. To show the generalization capability of our model, we further perform quantitative experiments, qualitative experiments, and a user study to examine FG-MDM’s ability to generate motions beyond the distribution of training datasets. Finally, to evaluate our method comprehensively, we design two additional ablation experiments in Sect. 4.3.

4.1 Experimental Details

Datasets. We utilize the HumanML3D [8] dataset and the KIT [29] dataset to train and evaluate our model. The HuMMan [4] dataset and Kungfu dataset from the Motion-X dataset [19, 22] are employed to assess the models’ zero-shot performance. HumanML3D is a recently proposed large-scale dataset of motion-text pairs. It consists of 14,616 motion sequences from the AMASS [23] and HumanAct12 [9] datasets, with multiple ways of describing each motion, resulting in a total of 44,970 text annotations. The KIT dataset, on the other hand, is relatively smaller and contains 3,911 motion sequences along with their corresponding 6,353 text descriptions. For both datasets, we follow the default settings, using 80% of the data for training and the remaining for testing. Motion-X is a large-scale dataset of whole-body motions and whole-body pose annotations, integrating several existing datasets and additional online videos. For zero-shot testing, we utilize 100% of the HuMMan and Kungfu subsets from it. HuMMan is a multi-modal human dataset, containing 744 motion sequences and their corresponding 744 texts descriptions. Kungfu encompasses many human motions related to martial arts, with a total of 1040 motion sequences and their corresponding 1040 texts descriptions.

We preprocess the 44,970 text descriptions from HumanML3D and 6,353 text descriptions from KIT using ChatGPT-3.5. This preprocessing extends these descriptions into fine-grained ones for our model training.

Evaluation Metrics. We employ three evaluation metrics for quantitative experiments to evaluate our model’s ability to fit the training data: FID, Multimodal Dist, and Diversity. Multimodal Dist assesses the correlation between

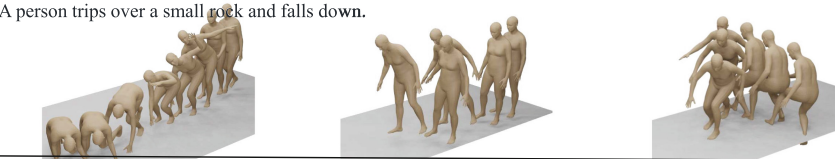
A person jumps happily, as they raise both hands in the air.



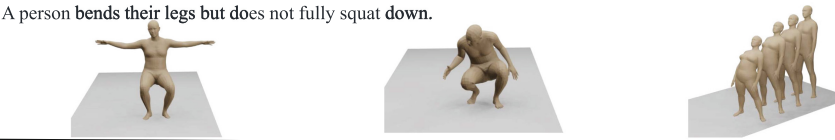
A person jumps sadly, their legs slightly bent but not fully embracing the leap with **enthusiasm**.



A person trips over a small rock and falls down.



A person bends their legs but does not fully squat down.



A person walks cautiously.



The dancer lifts one leg with their hands and maintains a one-legged standing position with both hands raised.



A person hops and jumps on one foot.



FG-MDM

MDM

MLD

Fig. 3. Qualitative results with unseen motions. We compare our FG-MDM with MDM [37] and MLD [5]. All three models are trained on HumanML3D. For better visualization, some pose frames are shifted to prevent overlap. Please refer to supplementary materials for more video demos.

generated motions and input text. Diversity is utilized to evaluate the diversity of generated motions. FID measures the difference in feature distribution between generated motions and ground truth in latent space, which is used to evaluate the quality of generated motions.

Implementation Details. In our study, the transformer accepts tokens whose feature dimension is 512 as input. We use four attention heads and apply a dropout rate of 0.1. The transformer encoder consists of 8 stacked encoder layers to capture complex relationships and hierarchies in the data. For ChatGPT, we adopt the gpt-3.5-turbo API provided by OpenAI. For text encoding, we employ the frozen CLIP-ViT-B/32 model as the encoder. Our batch size is set to 64. Additionally, we set the diffusion step to 1000. On a single NVIDIA GeForce RTX3090 GPU, it takes about six days to train our model.

Table 1. Quantitative results on HumanML3D and HuMMan. The model marked with * indicates that both the ChatGPT-Refined text and the manually annotated text provided by HumanML3D are used during training. **Bold text** is the best result, underlined text is the second-best result. Zero-Shot means that the models are evaluated directly on HuMMan after training on HumanML3D.

Methods	HumanML3D			HuMMan(zero-shot)		
	FID↓	MM Dist↓	Diversity↑	FID↓	MM Dist↓	Diversity↑
Real	0.002 \pm .000	2.974 \pm .008	9.503 \pm .065	0.032 \pm .002	23.019 \pm .042	4.709 \pm .097
MAA [3]	0.774 \pm .007	—	8.230 \pm .064	—	—	—
T2M-GPT [44]	0.116 \pm .004	3.118 \pm .011	9.761 \pm .081	9.631 \pm .203	27.582 \pm .073	5.149 \pm .145
MLD [5]	0.473 \pm .013	3.196 \pm .010	<u>9.724</u> \pm .082	14.970 \pm .472	27.104 \pm .024	5.493 \pm .101
MotionDiffuse [45]	0.630 \pm .001	3.113 \pm .001	9.410 \pm .049	30.138 \pm .712	28.747 \pm .041	5.357 \pm .045
MDM [37]	0.544 \pm .044	5.566 \pm .027	9.559 \pm .086	13.375 \pm .408	27.689 \pm .055	5.585 \pm .089
FG-MDM	0.663 \pm .012	5.649 \pm .024	9.476 \pm .068	17.180 \pm .272	<u>26.867</u> \pm .030	<u>5.589</u> \pm .124
FG-MDM*	0.618 \pm .009	5.274 \pm .048	9.563 \pm .097	<u>12.460</u> \pm .330	26.814 \pm .019	5.626 \pm .100

4.2 Comparison with Prior Work

To evaluate the performance of FG-MDM in handling zero-shot text-conditioned motion generation, we compare our work with five recent motion generation approaches: MAA [3], T2M-GPT [44], MLD [5], MotionDiffuse [45], and MDM [37]. In Table 1 and Table 2, we provide experimental results on the HumanML3D, HuMMan, and Kungfu datasets, respectively. For all experiments, We run the evaluation five times, and “ \pm ” indicates the 95% confidence interval. For the six SOTA methods, on HumanML3D, we directly cite their results reported in their original papers. To examine the generalization ability of the methods, we use HumanML3D as the training set and HuMMan and Kungfu as the test sets. To

Table 2. Quantitative results on Kungfu. Zero-Shot means that the models are evaluated directly on Kungfu after training on HumanML3D.

Methods	Kungfu(zero-shot)		
	FID↓	MM Dist↓	Diversity↑
Real	0.133 \pm .010	22.164 \pm .041	5.351 \pm .312
MAA [3]	—	—	—
T2M-GPT [44]	12.652 \pm .429	<u>25.826</u> \pm .041	5.702 \pm .428
MLD [5]	18.524 \pm .352	27.182 \pm .020	5.598 \pm .356
MotionDiffuse [45]	26.363 \pm .337	26.320 \pm .035	6.117 \pm .691
MDM [37]	16.396 \pm .466	26.280 \pm .095	5.468 \pm .590
FG-MDM	19.340 \pm .797	26.845 \pm .052	5.142 \pm .759
FG-MDM*	<u>15.892</u> \pm .567	25.325 \pm .035	<u>5.814</u> \pm .479

do so, we train TMR [27,28] using the HuMMan and Kungfu datasets to obtain a pair of text encoder and motion encoder for calculating the MM Dist metric. For SOTA methods, we apply their released pre-trained models on HumanML3D to HuMMan and Kungfu to evaluate their zero-shot generation performance. Since MAA [3] does not release the pre-trained model, we cannot test its zero-shot generation performance.

When evaluated on the test set of HumanML3D, all five methods achieve state-of-the-art performance. For FG-MDM, the ChatGPT-Refined fine-grained textual description doesn't match the manually annotated textual description well. Therefore, under within-dataset settings, our model does not exceed those SOTA models on HumanML3D. However, on the HuMMan and Kungfu datasets, FG-MDM captures most of the best and second-best results. Note that the size of our training dataset is much smaller than some SOTA methods like [3], but we still demonstrate solid zero-shot capabilities.

In addition, we provide some qualitative results to let readers intuitively feel the superiority of our method. In Fig. 3, we show motions generated by MDM [37], MLD [5] and our FG-MDM. Note that for all three methods, we use models trained on HumanML3D to generate motions. In comparison, our method generates motions more consistent with the details described in the fine-grained textual descriptions. This shows that our divide-and-conquer method works. Motion generation models require clear and specific conditions to generate the motions needed.

4.3 Ablation Study

To validate our contribution, we conduct two ablation studies. As shown in Table 3, the first row shows our baseline. The first study examines the contribution of ChatGPT-Generated fine-grained texts, which is performed by replacing the original short text with the fine-grained description. The improvement can be said to be huge. We cleverly utilize the powerful reasoning capabilities of

Table 3. Ablation study results on HumanML3D and KIT. “Fine-Grained” denotes using ChatGPT-Generated fine-grained descriptions. “Part” represents adopting part tokens. Note that the models are trained on HumanML3D.

Fine-Grained	Part	HumanML3D			KIT		
		FID↓	MM Dist↓	Diversity↑	FID↓	MM Dist↓	Diversity↑
		4.363	7.298	8.432	16.372	10.502	8.758
✓		1.050	6.778	9.509	0.549	9.826	10.829
✓	✓	0.663	5.649	9.476	0.344	9.352	10.707

LLMs and let them help our generative model better understand the nature of text conditions, bringing a leap to the zero-shot performance. The second study checks the contribution of part tokens when fine-grained descriptions are used. As observed, a reasonable framework also improves the quality of generated motions. However, perhaps more conditions bring more constraints to generations, leading to a decrease in diversity. But this drop is acceptable. So, we finally adopt the design of part tokens.

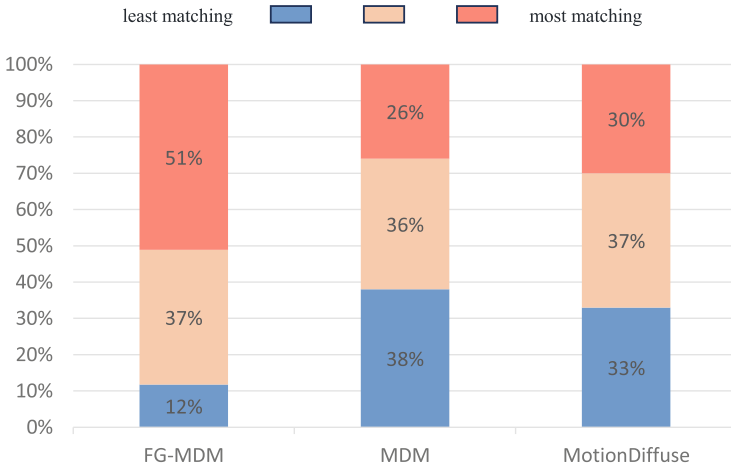


Fig. 4. User study results. For each method, a color bar ranging from blue to red represents the percentage of text-to-motion match levels, with blue indicating the least match and red indicating the most match. (Color figure online)

4.4 User Study

To further examine FG-MDM’s generalization capability, we conduct a user study to evaluate the quality of motions generated by our model based on human visual perception. We customize a total of 40 textual descriptions beyond the distribution of the dataset. With these descriptions, we generate motions by using

MDM [37], MotionDiffuse [45], and our FG-MDM. We then recruit 10 users for the study. In each question, participants are asked to rate the degree of matching between the generated motion and the textual description on a scale of 0 to 2. The results are given in Fig. 4. Apparently, FG-MDM matches texts much better in generating motions beyond the distribution of the dataset than the other two methods. Nearly half of the generated motions get the highest score. In contrast, MDM and MotionDiffuse perform poorly. Most of the generated motions are not satisfactory.

5 Conclusion

In this study, we used LLMs to perform fine-grained paraphrasing on the textual annotations of HumanML3D and KIT. With these fine-grained descriptions, we explored a Fine-Grained Human Motion Diffusion Model. It utilizes fine-grained descriptions of different body parts to guide the training of a diffusion model. This enables it to learn the essence of motions and thus generate motions beyond the distribution of training datasets. In the future, we would like to improve the quality of fine-grained annotations of human motions. Having high-quality text labels will greatly promote research on human motion generation.

References

1. Ahn, H., Ha, T., Choi, Y., Yoo, H., Oh, S.: Text2action: generative adversarial synthesis from language to action. In: IEEE International Conference on Robotics and Automation, pp. 5915–5920. IEEE (2018)
2. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: SINC: spatial composition of 3d human motions for simultaneous action generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9984–9995 (2023)
3. Azadi, S., Shah, A., Hayes, T., Parikh, D., Gupta, S.: Make-an-animation: large-scale text-conditional 3D human motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
4. Cai, Z., et al.: HuMMan: multi-modal 4D human dataset for versatile sensing and modeling. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13667, pp. 557–577. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20071-7_33
5. Chen, X., et al.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18000–18010 (2023)
6. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. *Adv. Neural. Inf. Process. Syst.* **34**, 8780–8794 (2021)
7. Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci.* **120**(30), e2305016120 (2023)
8. Guo, C., et al.: Generating diverse and natural 3D human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5152–5161 (2022)
9. Guo, C., et al.: Action2motion: conditioned generation of 3D human motions. In: Proceedings of the ACM International Conference on Multimedia, pp. 2021–2029 (2020)

10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. [arXiv:2207.12598](https://arxiv.org/abs/2207.12598) (2022)
12. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: AvatarClip: zero-shot text-driven generation and animation of 3D avatars. *ACM Trans. Graph.* **41**(4), 1–19 (2022)
13. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: MotionGPT: human motion as a foreign language. *Adv. Neural Inf. Process. Syst.* **36** (2024)
14. Kalakonda, S.S., Maheshwari, S., Sarvadevabhatla, R.K.: Action-GPT: leveraging large-scale language models for improved and generalized action generation. In: *IEEE International Conference on Multimedia and Expo* (2023)
15. Kim, J., Kim, J., Choi, S.: Flame: free-form language-based motion synthesis & editing. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 8255–8263 (2023)
16. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: AI choreographer: music conditioned 3D dance generation with AIST++. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13401–13412 (2021)
17. Liang, H., et al.: OMG: towards open-vocabulary motion generation via mixture of controllers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024)
18. Lin, A.S., Wu, L., Corona, R., Tai, K., Huang, Q., Mooney, R.J.: Generating animated videos of human activities from natural language descriptions. *Learning* **2018**(1) (2018)
19. Lin, J., et al.: Motion-X: a large-scale 3D expressive whole-body human motion dataset. *Adv. Neural Inf. Process. Syst.* (2023)
20. Lin, J., et al.: Being comes from not-being: open-vocabulary text-to-motion generation with wordless training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23222–23231 (2023)
21. Lin, X., Amer, M.R.: Human motion modeling using DVGANs. [arXiv:1804.10652](https://arxiv.org/abs/1804.10652) (2018)
22. Lu, S., et al.: HumanTOMATO: text-aligned whole-body motion generation. [arXiv:2310.12978](https://arxiv.org/abs/2310.12978) (2023)
23. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: archive of motion capture as surface shapes. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5442–5451 (2019)
24. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2891–2900 (2017)
25. Ouyang, L., et al.: Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022)
26. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10985–10995 (2021)
27. Petrovich, M., Black, M.J., Varol, G.: TEMOS: generating diverse human motions from textual descriptions. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) *ECCV 2022*. LNCS, vol. 13682, pp. 480–497. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20047-2_28
28. Petrovich, M., Black, M.J., Varol, G.: TMR: text-to-motion retrieval using contrastive 3D human motion synthesis. In: *International Conference on Computer Vision* (2023)

29. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. *Big data* **4**(4), 236–252 (2016)
30. Raab, S., et al: MoDI: unconditional motion synthesis from diverse data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13873–13883 (2023)
31. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*, pp. 8748–8763. PMLR (2021)
32. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. **1**(2), 3 (2022). [arXiv:2204.06125](https://arxiv.org/abs/2204.06125)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695 (2022)
34. Shi, M., et al.: MotioNet: 3D human motion reconstruction from monocular video with skeleton consistency. *ACM Trans. Graph.* **40**(1), 1–15 (2020)
35. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: *International Conference on Machine Learning*, pp. 2256–2265. PMLR (2015)
36. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: MotionLIP: exposing human motion generation to clip space. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) *ECCV 2022*. LNCS, vol. 13682, pp. 358–374. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20047-2_21
37. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: *The Eleventh International Conference on Learning Representations* (2023)
38. Tian, Y., Zhang, H., Liu, Y., Wang, L.: Recovering 3D human mesh from monocular images: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 15406–15425 (2023)
39. Tseng, J., Castellon, R., Liu, K.: EDGE: editable dance generation from music. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 448–458 (2023)
40. Vaswani, A., et al.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017)
41. Yao, W., Zhang, H., Sun, Y., Tang, J.: STAF: 3D human mesh recovery from video with spatio-temporal alignment fusion. *IEEE Trans. Circ. Syst. Video Technol.*, 1 (2024). <https://doi.org/10.1109/TCSVT.2024.3410400>
42. Zhang, H., et al.: PyMAF-X: towards well-aligned full-body model regression from monocular images. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 12287–12303 (2023)
43. Zhang, H., et al.: PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11446–11456 (2021)
44. Zhang, J., et al.: T2M-GPT: generating human motion from textual descriptions with discrete representations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023)
45. Zhang, M., et al.: MotionDiffuse: text-driven human motion generation with diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 4115–4128 (2024)
46. Zhang, Y., et al.: MotionGPT: finetuned LLMs are general-purpose motion generators. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 7368–7376 (2024)

47. Zhao, R., Su, H., Ji, Q.: Bayesian adversarial human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6225–6234 (2020)
48. Zhu, W., et al.: Human motion generation: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 2430–2449 (2023)

Author Index

A

Abdelhalim, Ibrahim 213
Abrol, Vinayak 367
Ahuja, Chirag 240
Ali, Khadiga M. 76
Azam, Mohamed T. 31

B

Balaha, Hossam Magdy 31, 76
Banerjee, Paromita 154
Baraas, Rigmor 168
Bart, Yakov 303
Ben-Artzi, Gil 17
Bhattacharyya, Riddhasree 124
Bhavsar, Arnav 240
Bolelli, Federico 108

C

Chanda, Sukalpa 224
Che, Hekuangyi 399, 415
Chen, Si 383
Chen, Tianxiang 46
Chowdhury, Pathikreet 1
Chung, Haejun 319

D

Daragma, Feras 17
Das, Sujit 58
Das, Surochita Pal 124
Deng, Yunteng 183
Dey, Shramana 124
Du, Chenxi 256
Duggal, Bhanu 154
Dutta, Pallabi 124

E

El-Baz, Ayman 31, 76, 213
Elsharkawy, Mohamed 213

F

Ficarra, Elisa 108

G

Gayen, Soumyajit 138
Ghazal, Mohammed 31, 76, 213
Gilson, Stuart 168
Goel, Anoushkrit 240
Gondim, Dibson D. 31
Grana, Costantino 108
Gupta, Akshat 367

H

Hao, Wanting 351
Hu, Junlin 183
Huang, Chao-Wang 431

J

Jang, Ikbeom 319
Jethava, Rutvik Narendrabhai 154
Jha, Ranjeet Ranjan 240
Jia, Zhenhong 92
Jiang, Nanfeng 383
Joshi, Ankita 240

K

Kaplun, Dmitrii 138
Keserwani, Prateek 198
Kulyabin, Mikhail 168
Kumar, Ajay 273

L

Li, Binyang 183
Li, Hang 399, 415
Li, Jiamao 399, 415
Li, Shuyi 288
Li, Yan 183
Lin, Minjin 415
Liu, Jing 256
Liu, Xiaomo 367

Luan, Lingfei 303
 Lumetti, Luca 108
 Luo, Chuanchen 446

M

Mahato, Anuradha 154
 Mahmoud, Ali 76, 213
 Mahpod, Shahar 17
 Maier, Andreas 168
 Mishra, Aakansha 198
 Mistry, Akshikumar 31
 Mitra, Sushmita 124
 Moon, Junho 319

N

Nadmid, Namuunaa 213
 Nigam, Aditya 240

O

Ostadabbas, Sarah 303

P

Pal, Umapada 224
 Palaiahnakote, Shivakumara 224
 Paul, Angshuman 154
 Pedersen, Hilde R. 168
 Peng, Junran 446
 Pipoli, Vittorio 108

R

Rajendiran, Vikram N. 198
 Raman, Rajiv 124
 Rather, Sajad Ahmad 58
 Ray, Amartya 138
 Ren, Dayong 92
 Roy, Ayush 224
 Roy, Partha Pratim 58

S

Sarkar, Ram 138
 Senapati, Ashok K. 198
 Shah, Sameena 367
 Shi, Fei 92
 Shi, Wenjun 399, 415
 Shi, Xu 446
 Sindel, Aline 168
 Singh, Bipanjit 240
 Singh, Richa 154

Srivastava, Gargi 1
 Su, Yanfei 383
 Sun, Yunlian 446

T

Thakur, Anshul 367

V

Vatsa, Mayank 154
 Vyas, Ritesh 273

W

Wang, Chien-Yao 431
 Wang, Da-Han 383
 Wang, Jia-Ching 431
 Wang, Jianyi 92
 Wang, Juelin 183
 Wang, Lei 399, 415
 Wang, Liejun 335
 Wang, Yuqi 288
 Wang, Ziyang 46
 Wu, Xiuying 256

X

Xie, Wen 303
 Xu, Miaomiao 335

Y

Yan, Yan 383
 Yang, Hao 288
 Yao, Wei 446
 Ye, Qinxian 383
 Ye, Zi 46
 Yu, Yinfeng 335
 Yuan, Chun 351
 Yuan, Rui 351

Z

Zhang, Bob 288
 Zhang, Dexin 256
 Zhang, Feifei 92
 Zhang, Guanghui 399, 415
 Zhang, Hongwen 446
 Zhang, Hui 256
 Zhang, Jiang 335
 Zhu, Dongchen 399, 415
 Zhu, Yanjun 303
 Zou, Hang 256