

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

LNCS 15333

# Pattern Recognition

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

33 Part XXXIII



# Lecture Notes in Computer Science

15333

## 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 XXXIII

*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 978-3-031-80135-8

ISBN 978-3-031-80136-5 (eBook)

<https://doi.org/10.1007/978-3-031-80136-5>

© 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-Elmageed	University of Southern California, USA
Maya Aghaei	NHL Stenden University, Netherlands
Alireza Alaei	Southern Cross University, Australia
Rajagopalan N. Ambasmudram	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  
 Anwesha 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 Otroschi 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  
Siladitty 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 XXXIII

PS-StyleGAN: Illustrative Portrait Sketching Using Attention-Based Style Adaptation .....	1
<i>Kushal Kumar Jain, J. Ankith Varun, and Anoop Namboodiri</i>	
Few-Shot and Portable 3D Manufacturing Defect Tracking with Enterprise Digital Twins Based Mixed Reality .....	17
<i>Yiyong Tan, Bhaskar Banerjee, and Rishi Ranjan</i>	
Augmented Reality-Assisted Environment for Medical Education: An Experience of Interactive and Immersive Learning .....	33
<i>Vikas Puthannadathil Reghunatha Kumar, Anurag Kujur, Bishnu Ganguly, Santosh Kumar Behera, and Ajaya Kumar Dash</i>	
P2A: Transforming Proposals to Anomaly Masks .....	48
<i>Huachao Zhu, Zhichao Sun, Zelong Liu, and Yongchao Xu</i>	
The 2D Shape Equipartition Problem Under Minimum Boundary Length .....	64
<i>Costas Panagiotakis</i>	
Random Frame: a Data Augmentation for Glass Detection .....	80
<i>Yiming Liang and Hiroshi Ishikawa</i>	
PolypSegDiff: Dynamic Multi-scale Conditional Diffusion Model for Polyp Segmentation .....	94
<i>Xiaogang Du, Yipeng Jiao, Tao Lei, Xuejun Zhang, Yingbo Wang, and Asoke K. Nandi</i>	
Exploiting Text-Image Latent Spaces for the Description of Visual Concepts ...	109
<i>Laines Schmalwasser, Jakob Gawlikowski, Joachim Denzler, and Julia Niebling</i>	
Back to Supervision: Boosting Word Boundary Detection Through Frame Classification .....	126
<i>Simone Carnemolla, Salvatore Calcagno, Simone Palazzo, and Daniela Giordano</i>	
Semi-supervised Cross-Lingual Speech Recognition Exploiting Articulatory Features .....	141
<i>Xinmei Su, Xiang Xie, Chenguang Hu, Shu Wu, and Jing Wang</i>	

Collaborative Transformer Decoder Method for Uyghur Speech Recognition in-Vehicle Environment .....	154
<i>Jiang Zhang, Liejun Wang, Yinfeng Yu, Miaomiao Xu, and Alimjan Mattursun</i>	
Audio-Visual Wake-up Word Spotting Under Noisy and Multi-person Scenarios .....	170
<i>Cancan Li, Fei Su, and Juan Liu</i>	
Missing Person Recognition Algorithms Based on Image Captioning and Visual Grounding .....	185
<i>Ayeong Jeong, Yeongju Woo, Han-young Kim, Gayun Suh, Chae-yeon Heo, Yeong-jun Cho, and Hieyong Jeong</i>	
Contrastive and Restorative Pre-Training for Medical VQA .....	198
<i>Vasudha Joshi, Pabitra Mitra, and Supratik Bose</i>	
MRCI: Multi-range Context Interaction for Boundary Refinement in Image Segmentation .....	211
<i>Yaqiang Wu, Wanjun Lyu, Xianchen Liang, Qinghua Zheng, Jin Wei, and Lianwen Jin</i>	
Cross Lingual Synopsis Generation in English, Dutch, Vietnamese, Indonesian, Russian, Portuguese, Korean, Hindi and French .....	227
<i>Sreejata Banerjee, Aditya Sadhukhan, Arijit Das, and Diganta Saha</i>	
<b>Author Index .....</b>	<b>243</b>



# PS-StyleGAN: Illustrative Portrait Sketching Using Attention-Based Style Adaptation

Kushal Kumar Jain<sup>(✉)</sup>, J. Ankith Varun, and Anoop Namboodiri

IIIT-Hyderabad, Gachibowli, India

{kushal.kumar, ankith.varun}@research.iiit.ac.in, anoop@iiit.ac.in

**Abstract.** Portrait sketching involves capturing identity specific attributes of a real face with abstract lines and shades. Unlike photo-realistic images, a good portrait sketch generation method needs selective attention to detail, making the problem challenging. This paper introduces **Portrait Sketching StyleGAN (PS-StyleGAN)**, a style transfer approach tailored for portrait sketch synthesis. We leverage the semantic  $W+$  latent space of StyleGAN to generate portrait sketches, allowing us to make meaningful edits, like pose and expression alterations, without compromising identity. To achieve this, we propose the use of Attentive Affine transform blocks in our architecture, and a training strategy that allows us to change StyleGAN's output without finetuning it. These blocks learn to modify style latent code by paying attention to both content and style latent features, allowing us to adapt the outputs of StyleGAN in an inversion-consistent manner. Our approach uses only a few paired examples ( $\sim 100$ ) to model a style and has a short training time. We demonstrate PS-StyleGAN's superiority over the current state-of-the-art methods on various datasets, qualitatively and quantitatively.

**Keywords:** Portrait Generation · Stylization · StyleGAN

## 1 Introduction

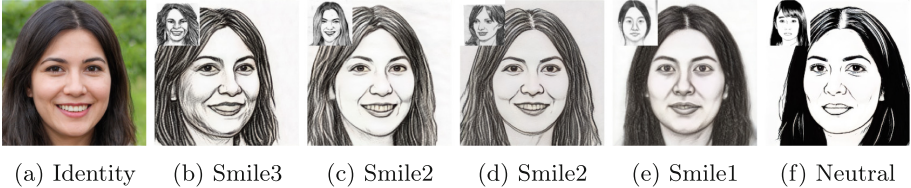
Drawing portrait sketches is an intricate but timeless form of artistic expression. It requires one to use minimalistic elements, such as lines, to encapsulate the distinctive features and the overall essence of an individual's identity. A lot of research has been done to understand how humans perceive 3D shapes through rough sketches or simple line drawings and why they effectively represent complex concepts like identity [4, 19]. While some theories exist, it is still not very well known as to how artists choose the lines that they draw [20, 46]. Hence

K.K. Jain and J.A. Varun—Equal contribution.

---

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/978-3-031-80136-5\\_1](https://doi.org/10.1007/978-3-031-80136-5_1).

the process of generation of such sketches remain a manual creation and time-consuming one. Many Non-Photorealistic Rendering methods tried to solve this artistic challenge [3, 8, 32, 37, 44]. However, they rely on ground truth geometry, which is noisy near detailed parts of the face like eyes, nose and lips. Humans are particularly sensitive to details in these regions as we have dedicated neural pathways [50] for face detection and identification (Fig. 1).



**Fig. 1.** Outputs of PS-StyleGAN for different sketching styles (inset) in specified poses and expressions while maintaining the input identity. A model trained on FS2K dataset was used for (b)-(d), while CUHK and APDrawing were used for the models in (e) and (f).

Deep learning [47] based approaches like style transfer [15] and image to image translation [24, 60] have been very successful in sketch generation. Innovations in style transfer [23, 38, 51] have made the generation process faster and more reliable. However, these methods only perform well for global texture transformations and do not consider local details, abstracting out crucial elements like eyes and lips. Following the development of Generative Adversarial Networks (GANs) [16] and cGANs [36], Isola *et al.* proposed a novel method for general image to image translation using cGANs [24, 60]. Even though training such generators is notoriously difficult, modern image to image translation methods [5] have shown impressive results and tremendous potential.

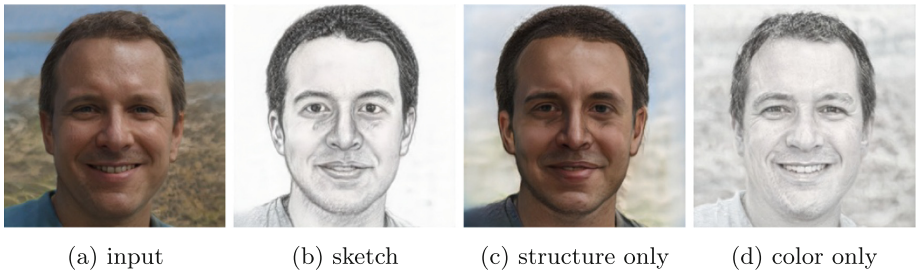
More recently, researchers have used a pretrained StyleGAN [27, 28] along with encoders that invert a given image into StyleGAN’s latent space to tackle the problem of general image to image translation [7]. Its highly semantic  $W+$  latent space allows one to make meaningful edits to the final output, like changing pose, facial expression or emotion, without affecting the identity. However, for portrait sketch generation, incorporating the sketch style into the original StyleGAN poses a significant challenge. It carries the risk of perturbing and changing the behaviour of the latent space of the pre-trained StyleGAN, making latent editing difficult. Yet a lot of works [6, 22, 30, 54, 55], have tried to solve the issue with different approaches. Some methods [54, 55] divide the latent space into dual spaces and others use attention [30] for better mapping features in the latent space.

DualStyleGAN [55] achieves portrait style transfer by disentangling the spatial resolution layers of StyleGAN to perform independent structure and color transfer between domains. They propose a ResBlock [18] based feature statistics alignment module using AdaIN [23] that incorporates *structure control* over

the coarse and middle layers of StyleGAN. Training their ResBlock does not change the latent distribution and thus it allows semantic editing. However, DualStyleGAN’s style blending might result in significant loss of identity. In our experiments we observed some extent of structure and color entanglement across all layers of StyleGAN, especially in the middle layers. Hence, it is difficult to decouple structure and color transformations without losing desirable artistic characteristics like pencil strokes and shading, see Fig. 2. Lastly, DualStyleGAN relies on a pre-training process to learn structure transfer in the source domain which is quite time consuming.

To this end, we propose Portrait Sketching StyleGAN (PS-StyleGAN), which converts real face photos into a portrait sketch while offering the semantic editability of StyleGAN without the need to finetune it. We use our novel attention [52] to bring the generator frozen. These blocks help us simulate the behaviour of a finetuned StyleGAN. We discard any form of structure transfer so as to ensure identity preservation and adopt a progressive training strategy to achieve a rapid but smooth domain transfer. We also run our model on different datasets to show that our model is inversion consistent. Our main contributions are:

1. We propose PS-StyleGAN, which can generate expressive portrait sketches from a photo-realistic face image. Specifically, our method can learn complex hairstyles and generate perfect eyes, nose and lips while preserving the subtleties of an artist’s style. Furthermore, our model converges quickly and can be trained on relatively small datasets.
2. We introduce a novel Attentive Affine transformation for better-transforming style latent codes based on style examples.
3. We perform experiments, conduct a user study and run ablations on various datasets to show the effectiveness of our method.



**Fig. 2.** Results of DualStyleGAN trained on CUHK [53] dataset. The generated sketch (b) is a result of complete structure and color transfer. Structure transfer (c) results in considerable loss of identity while color transfer (d) does not yield stylization.

## 2 Related Works

### 2.1 Image to Image Translation

Image-to-image translation techniques aim to learn a mapping function that can convert an input image from one domain into the corresponding image in another domain. This approach was initially introduced by Isola *et al.* using conditional GANs [24] and has since seen significant development. Recent methods like [14] use Dynamic Normalization (DySPADE) in the generator architecture along with depth maps to supervise the generation with encouraging results. An unsupervised version of Pix2Pix called the CycleGAN [29, 33, 60] sparked the creation of some fascinating sketch generation methods. In AP-Drawing GAN, Yi *et al.* [57] used dedicated GANs to generate difficult-to-sketch features like eyes, nose and lips. FSGAN [11] extends their approach and introduces a new dataset called FS2K, which has three styles and paired sketch examples. We use this dataset for training and comparison with other methods. New approaches like [5, 25], use CLIP [39] embeddings. In [5] the authors use CLIP along with a geometry-preserving loss to achieve line drawings that respect the scene’s geometry. These methods require training the generator, which is difficult and necessitates large datasets, which is not feasible for face sketches.

Diffusion models [10, 21] have made significant progress in text-guided image generation [40, 42, 45], in the past few years. Personalised sketch generation in the context of diffusion models has been achieved by either finetuning the generator itself [43] or by learning personalised word or image embeddings for the generator [13, 56], or by using ID embeddings as condition [25, 31]. Our method instead relies on a StyleGAN generator trained on realistic human faces [27]. We modify the generator’s output by learning crucial aspects of each style using only a few examples.

### 2.2 StyleGAN Latent Space Inversion

The exceptional image quality and semantic richness of StyleGAN [27, 28] has made it very attractive for directed image generation. GANs synthesize images by sampling a vector (latent code) from the latent space distribution. GAN inversion tackles the problem of finding the latent code that best recreates a given image. This can be done by direct optimization, learning encoders or a mix of both [1, 7]. For latent editing, some methods take the supervised approach by finding latent directions for labelled attributes, while others take a more unsupervised approach [17, 49].

Although there are many approaches for latent space manipulation of realistic images, they do not work for stylized generators as it would change the latent distribution, making latent manipulation inconsistent [48]. JoJoGAN [6] achieves style transfer by training a new mapper for every finetuned StyleGAN, but it comes at the cost of identity. DualStyleGAN [55] solves this issue by disentangling the spatial resolution layers of StyleGAN to perform independent structure and color transfer between domains. They propose a ResBlock [18]

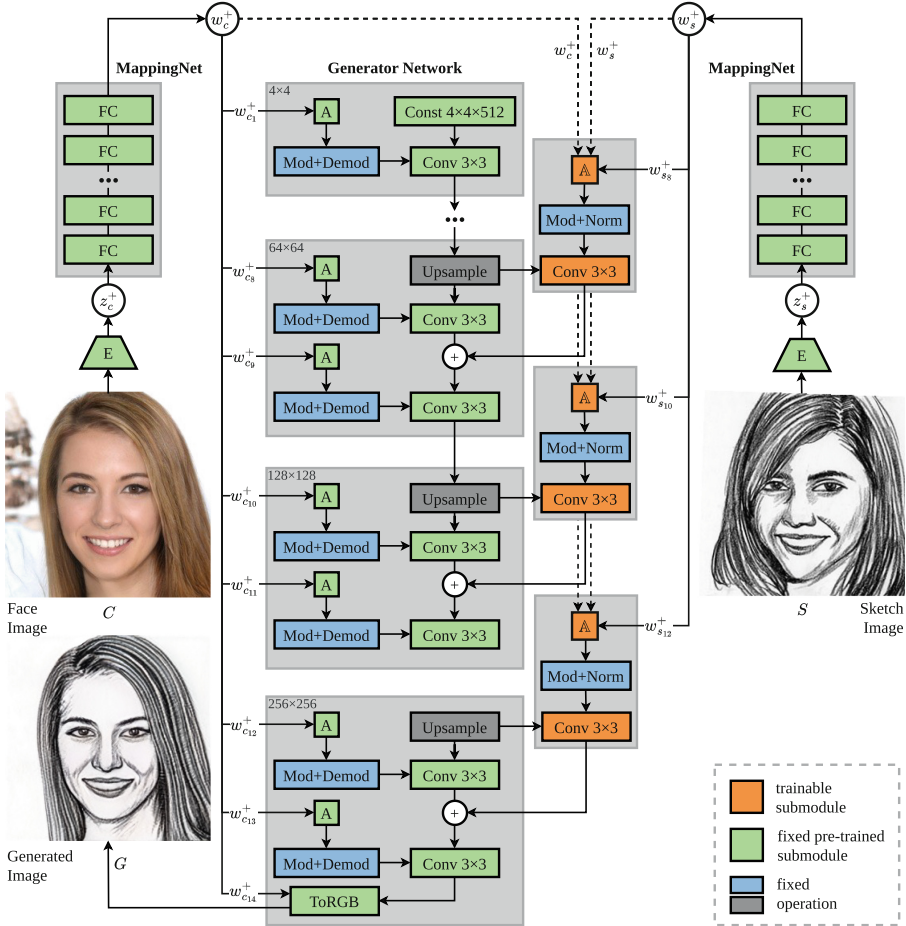
based feature statistics alignment module using AdaIN [23] that incorporates *structure control* over the coarse and middle layers of StyleGAN. Training their ResBlock does not change the latent distribution and thus it allows semantic editing. Our method differs from their approach as we investigate how to better preserve content structure, as portrait sketches have well-defined lines that need spatial consistency. We use attention based style adaption blocks to smoothly transform the generative space by aligning it to the feature statistics of the style examples.

### 2.3 Attention in Latent Space Manipulation

Following the development of  $W+$  space [1] many methods have tried to find new latent spaces that can offer better reconstruction ability while retaining the editing ability. In StyleTransformer [22], the authors use cross and self attention layers to aid in the inversion task of  $W+$  latent space, showing that transformers can be a useful addition here. In TransStyleGAN [30], the authors introduce a new  $W++$  latent space by replacing the MLP layers in the mapper network with transformer layers, resulting in better reconstruction and editing abilities. DualStyleGAN [55] proposes splitting the latent space into dual spaces effectively disentangling style and content spaces, allowing existing editing approaches to work on stylized spaces. A similar approach dubbed TransEditor [54] also divides the latent space into  $P$  and  $Z$  spaces but crucially also uses cross attention based interaction module to correlate between the separated spaces. In [2], the authors find that editing the style codes in early stages of the generation process affects the structural properties of the image, resulting in artefacts in the final results. TransEditor mitigates this issue by increasing collaboration between the two spaces. In our approach we use attention based style adaption blocks to transform the style codes only in the later stages of the generation process.

## 3 Method

We propose an end-to-end method for facial sketch synthesis using our model PS-StyleGAN  $g'$ , whose architecture is outlined in Fig. 3. Given a content image  $C$  and sketch image  $S$  of a particular style  $\mathbb{S}$ , we invert both images to the  $Z+$  latent space of a pre-trained StyleGAN generator  $g$  using a pSp-based encoder  $E$  [41, 49]. We train the encoder  $E$  on  $256 \times 256$  resolution of FFHQ dataset [27] and modify it to embed face images to the  $Z+$  latent space, which is more resilient to background details than the standard  $W+$  space as observed in [48]. Using StyleGAN’s mapping network  $f$ , we transform them into latent codes  $w_c^+$  and  $w_s^+$ , respectively, in the shared  $W+$  latent space of  $g$ . Finally, we pass the latent codes through our novel synthesis network  $g'$  to obtain the generated image  $G$ , successfully capturing the style of  $\mathbb{S}$ . In the following sections, we give a detailed description of our model architecture and training procedure.



**Fig. 3.** An overview of our model architecture. We use a pretrained  $256 \times 256$  resolution StyleGAN2 [28] generator  $g$  fitted with three style adaptation blocks at the fine resolution layers. Each block consists of a novel Attentive Affine transform module (A) that predicts affine parameters from attention-weighted latent codes of  $S$  using supervision from  $w_c^+$  and  $w_s^+$ . These parameters are then used to modulate and normalize the spatial features of  $g$  at different scales to imbibe the style  $S$  into  $C$ .

### 3.1 Hierarchical Style Control in StyleGAN

As described in [27], the style blocks/layers of StyleGAN of different spatial resolutions controlled specific aspects of face generation. *Coarse layers* ( $4 \times 4$ – $8 \times 8$  resolution) affect high-level aspects such as pose, hair texture, face structure and accessories. *Middle layers* ( $16 \times 16$ – $32 \times 32$  resolution) generate smaller-scale features like eyes, smile, hairstyle, etc. *Fine layers* ( $64 \times 64$ – $256 \times 256$  resolution) mainly control the general color scheme and microstructure of the generated image.



To tackle the challenges pointed out in Sect. 1, we use attention-based style adaptation blocks in the fine layers of the generator network that perform feature transformations by considering both global and local style patterns. Each block consists of a novel Attentive Affine transform module ( $\mathbb{A}$ ) and StyleGAN’s modulative convolution layer, which provide instance-wise style conditioning to the content features. We choose to modulate just the fine layer features of the generator so as to preserve the overall structure of the content image. The adapted features are then fused with the original content features at each layer to allow a smooth transition of the generative space from the photo-realistic domain to the sketch domain. We show experimentally in Sect. 3.1 of the supplementary that the latent space of StyleGAN remains consistent, allowing us to manipulate sketches using methods designed for realistic images.

We use the following notation for subsequent analysis -  $w_{x_i}^+$  denotes the  $i^{\text{th}}$  segment of the latent code of an input image  $X$ ,  $F_i^X$  denotes the feature maps of  $X$  that go into the  $i^{\text{th}}$  convolution layer of the synthesis network and  $y_i^X = (y_{s,i}^X, y_{b,i}^X)$  denotes the corresponding affine parameters computed at that layer.

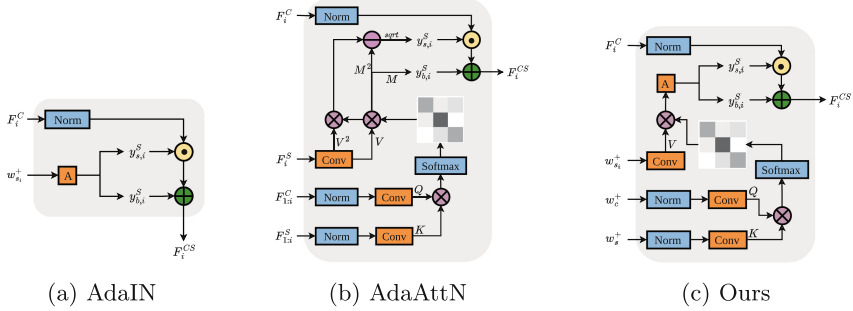
### 3.2 Paying Attention in Latent Space

Inspired by AdaAttN [34], we introduce *attentive affine* transformations to obtain improved affine parameters  $y_i^S$  at the fine layers, which encapsulate the complete feature distribution of the style image. These parameters are then used by the AdaIN operation to achieve style transfer. As shown in Fig. 4c, the style adaptation process works in three steps.

1. Computing attention maps with content and style latent codes  $w_c^+$  and  $w_s^+$ , respectively.
2. Calculating weighted segment of the style latent code and obtaining improved affine parameters  $y_{s,i}^S$  and  $y_{b,i}^S$  of the style features.
3. Adaptively normalizing the content features for instance-wise feature distribution alignment.

**Attention Map Generation:** Different from standard style transfer methods, we use the attention mechanism to measure the similarity between the content and style latent codes instead of the corresponding features themselves. Due to the highly disentangled nature of the  $W+$  latent space of StyleGAN, similarity in the latent space extrapolates well to that in the feature space. The relatively low dimensionality of the latent space keeps the model lightweight and cuts down on the computational costs of calculating attention maps. To compute the attention map  $A$  corresponding to the fine layer  $i$ , we formulate query ( $Q$ ), key ( $K$ ) and value ( $V$ ) as given below.

$$\begin{aligned} Q &= f(\text{Norm}(w_c^+)) \\ K &= g(\text{Norm}(w_s^+)) \\ V &= h(w_{s_i}^+) \end{aligned} \tag{1}$$



**Fig. 4.** (a) The structure of AdaIN [23] module used in StyleGAN [27]. (b) The structure of AdaAttN [34] module. (c) The structure of our proposed design showing *attentive affine* transform blocks. Here,  $A$  denotes a basic affine transform block consisting of a single trainable fully-connected layer and  $Norm$  denotes channel-wise mean-variance normalization.

where  $f$ ,  $g$ , and  $h$  are standard trainable  $1 \times 1$  convolution layers while  $Norm$  is the instance normalization operation carried out channel-wise. We compute attention map  $A$  as:

$$A = Softmax(Q^T \otimes K) \quad (2)$$

where  $\otimes$  represents matrix multiplication.

**Improved Affine Parameters:** In AdaAttN [34], applying the attention map to the style feature  $F_i^S$  is interpreted as observing a target style feature point as a distribution of all the weighted style feature points by attention. Then, statistical parameters are calculated from each distribution for subsequent modulation. In our case, the style latent code segment  $w_{s_i}^+$  is multiplied with the attention score matrix to represent it as a distribution of all style points in the latent space. We term this as *attention-weighted* latent code segment  $x_{s_i}^+ \in \mathbb{R}^{512}$  from which we learn improved affine transformations to get better representative affine parameters  $y_{s,i}^S$  and  $y_{b,i}^S$  as follows.

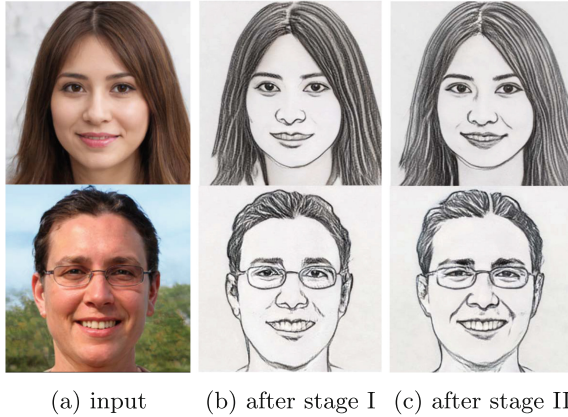
$$x_{s_i}^+ = V \otimes A^T \quad (3)$$

$$(y_{s,i}^S, y_{b,i}^S) = Affine(x_{s_i}^+) \quad (4)$$

where  $Affine$  is a learnable single fully-connected layer identical to the traditional StyleGAN’s affine transform. The output dimensionality of the layer is twice the number of feature maps on the corresponding spatial resolution of the generator.

**Adaptive Normalization:** Finally, we use the obtained affine parameters to modulate the normalized content feature map point-wise for each channel to generate the transformed feature map. Thus, the AdaIN operation in our case would become

$$F_i^{CS} = y_{s,i}^S \frac{F_i^C - \mu(F_i^C)}{\sigma(F_i^C)} + y_{b,i}^S \quad (5)$$



**Fig. 5.** Results after each stage of progressive transfer learning. At the end of stage I, the model converges to an average representative style as seen in (b) where the eyes, nose and mouth are sketched in a similar manner. Stage II widens the model’s generative space to capture subtle style variations resulting in better identity preservation as shown in (c).

The transformed feature maps  $F_i^{CS}$  go into a trainable convolution layer whose outputs are selectively fused with those of the fine layers of the pre-trained synthesis network  $g$  to complete the style adaptation process. We notice that omission of the mean affine parameter i.e.  $y_{b,i}^S$  during modulation does not affect the generated results. Therefore, like StyleGAN2 [28], we combine the modulation and convolution operation by scaling the convolution weights and effectively reduce the output dimensionality of the affine transform blocks.

To summarize, we perform feature statistics alignment using *attentive affine* transformations by generating attention-weighted latent code that better represents the target style feature distribution in the fine layers ensuring that middle and coarse layer features are not lost.

### 3.3 Training Strategy

We adopt a progressive transfer learning scheme using a pretrained StyleGAN to smoothly refine its generative space to align with the target style distribution  $\mathbb{S}$  comprising of limited samples. The scheme consists of two stages as illustrated in Fig. 5.

**Stage I - Domain Transfer:** Similar to fine-tuning, we seek to achieve a general transformation from the photo-realistic domain to the sketch domain defined by  $\mathbb{S}$ . We randomly generate a latent code  $Z+$  and sample a sketch image  $S$  and its corresponding style latent code  $z_s^+$ . Using StyleGAN’s mapping network  $f$ , we obtain the  $W+$  latent space embeddings for the content and style images as  $w^+ = f(z^+)$  and  $w_s^+ = f(z_s^+)$ , respectively. Subsequently, we pass the latent codes through our synthesis network  $g'$  to obtain the generated sketch as

$G = g'(w^+, w_s^+)$ . Following standard style transfer practices, we employ a style loss to fit the style of the generated sketch  $G$  to  $S$  which is given by

$$\mathcal{L}_{\text{sty}} = \lambda_{\text{CX}}\mathcal{L}_{\text{CX}}(G, S) + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}}(G, S) \quad (6)$$

where  $\mathcal{L}_{\text{CX}}$  denotes contextual loss [35] and  $\mathcal{L}_{\text{FM}}$  denotes feature matching loss [23]. To preserve the content features we use an identity loss [9] between  $G$  and the reconstructed content image  $g(w^+)$  thus constituting a content loss as follows.

$$\mathcal{L}_{\text{cont}} = \lambda_{\text{ID}}\mathcal{L}_{\text{ID}}(G, g(w^+)) \quad (7)$$

where  $\mathcal{L}_{\text{ID}}$  represents identity loss. Adding the standard StyleGAN adversarial loss  $\mathcal{L}_{\text{adv}}$ , our complete objective function takes the form of

$$\min_G \max_D \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{sty}} + \mathcal{L}_{\text{cont}}$$

**Stage II - Conditional Refinement:** Stage I transforms StyleGAN’s generative space to a narrow domain, failing to capture the diversity of styles contained in  $\mathbb{S}$  as shown in Fig. 5b. We use paired data of ground truth sketches and their photo-realistic counterparts as conditional supervision to broaden the generative domain. Given a sketch image  $S$  and corresponding photo  $P$ , we get the  $W+$  latent space embeddings as  $w_s^+ = f(E(S))$  and  $w_p^+ = f(E(P))$ , and use them to obtain the generated sketch  $G = g'(w_p^+, w_s^+)$ . In addition to the losses used in stage I, we use perceptual loss [26] for  $G$  to reconstruct  $S$  thereby learning a varied set of style specific transformations. We also introduce a regularization term in  $\mathcal{L}_{\text{cont}}$  which is the  $L_2$  norm of the convolution weights comprising our style adaptation blocks. Therefore, Eq. 7 changes to

$$\mathcal{L}_{\text{cont}} = \lambda_{\text{ID}}\mathcal{L}_{\text{ID}}(G, g(w_p^+)) + \lambda_{\text{reg}}\|W\|_2 \quad (8)$$

where  $W$  represents the weight matrices of the trainable convolution layers. This regularization term controls the degree of style adaptation and helps prevent overfitting. Thus, the objective function modifies to

$$\min_G \max_D \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \mathcal{L}_{\text{sty}} + \mathcal{L}_{\text{cont}}$$

## 4 Experiments

In this section, we assess the effectiveness of our proposed method by conducting comprehensive evaluations, which include qualitative and quantitative comparisons.

**Datasets:** We carry out our experiments on the FS2K dataset [11], which stands as the most extensive publicly available FSS (Face Sketch Synthesis) dataset to date. This dataset comprises a substantial collection of 2,104 photo-sketch pairs, featuring a wide diversity of image backgrounds, skin tones, sketch styles, and lighting conditions. These sketches are classified into three distinct artistic styles.

**Table 1.** Quantitative comparison of AdaAttN [34], FSGAN [11], HIDA [14] and DualStyleGAN [55] with our method based on SCOOT, LPIPS, FSIM and ID loss. Our method shows considerably better SCOOT and ID loss values indicating more visually appealing and recognizable results.

Method	SCOOT $\uparrow$	LPIPS $\downarrow$	FSIM $\uparrow$	ID $\downarrow$
HIDA	0.4433	0.3214	0.3660	0.0241
FSGAN	0.3621	0.2890	0.3692	0.0424
AdaAttN	0.4670	0.2600	0.3806	0.0233
DualStyleGAN	0.4490	0.3012	0.3631	0.0247
Ours	<b>0.5603</b>	<b>0.2303</b>	<b>0.4283</b>	<b>0.0206</b>

We also use the CUHK dataset [53], which comprises mostly of asian faces, to measure our method against DualstyleGAN, a technique that introduces a bias of shape characteristics within the results. We further experiment with AP-Drawing dataset [57] to evaluate our method’s ability to generalize and adapt to challenging sketching scenarios.

**Comparison Methods:** We compare our method to other state of the art methods that have shown good performance in facial sketch synthesis, like HIDA [14], FSGAN [11], DualStyleGAN [55] and AdaAttN [34].

#### 4.1 Quantitative Analysis

To quantitatively compare our method with others, we utilize four performance metrics: Learned Perceptual Image Patch Similarity (LPIPS) [59], Structure Co-Occurrence Texture (SCOOT) [12], Feature Similarity Measure (FSIM) [58] and ID loss [9]. Lower LPIPS and ID loss value suggests a more realistic synthesized sketch, while higher SCOOT and FSIM values indicate better similarity with artist-drawn sketches. We present the average SCOOT, LPIPS, FSIM and ID loss values across all test samples in Table 1. More details on quantitative evaluations can be found in supplementary material.

#### 4.2 Qualitative Analysis

Visually comparing our PS-StyleGAN with leading methods, namely FSGAN [11], HIDA [14], DualStyleGAN [55], and AdaAttN [34], we observe that our method excels in rendering eyes and lips, showcasing sharper details and enhanced realism, see Fig. 6. Our results are visually most similar to DualStyleGAN but their method also learns shape biases in the dataset hence affecting recognizability. DualStyleGAN often changes the gaze direction and shape of lips too. The Attentive Affine transform blocks in PS-StyleGAN contribute to a superior balance between artistic expression and accuracy, resulting in more visually appealing and faithful representations of facial features.



**Fig. 6.** Comparison of our method with other state of the art methods on the 3 styles (inset) of FS2K: style 1 (row 1), style 2 (row 2), style 3 (row 3). From left to right : Input identity image, Ours, DualStyleGAN [55], HIDA [14], FSGAN [11], AdaAttN [34].

## 5 Conclusion

We introduced **Portrait Sketching StyleGAN (PS-StyleGAN)**, an approach tailored specifically for the intricate color transformation demands in portrait sketch synthesis. Leveraging the semantic  $W+$  latent space of StyleGAN, our method not only generates portrait sketches but also allows meaningful edits, such as pose and expression alterations, while preserving identity. The incorporation of Attentive Affine Transform blocks, fine-tuned through extensive experimentation, allows us to adapt StyleGAN outputs in an inversion-consistent manner by considering both content and style latent features. The model demonstrates efficacy with minimal paired examples (approximately 100) and boasts a short training time, contributing to its practical applicability. However, our method may be susceptible to data bias, and performance could vary across datasets. Additionally, one noteworthy limitation is the current inability to generate realistic accessories in the synthesized sketches. Future work could focus on addressing these limitations to enhance the utility of the proposed PS-StyleGAN further.



## References

1. Abdal, R., Qin, Y., Wonka, P.: Image2StyleGAN: how to embed images into the StyleGAN latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
2. Alharbi, Y., Wonka, P.: Disentangled image generation through structured noise injection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5133–5141 (2020). <https://api.semanticscholar.org/CorpusID:216553495>
3. Berger, I., Shamir, A., Mahler, M., Carter, E.J., Hodgins, J.K.: Style and abstraction in portrait sketching. *ACM Trans. Graph. (TOG)* **32**, 1 – 12 (2013). <https://api.semanticscholar.org/CorpusID:17238299>
4. Biederman, I., Ju, G.: Surface versus edge-based determinants of visual recognition. *Cogn. Psychol.* **20**, 38–64 (1988). <https://api.semanticscholar.org/CorpusID:14269563>
5. Chan, C., Durand, F., Isola, P.: Learning to generate line drawings that convey geometry and semantics. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7905–7915 (2022). <https://api.semanticscholar.org/CorpusID:247628105>
6. Chong, M.J., Forsyth, D.A.: JoJoGAN: one shot face stylization. *arXiv arXiv:2112.11641* (2021). <https://api.semanticscholar.org/CorpusID:245385527>
7. Collins, E., Bala, R., Price, B., Süssstrunk, S.: Editing in style: uncovering the local semantics of GANs. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5770–5779 (2020)
8. DeCarlo, D., Finkelstein, A., Rusinkiewicz, S., Santella, A.: Suggestive contours for conveying shape. In: *ACM SIGGRAPH 2003 Papers* (2003). <https://api.semanticscholar.org/CorpusID:1485904>
9. Deng, J., Guo, J., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4685–4694 (2018)
10. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems* (2021). <https://openreview.net/forum?id=AAWuCvzaVt>
11. Fan, D.P., Huang, Z., Zheng, P., Liu, H., Qin, X., Gool, L.V.: Facial-sketch synthesis: a new challenge. *Mach. Intell. Res.* **19**, 257–287 (2021). <https://api.semanticscholar.org/CorpusID:248987735>
12. Fan, D.P., et al.: Scoot: a perceptual metric for facial sketches. In: *The IEEE International Conference on Computer Vision (ICCV)* (2019)
13. Gal, R., et al.: An image is worth one word: personalizing text-to-image generation using textual inversion. *arXiv arXiv:2208.01618* (2022). <https://api.semanticscholar.org/CorpusID:251253049>
14. Gao, F., Zhu, Y., Jiang, C., Wang, N.: Human-inspired facial sketch synthesis with dynamic adaptation. In: *Proceedings of the International Conference on Computer Vision (ICCV)* (2023)
15. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423 (2016). <https://api.semanticscholar.org/CorpusID:206593710>

16. Goodfellow, I.J., et al.: Generative adversarial nets. In: Neural Information Processing Systems (2014). <https://api.semanticscholar.org/CorpusID:261560300>
17. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: GanSpace: discovering interpretable GAN controls. arXiv [arXiv:2004.02546](https://arxiv.org/abs/2004.02546) (2020). <https://api.semanticscholar.org/CorpusID:214802845>
18. 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 (2015). <https://api.semanticscholar.org/CorpusID:206594692>
19. Hertzmann, A.: Why do line drawings work? A realism hypothesis. *Perception* **49**, 439–451 (2020). <https://api.semanticscholar.org/CorpusID:211132554>
20. Hertzmann, A.: The role of edges in line drawing perception. *Perception* **50**(3), 266–275 (2021). <https://api.semanticscholar.org/CorpusID:231698870>
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv [arXiv:2006.11239](https://arxiv.org/abs/2006.11239) (2020). <https://api.semanticscholar.org/CorpusID:219955663>
22. Hu, X., et al.: Style transformer for image inversion and editing. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11327–11336 (2022). <https://api.semanticscholar.org/CorpusID:247450902>
23. Huang, X., Belongie, S.J.: Arbitrary style transfer in real-time with adaptive instance normalization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1510–1519 (2017). <https://api.semanticscholar.org/CorpusID:6576859>
24. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
25. Jain, K.K., Grosz, S., Namboodiri, A.M., Jain, A.K.: Clip4Sketch: enhancing sketch to mugshot matching through dataset augmentation using diffusion models (2024). <https://arxiv.org/abs/2408.01233>
26. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46475-6\\_43](https://doi.org/10.1007/978-3-319-46475-6_43)
27. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4396–4405 (2018). <https://api.semanticscholar.org/CorpusID:54482423>
28. 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 (CVPR) (2019)
29. Kim, J., Kim, M., Kang, H., Lee, K.H.: U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: International Conference on Learning Representations (2019)
30. Li, H., Liu, J., Bai, Y., Wang, H., Mueller, K.: Transforming the latent space of StyleGAN for real face editing. *Vis. Comput.* **40**(5), 1–16 (2021). <https://api.semanticscholar.org/CorpusID:235254637>
31. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: PhotoMaker: customizing realistic human photos via stacked ID embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
32. Liu, D., Fisher, M., Hertzmann, A., Kalogerakis, E.: Neural Strokes: stylized line drawing of 3D shapes. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14184–14193 (2021). <https://api.semanticscholar.org/CorpusID:238531682>



33. Liu, M.Y., Breuel, T.M., Kautz, J.: Unsupervised image-to-image translation networks. In: *Neural Information Processing Systems* (2017). <https://api.semanticscholar.org/CorpusID:3783306>
34. Liu, S., et al.: AdaAttN: revisit attention mechanism in arbitrary neural style transfer. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6629–6638 (2021). <https://api.semanticscholar.org/CorpusID:236956663>
35. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *Computer Vision – ECCV 2018*. LNCS, vol. 11218, pp. 800–815. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01264-9\\_47](https://doi.org/10.1007/978-3-030-01264-9_47)
36. Mirza, M., Osindero, S.: Conditional generative adversarial nets. *arXiv arXiv:1411.1784* (2014)
37. Ohtake, Y., Belyaev, A.G., Seidel, H.P.: Ridge-valley lines on meshes via implicit surface fitting. In: *ACM SIGGRAPH 2004 Papers* (2004). <https://api.semanticscholar.org/CorpusID:8500135>
38. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5873–5881 (2018). <https://api.semanticscholar.org/CorpusID:54447797>
39. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning* (2021). <https://api.semanticscholar.org/CorpusID:231591445>
40. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. *arXiv arXiv:2204.06125* (2022). <https://api.semanticscholar.org/CorpusID:248097655>
41. Richardson, E., et al.: Encoding in style: a StyleGAN encoder for image-to-image translation. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2287–2296 (2020)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685 (2021). <https://api.semanticscholar.org/CorpusID:245335280>
43. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dream-Booth: fine tuning text-to-image diffusion models for subject-driven generation. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510 (2022). <https://api.semanticscholar.org/CorpusID:251800180>
44. Rusinkiewicz, S., DeCarlo, D., Finkelstein, A.: Line drawings from 3D models. In: *International Conference on Computer Graphics and Interactive Techniques* (2005). <https://api.semanticscholar.org/CorpusID:10994464>
45. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *arXiv arXiv:2205.11487* (2022). <https://api.semanticscholar.org/CorpusID:248986576>
46. Sayim, B., Cavanagh, P.: What line drawings reveal about the visual brain. *Front. Hum. Neurosci.* **5**, 118 (2011). <https://api.semanticscholar.org/CorpusID:263708652>
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014). <https://api.semanticscholar.org/CorpusID:14124313>

48. Song, G., et al.: AgileGAN: stylizing portraits by inversion-consistent transfer learning. *ACM Trans. Graph.* **40**, 117:1–117:13 (2021). <https://api.semanticscholar.org/CorpusID:236006017>
49. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for styleGAN image manipulation. *ACM Trans. Graph. (TOG)* **40**, 1–14 (2021). <https://api.semanticscholar.org/CorpusID:231802331>
50. Tsao, D.Y., Livingstone, M.S.: Mechanisms of face perception. *Ann. Rev. Neurosci.* **31**, 411–37 (2008). <https://api.semanticscholar.org/CorpusID:14760952>
51. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.: Texture networks: feed-forward synthesis of textures and stylized images. In: 33rd International Conference on Machine Learning, ICML 2016, pp. 2027–2041 (2016)
52. Vaswani, A., et al.: Attention is all you need. In: *Neural Information Processing Systems* (2017). <https://api.semanticscholar.org/CorpusID:13756489>
53. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(11), 1955–1967 (2009). <https://doi.org/10.1109/TPAMI.2008.222>
54. Xu, Y., et al.: TransEditor: transformer-based dual-space GAN for highly controllable facial editing. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7673–7682 (2022). <https://api.semanticscholar.org/CorpusID:247839345>
55. Yang, S., Jiang, L., Liu, Z., Loy, C.C.: Pastiche master: exemplar-based high-resolution portrait style transfer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7683–7692 (2022). <https://api.semanticscholar.org/CorpusID:247627720>
56. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-adapter: text compatible image prompt adapter for text-to-image diffusion models. *arxiv* [arxiv:2308.06721](https://arxiv.org/abs/2308.06721) (2023). <https://api.semanticscholar.org/CorpusID:260886966>
57. Yi, R., Liu, Y.J., Lai, Y.K., Rosin, P.L.: APDrawingGAN: generating artistic portrait drawings from face photos with hierarchical GANs. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10735–10744 (2019). <https://api.semanticscholar.org/CorpusID:194358484>
58. Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: a feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **20**, 2378–2386 (2011). <https://api.semanticscholar.org/CorpusID:10649298>
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018). <https://api.semanticscholar.org/CorpusID:4766599>
60. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251 (2017)



# Few-Shot and Portable 3D Manufacturing Defect Tracking with Enterprise Digital Twins Based Mixed Reality

Yiyong Tan<sup>(✉)</sup> , Bhaskar Banerjee , and Rishi Ranjan 

Gridraster Inc, Mountain View, CA 94043, USA  
yiyong.tan@gridraster.com

**Abstract.** We present a time of flight (TOF) mixed reality (MR) digital twin mobile system supporting three-dimensional (3D) tracking and defect detection using recursively fused multimodal segmentation paradigm. Simplified machine learning can be used for clustering multimodal 3D semantic label distribution (output of generic data trained segmentation deep learning model) and to reduce the need to obtain high cost and extremely scarce non-generic training data to flexibly customize segmentation for non-generic enterprise defect inspection applications. The fused model first segments with 3D physics properties (reflection, curvature, materials etc.) obtained from TOF and tracks objects with defect from a 3D scene and then further segments recursively on different level of details to detect defects with quantification analysis based on segmentation distribution statistic distance. This method also removes the need to do compute intensive non-real-time algorithms (3D mesh generation, SLAM bundle adjustment and cross source 3D alignment) needed for 3D defect detection. User can do portable free hand acquisition to track and quantify the severity of 3D anomaly defects and categories of 3D configuration without the need to follow strict data capture guidance and 3D point cloud alignment registration as required by other state of the art enterprise MR systems.

**Keywords:** Mixed Reality · 3D Manufacturing Defect · Digital Twins · Portable · 3D Recursive Fusion

## 1 Introduction

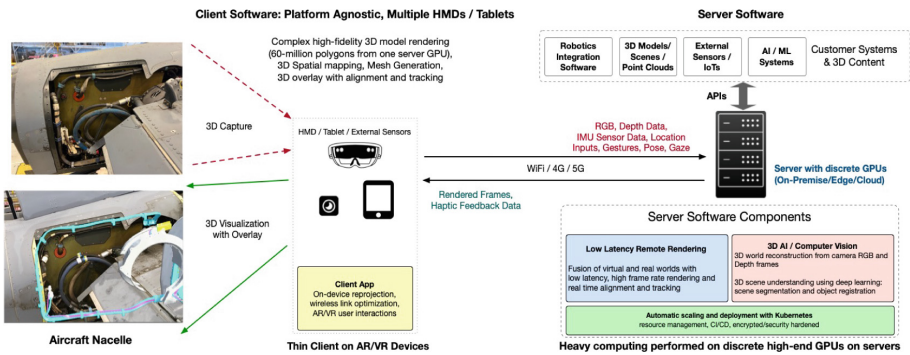
Human visual inspection is still considered as the most widely used method for large area surface inspections in manufacturing and maintenance operations, such as aircrafts, rockets, construction/manufacture sites, etc. Mobile drone, wearable mobile AR glass and robot assisted inspection has also emerged in recent years [1, 2]. Designing reliable portable mobile inspection system and flexible robotics systems are challenging, facing difficulties such as inadequate coverage of inspection area of certain view angle, false positive defects identification of pure 2D video based mobile wearable AR inspection system, requiring reconfigurable robotics path to be found in cluttered repair plants, and low resolution from noncontact or far sight distant camera-based drone inspection

system [3, 4]. Deep learning-based methods are attracting more and more interest at micro-, meso-, and macro-scale level although these methods are primarily 2D defect detection and are usually compromised by the lack of quality and quantity of relevant data that can truly represent the industry non-generic defects [5–9].

This paper introduces a practical robust procedure for 3D defect inspection based on the matching feature profile of the digital twins to their real-world counterparts. For productions in large scale, quick and early identification of defects is the crucial step. This MR defect detection system is designed to be practically implemented into a user-friendly enterprise product both for the developers and the end users. The versatile, learning based workflow is robust and adaptable for various industry inspection projects with reduced few-shot of non-generic 3D training data by recursive fusing different segmentation algorithms or deep learning models trained with different related generic 3D benchmark training datasets. The method presented here is majorly focused on detecting the defect entity through an industrial grade mobile wearable MR head mounted device (HMD), such as the Microsoft HoloLens2, with enhanced portable MR and 3D AI features supported by a local server or remote cloud services, as in Fig. 1.

The paper is organized as follows: Sect. 2 describes and summarizes related work and concept; in Sect. 3, the enhanced fused inspection model is described in detail; in Sect. 8 the results are presented followed by additional discussions are in Sect. 13. Finally, Sect. 14 provides conclusion and future work.

## 2 Related Work and Concept



**Fig. 1.** Illustration of our mixed reality system with remote rendering, and 3D AI/Computer Vision based object detection and precise overlay of 3D digital twin over the object. In this example, the wiring harness of an aircraft nacelle is precisely overlaid and rendered when viewed through the Microsoft HoloLens 2 by detecting the nacelle and estimating its pose accurately.

Digital twin-based inspection is considered a key component for quality assurance in industry 4.0. The application is already in many industry fields in a stage of standardization and attracting broad interests [10–12]. With new generation of MR HMDs with on-device 2D and 3D cameras, digital twin-based RGB/D method can achieve 3D geometric related defect detection that 2D defect inspections cannot, such as: 1) Robustness

to different lighting conditions, 2) Categorize defect severity with the depth information in the detected regions, 3) Avoid false positives of 2D defect detection without indication of uncertainty due to color, dusts and stains etc., all of which can be mitigated by additional depth information and geometry consistency of different view angles. Sect. 8 will provide comparison to the current state-of-the-art solution and further discuss in detail with real world mobile MR use case applications.

The proposed method is a multimodal system that utilizes multiple machine learning and artificial intelligence systems on visual, spatial and gestural signals, such as a plurality of neural networks wherein each neural network has its own unique topology network structure which inherently exhibits different numerical feature extraction behaviors when learning 3D scene features from a publicly available benchmark 3D training dataset. The distribution of features and scene context learned in a certain pre-trained model can probe certain aspects in the higher dimensional feature space of real-world objects and scene point clouds so that a pre-trained model trained by general benchmark data can be used as a weak classifier for specific applications. Combining inference results of multiple pre-trained models can yield a full spectrum of properties which are defined in the features extracted from generic benchmark datasets by individual pre-trained deep learning models. This uncertainty reduction concept is like sensor fusion in autonomous driving to understand the real driving environment and can also be seen as painting objects and scenes by using multiple colors to maintain high fidelity.

### 3 Methodology

The paper is particularly applicable to a mixed reality system with 3D object tracking of defects and anomalies that overcomes technical problems and limitations of existing deep learning systems by reducing training data requirements to few-shot 3D scan and employing a simpler fusion machine learning model to learn from feature distributions already extracted from complicated deep learning models.

### 4 System and Preliminary

The system, as illustrated in Fig. 2 below, receives the 3D data with a complicate 3D points cloud using two or more machine learning and/or deep learning systems, each of them generates a histogram based on public generic 3D training data reducing the complexity of the initial 3D data to a vector of hundreds of values. The system then trains a simpler machine learning model (since the 3D data is now less complex – hundreds of histogram values vs. millions of 3D point values) that: 1) requires less training data; and 2) can solve the 3D inspection and tracking problem without both the complex non-generic 3D scene data (training data is often not available) and complicated big deep learning networks training. In the paper, several public benchmark 3D datasets are chosen for a certain typical use case. To apply to a different defect inspection use case, the system and method can be applied to different public datasets, public and private datasets or only private datasets that can be similar to the objects of interest to train two or more deep learning models. The features can then be extracted from them, allowing the system to significantly reduce the complexity of AI model from deep learning models

to machine learning models and allow to only use few-shot 3D scanning of the scarce costly non-generic 3D digital twin and mobile AR glass scanned datasets.

### 5 Tracking a Component Part of a 3D Scene

The proposed method solves these technical problems by providing a hybrid 2D/3D tracking, as illustrated in the workflow in Fig. 3. For example, for 3D object tracking, the system performs a detailed 3D scene understanding following the workflow discussed below. Given the computation limitations of the MR HMDs, the entire 3D processing is done on the backend server with discrete high-end GPUs, where the color (RGB) and depth (D) data (RGB/D) from the camera of the computing device may be used to reconstruct a full 3D point cloud with complete texture mapping. A fine mesh is then generated using this 3D depth map and the relation between different parts of the scene is established. Both the RGB features and the depth geometry are used to segment the scene and establish association as discussed below. In the example in Fig. 3, the object of interest is the aircraft nacelle. The system isolates the nacelle from the rest of the scene by identifying its 3D/2D features using our deep learning-based inference engine (for example by histogram distribution-matching based cluster labeling in Fig. 2) that matches the object scanned by mobile AR glass to the 3D digital-twin.

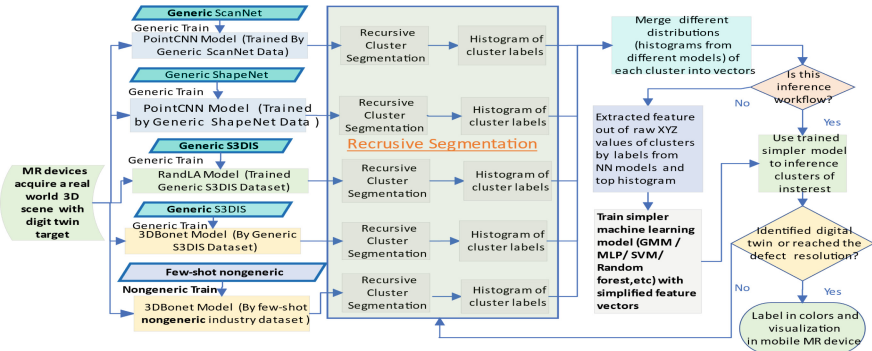
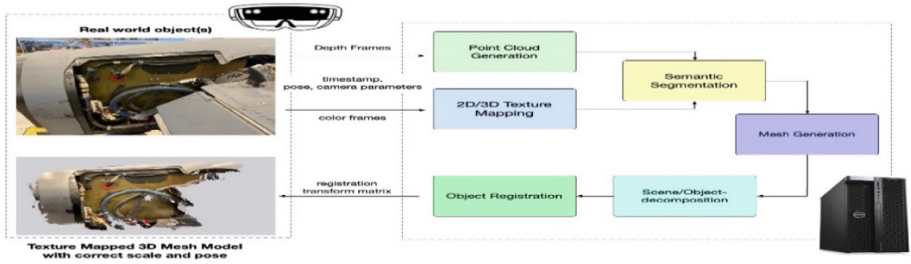


Fig. 2. Flowchart for 3D object/anomaly tracking with reduced training data.

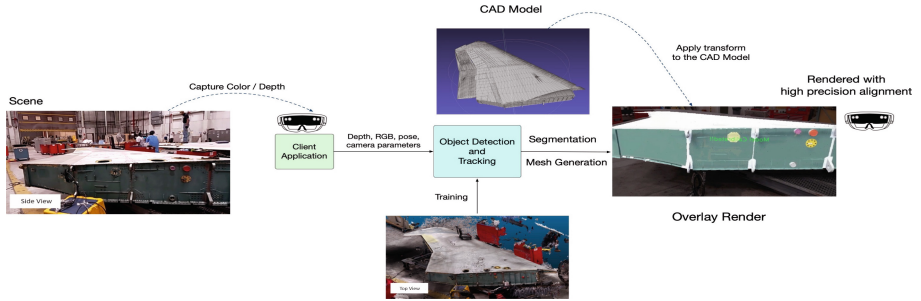
### 6 Tracking in Cluttered Environment

As shown in Fig. 4, the system provides real time object tracking while wearing a mobile MR HMD and overlaid rendering in a cluttered manufacture environment among which there is a model of aircraft wing. The deep learning-based recursive segmentation allows the system to identify and track 3D objects of arbitrary shape and size in various orientations with high accuracy in the 3D space. This approach is scalable with any arbitrary shape and is amenable to use in enterprise use cases requiring rendering overlay of complex 3D models and digital twins with their real-world counterparts. This can also

be scaled to register with partially completed structures with the complete 3D models, allowing for on-going construction and assembly. The system and method achieve an accuracy (relatively  $< 1\%$  error) of 1 ~ 10mm depending on the object's dimension size during recursive segmentation-based tracking and rendering using the system that illustrates the improvement over conventional systems that cannot achieve that accuracy. This approach to 3D object tracking will allow the system to truly fuse the real and virtual worlds, enabling many applications including but not limited to training with work instructions, defect detection, manufacturing error inspection in construction and assembly and 3D engineering design with life size rendering and overlay.



**Fig. 3.** 3D object detection and pose estimation workflow using hybrid 2D/3D (RGB/D) data.



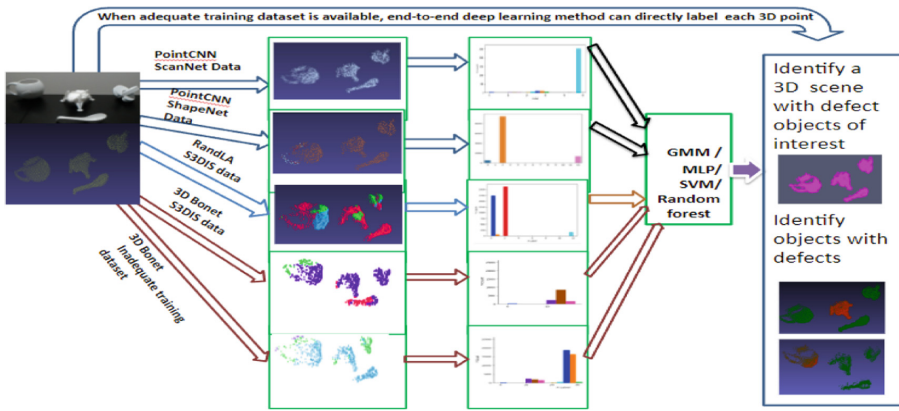
**Fig. 4.** 3D object recognition and tracking using the workflow from Fig. 3 and can also reduce the reliance on training data with the multimodal fusion from Fig. 2. In this example scene, RGB/D data (color and depth), pose data and camera parameters of the HMD is sent to the back-end server to enable precise 3D object tracking and inverse rendering overlay.

## 7 Implementation Details

For 3D semantic segmentation, the method (shown in detail in Fig. 2 and Fig. 5) uses multiple deep neural network structures (such as PointCNN [13], 3D Bonet [14], RandLA [15], etc. in one fused system) trained by different benchmark generic 3D datasets (ScanNet [16], ShapeNet [17], S3DIS [18], inadequate few-shot nongeneric enterprise training datasets, etc.) to perform 3D semantic segmentation of 3D scenes not seen by the fused recursive segmentation workflow.



For each cluster of a point cloud, each pre-trained model will label 3D objects in different distributions (histogram of object labels existing in generic 3D benchmark datasets, sharing some geometric similarity with different objects in the current non generic 3D scene). The labeled distribution can be used as the fingerprint of the 3D point clustering so that object/scene can be understood. Combining using different approaches, such as Gaussian mixture modeling (GMM), multilayer perceptron (MLP), support vector machine (SVM), random forest, k-nearest neighbours (KNN), distribution distance-based clustering etc., these specific distributions of multiple pre-trained models are merged into a stronger classifier. The major advantage of this approach is to minimize the non-generic labelled training data requirement for a specific enterprise use case whose dataset usually is not public available in generic 3D bench mark datasets and improve the generalizability of combined deep neural networks.



**Fig. 5.** An example of the 3D object tracking with reduced training data (few-shot nongeneric) with multimodal fusion.

Figure 2 illustrates a process for 3D object tracking with reduced training data (only need few-shot nongeneric dataset) and Fig. 5 illustrates an example of the 3D object tracking with few-shot non-generic data. As illustrated in Fig. 5, we can identify the two objects with complicated geometry/components/volume matching the objects we want to track in a 3D scene, and merge all other objects (simple geometry, volume mismatch objects like support table, wall, spoon, simple objects with two cylinder) as background to automatically remove them.

To adapt system for better results in an unseen use case, we can replace the exemplified generic dataset in Fig. 2 with public generic dataset more similar and relevant to enterprise targeted use case applications. For easy maintenance and comparison, loss functions optimization is not conducted for all the results presented. If trained by different dataset (other generic data or inadequate few-shot non-generic enterprise training datasets), same loss functions provided by original authors of deep neural networks in Fig. 2 are used and default loss functions of simpler machine learning models (GMM, MLP, SVM etc. For our use case, GMM is the method heuristically recommended to use) from opensource packages can be used without changes.

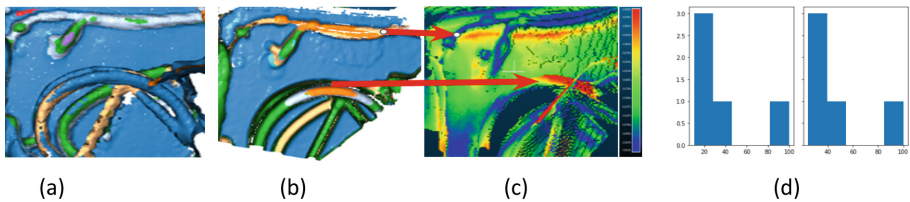


## 8 Results

After using methods described to track the defect region or objects, we can further analyze the region under different sizes of defected objects by applying a fused model recursively to reach different level of detail. As different scans will have different output points from different portions of defect region due to scan conditions like scanner configuration, distance, and view angle, etc., we normalize the biggest segmented region in terms of numbers of points, the histogram of segments represents relative size in point numbers with respect to the region with biggest points, to make different batches of scan comparable to each other. In addition, as histogram distribution of different segmented surface area of different parts of objects is not sensitive to the orientation and arrangement of the components, the defect detection is relatively robust to various view angles, moving components (cables or fixed region), and arrangement.

Due to the point geometry changes caused by defects, the distribution of segments with defects will be different in distribution from segments without defects. For example, the defects caused by bending and impact will cause original one segment to become multiple segments with high curvature crack/defect as boundary; defects due to wearing off will blur the boundary and merge multiple components into one bigger segment; once we select detected defect regions, by applying recursively into different scales of regions of interest, defects in different level of detail and resolution can be revealed according to the requirements of use cases. Based on histogram's statistical distance with/without defect, semi quantification of severity of defect can be provided.

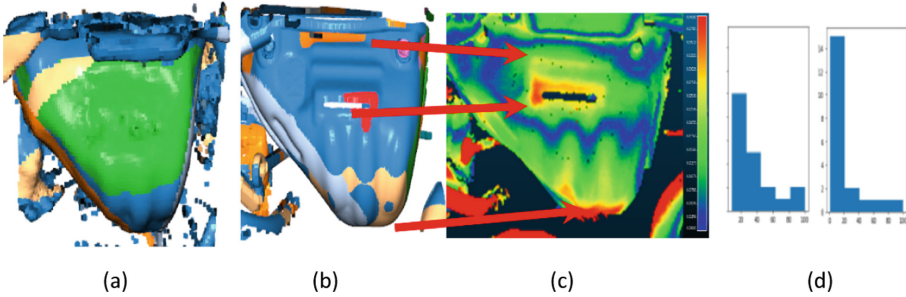
## 9 3D Defect Detection without Alignment



**Fig. 6.** A typical defect detection case of a big size subject. (a) 3D segmentation of region of interest without defects. (b) 3D segmentation of components (with defect regions segmented in orange color). (c) Mesh 3D geometry Hausdorff distance: blue indicates normal region, red indicates bigger difference/defect, which corresponds to a segmentation difference between (a) and (b). (d) Histogram of segmentation of corresponding scan (x axis is the segmentation size measured by number of 3D points normalized to biggest segmented region as 1 or 100, y axis is the number of segmentations within a range of x labeled size).

For big size subjects as in Fig. 6, point clouds (a) and (b) are viewed from different angles. The results show the robustness of our method to detect defects in various view angles without alignment. As the defect region is only a small portion of the whole region, the overall shape of histograms is similar, while noticeable difference can be observed

between the segmentation size of each segment and the gap between the biggest segment and second biggest segment. The histogram similarity is computed as  $(1 - \text{distance})$  between the two probabilities or frequency distributions with two methods (Hellinger:  $1 - 0.14 = 0.86$ , Wasserstein:  $1 - 0.22 = 0.78$ ). In both the cases, the similarity is lower than a preset defect threshold of 0.9, which is further confirmed by an aligned mesh difference map: two red regions (potentially defect regions).



**Fig. 7.** A typical defect detection case of small size subject. (a) 3D segmentation of components around an engine (without defect). (b) 3D segmentation of components around an engine (with defects). (c) Mesh Hausdorff distance comparison result. (d) histogram of segmentation of corresponding scan: left side calculated from (a) and right-side plot calculated from (b).

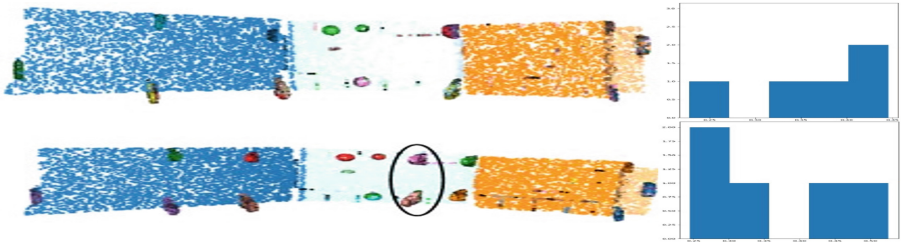
For small size subjects as in Fig. 7, the major component of point clouds (a) and (b) can be viewed completely even from different angles but there are some surrounding components to add quite some more small segments. The results show the robustness of our method to detect defects without alignment in this case. As the defect region occupies more percentage of region, the histogram shape also becomes different from each other with noticeable difference in all sizes of segments.

The histogram similarity is computed (Hellinger:  $1 - 0.16 = 0.84$ , Wasserstein:  $1 - 0.15 = 0.85$ ). In both the cases, the similarity is lower than a preset defect threshold of 0.9 and is reported as a defect, which is again verified by difference map of aligned paired mesh: three regions (potentially defect regions) within the component of interest. The right corner red segment outside region of interest components is due to adjacent parts being movable with relative position changes during the two scans. (Won't affect the histogram of region of interest selected by recursive segmentation).

## 10 3D Configure Detection

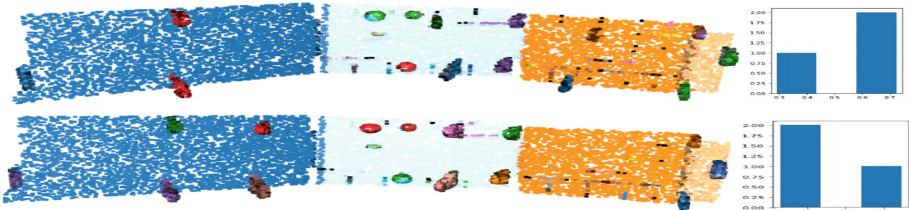
In the airplane industry the cabin interior design is configured differently for different airlines. For mobile AR devices to be applicable during interior assembly, design and training applications, airline configuration needs to be correctly identified. Take the mounting brackets as an example – our method can quickly differentiate both the differences in the number of brackets and their location configuration inside an airplane cabin.

**Different Number of Brackets.** The cyan region has a different number of brackets (above: base line typical brackets configuration; bottom: two additional brackets circled on the right side of three circle screws added for a new configuration).



**Fig. 8.** Mounting brackets on a typical structure. The top and bottom show different numbers of brackets in two different configurations.

**Same Number of Brackets but at Different Locations.** When the same brackets have different locations, we can subdivide the histogram in different regions of the original model. In the following example, we have four major regions segmented out (blue, cyan, yellow, and brown). For two subregions (blue and brown), we can detect changes in the number of brackets, although the total bracket number is same.

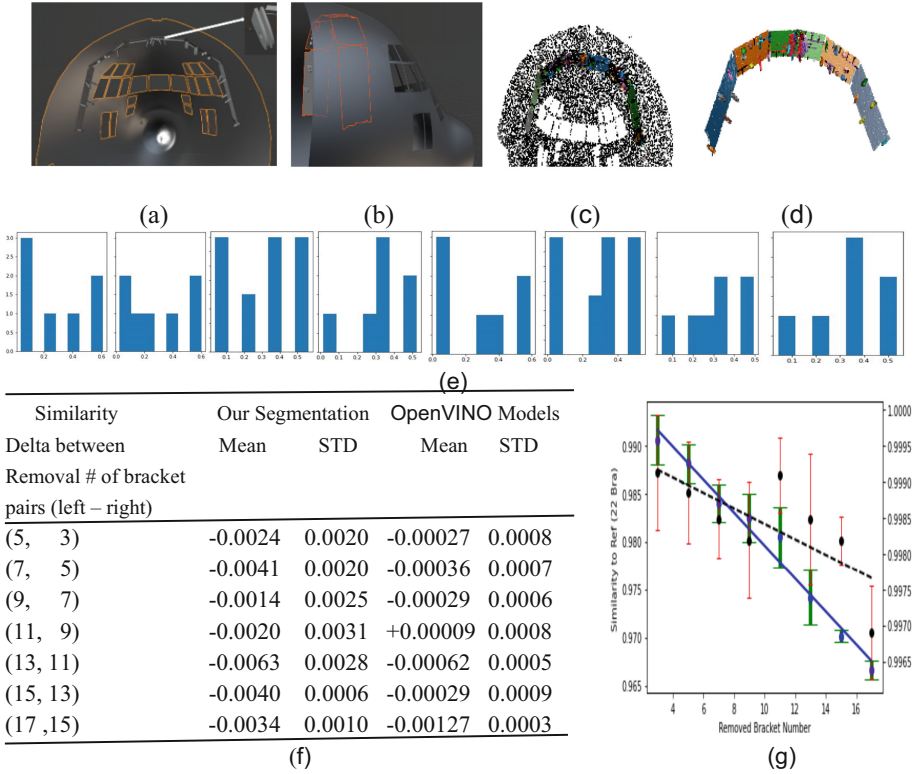


**Fig. 9.** Same number of brackets with different distribution in the subregion.

As the Figs. 8 and Fig. 9 present, the two types of configuration changes (different number of instances and same number of instances but with different distribution in subregions) can be detected with the similarity score between two histograms.

**A Real-World Quantitative Use Case.** AR training or maintenance inside an airplane (for example, the front cabin of the C130 aircraft head) requires the detailed content to be automatically selected and loaded using the geometry fingerprint of the interior 3D scene of the airplane.

As in Fig. 10, we first remove the unchanged background from the arc structure using strong features of windows; then calculate the Wasserstein distance between the segmentation histogram distribution of design configuration of different number of brackets on arc support structure. For example, the lowest 0.967 similarity score (1- Wasserstein distance) in Fig. 10 (x axis = 17) is calculated between segmentation histogram of original arc point cloud with all 24 brackets and segmentation histogram of arc with



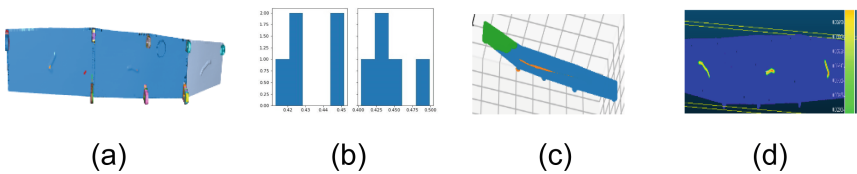
**Fig. 10.** Quantitative study of configuration of variation in number of brackets. (a) internal view of C130 head with arc structure. (b) external side view of C130 head. (c) segmentation results of point cloud, black color point cloud can be recursively removed as background. (d) segmented arc structure. (e) histogram of segmentation from 8 different configurations with different number of brackets. (f) comparison table between the change of correlation of bracket numbers by our method and latest 3D segmentation deep learning models of openVINO. (g) linear regression of similarity score calculated between histograms of the various number of brackets removed from arc structure with respect to original arc with all brackets: our method is in solid line and existing reference method is in dashed line As table shows, under different configuration (bracket numbers), deformation and noise, our fused recursive segmentation method show ~ 1 magnitude better sensitivity and robustness (STD/Mean ratio) than industry state of art 3D deep learning segmentation models.

17 brackets removed. The similar score (1-Wasserstein distance, more information as in supplementary materials) is proportional to the removed bracket number between the paired point clouds. Each error bar is calculated by mean and std from 5 different experiments (applied noise and deformation perturbation to point cloud) and can be linear regression fit to  $Y = 0.99675 + - 0.00172X$  with R-squared = 0.977. In contrast, results from openVINO [19] state of art 3D segmentation models (plot in thinner red error bar and data in black dots with respect to right side Y axis) has much worse variance with respect to mean and worse linear correlation (dashed line) to configuration change

( $Y = 0.99949 + - 0.00010793X$ , with R-squared = 0.547). Therefore, with the method discussed in this paper, we can automatically identify different configuration/design layout and number of brackets when changes of bracket number on the arc structure are quantitatively calibrated as in Fig. 10. The use case is tested and run on Window 11 desktop with RTX 3080 GPU, Intel i7,32G memory and 256 G SSD storage. As recursively hierarchy segmentation can split the 3D point cloud into multiple independent regions which can further leverage multiple parallelly GPU servers through our well established sever based HMD MR pipeline. To achieve close to real time performance, our system can conveniently dispatch recursively segmented clusters to different GPU servers for heavy spatial computation so that the network socket communication speed and latency between servers and HMD device become the up-limit bottleneck when target to solve more challenging point clouds with million vertices. As more and more spatial computation pretrained generic models get published, computation power ramps up, and improvement in loss functions customized in different recursive stage, our method can further support unseen specified or novel defect in heavy 3D models.

## 11 Sensitivity Study

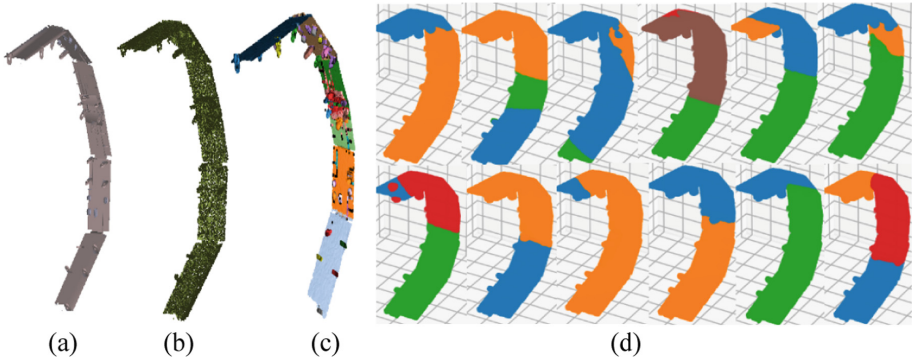
As in 4.2, we can get good quantitative detection on number of the brackets. Each bracket roughly  $\sim 1\%$  of total point cloud ( $\sim 300$  points out of total 22k point cloud). To understand the specification of the qualitative limit of defect detection, a sensitivity analysis for 3D defect detection is also conducted. As expected, the sensitivity is dependent on the whole model dimension, performance requirement and geometry complexity. For a typical use case in manufacturing and sustainment, our method can detect relative change of model within 0.2%. As the following picture shows for a 4-m-long panel, the lowest level we can segment out and detect in histogram similarity comparison is 5 mm for the full width at half maximum (FWHM) in depth direction as in Fig. 11.



**Fig. 11.** Segmentation sensitivity in an end2end MR (with HL2) study. (a) 5 mm FWHM defect detected by segmentation without alignment as the defect can result in new histogram bin in the cluster with 0.2% changes. (b) comparison of segmented distribution histogram with and without defects; Left side: the side panel without defects, Right side with defect; segmentation distribution can identify a new instance around 0.45 (normalized to biggest size of ROI clusters recursively detected for defects/brackets). (c) best defect segmentation results from 12 openVINO latest 3D segmentation models, which are still not able to segment out correct position of challenging trace level defects. (d) defect ground truth verified by rigid alignment based mesh Hausdorff distance.

## 12 Comparative Study with Other Opensource Deep Learning Models

Recently 3D segmentation leverage 2D RGB segmentation, multi-frame projection/pose estimation transformer, point cloud and 3D mesh connection graph, along with computationally expensive deep learning framework to get semantic point/voxel segmentation, which in general requires enormous training data and does not generalize well to get satisfactory results for unpublic enterprise 3D scenes and nongeneric targets [20]. For example, even for a 2D model, segment everything (SAM) is trained by 11M images and 1.1 billion segmentation masks [21], and still majorly rely on color and does not work well for 3D geometry deflection segmentation on uniform texture surface. The benchmark 3D datasets (S3DIS [18], LLFF [22], Co3D [23] etc.) are usually composed of good quality point cloud from high end scanner of common everyday objects, which are not applicable well to enterprise 3D use cases and not from AR edge devices' (Hololens2 etc.) noisy scanning. Generally, for 2D SAM based 3D segmentation model and 2D anomaly detection without the need of training dataset, accuracy of segmentation is dependent on input texture/RGB contrast of surface, which usually not available in geometry-based 3D deflection as there are no vivid color contrast on most surface of metal components and defects [24–26]. For enterprise applications, speed, repeatability, and practical flexibility are crucial to land a user-friendly quality product to market. The existing pretrained 3D geometry model does not generalize and work well in specific real world nongeneric enterprise use cases as we compared in this section. On the other hand, we can flexibly fine tune both deep learning model's feature weight and recursive level balance responsive time requirement and the level of details.



**Fig. 12.** Segmentation comparison between our recursive method and OpenVINO pretrained 3D segmentation models. (a) Digital twin CAD model ground truth. (b) point cloud for 3D segmentation. (c) segmented results from our recursive fusion methods. (d) 3D segmentation results from 12 different OpenVINO deep learning pretrained models in open-source model zoo.

As an industry enterprise deep learning framework, OpenVINO (developed by Intel) is optimizing and deploying open cross platform framework for enterprise AI inference applications. Targeting real world enterprise applications, here we use OpenVINO and



its open model zoo as a comparison. As Fig. 12 shows (also two extra examples are in the supplementary material), 12 off-the-shelf deep learning models from open model zoo majorly focus on common 3D features extracted from internet accessible 3D benchmark training datasets [19, 27–29], for unseen data, especially unseen private non-generic enterprise 3D subjects, the generalization is not capable enough to get satisfactory 3D segmentation in enterprise applications.

### 13 Discussion

From practical implementation point of view, different level of detail in training datasets and segmentation can be achieved based on the need of different level of accuracy requirements in applications. For HoloLens 2, the absolute highest sensitivity can be achieved is 0.5 ~ 1mm, and defects can be detected in most scenarios when the geometry variance toward adjacent objects is relatively 0.2% or higher (5mm defect can be detected if region of interest is a 3000mm side panel). When we use the HoloLens 2 to inspect large targets (airplanes, space craft components, etc.), hierarchical scan with recursive segmentation and tracking can achieve desirable accuracy and sensitivity.

If, based on the accuracy of the merged label data, more training is needed, the method may perform additional processes to reduce data complexity so that we can further minimize the need for training data of the digital twin target. The two processes may be: 1) extract features (line, corners, and primitive shapes) out of raw XYZ and RGB values of the clusters to label from the multiple DL models, machine learning algorithm or 3D vision segmentation methods in opensource library like open3D, Point Cloud Library (PCL) etc.; and 2) to further reduce labels vector into histogram and count point number of each label. In more detail, the processes select top predicted labels (label existing in benchmark datasets) to filter out noise and the reduced labels (keep only top dominant clusters like salient target and background) of the clusters as the training binary positive and negative datasets and further perform the two-step data feature extraction with the much smaller training datasets as the input is already extracted features from pre-trained models trained by public generic benchmark datasets.

In our system, the method uses the trained simpler machine learning models to infer a group of clusters or a single cluster. The method provided can be extended to detect huge 3D scenes (an airplanes, tunnel construction, etc.) in a recursive paradigm for different level of detail. To detect defects in a huge point cloud allowed by the HMD's effective range and computation power of GPU server, the trained simpler machine learning model can first identify the digital twin target or the background first, and then configured. The workflow specific for the current use case of identified targets. To classify or do segmentation of different level objects, we can recursively apply our workflow to establish a database of segmentation histograms that can be easily mapped. To label a group of clusters, the method identifies whether there is a specific 3D scene by using the histogram of all the clusters to determine the label of the scene. If the use case to identify the digital twin target is successful, as shown in Fig. 5 and Fig. 2, the hierarchy recursive segmentation and tracking can retrieve the normalized object histogram distribution of each cluster in different level of details (either the digital twin target or background).

## 14 Conclusion

This study presented an automated 3D digital twin learning based inspection system that can track objects and detect defects for enterprise applications, aiming to facilitate both development and user experience:

1) with minimum few-shot or no need of training datasets for specific non generic defects of interest, 2) free hand acquisition which tolerates various capture poses and lighting conditions without the overlap strictness of the same capture region between 3D point clouds with and without defects, 3) without the need to conduct 3D alignment during defect detection, and 4) without the need to create high quality 3D mesh.

The process first tracks the region or the object of interest and do segmentation with pre-trained models by benchmark generic datasets, machine learning algorithm or 3D vision segmentation methods in opensource library like open3D, Point Cloud Library (PCL), and then classify and compare segmentation profile distribution to provide a similarity score with respect to the original normal digital twin counterpart, which represents the severity of the defects. The Hausdorff distance mapping of overlap aligned mesh pairs from two different typical sizes of defect objects further confirmed the correctness of our 3D defect detect methods for real world point clouds of 3D scene.

We hope our work can inspire further improvement in few-shot learning for time-of-flight sensors-based mixed reality AR/VR edge devices to address challenging defect detection requirements like reflection removal, missing components remind and foreign objects alert in both forward/inverse and cloud/devices renderings. In further research and product development, we plan to adapt the existing standard loss function to meet broader non generic use cases and improve automatic digital twins guided alignment besides doing a binary qualitative classification and instance quantification between defected subjects and their digital twin for a chosen level of detail. We plan also to combine recursive defect & configuration detection and alignment in different levels of accuracy so that we can do adaptive real-time identification, annotation and quantitative analysis of the defect severity adaptively in various levels of details.

## References

1. Yuri, D.V.Y., Fabio, A.M.C., Luiz, E.G.M., Jorge, A.B.G.: Aircraft visual inspection: a systematic literature review. *Comput. Ind.* **141**, 103695 (2022)
2. Yanjuan, H., Feifan, Z., Lin, Z., Yongkui, L., Zhanli, W.: Scheduling of manufacturers based on chaos optimization algorithm in cloud manufacturing. *Rob. Comput. Integr. Manuf.* **58**, 13–20 (2019)
3. Ali, M., et al.: Augmented reality-computer vision combination for automatic fatigue crack detection and localization. *Comput. Ind.* **149**, 103936 (2023)
4. Anh, V.L., Veerajagadheswar, P., Vinu, S., Rajesh, M.: Modified a-star algorithm for efficient coverage path planning in tetris inspired self-reconfigurable robot with integrated laser sensor. *Sensors* **18**(8), 2585 (2018)
5. Xian, T., Dapeng, Z., Wenzhi, M., Xilong, L., De, X.: Automatic metallic surface defect detection and recognition with convolutional neural networks. *Appl. Sci.* **8**(9), 1575 (2018)
6. Rafia, M., Mustafa, M., Atif, Bin, M., Hassan, M.: Computer aided visual inspection of aircraft surfaces. *Int. J. Image Process* **6**(1), 38–53 (2012)





7. Xiangwen, S., Shaobing, Z., Miao, C., Lian, H., Xianghong, T., Zhe, C.: Few-shot semantic segmentation for industrial defect recognition. *Comput. Ind.* **148**, 103901 (2023)
8. Wendy, F., et al.: 3D spatial measurement for model reconstruction: a review. *Measurement*, 112321 (2022)
9. Yadong, X., Peizhe, S., Fei, J., Hongwei, H.: 3D reconstruction and automatic leakage defect quantification of metro tunnel based on SfM-Deep learning method. *Undergr. Space* **7**(3), 311–323 (2022)
10. Kamil, Ž., Ján P., Milan, A., Peter, L., Alexander, H.: Digital, twin of experimental smart manufacturing assembly system for industry 4.0 concept. *Sustainability* **12**(9), 3658 (2020)
11. Sepehr, A., Ibrahim, Y.: Digital twin-based progress monitoring management model through reality capture to extended reality technologies (DRX). *Smart Sustain. Built Environ.* **12**(1), 200–236 (2023)
12. Hosamo, H.H., Nielsen, H.K., Alnmr, A.N., Svennevig, P.R., Svidt, K.: A review of the digital twin technology for fault detection in buildings. *Front. Built Environ.* **8**, 1013196 (2022)
13. Yangyan, L., Rui, B., Mingchao, S., Wei, W., Xinhan, D., Baoquan, C.: PointCNN: convolution On X-transformed points. In: *Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 828–838 (2018)
14. Bo, Y., et al.: Learning object bounding boxes for 3D instance segmentation on point clouds. In: *Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 6740–6749 (2019)
15. Qingyong, H., et al.: Randla-Net: efficient semantic segmentation of large-scale point clouds. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11108–11117 (2020)
16. Angela, D., et al.: Richly-annotated 3D reconstructions of indoor scenes. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839. (2017)
17. Chang, A.X., et al.: ShapeNet: an information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015)
18. Iro, A., et al.: 3D semantic parsing of large-scale indoor spaces. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1534–1543 (2016)
19. OpenVINO™ Toolkit (2024). <https://github.com/openvinotoolkit/openvino>
20. Prasoon, K.V., Dogus, K., Egils, A., Cagri, O., Gholamreza, A.: A survey on deep learning based segmentation, detection and classification for 3D point clouds. *Entropy* **25**, 635 (2023)
21. Alexander, K., et al.: Segment anything. In: *International Conference on Computer Vision (ICCV)*, pp. 4015–4026 (2023)
22. Ben, M., Pratul.P, S., Matthew, T., Jonathan, T.B., Ravi, R., Ren N.: NERF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, **65**, 99–106 (2021)
23. Jeremy, R., Roman, S., Philipp, H., Luca, S., Patrick, L., David, N.: Common objects in 3D: large-scale learning and evaluation of real-life 3D category reconstruction. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10901–10911 (2021)
24. Jiazhong, C., et al.: Segment anything in 3D with NERFS. In: *Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 25971–25990 (2023)
25. Yunhan, Y., Xiaoyang, W., Tong, H., Hengshuang, Z., Xihui, L.: SAM3D: segment anything in 3D scenes. *arXiv preprint arXiv:2306.03908* (2023)
26. Yunkang, C., Xiaohao, X., Chen, S., Yuqi, C., Zongwei, D., Liang, G., Weiming, S.: Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*. (2023)
27. Guocheng, Q., et al.: PointNeXt: revisiting PointNet++ with improved training and scaling strategies. In: *Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 23192–23204 (2022)

28. Christopher, C., JunYoung, G., Silvio, S.: 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3075–3084 (2019)
29. Songyou, P., Kyle, G., Chiyu, J., Andrea, T., Marc, P., Thomas, F.: Openscene: 3D scene understanding with open vocabularies. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–824 (2023)



# Augmented Reality-Assisted Environment for Medical Education: An Experience of Interactive and Immersive Learning

Vikas Puthannadathil Reghunatha Kumar<sup>1</sup>, Anurag Kujur<sup>1</sup>,  
Bishnu Ganguly<sup>1</sup>, Santosh Kumar Behera<sup>1</sup>, and Ajaya Kumar Dash<sup>2</sup>

<sup>1</sup> National Institute of Technology Calicut, Kozhikode, Kerala, India  
{vikas\_p230666cs, anurag\_m210691ca, bishnu\_m230338cs,  
skbehera}@nitc.ac.in

<sup>2</sup> International Institute of Information Technology Bhubaneswar,  
Bhubaneswar, Odisha, India  
ajaya@iiit-bh.ac.in

**Abstract.** Traditional methods of medical education often face challenges such as limited visual representation, lack of interactive experiences, and outdated technological integration. Generally, medical education often relies on static images or diagrams to represent complex anatomical structures, which can hinder students' understanding and retention of information. Additionally, the lack of interactive experiences restricts students' interaction and manipulation of anatomical models, creating a less engaging learning environment. This research offers a real-time, seamless method that tackles the noted difficulties by utilising AR technology. To give students a dynamic and engaging learning experience, the suggested AR-based solution seeks to project 3D human organs onto a tracked human body. Apart from that the user can interact with the superimposed organs in real-time using natural gestures and image markers. The proposed AR system leverages advanced technologies including a vision transformer for precise image recognition, deep learning techniques for human hand gesture recognition, and human pose tracking. By integrating these components, the system enables the accurate projection of organs onto a real-world environment, synchronized with the user's movements. Interactions with the virtual organs are facilitated through intuitive gestures, allowing users to manipulate, scale, and rotate the projected organs effortlessly. To sum up, incorporating augmented reality into medical education presents a viable way to get around current obstacles and give students a more dynamic, immersive, and interesting educational experience. By promoting a greater comprehension and recall of intricate anatomical concepts, this method has the potential to transform medical education.

**Keywords:** Learning Technology · Augmented Reality · Human Computer Interaction · Medical Education

# 1 Introduction

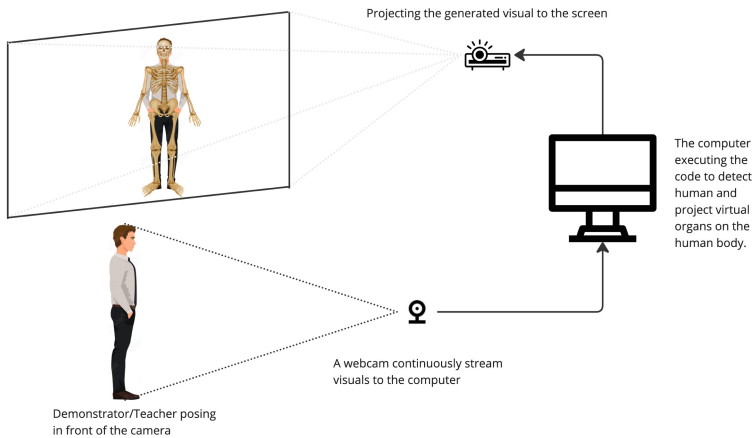
In the realm of medical education, a comprehensive understanding of human anatomy stands as the cornerstone for aspiring healthcare professionals. Proficiency in anatomy not only forms the basis of clinical practice, but also fosters the development of critical thinking and diagnostic skills. However, traditional approaches to teaching anatomy often fall short of providing students with the depth of understanding required to navigate the complexities of the human body. Recent revolutionary changes in technology, particularly in the field of augmented reality (AR), have opened new avenues for transforming medical education [17,20]. AR offers a dynamic platform that seamlessly integrates virtual elements into the real world, providing users with immersive and interactive experiences [1]. Johnson has stated that AR has strong potential to provide powerful contextual learning experiences [11]. Kirkpatrick has asserted that if visuals are shown to students, they can obtain a genuine knowledge of things more readily than they can be crammed with the verbal appearance of knowledge [14]. Along with providing immersive educational experiences for medical students, interactive AR allows them to engage with 3D models of anatomical structures, surgical procedures, and medical simulations. This hands-on approach enhances learning outcomes, as students can visualize complex concepts in a realistic context, leading to better retention and understanding.

The focus of recent advancements in interactive AR technology has been on enhancing engagement with virtual objects through various modalities. The existing modalities for interactive-AR technology can be broadly categorized into input device-based [15] (e.g. mouse [2] and keyboard [7], digital gloves, digital pens etc.), tangible user interfaces (TUI) [3], action-based (e.g. gesture [12], gaze [19], haptic [18] etc.), and multi-modal interfaces [13]. This convergence of advanced interaction modalities makes augmented reality an ideal platform for educational purposes, particularly in the field of medicine. The ability to incorporate multiple modalities allows for a smooth and intuitive user experience, enabling educators to convey complex concepts with clarity and precision [8]. In medical education, these technologies hold immense potential to benefit both teachers and students. However, the usability of the aforementioned modalities is confined to some constrained setups that limit the possibility of interactive augmentation anywhere. Researchers have argued that interactions with virtual objects using hand gestures are more natural than interactions using other devices [5,9].

Gesture-based interaction techniques can be broadly categorized into vision-based and sensor-based systems. While vision-based approaches are affordable, most existing systems are not robust as they are prone to background complexity, illumination, colour, the shape of the hand, finger movement, occlusion etc. On the contrary, sensor-based techniques use sophisticated sensors like Kinect or Leap Motion Controller to handle gestures. However, sensors often have some limitations. Firstly, sophisticated sensors are generally expensive. Secondly, the life-cycle of a sensor is always unforeseeable, e.g., Microsoft Corporation discontinued Kinect in few years after its successful launch. Thirdly, most sensors need

an uninterrupted power source and are not portable enough to support interactive AR anywhere. Apart from the above limitations, some of the available mixed reality interactive devices, such as Magic Leap 1, Microsoft HoloLens 2, Mira Prism Pro, or Apple Vision Pro, are expensive and not affordable to the masses.

Thus, after aligning the thoughts in a similar direction, a feasible, real-time, and inexpensive gesture-based interactive AR framework has been proposed in this article. Using the presented idea, teachers can utilize simple and natural hand gestures to elucidate intricate anatomical structures, while students gain access to visual representations that closely mimic real-world objects. The primary focus of this article is to harness the power of AR to enable educators to superimpose virtual human organs onto real human subjects with proper alignment. By doing so, audiences can observe the size, shape, and position of organs relative to an actual human body, enhancing their understanding of anatomical relationships and spatial orientation. This approach has the potential to significantly improve the teaching of anatomy, providing students with a more engaging and comprehensive learning experience.



**Fig. 1.** The setup process involves connecting the webcam and projector to the computer, positioning the webcam to capture the demonstrator, and aligning the projector to ensure a clear display

## 1.1 Technical Setup

The proposed AR system has an easy-to-use technical setup. It is designed to enhance educational experiences in anatomy at an affordable cost. The setup of the proposed system is demonstrated in Fig. 1. The system requires basic hardware: a computer, a webcam, and a projector. The computer manages video processing, image recognition, and AR overlay tasks. The webcam captures live

video of the demonstrator standing in front of it and streams it to the computer, which then processes the visual input and superimposes anatomical structures onto the video feed. The projector, connected to the computer, displays these augmented visuals on a large screen, making it easier for participants to follow the presentation.

The software workflow begins with the webcam capturing real-time video of the demonstrator. The computer then uses pose detection techniques to track key points on the body, allowing for an accurate overlay of 3D anatomical structures. This augmented video feed is projected onto a large screen for audience viewing. The system also supports gesture-based interaction, enabling teachers to manipulate virtual organs through simple hand movements. This includes changing, scaling, or repositioning the organs, allowing for an interactive presentation.

## 2 Proposed System

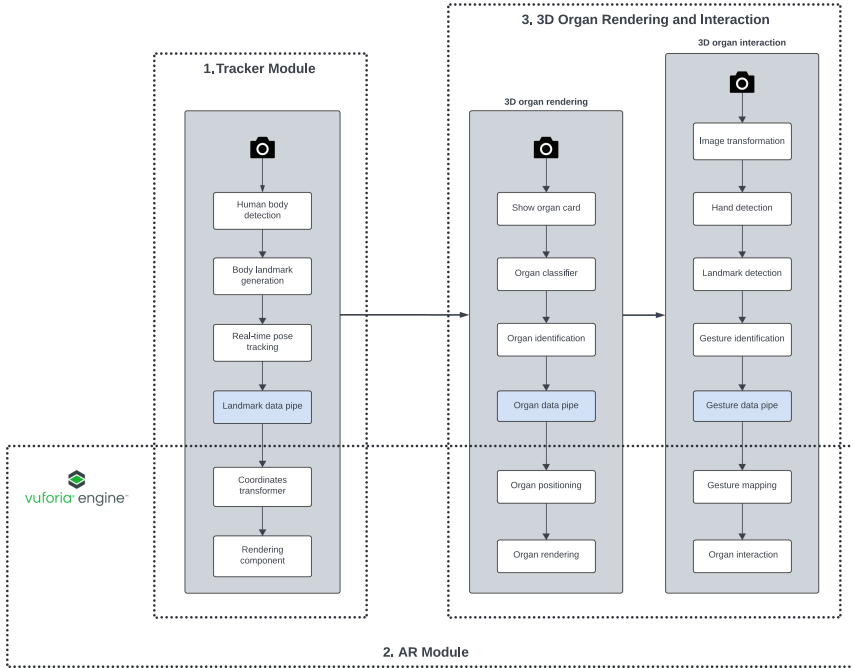
The proposed system uses AR technology with advanced Artificial Intelligence (AI) algorithms to enhance medical education through immersive and interactive learning experiences. As seen in Fig. 2, the proposed system is composed of three primary modules: the *tracker module*, the *AR module*, and the *3D rendering and interaction module*. The ensuing subsections explain the detailed description of these modules for a comprehensive AR-powered medical education system.

### 2.1 Tracker Module

This module generates the metadata that will be required by the other modules for the seamless integration of the whole system. It serves as the computational backbone of the system, leveraging computer vision algorithms to analyse real-world data and channel relevant information to the AR module and 3D rendering modules. This metadata includes the tracking components as well as the approximate landmark coordinate values of a human standing in front of a camera. The subsequent processing and portrayal of the anatomical organs in the proper locations on a real human body depend on the landmark values generated by the tracker module. This module consists of various sequential components, including human body detection, body landmark generation, real-time pose tracking and landmark data pipe.

We have used MediaPipe [16] baseline architecture to build the pipeline to detect the human body as well as the pose landmarks. This information can be used to track the position and movement of the human body within the environment accurately. The pose landmarks are the 3D points that can be used to enable the precise placement of virtual organs relative to the human body.

This proposed system uses the MediaPipe's pose landmarker, an advanced deep learning (DL) solution for high-fidelity human pose tracking which uses the blazePose model to predict 33 human landmark points as depicted in Fig. 3 [16]. The main feature of this model is its ability to process each frame of the input live feed from the camera in real time and produce continuous pose estimation,



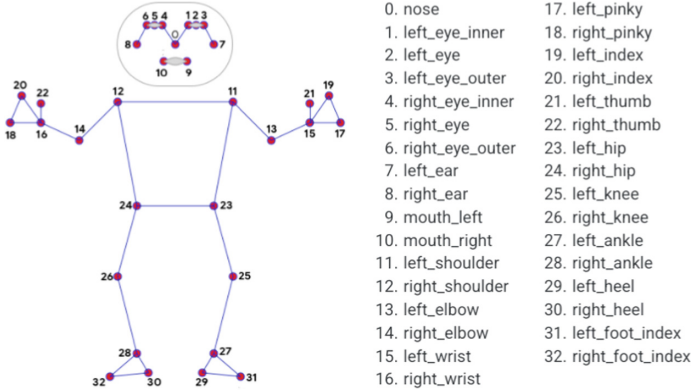
**Fig. 2.** Complete methodology architecture of the proposed system consists of three modules: Tracker module, AR module, and 3D organ rendering and interaction module. The interaction between these three modules builds our proposed medical AR system

as seen in Fig. 4. These pose estimation landmarks are sent to the AR module to render a 3D virtual skeleton over the real human body.

**Landmark Data Pipe:** This pipe acts as an interface between the tracker and the AR modules, as mentioned in Fig. 2. The metadata generated in the tracker module is transmitted through this data pipe to the AR module for the accurate placement of the virtual skeleton over the human body.

## 2.2 AR Module

The purpose of the AR module is to overlay computer-generated 3D virtual objects over the actual environment after obtaining the required data from the tracker module. This allows for the easy integration of anatomical models into real-world environments. However, the data in the tracker module uses OpenCV's [4] coordinate frame-of-reference while the AR module refers the Unity's coordinate frame-of-reference as shown in Fig. 5. To maintain the synchronization between these two frame-of-references, we need a transformation before the actual rendering in the AR module. Hence, the AR module consists of two main parts: the *coordinate transformer* and the *rendering component*.



**Fig. 3.** Pose landmark detector [16] gives the 33 landmark points mapped to specific parts of the body



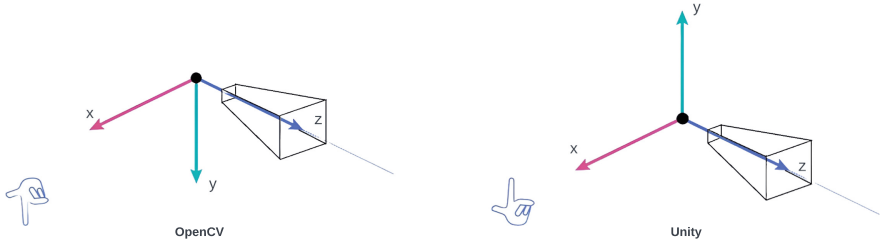
**Fig. 4.** Camera feed takes the human body as input and pose landmarker estimates the body pose using human landmark points

Together, these two components of the AR module, are employed to superimpose virtual organs onto the human image.

**A. Coordinate Transformer.** As specified earlier, the tracker module and the AR module use different coordinate systems. Thus, the landmark data points generated in the tracker module need to be transformed before they can be used inside the AR module. By looking at the coordinate frame-of-references in Fig. 5, we can say, they only differ by the direction of the  $y$ -axis. The task of the coordinate transformer is to ensure the aforementioned necessary transformation which can be used to develop an AR environment in Unity. If  $P_t (x_t, y_t, z_t)$  and  $P_a (x_a, y_a, z_a)$  refer to the point in the tracker module and AR module, respectively, they can be synchronized by a transformation matrix  $T$  as in Eq. (1),

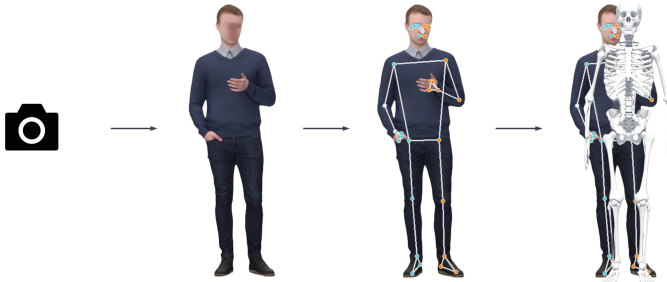
$$P_a = TP_t \implies \begin{bmatrix} x_a \\ y_a \\ z_a \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} \quad (1)$$





**Fig. 5.** The different coordinate frame-of-reference used by tracker module (OpenCV framework) and AR module (Unity framework)

**B. Rendering Component.** After the 3D landmark points from the tracker module are converted to the coordinate system that Unity uses, the rendering component uses the transformed coordinate points to project the virtual organ into a real-world setting. The rendering component dynamically adjusts the position, orientation, and appearance of virtual organs based on the user's position. Vuforia engine in Unity is used for rendering the virtual 3D object in the AR environment setup. It uses the device camera to view and understand the real-world environment. The transformed coordinate points act as reference points to place virtual objects in the AR environment. The virtual human skeleton has been superimposed over the human body using reference points as shown in Fig. 6.



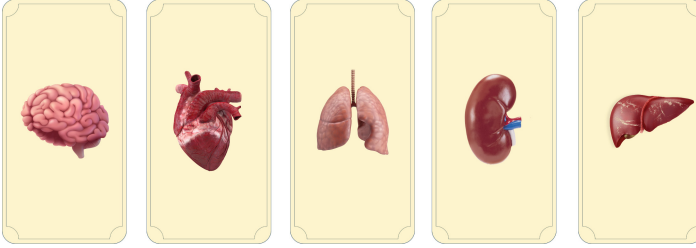
**Fig. 6.** Camera feed takes the human body as input; pose landmarker generates the landmark points and estimates the body pose; and finally human skeleton is augmented over the human body

Rendering virtual organs accurately on a detected human presents significant challenges, particularly in terms of ensuring precise alignment and scaling relative to the individual's pose and dimensions. Accurate human pose detection is crucial for this process, as it forms the foundation for scaling, rotating, and transforming virtual organ objects to match the human body's dimensions and orientation. Given that the entire body might not be visible in the camera frame, the system relies on landmark points to infer body position and proportions. By dynamically adjusting the virtual organs based on the detected landmark points,

the system achieves accurate placement and scaling, resulting in a realistic and educationally effective augmented reality experience.

### 2.3 3D Organ Rendering and Interaction Module

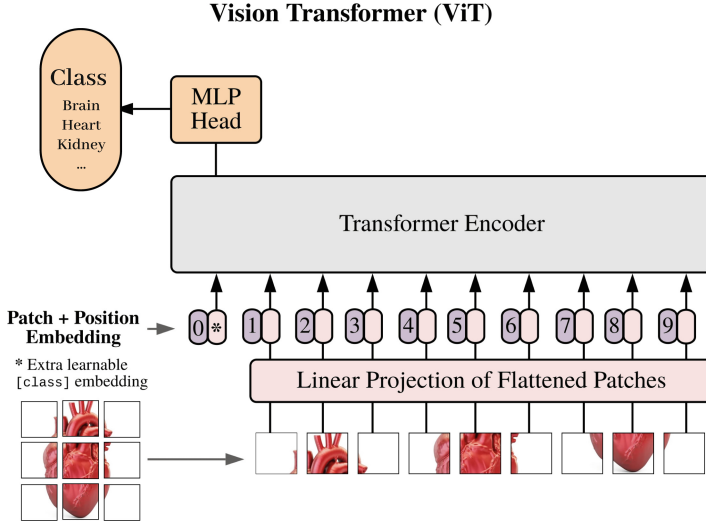
This module provides an extension for overlaying and interacting with 3D virtual anatomical organs after the tracker module and AR module tracked and superimposed the 3D virtual skeleton over the human subject. The two main sub-parts of this module are the *3D organ rendering* and *3D organ interaction*.



**Fig. 7.** Physical organ cards which will be identified by the organ classifier for appropriate organ activation and rendering

**A. 3D Organ Rendering:** This part uses the coordinate information gathered from the tracker module to render 3D organs over the human body at the correct location and interact with them seamlessly. AR module helps in the proper rendering and interaction of the organs by creating an AR environment set up for all the virtual structures. Each organ is associated with the organ card which acts as a fiducial marker for that organ. Firstly, physical organ cards, as shown in Fig. 7, are brought in front of the camera for recognition. These physical cards go through an organ classifier algorithm for the accurate identification of the organ. Depending on the identification result, the respective organ should be generated in real-time based on the pose of the human subject and superimposed over the human body. This approach will help the presenter to give more visual clues to the audience and it is more robust than any other forms of teaching like verbal instructions. For accurate identification of the organ cards, the Vision Transformer (ViT) algorithm is used with a custom dataset to train the classifier.

The Vision Transformer (ViT) algorithm can learn from large diverse datasets, uniform feature representation and strong information propagation to achieve remarkable performance. Due to its high performance and less need for vision-specific inductive bias, the transformer has gained much popularity in the computer vision community. Research studies have also shown that ViT can perform better than convolutional and recurrent neural networks on some visual benchmarks [10]. Thus here, the ViT algorithm is used to recognize the organ card shown in front of the camera. A detailed explanation of how to produce virtual 3D organs from physical organ cards using ViT is described below:



**Fig. 8.** The organ image goes through a patch embedding process to generate image patches which are then flattened and transformed and then fed into the transformer encoder which consists of self-attention layers and feed-forward neural networks. Finally, the embeddings are sent to the classification head for organ classification

**Vision Transformer.** ViT transforms images into sequences of fixed-size patches, enabling efficient processing and scalability. we resize the input image to  $224 \times 224$  pixels and extract patches of size  $16 \times 16$  as shown in Fig. 8 [6].

- **Patch Embedding:** Each  $16 \times 16$  patch is linearly embedded into a token representation. Let  $X$  denote the input image and  $P$  represent the set of patches extracted from  $X$ . Each patch  $p_i$  is embedded into a token  $x_i$  using a learnable linear projection:

$$x_i = W \cdot p_i + b \quad (2)$$

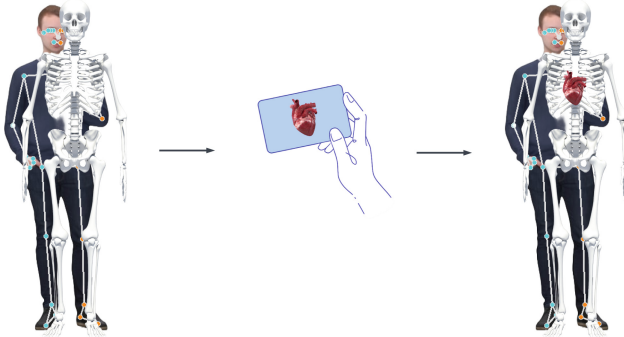
where,  $W$  and  $b$  are the weight matrix and bias vector, respectively.

- **Transformer Encoder:** The token sequence undergoes multiple Transformer encoder layers, capturing global dependencies and learning hierarchical representations. Mathematically, at layer  $l$ ,  $H^l = \text{Transformer Layer}(H^{l-1})$ , where  $H^l$  represents the hidden representations.
- **Classification Head:** The pooled token embeddings are passed through a classification head to predict organ class probabilities. Let  $z$  denote the pooled token embeddings. The predicted probability distribution over organ classes is obtained as

$$\hat{y} = \text{Softmax}(W_c \cdot \text{Pool}(z) + b_c), \quad (3)$$

where  $W_c$  and  $b_c$  are the weight matrix and bias vector of the classification head.

We fine-tune the pre-trained ViT model on the dataset of organ images to adapt it to our task requirements.



**Fig. 9.** Organ classifier identifies the organ card and the corresponding organ is activated in the accurate body location

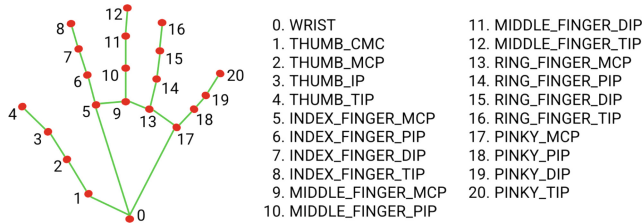
Once the organ is classified and identified it is sent to the AR module through the organ data pipe for further organ rendering. The body coordinates generated from the tracker module will be utilised to find the accurate position of the organ in the human body. Finally, the 3D organ will be activated and rendered in the correct location of the human body as shown in Fig. 9.

**B. 3D Organ Interaction:** This component interprets hand gestures made by presenters or demonstrators in the field of view (FOV) of the camera. It translates these gestures into simple text commands, which are then passed to the AR module for action. The gesture recognizer is thoroughly discussed below.

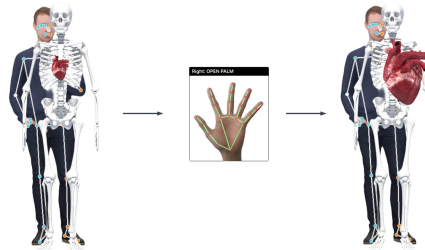
***Gesture Recognizer.*** The gesture recognizer component is designed for real-time hand gesture recognition. Its architecture comprises multiple layers, each serving a specific function to enable accurate and efficient recognition of hand gestures. The component operates within a framework tailored for inference from sensory data, with a focus on modularity and scalability. Below is a breakdown of the key elements and their functionalities:

- **Image Transformation Layer:** This initial layer preprocesses the input image, transforming it into a format suitable for further analysis. This includes resizing, normalization, and noise reduction, ensuring that the subsequent layers receive clean and standardized input data.
- **Hand Detection Layer:** Following preprocessing, the hand detection layer identifies and segments the region of interest corresponding to the hand within the image. This isolates the hand from the background and other irrelevant elements, thereby enhancing the accuracy of gesture recognition.
- **Landmark Detection Layer:** Once the hand region is detected, the landmark detection layer identifies the position of predefined points on the hand, typically represented as landmarks or key points. These points capture important anatomical features of the hand as shown in Fig. 10 and are subsequently translated into three-dimensional coordinates to provide spatial information.

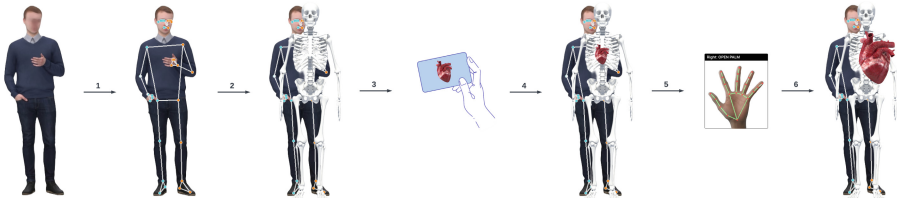
- o **Gesture Identification Layer:** Building upon the landmark coordinates, the gesture identification layer classifies the current hand configuration into predefined gestures. This classification process involves comparing the detected landmarks with reference points corresponding to known gestures, utilizing techniques such as machine learning or pattern recognition. The identified gesture, along with the landmark points, is then sent to the output layer for further processing or visualization. Figure 11 shows how the OPEN PALM gesture is detected and used to increase the size of the heart object.



**Fig. 10.** Hand landmarker generates 21 hand landmark points mapped to specific parts of the hand



**Fig. 11.** The gesture recognizer recognizes the OPEN PALM gesture, and the organ size increases until the gesture is recognized



**Fig. 12.** Pipeline of the complete methodology: 1. Body pose detection using landmarks, 2. Skeleton projection over the body using the body pose data, 3. Organ card recognition using organ classifier, 4. Corresponding organ projection at accurate body location, 5. Gesture recognition using gesture recogniser, 6. Organ interaction according to gesture



**Fig. 13.** Real-time tracking of the human body and overlaying human skeleton over the human body in AR environment

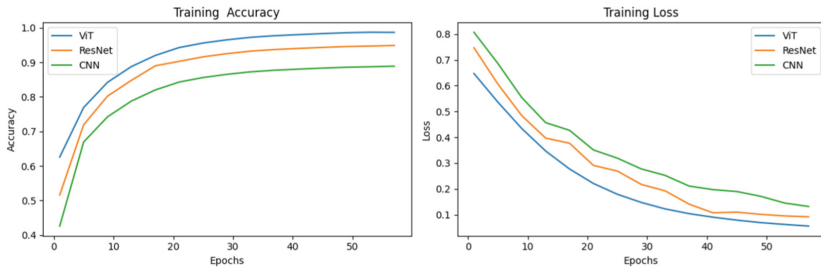
### 3 Results and Discussion

The overall insight into the pipeline of the complete methodology is depicted in Fig. 12. The proposed AR system addresses the challenges encountered in traditional medical education methods by providing immersive visual representations and interactive experiences. Through rigorous testing and evaluation, the following key findings emerged:

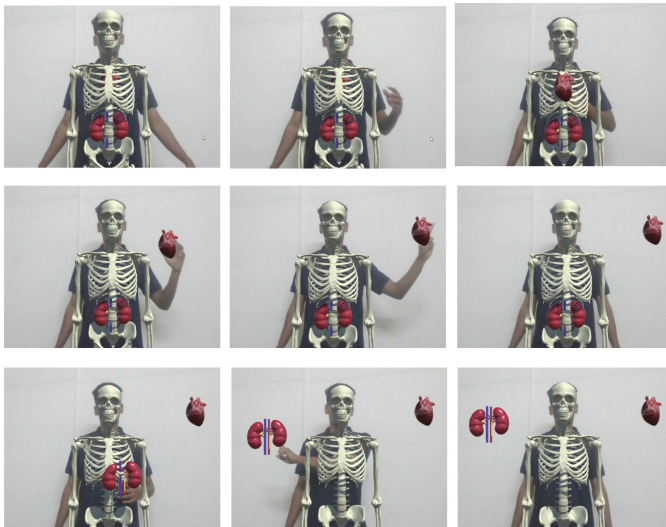
- (i) **Real-time Tracking and Visualization:** By incorporating Mediapipe human pose tracking, the proposed system achieves real-time tracking of the user's movements, ensuring dynamic alignment and visualization of organs relative to the user's body position, as demonstrated in Fig. 13. This real-time tracking enhances the realism and immersion of the educational experience, facilitating a deeper understanding of anatomical relationships and spatial dynamics.
- (ii) **Specialized Vision Transformer for Organ Recognition:** A specialized Vision Transformer model for organ recognition is trained on a dataset comprising 400 images each of heart, kidney, brain, and other organ classes, totalling 2400 images across four classes. 80–20% split is used to train and validate the model. Additionally, 100 images per class are reserved for testing, resulting in a balanced dataset of 400 test images. For model evaluation, we conduct the training for 50 epochs, monitoring both training and testing accuracy in 10 separate runs. The average training and testing accuracy have been noted as 98.7% and 96.9%, respectively. The trajectory of the loss function and corresponding accuracy over the epochs is plotted in Fig. 14. These visualizations show satisfactory results on model performance. Additionally, we use a set of organ cards, as depicted in Fig. 7, to further test the model's recognition capabilities.
- (iii) **Accurate Organ Projection:** An achievement of our project is the seamless projection of human organs onto a tracked human body shown in Fig. 15, ensuring impeccable alignment and placement within the virtual environment. By utilising Vision Transformer technology for image recognition, the system showcases high accuracy in identifying and overlaying anatomical structures onto the user's body, marking it very useful in advancing medical education through augmented reality.
- (iv) **Intuitive Gesture Interaction:** Harnessing the power of Mediapipe for human hand gesture recognition, our system proudly enables intuitive interaction with the projected organs. Users effortlessly control virtual objects

through natural hand movements, seamlessly executing gestures for moving, scaling, and rotating the organs. This intuitive interface is a testament to our achievement, elevating user engagement and fostering immersive hands-on learning experiences.

- (v) **Enhanced Learning Engagement:** Through user testing and feedback, it was observed that the AR-based approach significantly enhances learner engagement and retention compared to traditional teaching methods. The interactive nature of the system promotes active participation and exploration, leading to improved comprehension and knowledge retention among students.



**Fig. 14.** Training metrics visualization over the epochs in the seventh run: (left) Loss vs. Epoch; (right) Accuracy vs. Epoch



**Fig. 15.** Accurate organ projection and repositioning of organ using the POINTING UP gesture

Overall, the results demonstrate the efficacy of the proposed AR solution in addressing the limitations of current medical education practices. By seamlessly integrating advanced technologies, the system offers a transformative approach to medical education, empowering learners with immersive, interactive experiences that bridge the gap between theory and practice.

## 4 Conclusion and Future Enhancements

The integration of augmented Reality technology into medical education is going to solve and overcome a lot of traditional challenges being faced in the current medical education system and make learning better for both students and teachers. Advanced technologies like artificial intelligence and human-computer interaction used in this system will increase the engagement and understanding of students. Several problems faced in the conventional methods of teaching medical science are successfully solved in this system. This system creates a digital space for the students where they can create, place and manipulate the complex anatomical body organs according to their needs and understanding. The students own this digital space and they get the power and freedom to learn medical science in their own way. It helps students to get a better understanding of spatial relationships and body movements by using real-time tracking and visualization.

The current version includes the basic functionalities necessary for studying organ structure and behaviour. The next versions will include the integration of multi-modal learning such as audio for much better immersive learning. Potential enhancement may also include the incorporation of animated organ capabilities into the AR system, so enabling users to engage in interactive manipulation. The addition and training of more user-defined gestures will also serve for better manipulation of the anatomical organs.

## References

1. Azuma, R., Bailiot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, B.: Recent advances in augmented reality. *IEEE Comput. Graph. Appl.* **21**, 34–47 (2001)
2. Besançon, L., Issartel, P., Ammi, M., Isenberg, T.: Mouse, tactile, and tangible input for 3D manipulation. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17*, pp. 4727–4740. Association for Computing Machinery, New York, NY, USA (2017). <https://doi.org/10.1145/3025453.3025863>
3. Bozgeyikli, E., Bozgeyikli, L.L.: Evaluating object manipulation interaction techniques in mixed reality: tangible user interfaces and gesture. In: *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 778–787. IEEE (2021). <https://doi.org/10.1109/VR50410.2021.00105>
4. Bradski, G.: *The OpenCV Library*. Dr. Dobb's Journal of Software Tools (2000)
5. Dash, A.K., Balaji, K.V., Dogra, D.P., Kim, B.G.: Interactions with 3D virtual objects in augmented reality using natural gestures. *Vis. Comput.* **40**(9), 1–14 (2023). <https://doi.org/10.1007/s00371-023-03175-4>



6. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
7. Froehlich, B., Hochstrate, J., Skuk, V., Huckauf, A.: The globefish and the globe-mouse: two new six degree of freedom input devices for graphics applications. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06, pp. 191–199. Association for Computing Machinery, New York, NY, USA (2006). <https://doi.org/10.1145/1124772.1124802>
8. George, O., Foster, J., Xia, Z., Jacobs, C.: Augmented reality in medical education: a mixed methods feasibility study. *Cureus* **15**(3) (2023). <https://doi.org/10.7759/cureus.36927>
9. Ha, T., Woo, W.: Bare hand interface for interaction in the video see-through HMD based wearable AR environment. In: Harper, R., Rauterberg, M., Combetto, M. (eds.) ICEC 2006. LNCS, vol. 4161, pp. 354–357. Springer, Heidelberg (2006). [https://doi.org/10.1007/11872320\\_48](https://doi.org/10.1007/11872320_48)
10. Han, K., et al.: A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(1), 87–110 (2023). <https://doi.org/10.1109/TPAMI.2022.3152247>
11. Johnson, L., Levine, A., Smith, R., Stone, S.: The 2010 Horizon Report. The New Media Consortium, Austin, Texas (2010)
12. Kerdvibulvech, C.: A review of augmented reality-based human-computer interaction applications of gesture-based interaction. In: Stephanidis, C. (ed.) HCII 2019. LNCS, vol. 11786, pp. 233–242. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-30033-3\\_18](https://doi.org/10.1007/978-3-030-30033-3_18)
13. Kim, J.C., Laine, T.H., Ahlund, C.: Multimodal interaction systems based on internet of things and augmented reality: a systematic literature review. *Appl. Sci.* **11**(4) (2021). <https://doi.org/10.3390/app11041738>
14. Kirkpatrick, E.A.: An experimental study of memory. *Psychol. Rev.* **1**(6), 602 (1894). <https://doi.org/10.1037/h0068244>
15. Krichenbauer, M., Yamamoto, G., Taketom, T., Sandor, C., Kato, H.: Augmented reality versus virtual reality for 3D object manipulation. *IEEE Trans. Vis. Comput. Graph.* **24**(2), 1038–1048 (2018). <https://doi.org/10.1109/TVCG.2017.2658570>
16. Lugaresi, C., et al.: MediaPipe: a framework for building perception pipelines. arXiv eprint [arXiv:1906.08172](https://arxiv.org/abs/1906.08172) (2019). <https://doi.org/10.48550/arXiv.1906.08172>
17. Mendoza-Ramírez, C.E., Tudon-Martínez, J.C., Félix-Herrán, L.C., Lozoya-Santos, J.d.J., Vargas-Martínez, A.: Augmented reality: survey. *Appl. Sci.* **13**(18), 1–35 (2023). <https://doi.org/10.3390/app131810491>
18. Pacchierotti, C., Sinclair, S., Solazzi, M., Frisoli, A., Hayward, V., Prattichizzo, D.: Wearable haptic systems for the fingertip and the hand: taxonomy, review, and perspectives. *IEEE Trans. Haptics* **10**(4), 580–600 (2017). <https://doi.org/10.1109/TOH.2017.2689006>
19. Pfeuffer, K., Mayer, B., Mardanbegi, D., Gellersen, H.: Gaze + pinch interaction in virtual reality. In: Proceedings of the 5th Symposium on Spatial User Interaction, SUI 2017, pp. 99–108. Association for Computing Machinery, New York (2017). <https://doi.org/10.1145/3131277.3132180>
20. Tang, K.S., Cheng, D.L., Mi, E., Greenberg, P.B.: Augmented reality in medical education: a systematic review. *Can. Med. Educ. J.* **11**(1), e81–e96 (2020). <https://doi.org/10.36834/cmej.61705>



# P2A: Transforming Proposals to Anomaly Masks

Huachao Zhu<sup>1</sup>, Zhichao Sun<sup>1</sup>, Zelong Liu<sup>1</sup>, and Yongchao Xu<sup>1,2</sup>

<sup>1</sup> School of Computer Science, Wuhan University, 430072 Wuhan, China  
{huachao.zhu, zhichaosun, zelong.liu, yongchao.xu}@whu.edu.cn

<sup>2</sup> Hubei LuoJia Laboratory, 430079 Wuhan, China

**Abstract.** Detecting anomalies in road scenes is essential for safe autonomous driving. Existing methods often consider the likelihood of pixels not belonging to a closed set of classes as the anomaly score. However, this approach lacks object-level understanding and frequently results in numerous false positives at boundaries and ambiguous regions. In this paper, we present a novel method that directly computes the probability of pixels being anomalous and outputs both anomaly segmentation results and score maps. Our approach utilizes the rich semantic information correlated to linguistic concepts in Stable Diffusion to compensate for the low coverage of anomalies caused by limited annotated samples. Using a query-based segmentation model, we transform the proposals into masks of both in-distribution and out-of-distribution objects. Additionally, we introduce an image-mask-image pipeline to generate various annotated data as outliers for supervised training. Extensive experiments across multiple benchmarks confirm that the proposed method outperforms previous state-of-the-art methods in road anomaly segmentation. Code is available at <https://github.com/huachao0124/P2A>.

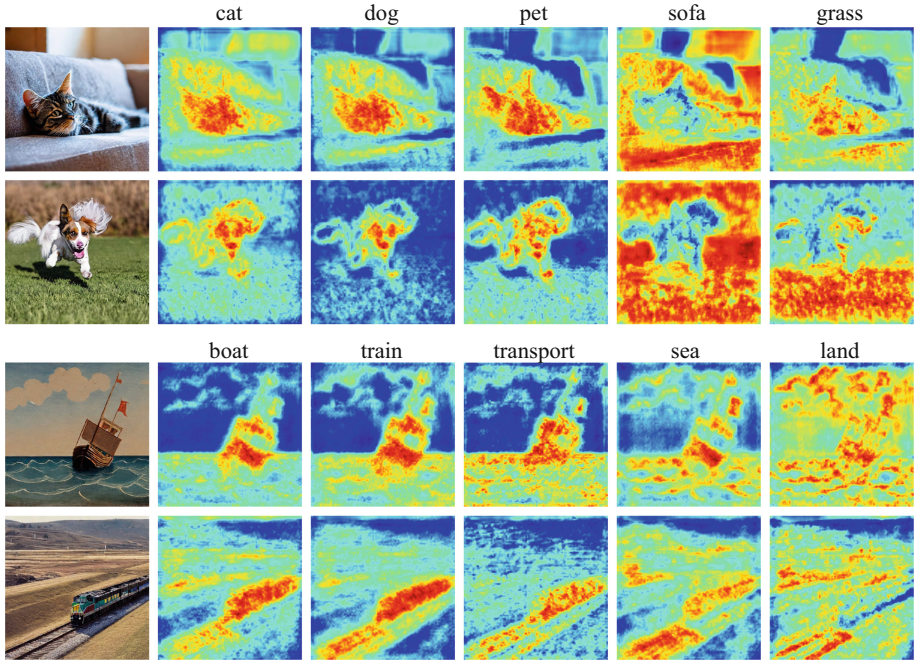
**Keywords:** Road anomaly detection · Stable diffusion · Semantic segmentation

## 1 Introduction

Most visual tasks assume a closed-world scenario with a fixed set of known categories. With sufficient annotations. However, for anomaly detection in driving scenarios, the road may present anomalies of varying sizes, locations, and types, including but not limited to animals, stones, and garbage. The limited number of annotated samples with anomalies makes supervised anomaly detection extremely challenging.

Existing methods hold a very intuitive reverse thinking that the less likely an object belongs to a known category, the more likely it is an anomaly. How to measure the probability that a pixel does not belong to a known category is crucial for determining the model performance. Some methods rely on that multiple models usually give the same prediction for known categories, while they often show

inconsistencies in anomalous regions. They employ segmentation module ensemble [12] or Monte Carlo sampling [20] to obtain multiple segmentation maps. However, this leads to much higher inference costs and suffers from unsatisfactory results from uncontrollable generalization. Other methods utilize Softmax probability entropy [3], maximum logit [19], or the sum of logits [16, 27, 33] for known categories as confidence scores. This reverse scoring lacks object-level comprehension, leading to high false-positive rates at boundaries between different semantic segments and ambiguous regions. Moreover, segmentation models, such as DeepLabv3 [4], typically approach segmentation as per-pixel classification, generating probability distributions for every pixel in the image, which may output masks that only partially cover anomalous regions, failing to capture the entire extent of the anomalies. Although some recent approaches [27, 30] adopt mask classification based segmentation model, Mask2Former [5] that transforms proposals to masks with classification distribution, their design still relies on supervision at the pixel level, suffering the same problem as mentioned above.



**Fig. 1.** Heat maps of cross attention weights between the text embedding of the noun and the visual representations. Linguistically related concepts (e.g., ‘cat’ and ‘dog’) can activate each other, potentially enabling the model to generalize to novel objects not present in the training data

Vision-language models have shown promise in zero-shot classification [29] and open-vocabulary segmentation [14, 22]. For instance, CLIP [29] trained by

contrast learning on billion-scale image-text pairs learns rich expressive multi-modal features that capture broad concepts. Stable Diffusion [31] uses the text encoder of CLIP [29] to control the generated image through cross attention, which also inherits the multi-modal representations. We use Stable Diffusion [31] to generate some images and extract the cross-attention maps between the text embedding of the noun and the visual representations. These maps are then averaged across all attention heads and layers to generate the heat map, as demonstrated in Fig. 1. Both ‘dog’ and ‘pet’ indicate the location of the dog. Besides, ‘cat’ and ‘dog’ can represent each other, while unrelated objects like ‘grass’ and ‘sofa’ cannot. The same observation is also reflected in ‘train’ and ‘ship’. This is because in linguistic concepts, ‘cat’ and ‘dog’ both belong to pets, while ‘train’ and ‘boat’ are related to transportation. This property benefits generalization ability, and helps competitive low-shot performances in both seen and unseen objects compared to full supervision.

In this paper, we propose P2A, a novel method for road anomaly detection that transforms the **Proposals** to masks of **Anomalies** and also directly measures the probability of objects being anomalies. This forward scoring considers anomalies on the object level rather than the pixel level, helping to eliminate false positives at boundaries. The primary challenge is how to enable the model to deal with various types of anomalies. We approach this in two ways: (1) Leveraging visual representations aligned with human conceptual understanding; and (2) Synthesizing a comprehensive range of objects with corresponding masks to mimic the anomalies. Specifically, we utilize the visual representations from Stable Diffusion [31], which is incorporated with a high-level conceptual understanding of human natural language. This enables P2A to generalize to previously unseen objects that share similar semantics with those in the training set, making directly measuring the probability of pixels belonging to anomalies feasible. Besides, to train the model on a sufficient variety of anomalies as in real-world scenarios, we design a novel image-mask-image pipeline that generates image-mask pairs according to specified categories. This approach offers two key advantages: first, it circumvents the need for costly manual annotations, and second, it liberates us from the constraints imposed by predefined categories in existing datasets.

While ensuring the generalization capability of visual representations, we then generate masks for every entity. Compared with per-pixel classification based segmentation approaches [4, 41], mask classification based segmentation models [5, 21] have supervision at the whole object level, enabling them to better segment entire objects. They first generate a set of candidate masks and then classify them into known categories or void. For instance, Mask2Former [5] employs randomly initialized queries to group pixels into proposals. Once trained, these queries capture both semantic and spatial information, being able to cluster surrounding pixels with similar semantics. This process is highly compatible with visual representations with higher-level conceptual understanding from natural language, and their combination can generalize effectively to segmenting unseen objects. With visual representations enhanced by natural lan-

guage and queries that group pixels with similar semantics as proposals, our P2A is able to directly obtain the probability of each mask being an anomaly, in addition to the likelihood of every pixel belonging to known categories. Experiments demonstrate that the former (forward thinking) works well for large anomalous objects but is insensitive to small ones, whereas the latter (reverse thinking) is overly sensitive to all anomalies as well as boundaries and ambiguous areas. The combination of both helps achieve state-of-the-art performance across several benchmarks for road anomaly detection. Notably, the closed-set segmentation performance remains unaffected. To summarize, the main contributions of this work are as follows:

- We present P2A, an innovative approach that leverages queries to cluster pixels with similar semantics into mask proposals. This enables us to obtain the probability of each proposal representing an anomaly.
- We design an image-mask-image pipeline to generate synthetic images containing objects along with corresponding masks, thereby alleviating the constraints imposed by the predefined categories present in existing datasets.
- We conduct comprehensive experiments across multiple datasets, showcasing the effectiveness and versatility of the proposed method in scenarios under different conditions.

## 2 Related Work

### 2.1 Road Anomaly Detection

Detecting anomalies in road scenes is crucial for the safety of autonomous driving. Current approaches can be broadly categorized into two lines: reconstruction-based and uncertainty-based methods. Reconstruction-based methods assume that the reconstruction networks trained on road scene images with only known categories may not generalize well to previously unseen objects. Consequently, areas containing obstacles will exhibit noticeable differences in appearance between the reconstructed image and the original image. Lis et al. [25] re-synthesize the image based on the semantic segmentation labels, which prevents the leakage of anomalous information from the input image. Uncertainty-based models operate on the principle that segmentation models will exhibit low confidence or high uncertainty when encountering unseen or anomalous regions, which is reflected by the output probability distribution or the logits before the softmax layer. Hendrycks et al. [18] proposes a simple baseline approach that applies a threshold over the maximum softmax probability to distinguish between in-distribution and out-of-distribution data. Jung et al. [19] notices that the distribution of max logits of each known category is significantly different from each other. Thus, they propose SML, the standardized max logits calibrated by respective mean and variance obtained from statistics on the training set. To increase the score difference between known regions and anomalies, Meta-OoD [3] pastes samples from the COCO [24] dataset as out-of-distribution

proxy and trains the model again to maximize the Softmax entropy on these samples.

All existing methods adopt a reverse thinking approach, *i.e.*, considering that the less likely a pixel belongs to a known category, the more likely it is anomalous. While this reasoning is fundamentally correct, it lacks object-level understanding, leading to numerous false positives at boundaries and areas with semantic ambiguity. In contrast, the proposed method can directly identify and mask out anomalous regions, thereby reducing the occurrence of such false positives.

## 2.2 Open-Vocabulary Segmentation

Open-vocabulary segmentation is an emerging task that aims to interpret an image by categorizing regions into arbitrary classes specified through textual descriptions. In human cognition, the concept of “anomaly” serves as a generalized term for any unknown or unrecognized entity in a scene. Therefore, we can view road anomaly segmentation as a specialized subset of the broader open-vocabulary segmentation task. Some methods [9, 23, 37–39] first generate class-agnostic mask proposals and then leverage pre-trained vision-language models to classify these masked regions. Other methods [14, 22] present the task as a zero-shot generalization task, taking advantage of the correlation between the visual and linguistic representations in the vision-language model. For instance, LSeg [22] aligns pixel embeddings to the text embedding of the corresponding semantic class, which is generated by the text encoder of CLIP [29]. Excitingly, the trained model exhibits good generalization capabilities for relevant concepts in language, such as “cat” and “dog”. This demonstrates that the semantics highly correlated with linguistic concepts hold great potential to enable the segmentation of previously unseen objects.

# 3 Method

In this work, we reframe road anomaly detection as the binary classification of masks. We first perform an analysis on how queries serve as proposals in mask classification based segmentation models. Then We introduce the proposed method in detail.

## 3.1 Queries as Proposals

Most of segmentation models predict the probability distribution for every pixel for the input image. Differently, mask classification based methods [5, 21] first group pixels into regions, and then associate each region as a whole with some distribution. Mask2Former [5] optimize randomly initialized queries to cluster pixels with similar semantics into regions. SAM [21] adopts points, boxes, and texts as prompts to get corresponding masks. We build our method on top of the Mask2Former [5] architecture. Below, we will briefly review how semantic segmentation is formulated as a mask classification problem.



Given an image  $\mathbf{x} \in \mathbb{R}^{3 \times H \times W}$ , the backbone extracts multi-resolution feature maps  $\{\mathbf{f}_s \in \mathbb{R}^{C_s \times H_s \times W_s}\}_{s=1}^S$ , where  $S$  represents the number of resolutions. Then the pixel decoder unifies the channel dimensions of these features and processes them except the highest resolution one with deformable attention for multi-scale semantic information and generates per-pixel embeddings  $\{\mathbf{e}_s \in \mathbb{R}^{C_p \times H \times W}\}_{s=1}^S$ . The transformer decoder holds  $N$  queries  $\mathbf{q} \in \mathbb{R}^{N \times C_q}$ . In each block of the transformer decoder, there is a self-attention on  $\mathbf{q}$ , a cross attention between  $\mathbf{q}$  and  $\{\mathbf{e}_s\}_{s=2}^S$ , followed by a feed-forward network. Through the updating process in the transformer decoder,  $\mathbf{q}$  are transformed to candidates  $\mathbf{q}_c \in \mathbb{R}^{N \times C_p}$ . The masks  $\mathbf{m} \in \mathbb{R}^{N \times H \times W}$  are generated by the dot product of  $\mathbf{q}_c$  and  $\mathbf{e}_1$ , with sigmoid  $\sigma$  for normalization:

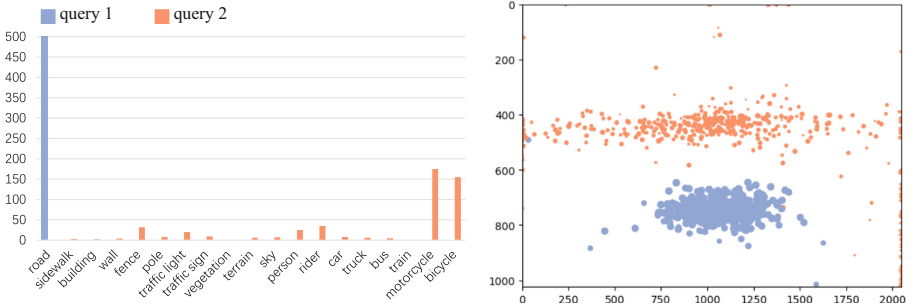
$$\mathbf{m} = \text{Upsample}(\sigma(\mathbf{q}_c \mathbf{e}_1)). \quad (1)$$

Meanwhile, the corresponding classification distribution  $\mathbf{c} \in \mathbb{R}^{N \times K}$  are generated by a simple MLP followed by a Softmax:

$$\mathbf{c} = \text{Softmax}(\text{MLP}(\mathbf{q}_c)). \quad (2)$$

The logits  $\mathbf{l} \in \mathbb{R}^{K \times H \times W}$  are the probability weighted masks over all queries:

$$\mathbf{l} = \sum_{n=1}^N \mathbf{c}_n \cdot \mathbf{m}_n. \quad (3)$$

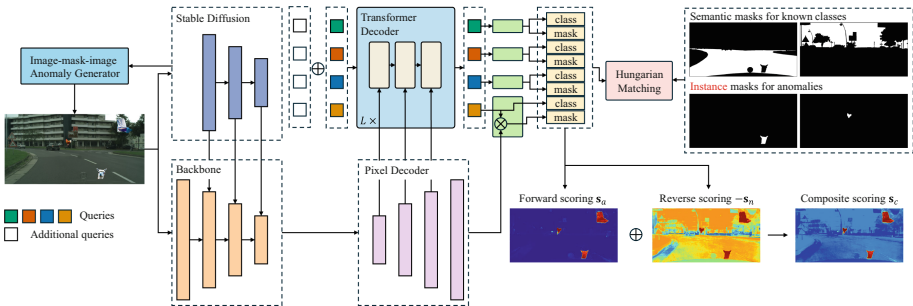


**Fig. 2.** We examine the masks and corresponding categories predicted by queries on the Cityscapes validation set. Left: The classification results of two queries. Right: The masks’ centers of two queries. Once trained, queries serve as proposals, indicating potential categories and locations. Each query consistently contributes to a specific semantic group, including one or more categories

During training, the predicted segments are matched with the ground-truth segments through bipartite matching. The corresponding queries gradually learn semantic and spatial information through this process. Once training is completed, these queries serve as proposals, indicating the locations where certain

objects may be found. It is noteworthy that a specific query may contribute not only to the prediction of a particular semantic segment but also to multiple semantic segments that share similar semantics. We randomly select 2 queries and calculate their classification results and centers of masks on Cityscapes [6] validation set. As illustrated in Fig. 2, query 1 consistently contributes to the segmentation of the “road” category, with the centers of the predicted masks always located on the road surface. On the other hand, query 2 is usually classified as the “motorcycle” and “bicycle” categories. This visualization verifies that the queries encode both semantic and spatial information. In other words, the queries, serving as proposals, are capable of grouping surrounding pixels with similar semantics into a cohesive whole.

Our analysis suggests that the queries can generalize to novel objects with visual representations semantically similar to those in the training set. To address the challenge of segmenting an unlimited range of anomalies, we implement two key strategies: (1) Improving generalization by utilizing visual representations that align with linguistic concepts (Sect. 3.2); (2) Synthesizing a diverse variety of images with corresponding masks as pseudo anomalies (Sect. 3.3).



**Fig. 3.** Overview of the proposed method. During training, we use Stable Diffusion to generate various anomalies with masks to train the model for masking out anomalies directly. During inference, Stable Diffusion helps extract visual representations with concepts from natural language. Then the queries are able to group pixels with similar semantics. Both forward scoring and reverse scoring help to detect the anomalies

### 3.2 Proposals to Masks of Anomalies (P2A)

The overview of the proposed method is illustrated in Fig. 3. We first extract the multi-scale feature maps using both ResNet-50 [17] and Stable Diffusion [31], and concatenate the feature maps of the same resolution across the channel dimension. To avoid increasing the cost, we freeze the parameters of the diffusion model and train the remaining components on Cityscapes [6] to fit road scene data. Once trained, the queries hold the ability to serve as proposals and the transformer decoder can transform the proposals to masks of known categories.

Next, we train the model for transforming the proposals to masks of anomalies. Specifically, we freeze the entire model except for the convolutional layer



that outputs the mask  $\mathbf{m}$  and add  $N$  additional zero-initialized queries  $\mathbf{q}_a$  on the original queries  $\mathbf{q}$ . This ensures that our approach does not compromise the original model’s segmentation performance. We treat the road anomaly segmentation as a mask binary classification task and adopt the strategy of part instance segmentation to train the additional queries for anomaly segmentation. Specifically, given an image-mask pair  $\{\mathbf{x} \in \mathbb{R}^{3 \times H \times W}; \mathbf{y} \in \mathbb{R}^{K \times H \times W}\}$  with  $K$  known categories, we randomly select  $K'$  anomalies and paste them on the image. Then the ground truth segmentation map is split into  $(K + K')$  masks, with label 0 for known categories and label 1 for anomalies. That is to say, we perform semantic segmentation on known categories, but instance segmentation on anomalies. The reason for the strategy of part instance segmentation is that the semantics of these anomalies are usually different, we do not expect to use one proposal to hold all kinds of anomalies.

According to the segmentation process introduce in Sect. 3.1, we can segment an input image to known categories and anomalies. We would obtain the forward score map, which is the segmentation logits of pixels being anomalies  $\mathbf{s}_p$ , as well as the reverse score map, which is the negatives of the segmentation logits of pixels belonging to known categories  $-\mathbf{s}_n$ .  $\mathbf{s}_p$  holds an object-level understanding, minimizing false positives at boundaries. Besides,  $\mathbf{s}_p$  provides a binary segmentation result, which is more practical in driving scenarios. However, it may struggle with tiny objects. Conversely,  $-\mathbf{s}_n$  effectively identifies most anomalies but also produces numerous false positives at boundaries. The verification experiment for two score maps is detailed in Sect. 4.3. For evaluation, we take both into account and get the composite score map:

$$\mathbf{s}_c = \mathbf{s}_p - \mathbf{s}_n. \quad (4)$$

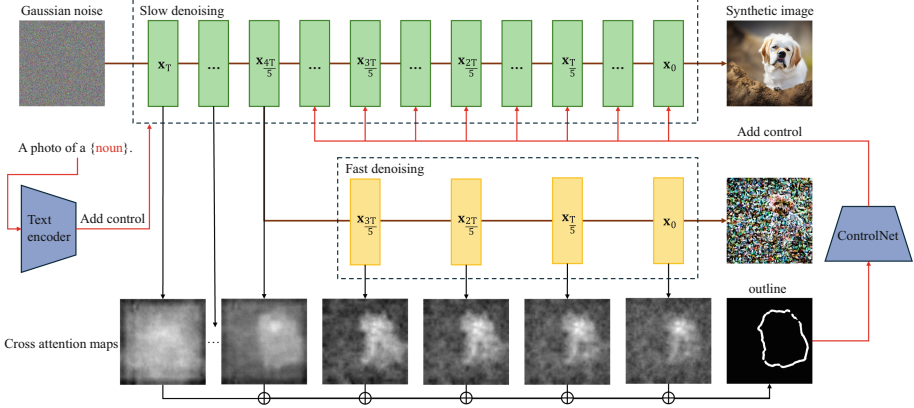
During training, The queries are matched with ground truth segments by Hungarian matching as in Mask2Former [5]. For matched queries, the predicted class distribution is supervised by the cross entropy loss, and the predicted mask is optimized towards the ground truth by a combination of cross entropy loss and dice loss. While those unmatched queries are classified as **void** and their masks are not supervised. Besides, we also utilize a contrastive loss to encourage the model to have a significant margin between the anomaly scores for known categories and that for anomalies:

$$\mathcal{L}_c = \sum_{\mathbf{y}(h,w)=0} (\mathbf{s}_c(h,w))^2 + \sum_{\mathbf{y}(h,w)=1} (m - \mathbf{s}_c(h,w))^2, \quad (5)$$

where  $h, w$  are the coordinates on the image, and  $m$  is the margin, which is set to 2 as default.

### 3.3 Image-Mask-Image Anomaly Synthesis

Most previous state-of-the-art methods crop objects from external segmentation datasets like COCO [24], PASCAL VOC [10], and then paste them on images



**Fig. 4.** Image-mask-image pipeline for synthesizing annotated samples. The first 10 denoising steps are shared by both the fast path and slow path. Then the fast path denoises the image for another 4 steps to generate a coarse mask according to the cross-attention maps. The outline of the mask serves as the conditioning control for the slow path to synthesize a high-quality image that matches the mask

from Cityscapes [6] as training data. The number of anomaly types is limited by the annotations of external datasets. To maximize the coverage of anomalies as much as possible, we propose a simple image-mask-image pipeline to generate various image-mask pairs, as shown in Fig. 4.

In addition to extracting visual representations, Stable Diffusion [31] is also employed to generate high-quality images in 50 steps with DDIM sampler [32]. While this process does not produce corresponding masks simultaneously, we utilize ControlNet [40] to obtain these masks with several additional denoising steps. As the heat maps illustrated in Fig. 1, the cosine similarity map between the representations of the image and the embedding of the noun masks out the object roughly. Based on this, we can average all the cross-attention maps to get a coarse mask when generating an image. Nevertheless, the mask is not good enough to fit the contour of the object well. Rather than apply complex post-processing like some methods [34, 36], we utilize ControlNet [40] to add the contour of the coarse mask as conditioning scribble control to generate an image. With this image-mask-image pipeline, we get an image-mask pair with the object specified by the input text. Besides, our pipeline is also compatible with other segmentation data synthesis methods [34, 36] to refine the image-mask pairs.

Fully executing both image-to-mask and mask-to-image processes (50 steps each) would be computationally prohibitive. To optimize this, we introduce a fast path for the image-to-mask process and a slow path for the mask-to-image process. Specifically, we share the initial 10 denoising steps between both processes. For the fast path, we apply 4 additional denoising steps and average the cross-attention maps derived from the attention layers in Stable Diffusion [31]. These maps captures interactions between visual representations and noun embed-

dings. Although the image quality produced by fast denoising is poor, the cross-attention maps sufficiently capture the relevant semantic to generate an effective mask. For the slow path, we use the mask contour as conditioning control and denoise the final shared result for 40 steps. This slow denoising produces a high-quality image that matches the mask.

Our P2A model is able to generalize to unseen objects with similar semantics. Additionally, we can generate ample diverse image-mask pairs for training. This enables the proposed P2A to perform forward scoring effectively.

This simple pipeline allows us to generate diverse image-mask pairs across unlimited categories. We With sufficient samples as pseudo anomalies, and the generalization capability to previously unseen objects with similar semantics, we can train the proposed P2A model to directly mask out anomalies, making forward scoring feasible.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** *Road Anomaly* [25] contains 60 images sourced from the internet, showing various kinds of unexpected elements in vehicle travel scenes, such as animals, rocks and cones. *Fishyscapes* benchmark [1] consists of two subsets: Fishyscapes Lost & Found (FS L&F) and Fishyscapes Static (FS Static). FS L&F comprises 100 images from the LostAndFound dataset [28] with fine labels. FS Static is constructed by blending anomalous objects from Pascal VOC [10] into Cityscapes [6] validation images. *SMIYC Anomaly Track* [2] consists of 100 images containing unknown objects of different sizes in different environments.

**Implementation Details.** We use a ResNet-50 [17] pre-trained on ImageNet-1K [7] as the base backbone. Throughout the training process, we freeze the weights of Stable Diffusion v1.5 [31] to save memory and training time. We first use the AdamW [26] optimizer to train the model on Cityscapes [6] for 90k iterations with learning rate of  $1e-4$  and batch size of 2. Then we train the model on images with pasted anomalies for another 5k iterations. When trained on synthetic anomalies, we freeze all weights of the model but the additional queries and the last convolutional layer that outputs the mask.

**Evaluation Metrics.** We report the average precision (AP), the area under ROC curve (AuROC), and the false positive rate at 95% true positive rate (FPR<sub>95</sub>) on all datasets. For SMIYC Anomaly Track [2] dataset, we additionally report averaged component-wise F1 (mean F1), positive predictive value (PPV), and the component-wise intersection over union (sIoU gt) as officially provided.

### 4.2 Main Results

**Road Anomaly and Fishyscapes.** Table 1 displays the quantitative evaluation results for the Road Anomaly [25] test set, FS L&F, and FS Static [1] validation set. The Road Anomaly dataset, derived from real driving scenes, presents

**Table 1.** Results on Road Anomaly and Fishyscapes validation set. For the Fishyscapes benchmark, we report results both on FS L&F and FS Static, with the best result in each column indicated in bold and the second-best result indicated with an underline

Methods	Road Anomaly [25]			FS L&F [1]			FS Static [1]		
	AUC ↑	AP ↑	FPR <sub>95</sub> ↓	AUC ↑	AP ↑	FPR <sub>95</sub> ↓	AUC ↑	AP ↑	FPR <sub>95</sub> ↓
SynthCP [35]	88.34	6.54	45.95	89.90	23.22	34.02	76.08	24.86	64.69
SML [19]	81.96	25.82	49.74	96.88	36.55	14.53	96.69	48.67	16.75
Meta-OoD [3]	-	-	-	93.06	41.31	37.69	97.56	72.91	13.57
SynBoost-WR38 [8]	81.91	38.21	64.75	96.21	60.58	31.02	95.87	66.44	25.59
MOoSe [12]	-	43.59	32.12	-	-	-	-	-	-
PEBAL [33]	87.63	45.10	44.58	<u>98.96</u>	58.81	<u>4.76</u>	<u>99.61</u>	92.08	1.52
ATTA [13]	92.11	59.05	33.59	<b>99.05</b>	65.58	<b>4.48</b>	<b>99.66</b>	<u>93.61</u>	<u>1.15</u>
Mask2Anomaly [30]	96.57	79.70	13.45	95.41	<u>69.46</u>	9.31	98.35	90.54	1.98
RbA [27]	<u>97.99</u>	85.42	<u>6.92</u>	98.62	<b>70.81</b>	6.30	98.96	75.43	3.52
cDNP [11]	-	<u>85.6</u>	9.8	-	-	-	-	-	-
P2A (ours)	<b>98.40</b>	<b>89.42</b>	<b>5.95</b>	97.24	65.15	13.98	<b>99.66</b>	<b>96.93</b>	<b>0.11</b>

a significant challenge due to its diverse range of anomaly scales and shapes. Despite this complexity, P2A demonstrates improvements across all three metrics compared to previous state-of-the-art methods, underscoring its effectiveness and robustness in detecting various anomaly morphologies. For the FS benchmark, P2A maintains state-of-the-art or competitive performance. Notably, on the FS Static dataset, P2A reduces the FPR<sub>95</sub> value by nearly 90% compared to the next best method. Overall, the proposed P2A effectively meets the detection requirements for obstacles of various categories in real driving scenarios, exhibiting a robust and comprehensive performance.

**SMIYC Anomaly Track.** Table 2 presents the quantitative evaluation results for SMIYC Anomaly Track [2]. The proposed P2A outperforms previous state-of-the-art methods on most metrics, achieving a mean F1 score above 50% for the first time on this dataset. Comparing Table 1 and Table 2, we observe that previous top-performing methods on the Road Anomaly dataset, such as ATTA [13] and PEBAL [33], show a significant decrease in performance on SMIYC Anomaly Track. This decline can be attributed to the domain gap between the training data (Cityscapes [6]) and the evaluation data (SMIYC), which presents a significant challenge for network generalization. In contrast, the proposed P2A maintains excellent performance across both datasets, demonstrating its effectiveness and robust generalization capability in the face of domain shifts.

**Table 2.** Results on SMIYC Anomaly Track. The best result and the second best result in each column are indicated in bold and with an underline, respectively

Methods	SMIYC Anomaly Track [2]					
	AP $\uparrow$	FPR <sub>95</sub> $\downarrow$	SIoU	gt $\uparrow$	PPV $\uparrow$	mean F1 $\uparrow$
Image Resynthesis [25] (ICCV'19)	52.28	25.93	39.68	10.95	12.51	
SML [19] (ICCV'21)	46.8	39.5	26.0	24.7	12.2	
SynBoost [8] (CVPR'21)	56.44	61.86	34.68	17.81	9.99	
Void Classifier [1] (IJCV'21)	36.61	63.49	21.14	22.13	6.49	
DenseHybrid [16] (ECCV'22)	77.96	<u>9.81</u>	54.17	24.13	31.08	
PEBAL [33] (ECCV'22)	49.14	40.82	38.88	27.20	14.48	
Mask2Anomaly [30] (ICCV'23)	88.72	14.63	55.28	51.68	<u>47.16</u>	
RbA [27] (ICCV'23)	<u>90.9</u>	11.6	<b>55.7</b>	<u>52.1</u>	46.8	
cDNP [11] (ICCV'23)	88.90	11.42	50.44	29.04	28.12	
ATTA [13] (NeurIPS'24)	67.04	31.57	44.58	29.55	20.64	
NFlowJS [15] (Sensors'24)	56.92	34.71	36.94	18.01	14.89	
P2A (ours)	<b>91.5</b>	<b>8.9</b>	<u>55.5</u>	<b>52.2</b>	<b>53.4</b>	

**Table 3.** Ablation study on the scoring function

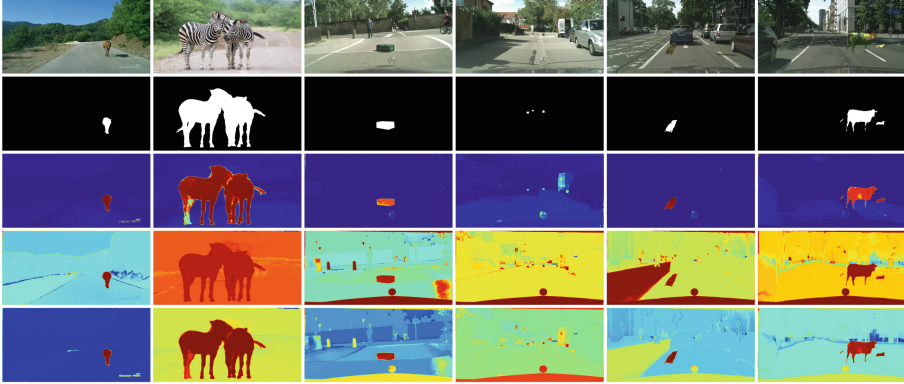
Scoring function	Road Anomaly	
	AP $\uparrow$	FPR <sub>95</sub> $\downarrow$
Forward scoring $s_a$	85.87	11.56
Reverse scoring $-s_n$	70.33	18.82
Composite scoring $s_c$	<b>89.42</b>	<b>5.95</b>

**Table 4.** Ablation study on the number of anomaly categories

#anomaly categories	Road Anomaly	
	AP $\uparrow$	FPR <sub>95</sub> $\downarrow$
10	71.22	11.73
50	79.5	7.73
150	<b>89.42</b>	<b>5.95</b>
COCO (163)	86.02	6.68

### 4.3 Ablation Study

**Scoring Function.** Compared to previous methods that adopt reverse thinking, We not only calculate the likelihood of pixels not belonging to known categories, but also make forward thinking feasible, *i.e.*, directly estimating the probability of pixels being anomalies. We evaluate three types of anomaly scores in Table 3 and show several qualitative anomaly score maps in Fig. 5. Forward scoring performs well on large anomalies but struggles with tiny ones, due to limited training data and feature map downsampling. Reverse scoring effectively identifies pixels outside known categories, but lacks object-level understanding, potentially introducing false positives at boundaries. In contrast, forward scoring directly predicts masks for objects, considering the presence of objects but potentially overlooking certain anomalies. By integrating both, the proposed method mit-



**Fig. 5.** Visualization of scoring functions. From top to down: input images, grounding truths, heat maps of forward scoring, reverse scoring, and composite scoring

igates their individual weaknesses, achieving more robust and precise anomaly detection.

**Number of Anomaly Categories.** The diversity of anomaly categories is crucial for effectively transforming proposals into anomaly masks. We conduct an ablation study on the number of anomaly categories and compare performance with anomalies generated from samples in COCO [24] dataset. Results are shown in Table 4. As expected, too few anomaly types fail to represent real-world scenarios adequately, while increasing the number of categories improves performance. Compared to a similar number of categories from COCO, the model trained with synthetic anomalies performs better. This is because many labels in COCO are related concepts, such as sheep and cow. We believe that more diverse types of anomalies could further unleash the potential of the proposed method.

**Types of Segmentation.** Query-based segmentors unify different types of segmentation within a single architecture. According to the analysis in Sect. 3.1, the queries contains semantic and spatial information. To accommodate the diverse semantics of different anomaly types, we implement a part-instance segmentation strategy for model training. To verify the effectiveness of the strategy, we also train a model with semantic segmentation for both known categories and anomalies. We find that models trained with the part instance strategy converge faster (5000 iterations v.s. 12000 iterations) and train more stably. This improvement is attributed to queries not being forced to learn disparate semantic and spatial information simultaneously.

## 5 Conclusion

In this work, we propose a novel method P2A for road anomaly detection by transforming proposals to masks of anomalies and directly measuring the probability of objects being anomalies to reduce false positives at boundaries. Specifically, we leverage semantic information, supervised by corresponding texts, to

help the model generalize to previously unseen objects. We train a query-based segmentation model using a part instance strategy, which performs semantic segmentation for known categories and instance segmentation for anomalies simultaneously. Moreover, we introduce an image-mask-image pipeline to generate annotated samples to mimic various anomalies in real-world scenarios. Experiments demonstrate that the proposed method achieves state-of-the-art performance on different benchmarks, validating the effectiveness and versatility of the proposed method across different road conditions.

**Acknowledgements.** This work was supported in part by the National Key Research and Development Program of China (2023YFC2705700), NSFC 62222112, and 62176186, the Innovative Research Group Project of Hubei Province under Grants (2024AFA017).

## References

1. Blum, H., Sarlin, P.E., Nieto, J., Siegwart, R., Cadena, C.: The Fishyscapes benchmark: measuring blind spots in semantic segmentation. *Int. J. Comput. Vis.* **129**, 3119–3135 (2021)
2. Chan, R., et al.: Segmentmeifyoucan: a benchmark for anomaly segmentation. In: *Proceedings of NeurIPS Datasets and Benchmarks* (2021)
3. Chan, R., Rottmann, M., Gottschalk, H.: Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In: *Proceedings of ICLR*, pp. 5128–5137 (2021)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of ECCV*, pp. 801–818 (2018)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: *Proceedings of CVPR*, pp. 1290–1299 (2022)
6. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: *Proceedings of CVPR*, pp. 3213–3223 (2016)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: *Proceedings of CVPR*, pp. 248–255 (2009)
8. Di Biase, G., Blum, H., Siegwart, R., Cadena, C.: Pixel-wise anomaly detection in complex driving scenes. In: *Proceedings of CVPR*, pp. 16918–16927 (2021)
9. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation. In: *Proceedings of CVPR*, pp. 11583–11592 (2022)
10. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal visual object classes challenge: a retrospective. *Int. J. Comput. Vis.* **111**, 98–136 (2015)
11. Galesso, S., Argus, M., Brox, T.: Far away in the deep space: dense nearest-neighbor-based out-of-distribution detection. In: *Proceedings of ICCV*, pp. 4477–4487 (2023)
12. Galesso, S., Bravo, M.A., Naouar, M., Brox, T.: Probing contextual diversity for dense out-of-distribution detection. In: *Proceedings of ECCV*, pp. 492–509 (2022)
13. Gao, Z., Yan, S., He, X.: ATTA: anomaly-aware test-time adaptation for out-of-distribution detection in segmentation. In: *Proceedings of NeurIPS* (2024)

14. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: Proceedings of ECCV, pp. 540–557 (2022)
15. Grcić, M., Bevandić, P., Kalafatić, Z., Šegvić, S.: Dense out-of-distribution detection by robust learning on synthetic negative data. *Sensors* **24**(4), 1248 (2024)
16. Grcić, M., Bevandić, P., Šegvić, S.: Densehybrid: hybrid anomaly detection for dense open-set recognition. In: Proceedings of ECCV, pp. 500–517 (2022)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of CVPR, pp. 770–778 (2016)
18. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint [arXiv:1610.02136](https://arxiv.org/abs/1610.02136) (2016)
19. Jung, S., Lee, J., Gwak, D., Choi, S., Choo, J.: Standardized max logits: a simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In: Proceedings of ICCV, pp. 15425–15434 (2021)
20. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint [arXiv:1511.02680](https://arxiv.org/abs/1511.02680) (2015)
21. Kirillov, A., et al.: Segment anything. In: Proceedings of ICCV, pp. 4015–4026 (2023)
22. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: Proceedings of ICLR (2022)
23. Liang, F., et al.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of CVPR, pp. 7061–7070 (2023)
24. Lin, T.Y., et al.: Microsoft coco: common objects in context. In: Proceedings of ECCV, pp. 740–755 (2014)
25. Lis, K., Nakka, K., Fua, P., Salzmann, M.: Detecting the unexpected via image resynthesis. In: Proceedings of ICCV, pp. 2152–2161 (2019)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint [arXiv:1711.05101](https://arxiv.org/abs/1711.05101) (2017)
27. Nayal, N., Yavuz, M., Henriques, J.F., Güney, F.: RbA: segmenting unknown regions rejected by all. In: Proceedings of ICCV, pp. 711–722 (2023)
28. Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., Mester, R.: Lost and found: detecting small road hazards for self-driving vehicles. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1099–1106 (2016)
29. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of ICML, pp. 8748–8763 (2021)
30. Rai, S.N., Cermelli, F., Fontanel, D., Masone, C., Caputo, B.: Unmasking anomalies in road-scene segmentation. In: Proceedings of ICCV, pp. 4037–4046 (2023)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of CVPR, pp. 10684–10695 (2022)
32. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: Proceedings of ICLR (2020)
33. Tian, Y., Liu, Y., Pang, G., Liu, F., Chen, Y., Carneiro, G.: Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In: Proceedings of ECCV, pp. 246–263 (2022)
34. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: DiffuMask: synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. arXiv preprint [arXiv:2303.11681](https://arxiv.org/abs/2303.11681) (2023)
35. Xia, Y., Zhang, Y., Liu, F., Shen, W., Yuille, A.L.: Synthesize then compare: detecting failures and anomalies for semantic segmentation. In: Proceedings of ECCV, pp. 145–161 (2020)



36. Xie, J., Li, W., Li, X., Liu, Z., Ong, Y.S., Loy, C.C.: MosaicFusion: diffusion models as data augmenters for large vocabulary instance segmentation. arXiv preprint [arXiv:2309.13042](https://arxiv.org/abs/2309.13042) (2023)
37. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of CVPR, pp. 2955–2966 (2023)
38. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of CVPR, pp. 2945–2954 (2023)
39. Xu, M., et al.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: Proceedings of ECCV, pp. 736–753. Springer (2022)
40. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of CVPR, pp. 3836–3847 (2023)
41. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of CVPR, pp. 2881–2890 (2017)



# The 2D Shape Equipartition Problem Under Minimum Boundary Length

Costas Panagiotakis<sup>(✉)</sup> 

Department of Management Science and Technology, Hellenic Mediterranean University, Agios Nikolaos 72100, Greece  
cpanag@hmu.gr

**Abstract.** In this paper, we present a general version 2D Shape Equipartition Problem (2D-SEP) under minimum boundary length. The goal of this problem is to obtain a segmentation into  $N$  equal area segments (regions), where the number of segments ( $N$ ) is given by the user, under the constraint that the boundaries between the segments have a minimum length. 2D-SEP is defined without any assumption or prior knowledge of the object structure and the location of the segments. In this work, we define the 2D-SEP and we propose a fast region growing based method that solves the general version of 2D-SEP problem. Additionally, we study the special case of the 2D-SEP in which the intrinsic boundaries are line segments, proving that it has at least one solution in convex shapes and presenting a sequential selection method that efficiently solves the problem. The quantitative results obtained on more than 2,800 2D shapes included in two standard datasets quantify the performance of the proposed methods.

**Keywords:** Shape analysis · Shape segmentation · Binary image · Image analysis

## 1 Introduction

Image segmentation is a key problem in computer vision and pattern recognition with several applications, including object recognition [8], remote sensing [6, 30] and medical image analysis [14, 21]. Image segmentation can be formulated as a classification problem of pixels with semantic labels (semantic segmentation) or partitioning of individual objects (instance segmentation). Semantic segmentation involves assigning pixel-level labels from a set of object categories (e.g., human, car, tree, sky) to all pixels in an image, making it generally more challenging than image classification, which assigns a single label to the entire image.

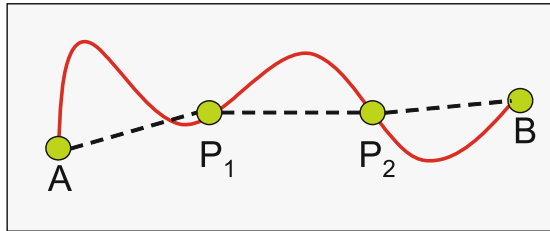
Costas Panagiotakis is also with the Foundation for Research and Technology-Hellas (FORTH), Institute of Computer Science, Greece.

---

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/978-3-031-80136-5\\_5](https://doi.org/10.1007/978-3-031-80136-5_5).

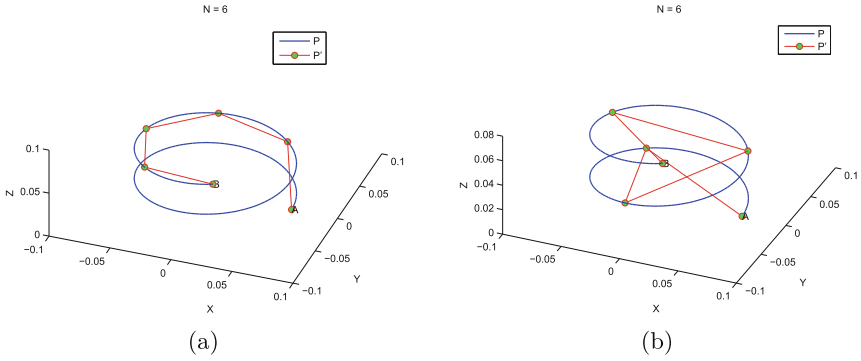
Instance segmentation takes semantic segmentation a step further by identifying and outlining each object of interest in the image (e.g., separating individual persons) [16]. Numerous image segmentation algorithms have been developed in the literature, such as thresholding [18], region growing [27], region merging [24], k-means clustering [4], watersheds [5], active contours [3], graph cuts [2], conditional and Markov random fields [6], and sparsity based methods [17]. Over the past few years, deep learning (DL) models have yielded a new generation of image segmentation models with remarkable performance improvements [16].

The curve equipartition problem has been defined and solved in [22]. It has several applications including polygonal approximation [25], signal modelling [26] and video summarization [22]. According to the curve equipartition problem, the goal is to locate  $N - 1$  consecutive curve points, so that the given curve can be divided into  $N$  segments with equal chords under a distance function (see Figs. 1 and 2). In [22], we adopt a level set approach to prove that for any continuous injective curve in a metric space and any number  $N$  there always exists at least one  $N$ -equipartition. An approximate algorithm, inspired from the level set approach is proposed for finding all solutions with high accuracy. In general, the number of solutions depends on the curve shape and  $N$ . There are special curves, where the number of solutions for some  $N$  is infinite. In [22], a geometric proof is given that the curve equipartition problem has at least one solution for every injective continuous curve and for any number of chords. Figure 2 depicts two solutions of curve equipartition problem with  $N = 6$ , which are projected on the curve  $c(t)$  (blue curve) with green color points connected with red line segments. In this problem instance, there exist four different solutions. A possible extension of the curve equipartition problem is to define and solve it under meshes [28], images [21] and shapes [20].

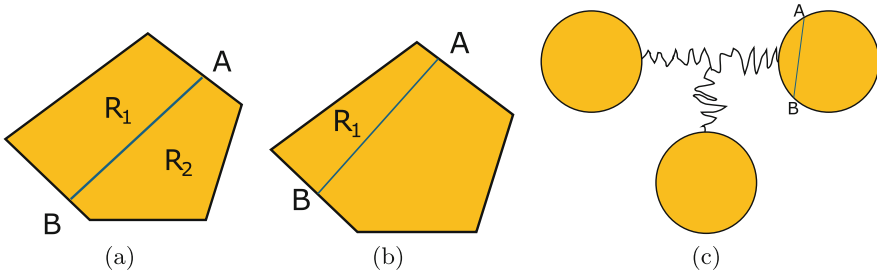


**Fig. 1.** A curve equipartition example for  $N = 3$ ,  $|AP_1| = |P_1P_2| = |P_2B|$ .

In this work, we extend the curve  $N$ -equipartition into 2D shapes, called 2D shape equipartition problem (2D-SEP). Therefore, we define and solve the 2D-SEP that can be considered as a special case of the image segmentation (shape segmentation) problem with equal segments area and minimum boundary length. According to the 2D shape equipartition problem, the goal is to compute a shape segmentation into  $N$  equal area segments, so that the length ( $L$ ) of the intrinsic boundary between the segments is minimized.



**Fig. 2.** Two solutions of curve equipartition problem with  $N = 6$ , are projected on the curve  $c(t)$  (blue curve) with the green color points connected with red line segments [22] (Color figure online).



**Fig. 3.** (a) A convex 2D shape and its 2D-SEP for  $N = 2$ . (b) A convex 2D shape and the region  $R_1$  2D-SEP for  $N = 4$ , ( $|R_1| = \frac{|S|}{4}$ ). (c) A non-convex 2D shape, where the 2D-SEP for  $N = 2$  has no solution under the constraint that the intrinsic boundary are line segments.

When a convex shape  $S$  is given, it is trivial to prove for  $N = 2$  that for each point  $A$  that belong on the boundary of  $S$ , there exists one point  $B$ , so that the shape is divided into two equal area segments via the line segment  $AB$ . Let  $R_1(S, AB)$  and  $R_2(S, AB)$  be the two regions and  $|\cdot|$  denotes the area of a region. Then it holds that

$$|R_1(S, AB)| = |R_2(S, AB)| = \frac{|S|}{2} \tag{1}$$

Figure 3(a) depicts a convex 2D shape and its 2D-SEP for  $N = 2$ . For any given  $N$ , it holds that the area of the first region  $R_1(S, AB)$  should be equal to  $\frac{|S|}{N}$ . Therefore, it is also trivial to prove that for each point  $A$  that belong on the boundary of  $S$ , there exists one point  $B$ , so that the convex shape is divided into two area segments via the line segment  $AB$  so that

$$|R_1(S, AB)| = \frac{|S|}{N} \tag{2}$$

and

$$|R_2(S, AB)| = \frac{N-1}{N} \cdot |S| \quad (3)$$

where  $R_2(S, AB)$  is the remaining region. The proof can be based on the analysis of the monotonicity of the function  $f(B) = |R_1(S, AB)|$ ,  $B \in BD(S)$ , where the boundary of shape  $S$  ( $BD(S)$ ) denotes the domain of the function  $f(B)$ . If we recursively apply the previous procedure for each remaining convex region of the previous step, in each step  $k$  we will get a region of area equal to  $\frac{|S|}{N}$  and a remaining region of area  $|S| - \frac{k \cdot |S|}{N}$ . Therefore, this is a constructive proof showing that for any given  $N$ , the 2D-SEP problem has a solution for each starting point  $A$  of the boundary of  $S$ , even if the intrinsic boundaries are line segments. Figure 3(b) depicts a convex 2D shape and the region  $R_1$  2D-SEP for  $N = 4$ . This means that the number of 2D-SEP solutions for any  $N$  is infinite, even if the intrinsic boundaries are line segments. However, when a non-convex 2D shape is given, there exist some cases where the 2D-SEP has no solution even for  $N = 2$  (see Fig. 3(c)).

Figure 4 presents examples of the proposed 2D-SEP for different number of segments ( $N \in \{2, 3, 4, 5\}$ ). In the first row, we depict the results of the proposed sequential selection method that efficiently solves the problem under the assumption that the intrinsic boundaries are line segments. In the second row, we depict the corresponding results of the proposed fast region growing based method that does not assume line segment boundaries. In any case, the segmentation consist of  $N$  equal area segments. However, the intrinsic boundary length ( $L$ ) differs by method. In Fig. 4(a), which shows a segmentation of an apple for  $N = 2$ , the proposed sequential selection method yields a lower intrinsic boundary length  $L = 46.1$ . Figure 4(e) depicts a corresponding segmentation using fast region growing method that yields a higher intrinsic boundary length  $L = 48.2$ . In the rest of the examples, the fast region growing based method yields lower intrinsic boundary length than the corresponding segmentation results of the sequential selection method.

Different error criteria have been proposed for image segmentation problems. The Intersection over Union (IoU) and F-measure are two of the most popular supervised methods to evaluate the quality of image segmentation, but it requires the ground truth [29]. Under unsupervised image (color or grayscale) segmentation methods, where the ground truth is completely unknown, clustering based criteria such as the heterogeneity of pixels between regions and the homogeneity within the region objectively can be used to evaluate the segmentation [10]. Under 2D-SEP problem, no ground truth is given. So we have to select an unsupervised criterion. Additionally, the given image is binary, so no color-grayscale is given. Similarly with the polygonal approximation [19] problem, the 2D-SEP problem can be formulated in two ways:

- The problem of minimum error, where the error (e.g. boundary length) is minimized given the number of segments  $N$ .

- The problem of minimum number of segments, where the approximation error is bounded and the goal is to find the minimum number of segments ( $N$ ) that gives error lower than the given error.

In this work, according to the proposed problem formulation, we select the first problem formulation of error minimization given the number of segments  $N$ , under the error criterion of minimum boundary length that may better divide the shape into  $N$  equal area segments. The boundary length criterion is selected, since in the given shape  $S$  there does not exist color information, model error, or weights for the boundaries to use a more complicated criterion. Additionally, the same idea called minimum cut has also been used in image segmentation [13].

In summary, the main contributions of our work are the following: To the best of our knowledge, this is the first work to define, study and solve the 2D-SEP problem under minimum boundary length. We proposed a fast region growing based method that solves the general version of 2D-SEP problem. Additionally, we study the special case of the 2D-SEP in which the intrinsic boundaries are line segments. We have also proposed a sequential selection method that efficiently solves the problem. The quantitative results obtained on more than 2,800 2D shapes included in two standard datasets quantify the performance of the proposed methods.

The rest of this paper is organized as follows. Section 2 presents the problem formulation of 2D-SEP. Sections 3 and 4 present the two proposed methods that solves 2D-SEP, respectively. The experimental results are given in Sect. 5. Finally, conclusions and future work are provided in Sect. 6.

## 2 Problem Formulation

The 2D Shape Equipartition Problem (2D-SEP) under minimum boundary length is formulated hereafter. Let  $S$  be a given shape and  $N$  be the given number of equal area segments (regions). Let  $R = \{R_1, R_2, \dots, R_N\}$  be a segmentation of  $S$ . Each region  $R_i$ ,  $i \in \{1, \dots, N\}$  should be connected, which means that the pixels of  $R_i$  segment belong to the same connected component. Let  $BD(R_i, R_j)$  be the common boundary between the regions  $R_i$  and  $R_j$ ,  $i, j \in \{1, \dots, N\}$ . Then, the optimal segmentation of 2D-SEP  $R^* = \{R_1^*, R_2^*, \dots, R_N^*\}$  should satisfy the following constraints:

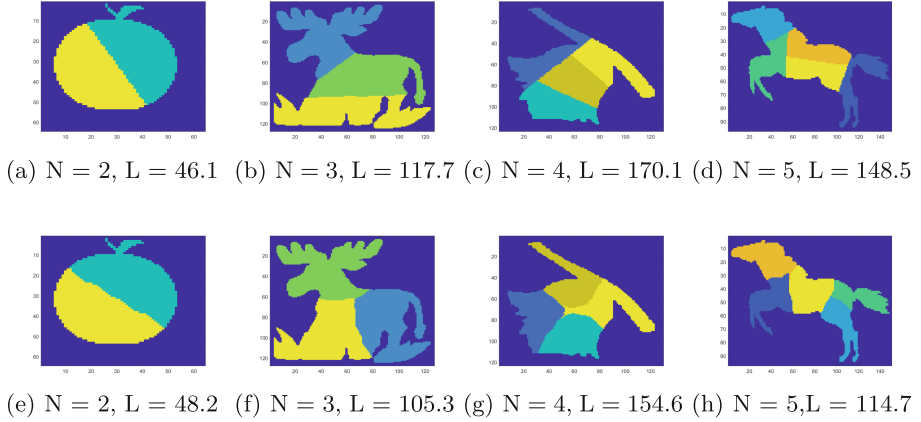
$$|R_1^*| = |R_2^*| = \dots = |R_N^*| = \frac{|S|}{N} \quad (4)$$

where  $|\cdot|$  denote the cardinality operation, e.g.  $|S|$  gives the area of shape  $S$  (number of pixels).

$$R^* = \underset{R}{\operatorname{argmin}} L(R) \quad (5)$$

where  $L(R)$  the total intrinsic boundaries' length of segmentation  $R$ :

$$L(R) = \sum_{i=1}^N \sum_{j=i+1}^N |BD(R_i, R_j)| \quad (6)$$



**Fig. 4.** Instances of the proposed 2D Shape Equipartition problem using line segments boundaries (first row) and without this assumption (second row) for different number of segments. In the first row, the results come from the proposed sequential selection method. In the second row, the corresponding results come from the proposed fast region growing based method. The number of segments ( $N$ ) and the intrinsic boundary length ( $L$ ) are reported in the caption of each shape.

where  $|BD(R_i, R_j)|$  denotes the length of boundary  $BD(R_i, R_j)$ .

In this work, we also study a variant of 2D-SEP. This includes the following additional constraint: The intrinsic boundaries between the regions  $R_i, R_j$  are line segments. We call this variant *2D-SEP-LS*. This variant makes sense due to the constraint of Eq. 5, since the simplest solution of a short-length boundary is the line segment. However, in non-convex shapes, this problem may not have a solution as depicted in Fig. 3(c). In this work, we have proposed two algorithms that solves 2D-SEP and 2D-SEP-LS that are described in the following sections. In order to be able to compare results under different image scales, in our experimental results we have used the normalized total intrinsic boundaries' length of segmentation  $R$  that is defined by the ratio of  $L(R)$  and the outer object boundary length  $|BD(S)|$ .

$$NL(R) = \frac{L(R)}{|BD(S)|} \quad (7)$$

### 3 SEP-Region Growing Based Method

This Section presents the proposed SEP-Region Growing based method (*SEP-RG*) The pseudo-code of the proposed *SEP-RG* method is given in Algorithm 1. The input of *SEP-RG* is the shape  $S$  (e.g. a binary image) and the number of the desired regions  $N$  of equipartition, and the output is the segmentation  $R$  according to the constraints of the problem as defined in Sect. 2. *SEP-RG* is an iterative method. In each iteration step, *SEP-RG* tries to find the most suitable neighbor pixel for each cluster (segment) to grow it.

```

input :  $S, N$ 
output:  $R$ 
1  $G = \text{bwGraph}(S)$ 
2  $C = \text{k-medoid}(S, N)$ 
3 foreach  $i \in \{1, \dots, N\}$  do
4   |  $D_i = \text{distances}(G, C_i)$ 
5 end
6  $F = O_{N \times |S|}$ 
7 foreach  $i \in \{1, \dots, N\}$  do
8   | foreach  $p \in S$  do
9     |  $F(i, p) = \frac{D_i(p)}{\min_{j \in \{1, \dots, N\} - i} D_j(p)}$ 
10  | end
11  |  $R_i = \emptyset$ 
12 end
13 while true do
14   | foreach  $i \in \{1, \dots, N\}$  do
15     |  $[m, p] = \text{getmin}(F(i, :))$ 
16     | if  $\text{isinf}(m)$  is true then
17       | return
18     | else if  $m \neq 1$  and  $\text{isconnected}(R_i, p)$  is false then
19       | continue
20     | else
21       |  $R_i = R_i \cup \{p\}$ 
22       | foreach  $j \in \{1, \dots, N\}$  do
23         |  $F(j, p) = \infty$ 
24       | end
25     | end
26   | end
27 end
28  $R_i = \text{correctEqualArea}(R)$ 

```

**Algorithm 1:** The proposed *SEP-RG* method.

In the following, we analytically present all the steps of the *SEP-RG* method:

- The graph  $G$  of the connected pixels in the 2D binary image of  $S$  is computed (see line 1 of Algorithm 1). Next, we compute  $N$  centroids ( $C = \{C_1, \dots, C_N\}$ ) using k-medoid method [9] (with computational cost  $O(|S|^2)$ ) that will be used for the region growing process. The centroid  $C_i$  corresponds on the region  $R_i$ .
- The lengths of shortest paths from each pixel  $p \in S$  to centroid  $C_i$  are then computed, stored in vectors  $D_i, i \in \{1, \dots, N\}$  (see line 4 of Algorithm 1). This can be computationally efficiently done in  $O(N \cdot |S| \cdot \log|S|)$  using Dijkstra's Algorithm with adjacency list<sup>1</sup>.
- For each region  $R_i, i \in \{1, \dots, N\}$  and pixel  $p \in S$ , the matrix  $F$  ( $N \times |S|$ ) stores the ratio between the distance between  $C_i$  and  $p$  ( $D_i(p)$ ) and the minimum distance between  $p$  and all the rest  $C_j, j \in \{1, \dots, N\} - \{i\}$  (see line 9 of Algorithm 1). If this ratio  $F(i, p)$  is lower than one, it means that for the pixel  $p$  the closest cluster center is  $C_i$ . Furthermore, the lower the ratio, the better the selection for the region growing process.
- Each region  $R_i, i \in \{1, \dots, N\}$  is initialized by the empty set (see line 11 of Algorithm 1).

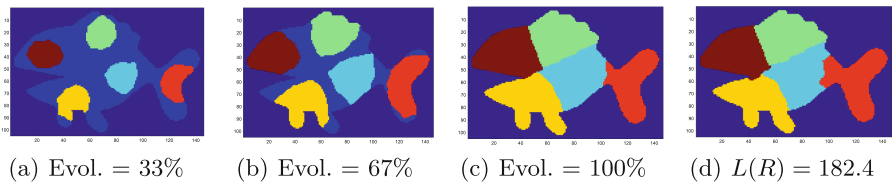
<sup>1</sup> <https://www.geeksforgeeks.org/dijkstras-shortest-path-algorithm-greedy-algo-7/>.



- Next, the main iterative process begins (see lines 13–27 of Algorithm 1). In each iteration step, for each cluster  $i \in \{1, \dots, N\}$ , we find the most appropriate pixel  $p$  for cluster  $i$  according to the distance ratio  $F$  (minimum of  $F(i, \cdot)$ ) that will be added to cluster  $i$ . In the case that all pixels of  $S$  have already been assigned, the method ends (see line 16 of Algorithm 1). In the case  $p$  is not connected pixel to  $R_i$  and  $C_i$  is not the closest center to  $p$ , we continue the process for the next cluster (see line 19 of Algorithm 1). Otherwise, the pixel  $p$  will be included in region  $R_i$  and we set the values  $F(j, p)$ ,  $j \in \{1, \dots, N\}$  to  $\infty$ , meaning that the pixel  $p$  can not selected again. It should be noticed that this iterative process does not guarantee that the resulting regions have exactly the same area. Therefore, we have proposed the following extra correction step to solve this problem. The computational cost of this step is  $O(|S|^2)$ .
- Finally, the iterative procedure *correctEqualArea* (see line 28 of Algorithm 1), reassigns pixels that belong on the boundaries of the regions. In each step of *correctEqualArea*, the region with the small area grows until its area is equal to  $\frac{|S|}{N}$ . The growing process is done in the direction of the larger neighbor regions.

Taking into account all the steps of the method, we get a total computation cost equal to  $O(N \cdot |S| \cdot \log|S| + |S|^2 + N^2 \cdot |S| + |S|^2) = O(N^2 \cdot |S| + |S|^2)$ .

Figure 5 depicts the evolution of the proposed *SEP-RG* method and the final correction step for  $N = 5$ . Figure 5(c) shows the segmentation of the iterative region growing procedure (before the correction step), which produces different segments sizes with areas in the range  $[1249, 1544]$  with a total intrinsic boundary length  $L(R) = 160.8$ . In the final segmentation shown in Fig. 5(d), the sizes of the segments are the same and the total intrinsic boundary length has increased to 182.4.



**Fig. 5.** (a), (b), (c) The evolution (see Evol. in captions that reports the percentage of the classified pixels in each instance) of the proposed *SEP-RG* method for  $N = 5$  and (d) the output of the final correction step.

## 4 SEP-Iterative Line Segment Selection Method

This Section presents the proposed SEP - Iterative Line Segment selection method (*SEP-ILS*) that sub-optimally solves the 2D-SEP-LS. The pseudo-code of the proposed *SEP-ILS* method is given in Algorithm 2. The input of *SEP-RG*

```

input :  $S, N$ 
output:  $R$ 
1  $\bar{A} = \frac{|S|}{N}$ 
2  $T = S$ 
3 foreach  $k \in \{1, \dots, N - 1\}$  do
4    $S_i = \text{getEquiAreaPartitionLines}(T, \bar{A})$ 
5    $T = S - S_i$ 
6 end
7  $S_N = T$ 

```

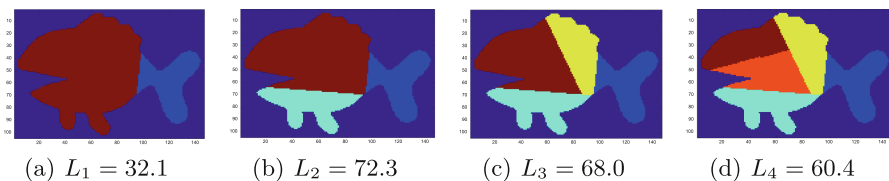
**Algorithm 2:** The proposed *SEP-ILS* method.

is the shape  $S$  (e.g. a binary image) and the number of desired regions  $N$  of equipartition, and the output is the segmentation  $R$  according to the constraints of the 2D-SEP-LS problem as defined in Sect. 2.

- *SEP-ILS* is an sequential method. Firstly, the shape  $T$ , which shows the remaining part of the shape, is initialized by  $S$ .
- In each iteration step, *SEP-ILS* tries to find the most suitable line segment that divides shape  $T$  into two regions with area  $\bar{A}$  and  $|T| - \bar{A}$ , so that the length of the line segment is minimized. The procedure *getEquiAreaPartitionLines* (see line 4 of Algorithm 2) computes this division by evaluating each pair of boundary point of  $T$ , that define a line segment according to the problem formulation (see Eq. 4 and 5). Firstly, the procedure detects a set of line segments  $S_{LS}$  that divides  $T$  into two compact regions with areas  $\bar{A} = \frac{|S|}{N}$  and  $|T| - \bar{A}$ , which satisfy the area constraint (see Eq. 4). Then, the line segment of minimum length is selected from the set  $S_{LS}$ . The computational cost of this procedure is  $O(|S|^2)$ , under the assumption that the number of boundary points of  $T$  is  $O(\sqrt{|S|})$ . The computational cost of *getEquiAreaPartitionLines* can be reduced, if we ignore some line segments that definitely do not satisfy the area constraint. This can be done by predicting the range of the two areas of the division taking into account similar line segments with known division areas that have already been examined by the method.
- After  $N - 1$  selections of line segments, the remaining area of  $T$  should be  $\bar{A}$ , so the last region  $S_N = T$  (see line 7 of Algorithm 2).

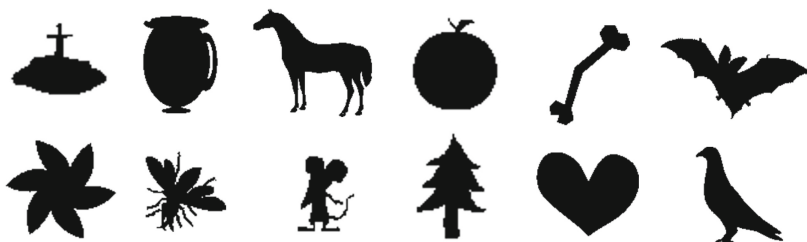
It should be noticed that the function *getEquiAreaPartitionLines* does not guarantee that it always finds a problem solution, since as explained in Sect. 1, there exist cases where no solution is found (see Fig. 3(c)). Taking into account all the steps of the *SEP-ILS* method, we get a total computation cost equal to  $O(N \cdot |S|^2)$ .

Figure 6 depicts the intermediate results for each iteration of the *SEP-ILS* method for  $N = 5$ . In the caption of each figure the length of the estimated line segment is depicted. In the final segmentation shown in Fig. 6(d), the total intrinsic boundary length is  $L(R) = 232.8$ .

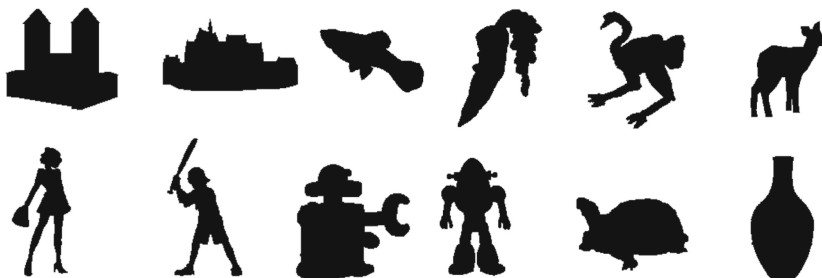


**Fig. 6.** Intermediate results for each iteration of the *SEP-ILS* method for  $N = 5$ . Captions show the corresponding total intrinsic boundaries' length  $L(R)$

## 5 Experimental Evaluation



(a) MPEG-7 dataset

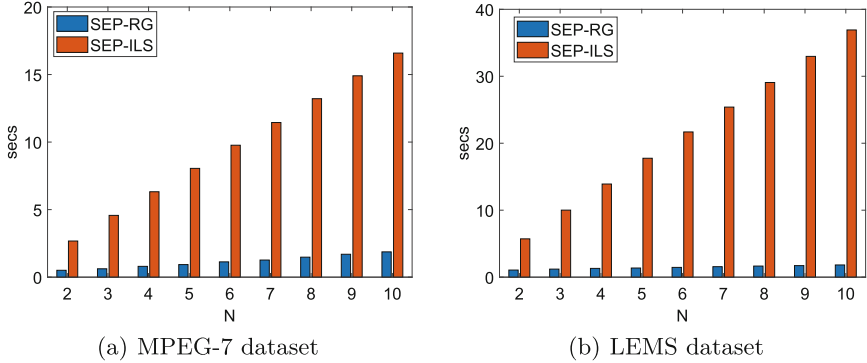


(b) LEMS dataset

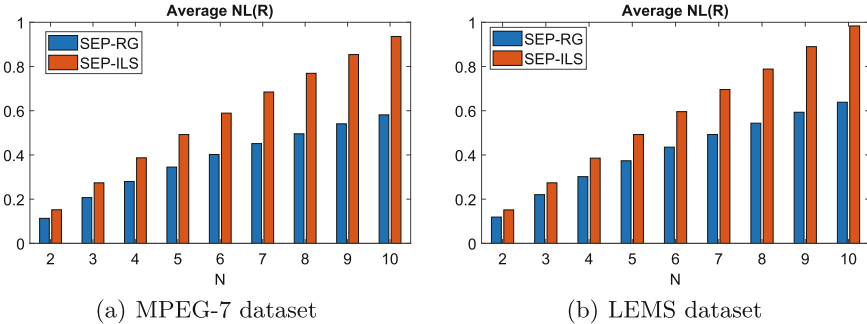
**Fig. 7.** Twelve sample images form (a) the MPEG-7 dataset and (b) the LEMS dataset.

The evaluation of the proposed approach was based on two standard datasets from the literature. More specifically, we employ:

- MPEG-7 [12], which consists of 1,400 binary shapes organised in 70 categories with 20 shapes per category. This dataset has been extensively used in shape tasks [1, 20].
- A subset of LEMS [11], that is, 1,462 shapes that come from the following categories of the original database: Buildings, Containers, Fish, Fruit and vegetables, Misc Animal, People, Robots, Toddlers, and Turtles [20].



**Fig. 8.** The average processing time of *SEP-RG* and *SEP-ILS* methods under different values of  $N$  for the (a) the MPEG-7 dataset and (b) the LEMS dataset.

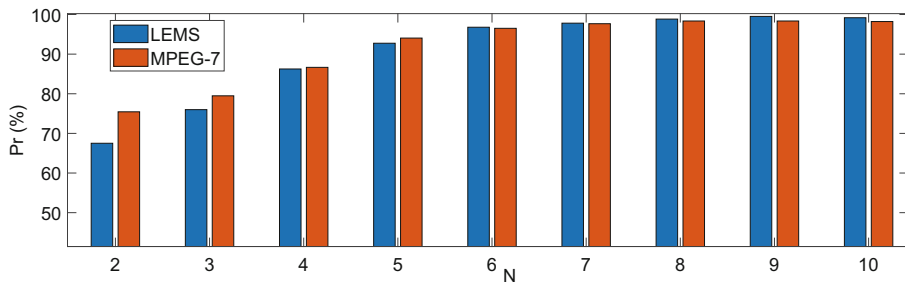


**Fig. 9.** The average value of  $NL(R)$  of *SEP-RG* and *SEP-ILS* methods under different values of  $N$  for the (a) the MPEG-7 dataset and (b) the LEMS dataset.

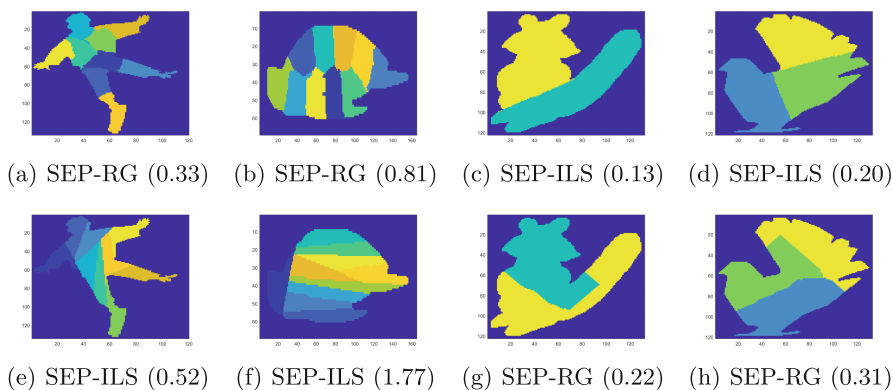
Figure 7 shows twelve sample images from the MPEG-7 and LEMS datasets.

The proposed methods have been implemented<sup>2</sup> using MATLAB and tested for each shape of the MPEG-7 and LEMS datasets using nine different number of segments  $N$ ,  $N \in \{2, \dots, 10\}$ . Therefore, we totally produced  $9 \times (1,400 + 1,462) = 25,758$  segmented images. All experiments were executed on an Intel I7 CPU processor at 2.3 GHz with 40 GB RAM. In Fig. 8, the average processing time for the execution of *SEP-RG* and *SEP-ILS* is depicted as a function of  $N$  without any speed optimization and parallelization. Taking into account the two datasets, the average processing time of *SEP-RG* and *SEP-ILS* methods is 1.30 and 15.61 s per shape, respectively. This result can also be explained by the higher computational complexity of *SEP-ILS* method.

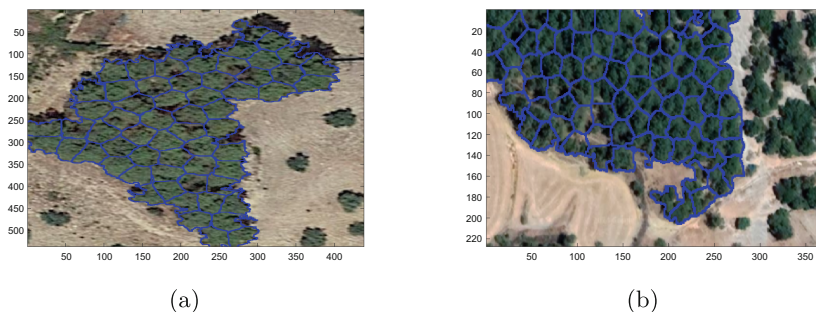
<sup>2</sup> The code implementing the proposed method together with the datasets will be publicly available at <https://sites.google.com/site/costaspanagiotakis/research/shape-equipartition>.



**Fig. 10.** The percentage  $Pr$  of shapes that  $SEP-RG$  clearly outperforms  $SEP-ILS$  in terms of  $NL(R)$  under different values of  $N$  for the MPEG-7 dataset and the LEMS dataset.



**Fig. 11.** (a), (b) Satisfactory and (g),(h) poor results of the proposed  $SEP-RG$  method. (c), (d) The corresponding (e),(f) poor and (c),(d) satisfactory results of the proposed  $SEP-ILS$  method. In the caption of each figure the  $NL(R)$  is depicted (in parenthesis).



**Fig. 12.** A promising result of the proposed  $SEP-RG$  method on the tree detection problem under low quality dense forest images.

We compared the proposed segmentation methods under normalized total intrinsic boundaries' length criterion  $NL(R)$  defined in Sect. 2. Figure 9 shows the average value of  $NL(R)$  of *SEP-RG* and *SEP-ILS* methods under different values of  $N$  for the (a) the MPEG-7 dataset and (b) the LEMS dataset. It holds that under any value of  $N$  and dataset *SEP-RG* outperforms *SEP-ILS*. It seems that the higher  $N$ , the higher outperformance of *SEP-RG*. This can be explained by the fact that as  $N$  increases, due to the sequential minimization of *SEP-ILS*, it is more possible to get a local minima that is also used for the next step of *SEP-ILS*. In contrast with the simultaneous region growing procedure of *SEP-RG*, that has not this effect of using a previous local minima solution. This is also validated by the Fig. 10 that shows the percentage ( $Pr$ ) of shapes that *SEP-RG* clearly outperforms *SEP-ILS* in terms of  $NL(R)$  under different values of  $N$  for the MPEG-7 dataset and the LEMS dataset. It holds that when  $N = 2$ , in 75.4% of the MPEG-7 dataset and in 67.5% of shapes of the LEMS dataset, *SEP-RG* outperform *SEP-ILS*. The outperformance *SEP-RG* increases for higher values of  $N$ , reaching the values 98.2% and 99.2% for the shapes of the MPEG-7 and LEMS dataset, respectively. On the average, *SEP-RG* outperforms *SEP-ILS* in 91.1% of the segmented examples of the MPEG-7 and LEMS dataset.

Figure 11 shows satisfactory and poor results of the proposed *SEP-RG* method. Furthermore, it shows the corresponding poor and satisfactory results of the proposed *SEP-ILS* method. In the first two examples *SEP-RG* clearly outperforms *SEP-ILS* in terms of  $NL(R)$  criterion, while in the second two examples *SEP-RG* clearly under-performs *SEP-ILS*. In these results, a complementary behavior of the proposed methods concerning their segmentation performance was observed. In the caption of each figure the normalized boundary length  $NL(R)$  is depicted (in parenthesis). Under any case, the poor result has at least 50% higher  $NL(R)$  than the corresponding satisfactory result.

The proposed method can be applied on segmentation applications, e.g., on tree detection problem [15], where the goal is to detect trees in aerial images. When the forest is very dense and the image quality is low, the unsupervised and deep learning methods is difficult to provide accurate segmentation. Figure 12 shows a promising result of the proposed *SEP-RG* method on the tree detection problem under low quality dense forest images. In these examples, even a human expert is almost impossible to detect the trees. *SEP-RG* has been applied on the largest region of the bitmap image derived by RGBVI index as used in [15]. The number of trees was given to the method by divided the area of the largest region by a typical tree size. The tree borders are depicted using blue color. Under the assumption that the trees are equal sized, in Figs. 12(a) and 12(b), 61 and 79 trees where detected by *SEP-RG* method.

## 6 Conclusions

In this work, we propose a fast region growing based method that solves the general version of the 2D Shape Equipartition Problem (2D-SEP) under minimum boundary length. We propose a fast region growing based method (*SEP-RG*)

that sub-optimally solves the general version of 2D-SEP problem. In addition, we study the special case of the problem in which the intrinsic boundaries are line segments, proving that it has at least one solution in convex shapes and presenting a sequential selection method (*SEP-ILS*) that efficiently solves 2D-SEP. The quantitative results obtained on more than 25000 segmentation instances included in two standard datasets and different number of segments, quantify the performance of the proposed methods. According to our experimental results, in most of the shapes *SEP-RG* outperforms *SEP-ILS* in terms of minimum boundary length criterion. However, complementary behavior of the proposed methods has also been observed.

In ongoing and future work, our aim is to apply 2D-SEP on real computer vision and pattern recognition problems where the goal is to provide segmentation of a given 2D shape. Finally, we plan to extend the proposed framework on 3D shapes and to explore real applications in which the proposed system may be useful.

**Acknowledgements.** This publication is financed by the Project "Strengthening and optimizing the operation of MODY services and academic and research units of the Hellenic Mediterranean University", funded by the Public Investment Program of the Greek Ministry of Education and Religious Affairs".

## References

1. Bai, X., Yang, X., Latecki, L.J., Liu, W., Tu, Z.: Learning context-sensitive shape similarity by graph transduction. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(5), 861–874 (2010)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(11), 1222–1239 (2001)
3. Chan, T.F., Vese, L.A.: Active contours without edges. *IEEE Trans. Image Process.* **10**(2), 266–277 (2001)
4. Dhanachandra, N., Mangleam, K., Chanu, Y.J.: Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Comput. Sci.* **54**, 764–771 (2015)
5. Grau, V., Mewes, A., Alcaniz, M., Kikinis, R., Warfield, S.K.: Improved watershed transform for medical image segmentation using prior information. *IEEE Trans. Med. Imag.* **23**(4), 447–458 (2004)
6. Grinias, I., Panagiotakis, C., Tziritas, G.: MRF-based segmentation and unsupervised classification for building and road detection in peri-urban areas of high-resolution satellite images. *ISPRS J. Photogramm. Remote. Sens.* **122**, 145–166 (2016)
7. Grinias, I., Panagiotakis, C., Tziritas, G.: Mrf-based segmentation and unsupervised classification for building and road detection in peri-urban areas of high-resolution satellite images. *ISPRS J. Photogramm. Remote. Sens.* **122**, 145–166 (2016)
8. Jiang, D., Li, G., Tan, C., Huang, L., Sun, Y., Kong, J.: Semantic segmentation for multiscale target based on object recognition using the improved faster-RCNN model. *Futur. Gener. Comput. Syst.* **123**, 94–104 (2021)

9. Kaur, N.K., Kaur, U., Singh, D.: K-medoid clustering algorithm-a review. *Int. J. Comput. Appl. Technol.* **1**(1), 42–45 (2014)
10. Khan, J.F., Bhuiyan, S.M.: Weighted entropy for segmentation evaluation. *Opt. Laser Technol.* **57**, 236–242 (2014)
11. Kimia, B.: A large binary image database, LEMS vision group at brown university (2002). <http://www.lems.brown.edu/~dmc/>
12. Latecki, L.J., Lakamper, R., Eckhardt, T.: Shape descriptors for non-rigid shapes with a single closed contour. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 424–429. IEEE (2000)
13. Lempitsky, V., Blake, A., Rother, C.: Image segmentation by branch-and-mincut. In: *Computer Vision—ECCV 2008: 10th European Conference on Computer Vision*, Marseille, France, October 12–18, 2008, Proceedings, Part IV 10, pp. 15–29. Springer (2008)
14. Li, H., Zhao, X., Su, A., Zhang, H., Liu, J., Gu, G.: Color space transformation and multi-class weighted loss for adhesive white blood cell segmentation. *IEEE Access* **8**, 24808–24818 (2020)
15. Markaki, S., Panagiotakis, C.: Unsupervised tree detection and counting via region-based circle fitting. In: *ICPRAM*, pp. 95–106 (2023)
16. 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 (2021)
17. Minaee, S., Wang, Y.: An ADMM approach to masked signal decomposition using subspace representation. *IEEE Trans. Image Process.* **28**(7), 3192–3204 (2019)
18. Ostu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62 (1979)
19. Panagiotakis, C.: Particle swarm optimization-based unconstrained polygonal fitting of 2D shapes. *Algorithms* **17**(1), 25 (2024)
20. Panagiotakis, C., Argyros, A.: Parameter-free modelling of 2D shapes with ellipses. *Pattern Recogn.* **53**, 259–275 (2016)
21. Panagiotakis, C., Argyros, A.: Region-based fitting of overlapping ellipses and its application to cells segmentation. *Image Vis. Comput.* **93**, 103810 (2020)
22. Panagiotakis, C., Doulamis, A., Tziritas, G.: Equivalent key frames selection based on ISO-content principles. *IEEE Trans. Circ. Syst. Video Technol.* **19**(3), 447–451 (2009)
23. Panagiotakis, C., Doulamis, A., Tziritas, G.: Equivalent key frames selection based on iso-content principles. *IEEE Trans. Circuits Syst. Video Technol.* **19**(3), 447–451 (2009)
24. Panagiotakis, C., Grinias, I., Tziritas, G.: Natural image segmentation based on tree equipartition, Bayesian flooding and region merging. *IEEE Trans. Image Process.* **20**(8), 2276–2287 (2011)
25. Panagiotakis, C., Tziritas, G.: Any dimension polygonal approximation based on equal errors principle. *Pattern Recogn. Lett.* **28**(5), 582–591 (2007)
26. Panagiotakis, C., Tziritas, G.: Simultaneous segmentation and modelling of signals based on an equipartition principle. In: *2010 20th International Conference on Pattern Recognition*, pp. 85–88. IEEE (2010)
27. Preetha, M.M.S.J., Suresh, L.P., Bosco, M.J.: Image segmentation using seeded region growing. In: *2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET)*, pp. 576–583. IEEE (2012)
28. Shapira, L., Shamir, A., Cohen-Or, D.: Consistent mesh partitioning and skeletonisation using the shape diameter function. *Vis. Comput.* **24**, 249–259 (2008)



29. Wang, Z., Wang, E., Zhu, Y.: Image segmentation evaluation: a survey of methods. *Artif. Intell. Rev.* **53**(8), 5637–5674 (2020)
30. Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W., Zhao, T.: Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network. *Remote Sens.* **11**(15), 1774 (2019)



# Random Frame: a Data Augmentation for Glass Detection

Yiming Liang<sup>(✉)</sup> and Hiroshi Ishikawa

Department of Computer Science and Communications Engineering, Waseda University, Tokyo, Japan

yiming.liang@toki.waseda.jp hfs@waseda.jp

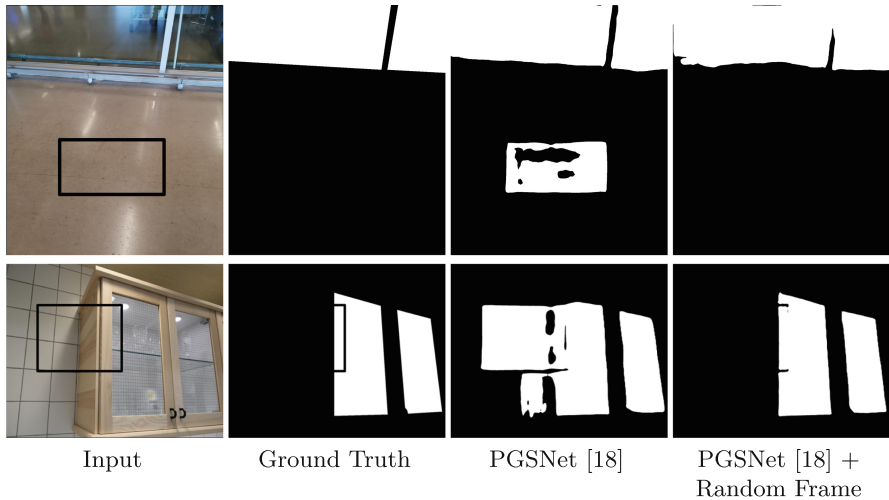
**Abstract.** Glass, though ubiquitous, is difficult to recognize in an image due to its transparency. Fine-grained low-level features indicating the presence of glass, such as refraction and reflection, are weak and subtle. This causes difficulties for existing glass detection models in learning those features, pushing them to rely on more overt cues, especially the frame surrounding the glass. Consequently, they can be fooled easily by frame-like objects. Here, we propose a simple data augmentation scheme called Random Frame to address this problem. Random Frame inserts a frame into an image to create an area with a frame but no glass. The model will receive a penalty if it only relies on the frame. The performances of existing models on various datasets improve when Random Frame is applied while being trained. Our comprehensive experiments demonstrate that our data augmentation can make models utilize more low-level features with more confidence in their predictions.

**Keywords:** Glass detection · Data augmentation · Image recognition

## 1 Introduction

Although glasses are ubiquitous in everyday scenes, their existence is ignored in many computer vision tasks, impacting the performances [13] (e.g., depth prediction, instance segmentation, reflection removal). Toward real-world application of computer vision, such as autonomous navigation for robots or drones, it is crucial to develop a method capable of detecting glass, addressing the limitations of current approaches.

Detecting Glass surfaces is challenging for mainly two reasons. First, they do not have their own semantic context [13]. Behind the same piece of glass, arbitrary objects can appear, which makes the detection of the glass more difficult. Second, visual cues such as refraction and reflection indicating the presence of the glass are both weak, due to its high transparency, and highly variable, depending on the illumination and the viewpoint. (Imagine standing in front of the window at night, in a room with lights on. You can see the reflections change as you walk around, but if you turn off the lights, the reflections will disappear, and all you can see is the scene outside the window.) Thus, the appearance of glass does not have a fixed pattern, which makes glass detection an ill-conditioned problem.



**Fig. 1.** Existing method [18] can be fooled by a simple black rectangle inserted into the input image (Left). Here, the Ground Truth and the two test outputs by PGSNet [18] trained with and without our proposed Random Frame data augmentation are shown. (White: glass; black: everything else.) PGSNet trained without Random Frame predicts the inside of the inserted rectangle as glass. By applying Random Frame during training, it becomes capable of making correct predictions. Note that the areas under inserted rectangle is not marked as glass in Ground Truth.

This paper stems from our observation: we noticed that it is not entirely correct to say that there is no semantic context to the glass. That is, since the glass is often held in place by a frame, if a frame is detected a glass is probably inside; and since frames are made of ordinary materials and have simple shapes such as rectangle, they are easier to recognize than transparent glass, which makes the frame a useful contextual cue to detect a glass.

Through our experiments, we found that existing methods can be fooled too easily by the presence of a frame-like object. As shown in Fig. 1, when even a simple black rectangle is inserted into the input image, PGSNet [18] incorrectly recognizes its inside as glass, even though the rectangle does not particularly resemble any kind of frame. It seems that the glass detection model has learned to use the one contextual cue that exists, which is the frame surrounding the glass, rather than the difficult-to-learn visual cues such as refraction and reflection. Though the strategy makes sense to some extent, relying on the frame *alone* is obviously not optimal, as there are cases where there is a frame-like object without a glass inside, such as a half-open sliding window, and a glass without a frame or with a frame outside of the image.

In this paper, we propose a novel data augmentation scheme called Random Frame to address this problem. It provides a simple and effective way to correct over-reliance on the contextual cues provided by frames. In Random Frame, we extract only the frames around the glasses and overlay them in the training

images. The augmented image now has a frame inserted, but the inside of the frame is unchanged, simulating a frame without glass. The ground-truth mask indicating the presence of glass is unchanged inside the overlaid frame. The models trained with our data augmentation showed a notable performance increase on GDD [13], HSO [18], and Trans10K-Stuff [17] datasets.

In summary, our contributions are as follows:

- We propose a novel data augmentation scheme, Random Frame, specific to the task of glass detection.
- We demonstrate that applying Random Frame at training time can improve the performance of existing methods by a large margin.
- We conduct comprehensive experiments to show the effect of Random Frame on what the model learns.

## 2 Related Work

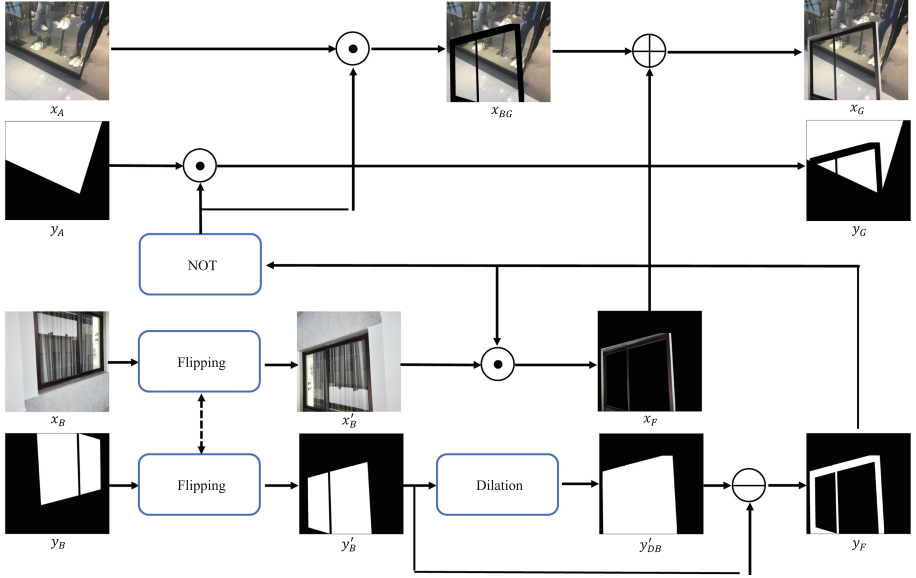
### 2.1 Glass Detection

The glass detection task was pioneered by Mei *et al.* [13], who proposed an atrous spatial pyramid pooling (ASPP) [2]-like module called the large-field contextual feature integration (LCFI) to utilize multi-scale features captured from a large receptive field. To improve the accuracy of boundary localization, [18] introduced an encoder-decoder architecture with skip connections [4], where the decoder progressively recovers the spatial resolution of the feature maps while highlighting the common and exploring the differences between features at different levels. While [8] used reflection as a prior, the semantic contexts are harvested more explicitly with a pre-train procedure using semantic labels in [10]. Other modalities, such as polarization image [12], thermal image [5], and depth map [9], have been incorporated to better resolve this hard problem.

### 2.2 Data Augmentation for Images

Data augmentation is a widely used regularization strategy when training deep neural networks. It generates a new sample by applying transformations, such as flipping, rotation, cropping, color jitter, and noise injection, to the original sample. It can increase the size and the variance of the datasets, which can help models generalize better. The effectiveness of data augmentation has been shown in image classification [3, 6, 7, 15, 19, 20], object detection [19], semantic segmentation [11], etc. The method we propose, Random Frame, shares similarities with Cutout [3], Mixup [20], and CutMix [19] in either partially occluding the original image or combining two different images. Cutout is inspired by Dropout [16], but it masks out a contiguous area of inputs rather than in a pixel-wise manner. Information about the masked-out area is completely unavailable, forcing the model to capture the global context rather than relying on specific features. Mixup selects two different images. The generated image and its label is the linear interpolation of selected images and their corresponding labels. The

motivation for Mixup is to smoothen the decision boundaries of the model by interpolating data points located between two classes into the distribution of the dataset. CutMix is a combination of Cutout and Mixup. Instead of masking out a region from the input image, CutMix replaces the region with a patch from another image. The label is computed by linear interpolation as in Mixup. The interpolation factor is the ratio of the size of the inserted patch to that of the original image.



**Fig. 2.** The processing flow of Random Frame. Flipping denotes random horizontal flipping and random vertical flipping. Flipping applied to  $x_B$  and  $y_B$  are identical.  $\odot$ ,  $\oplus$ , and  $\ominus$  denote element-wise multiplication, addition, and subtraction, respectively. NOT denotes an operation that converts 0 to 1 and 1 to 0.

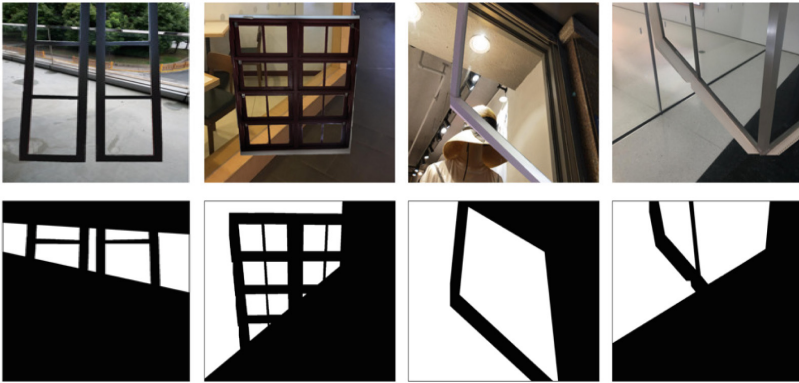
## 3 Method

### 3.1 Motivation

Our proposed method, Random Frame, is motivated by observations that existing method [18] incorrectly predicts the interior of a rectangle inserted in the input image (Fig. 1). Humans would never make this kind of mistake. From this, we surmise that the model relies too much on the co-occurrence of glasses and frames surrounding them.

We theorize as follows. Since glass is often surrounded by frames (e.g., window panes, glass guardrails, and showcases), the location of frames can be a useful

hint. Being able to recognize frames is also essential for predicting the boundaries of glass accurately. When humans recognize glass, they first utilize the location of frames or frame-like objects with the global context (high-level features) to find where glass is likely to be. Then, they check the refractions and reflections of light (low-level features) produced by the glass and make a final prediction. In this process, low-level features are equally or even more important than high-level features. However, as described in Sec. 1, such low-level features are extremely fine-grained and can be easily affected by surroundings. It is more difficult for the model to learn those features. Therefore, the model failed to capture low-level features and make the prediction mainly based on high-level features provided by frames. This is the reason for the failures shown in Fig. 1. Random Frame artificially creates areas in the image where there is a frame but no glass inside, by inserting a frame into the image. If the model relies too heavily on features provided by frames, it will predict the inside of the inserted frame wrongly and the loss will increase. Hence, Random Frame can force the model to learn to avoid this by relying more on low-level features provided by the glass itself. We intend Random Frame to be a regularization method, which can improve the test accuracy of the model and make the model more confident with its predictions.



**Fig. 3.** Examples of Random Frame. The first row is the generated image. The second row is the ground truth of the generated image.

### 3.2 Random Frame

Let  $(x_A, y_A)$  and  $(x_B, y_B)$  be two samples randomly selected from the training data, where  $x_A, x_B \in \mathbb{R}^{W \times H \times C}$  and  $y_A, y_B \in \{0, 1\}^{W \times H}$  denote images and ground truths, respectively. Ground truth is a binary mask in which 1 indicates glass while 0 indicates non-glass. First, we apply a random horizontal flip and a random vertical flip to  $(x_B, y_B)$  in order to increase the variety. Let  $(x'_B, y'_B)$  be the resulting image-ground truth pair. Second, we extract the frame from

$x'_B$ . Though it is possible to extract the frame using a detection algorithm, it is computationally expensive. Here, we utilize  $y'_B$  to extract the frame heuristically by assuming all four sides of a glass area are surrounded by a frame. Based on this assumption, we dilate the region with values of 1 in  $y'_B$  as:

$$y'_{DB} = D^n(y'_B), \quad (1)$$

where  $D$  denotes the dilation operation, and  $n$  is the number of iterations depending on the size of the input. The dilation operation is performed by a  $3 \times 3$  kernel filled with 1. Then, subtracting  $y'_B$  from  $y'_{DB}$ :

$$y_F = y'_{DB} - y'_B, \quad (2)$$

we use  $y_F$  as the binary mask of frame in  $x'_B$ . Third, we use  $y_F$  to extract the frame  $x_F$  from  $x'_B$ , and the background  $x_{BG}$  from  $x_A$ .  $x_{BG}$  is added to  $x_F$  to obtain the generated image  $x_G$ . The computations can be written as:

$$x_G = x'_B \odot y_F + x_A \odot (1 - y_F) = x_F + x_{BG}, \quad (3)$$

where  $\odot$  denotes element-wise multiplication. Finally, we generate the ground truth  $y_G$  of  $x_G$  as:

$$y_G = y_A \odot (1 - y_F). \quad (4)$$

In this manner,  $(x_G, y_G)$  can be generated from  $(x_A, y_A)$  and  $(x_B, y_B)$ . No operation used in Random Frame is computationally expensive. Loading data with multi-process can make the introduced overhead negligibly small. Figure 2 shows the whole processing flow. Examples are shown in Fig. 3.

A data augmentation called FakeMix [1] was proposed for transparent object detection. Both motivation and implementation are different from Random Frame. FakeMix aims to solve the boundary-related imbalance problem while Random Frame is proposed to force the model to learn fine-grained low-level features. FakeMix inserts boundaries with a width of 8 pixels, and the ground truth remains unchanged. In contrast, the width of the frame inserted by Random Frame varies depending on image resolution and the shape of the frame, and the ground truth is modified accordingly. We include comparisons with FakeMix in our experiments.

## 4 Experiments

### 4.1 Experiments Settings

**Implementation Details.** Random Frame is applied after resizing with 50% probability. The probability of performing horizontal flipping and vertical flipping on the image from which the frame is to be extracted is also set to be 50%. The iteration of dilation operation is  $\max(W, H)/10$ , where  $W$  and  $H$  denote the width and height of the image.

**Table 1.** Evaluation results on datasets GDD [13], GSD [8], HSO [18], and Trans10K-Stuff [17]. RFrame denotes our Random Frame. “ $\uparrow$ ” means larger values are better, while “ $\downarrow$ ” means smaller values are better. Random horizontal flipping is applied to GDNet [13] and PGSNet [18] by default. All results reported are the best among three runs.

Dataset	Method	IoU $\uparrow$	MAE $\downarrow$	$F_\beta$ $\uparrow$	BER $\downarrow$
GDD [13]	GDNet [13]	84.37	0.085	0.922	7.65
	GDNet [13] + RFrame	<b>85.45</b>	<b>0.076</b>	<b>0.937</b>	<b>7.02</b>
	PGSNet [18]	86.96	0.066	0.933	6.04
	PGSNet [18] + RFrame	<b>88.01</b>	<b>0.060</b>	<b>0.949</b>	<b>5.71</b>
GSD [8]	GDNet [13]	<b>78.89</b>	0.071	<b>0.896</b>	<b>8.22</b>
	GDNet [13] + RFrame	78.82	<b>0.070</b>	<b>0.896</b>	8.55
	PGSNet [18]	<b>81.51</b>	<b>0.059</b>	0.891	<b>7.06</b>
	PGSNet [18] + RFrame	81.42	<b>0.059</b>	<b>0.903</b>	7.75
HSO [18]	GDNet [13]	75.56	0.118	0.868	10.6
	GDNet [13] + RFrame	<b>78.24</b>	<b>0.101</b>	<b>0.888</b>	<b>9.48</b>
	PGSNet [18]	78.18	0.099	0.881	9.72
	PGSNet [18] + RFrame	<b>80.20</b>	<b>0.085</b>	<b>0.910</b>	<b>8.89</b>
Trans10K-Stuff [17]	GDNet [13]	86.30	0.058	0.936	5.53
	GDNet [13] + RFrame	<b>87.10</b>	<b>0.056</b>	<b>0.941</b>	<b>5.42</b>
	PGSNet [18]	88.20	0.047	0.938	4.95
	PGSNet [18] + RFrame	<b>89.83</b>	<b>0.041</b>	<b>0.953</b>	<b>4.36</b>

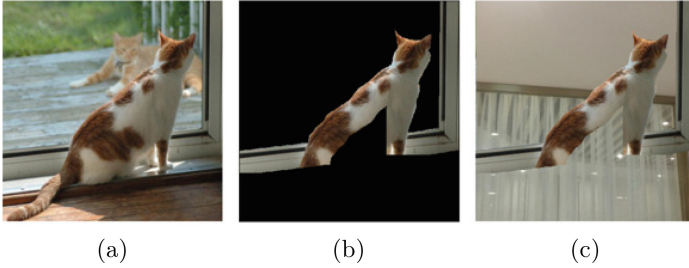
**Evaluation Datasets.** We evaluate the proposed method on the GDD [13], GSD [8], HSO [18], and Trans10K-Stuff [17] datasets. GDD and GSD datasets consist of both outdoor scenes and indoor scenes, whereas the HSO dataset only includes indoor home scenes. The Trans10K dataset is a dataset for transparent object detection. Following [18], we do the evaluations on the Trans10K-Stuff dataset, which is a subset of the Trans10K dataset.

**Evaluation Metrics.** We adopt intersection over union (IoU), mean absolute error (MAE), F-measure ( $F_\beta$ ), and balanced error rate (BER) as evaluation metrics, following [13].

## 4.2 Comparison by Architectures and Datasets

We evaluate the effectiveness of Random Frame in comparison with existing glass detection methods GDNet [13] and PGSNet [18], using GDD [13], GSD [8], HSO [18], and Trans10K-Stuff [17] datasets. Since the original codes for the models are not publicly available except for the inference code for GDNet, we re-implement GDNet and PGSNet by closely following the descriptions provided in their original papers [13, 18]. We re-train the models using the respective training





**Fig. 4.** An example of frames extracted from GSD dataset [8]. (a): Reference image from which the frame is extracted. (b): Extracted frame. (c): Image generated by Random Frame.

**Table 2.** Using frames extracted from different dataset. Inside the parentheses are the dataset from which frames are taken.

Method	GSD [8]			
	IoU $\uparrow$	MAE $\downarrow$	$F_{\beta}$ $\uparrow$	BER $\downarrow$
GDNet [13]	78.89	0.071	0.896	8.22
+ RFrame (GSD [8])	78.82	0.070	0.896	8.55
+ RFrame (GDD [13])	78.81	0.072	0.894	8.53
+ RFrame (HSO [18])	<b>79.60</b>	<b>0.069</b>	<b>0.899</b>	<b>7.97</b>
PGSNet [18]	81.51	<b>0.059</b>	0.891	7.06
+ RFrame (GSD [8])	81.42	<b>0.059</b>	0.903	7.75
+ RFrame (GDD [13])	81.64	<b>0.059</b>	<b>0.906</b>	7.62
+ RFrame (HSO [18])	<b>82.28</b>	<b>0.059</b>	0.905	<b>6.99</b>

set and then calculate the metrics on test set. Additional details about the re-implementation and training can be found in the supplementary materials. As for GSD [8] and GlassSemNet [10], which also take RGB image as input, not only are their training codes unavailable but the descriptions in their respective papers [8, 10] seem insufficient for re-implementation. Therefore, we do not include them into the comparison.

The results are shown in Table 1. All four metrics for GDNet [13] and PGSNet [18] improve on GDD, HSO, and Trans10K-Stuff datasets. On the other hand, the models do not benefit from our data augmentation on GSD dataset. However, as shown in Table 2, the performance improves when the inserted frames are taken from HSO dataset instead of GSD dataset. This indicates that HSO dataset contains frames more suitable for our purpose. When an object is in front of the glass, a portion of that object is extracted as a frame. Inserting such frames can affect the semantic context in the original image, which makes Random Frame behave similarly to CutMix [19] (example: Fig. 4). The frame from HSO dataset has the simplest shape of the three datasets, so its impact on semantic context is the smallest, which may be why the performance improves. Also, the photos in

**Table 3.** Comparison with other data augmentations. HFlip, VFlip, and RFrame denote random horizontal flipping, random vertical flipping, and Random Frame, respectively. Resizing and Normalizing are applied in Baseline as preprocessing. “+” means that the data augmentation is applied additionally to the Baseline. All results reported are of three runs, in “max (average  $\pm$  standard deviation)” format.

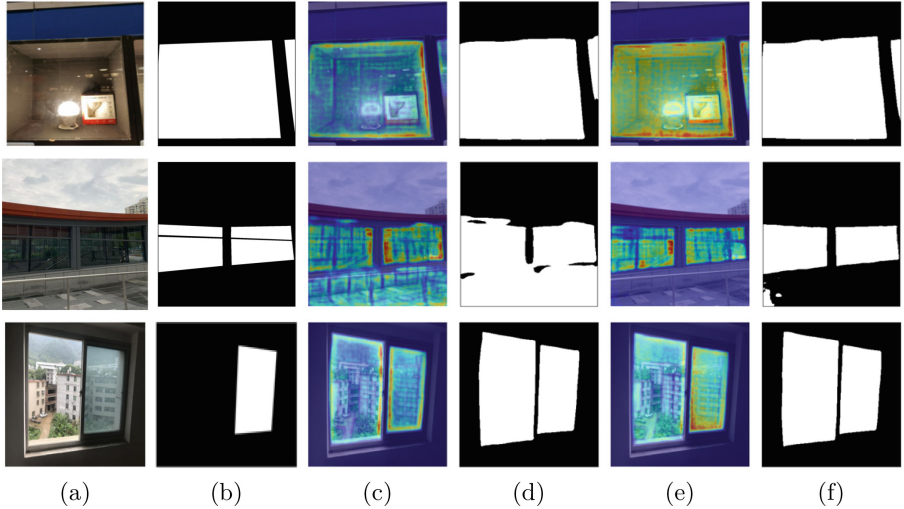
PGSNet [18]	IoU $\uparrow$
Baseline	85.42 (85.24 $\pm$ 0.16)
+ HFlip	86.96 (86.76 $\pm$ 0.18)
+ VFlip	86.32 (85.94 $\pm$ 0.33)
+ Color Jitter	85.29 (85.18 $\pm$ 0.12)
+ Cutout [3]	86.39 (86.13 $\pm$ 0.23)
+ CutMix [19]	85.81 (85.46 $\pm$ 0.33)
+ Rect	86.10 (86.00 $\pm$ 0.09)
+ RFrame	86.94 (86.74 $\pm$ 0.22)
+ FakeMix [1]	85.64 (85.32 $\pm$ 0.42)
+ HFlip + Cutout [3]	87.67 (87.11 $\pm$ 0.57)
+ HFlip + Rect	87.61 (87.45 $\pm$ 0.26)
+ HFlip + RFrame	<b>88.01 (87.69 <math>\pm</math> 0.33)</b>
+ HFlip + FakeMix [1]	87.42 (87.02 $\pm$ 0.39)
+ VFlip + RFrame	87.11 (86.94 $\pm$ 0.23)
+ Cutout [3] + RFrame	86.53 (86.40 $\pm$ 0.22)
+ CutMix [19] + RFrame	86.19 (85.84 $\pm$ 0.46)

GSD dataset are taken relatively far from the glass, so features of the glass are weaker. And there are more of other objects in the image that may be implicitly playing the role of frames.

### 4.3 Comparison with Other Data Augmentations

We compare our method with random horizontal flipping, random vertical flipping, color jitter, Cutout [3], CutMix [19], and FakeMix [1]. We adapt Cutout and CutMix for glass detection as follows. For Cutout, the erased areas are considered as non-glass. For CutMix, if the inserted patch contains glass areas, then those areas are considered as glass. We also include a simple variant (Rect) of Random Frame that inserts a rectangle with a random color as Fig. 1.

The evaluation results are shown in Table 3. When applied alone, Random Frame achieves comparable performance as random horizontal flipping and outperforms the others with large margins. Particularly, the superiority of Random Frame over Rect demonstrates the importance of using *real* frames. When Random Frame is applied with random horizontal flipping, the best result is achieved. This shows that our method is complementary to random horizontal flipping.



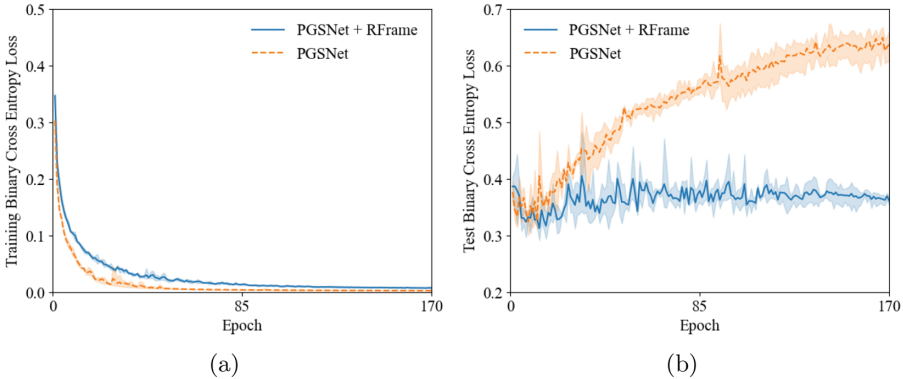
**Fig. 5.** Attention map visualization by Grad-CAM [14]. (a): Input. (b): Ground Truth. (c)-(d): PGSNet [18] *without* Random Frame. (e)-(f): PGSNet *with* Random Frame. Attention maps are generated from the feature maps after FEBF-1 [18] module. Red-yellow colors indicate stronger response. Note that in the bottom row, the left half of the window has no glass.

#### 4.4 Visualization by Grad-CAM

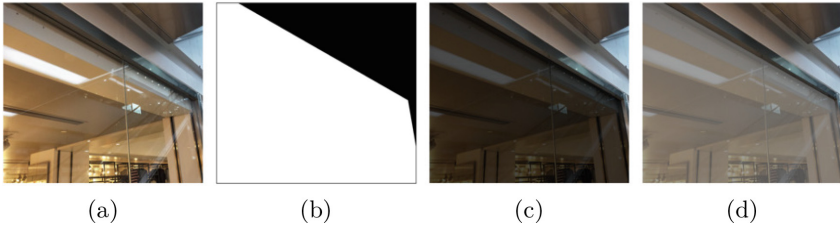
In Fig. 5, we show the attention map visualization by Grad-CAM [14] of PGSNet [18] trained with and without Random Frame. When Random Frame is not applied, the response is high near the edge of the glass. When Random Frame is applied, the overall response is higher, especially at the interior of the glass area (e.g., the first row). The second row shows that the model pays less attention to the handrail in front and produces a more accurate prediction. The bottom row is an example showing a half-open sliding window. Although both models (incorrectly) predict the left side as glass, the difference in the response between the left side and the right is larger when the model is trained with Random Frame. Overall, we can see that Random Frame encourages the model to pay more attention to the interior of the glass area and make the model become more confident with its predictions.

#### 4.5 Analysis of the Effect on BCE Loss

We investigate the effect of Random Frame on the binary cross entropy (BCE) loss while training PGSNet [18], with respect to the training and test data. As illustrated in Fig. 6(a), the training BCE loss becomes a little larger when Random Frame is applied. This indicates that the Random Frame can deceive the model and provide the model with more chances to learn. The test BCE loss is shown in Fig. 6(b). When Random Frame is not applied, the test loss



**Fig. 6.** Average BCE loss of PGSNet [18] in training with 95% confidence intervals (three runs) on the (a) training and (b) test data of the GDD dataset [13]. Note that the IoU loss decreases during training.



**Fig. 7.** Example of the edited image. (a): Input. (b): Ground Truth. (c): Brightness of the glass area is edited to be less. (d): Contrast of the glass area is edited to be less.

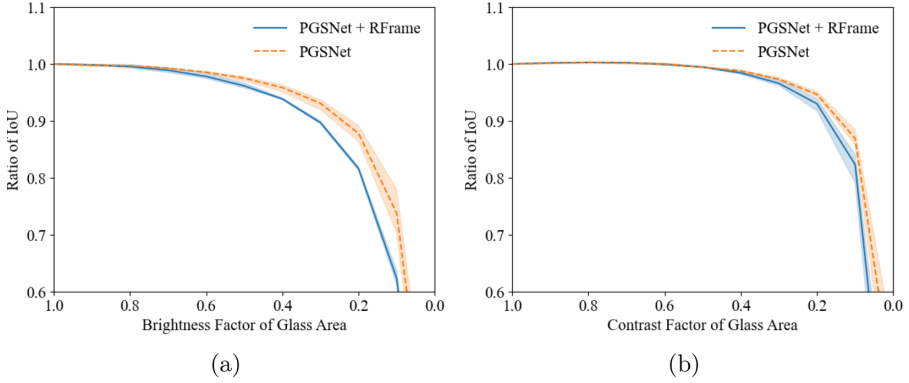
keeps increasing. This phenomenon usually suggests that the model overfits the training data. However, as the test IoU loss decreases during training, it can be interpreted as the model sacrificing the BCE loss to optimize IoU loss. On the other hand, Random Frame can keep the test BCE loss flat. Since the BCE loss reflects the confidence in predictions, we can tell that Random Frame can improve the robustness of the model. We also suspect that the reason the test BCE loss does not *decrease* may be that accurate pixel-wise prediction is too challenging for this model.

#### 4.6 The Usage of Features from the Glass

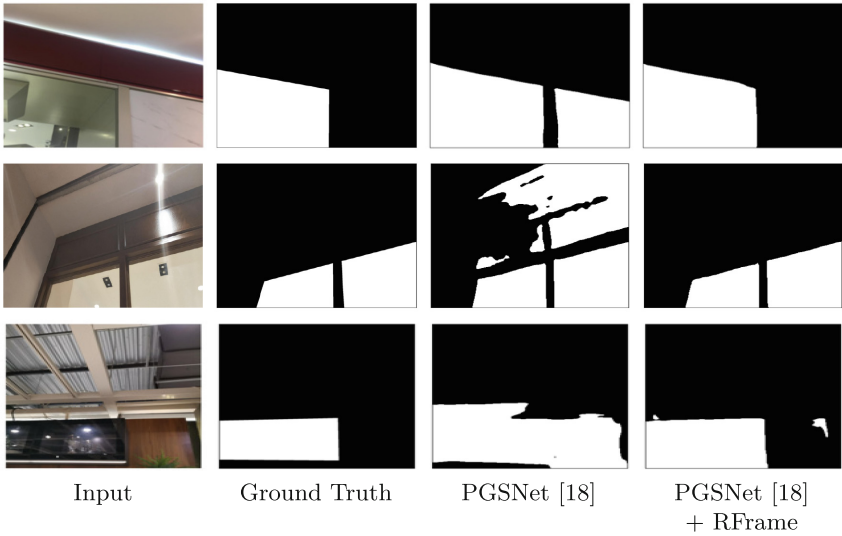
We wish to investigate the effect of Random Frame on the model’s reliance on features from the glass itself, rather than the surrounding area. We artificially adjust the brightness and the contrast in the glass area in test images and watch the effect. If the model relies on these cues, the performance would be affected.

**Brightness:** since the glass area is often brighter than the non-glass area because of reflection, we modify the brightness inside the glass area (example: Fig. 7(c)).

**Contrast:** as reflections generally increase the contrast inside the glass area, we



**Fig. 8.** Change of average IoU by PGSNet [18] with 95% confidence intervals (three runs) on GDD dataset [13], when (a) brightness or (b) contrast inside the glass area is edited. Y-axis is normalized by the IoU when test images are not edited.



**Fig. 9.** Qualitative results on the GDD dataset [13].

also change it in test images (example: Fig. 7(d)). Note that, in each experiment, the area outside of the glass area is left unchanged.

As shown in Fig. 8(a), IoU by the PGSNet [18] trained with Random Frame drops faster as the brightness in the glass areas is decreased. This suggests that Random Frame encourages the model to utilize more the brightness, or rather the contrast between glass and non-glass areas. From Fig. 8(b), we can see that when the contrast inside the glass area becomes less, which makes the reflections harder to recognize, the IoU also drops faster if Random Frame has been

applied. Thus, the model seems to become more sensitive to reflections by Random Frame. Comparing the two, the gap between the model trained with and without Random Frame is smaller in the latter figure. We theorize this is because the reflections only show in a limited region inside the glass area.

#### 4.7 Qualitative Comparison

We show qualitative results by PGSNet [18] on GDD dataset [13] in Fig. 9. Model trained with Random Frame is less likely to be deceived by frame-like objects and performs better.

## 5 Conclusions

We have presented the first data augmentation method specific to glass detection. Our proposed Random Frame inserts frames into training images to penalize the model for relying too heavily on features provided by the frames surrounding glasses, so that it learns to pay more attention to the interior of the glass area. We applied Random Frame to existing glass detection models and evaluated them on various datasets. The models received a noticeable performance increase and became more confident with predictions. Comprehensive experiments showed that it can make models better utilize the low-level features of glass. The limitation is that it is still difficult for models to make correct predictions in the case of an open window. As future work, we would like to address this problem by designing a new architecture.

**Acknowledgements.** This work was partially supported by JSPS KAKENHI Grant Number JP20H00615.

## References

1. Cao, Y., et al.: FakeMix augmentation improves transparent object detection. arXiv preprint [arXiv: 2103.13279v2](https://arxiv.org/abs/2103.13279v2) (2021)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. *PAMI* **40**(4), 834–848 (2018)
3. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint [arXiv: 1708.04552](https://arxiv.org/abs/1708.04552) (2017)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR*, pp. 770–778 (2016)
5. Huo, D., Wang, J., Qian, Y., Yang, Y.H.: Glass segmentation with RGB-thermal image pairs. *TIP* **32**, 1911–1926 (2023)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *NeurIPS*, pp. 1097–1105 (2012)
7. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)

8. Lin, J., He, Z., Lau, R.W.H.: Rich context aggregation with reflection prior for glass surface detection. In: CVPR, pp. 13415–13424 (2021)
9. Lin, J., Yeung, Y.H., Lau, R.W.H.: Depth-aware glass surface detection with cross-modal context mining. arXiv preprint [arXiv:2206.11250](https://arxiv.org/abs/2206.11250) (2022)
10. Lin, J., Yeung, Y.H., Lau, R.W.H.: Exploiting semantic relations for glass surface detection. In: NeurIPS, pp. 22490–22504 (2022)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
12. Mei, H., et al.: Glass segmentation using intensity and spectral polarization cues. In: CVPR, pp. 12622–12631 (2022)
13. Mei, H., et al.: Don’t hit me! Glass detection in real-world scenes. In: CVPR, pp. 3687–3696 (2020)
14. 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: ICCV, pp. 618–626 (2017)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
16. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
17. Xie, E., Wang, W., Wang, W., Ding, M., Shen, C., Luo, P.: Segmenting transparent objects in the wild. In: ECCV, pp. 696–711 (2020)
18. Yu, L., et al.: Progressive glass segmentation. *TIP* **31**, 2920–2933 (2022)
19. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: CutMix: regularization strategy to train strong classifiers with localizable features. In: ICCV, pp. 6023–6032 (2019)
20. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: beyond empirical risk minimization. In: ICLR (2018)



# PolypSegDiff: Dynamic Multi-scale Conditional Diffusion Model for Polyp Segmentation

Xiaogang Du<sup>1</sup>, Yipeng Jiao<sup>1</sup>, Tao Lei<sup>1</sup>(✉), Xuejun Zhang<sup>2</sup>, Yingbo Wang<sup>1</sup>,  
and Asoke K. Nandi<sup>3</sup>

<sup>1</sup> Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, China

leitao@sust.edu.cn

<sup>2</sup> School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, China

<sup>3</sup> Department of Electronic and Computer Engineering, Brunel University London, London, UK

**Abstract.** Diffusion models have demonstrated impressive potential in semantic segmentation tasks. However, these models cannot accurately segment hidden polyps with complex structures owing to the absence of multi-scale conditional features for guiding the reverse process of diffusion models. To address this issue, we propose a dynamic multi-scale conditional diffusion model (PolypSegDiff). First, we design a dynamic multi-scale integration module to fuse the noise segmentation mask and the original image, dynamically extract multi-scale conditional features, and strengthen the network's ability of identifying polyp areas. Second, we design a hierarchical feature enhancement module to extract and combine image features at different levels. This module significantly enriches the semantic diversity of conditional features, enabling the denoising network to more accurately understand the semantic relationships between polyps and the surrounding normal tissues. Experimental results across five publicly available polyp segmentation datasets demonstrate that PolypSegDiff outperforms existing popular methods in segmentation accuracy, achieving outstanding performance and robust generalization.

**Keywords:** Diffusion model · Polyp segmentation · Multi-scale features

## 1 Introduction

At present, colorectal cancer ranks as the fourth deadliest cancer worldwide, accounting for approximately 9.4% of all cancer-related deaths [19]. As major

---

This work is partly supported by National Natural Science Foundation of China (Nos. 61861024, 62271296, and 62201334), and Scientific Research Program Funded by Shaanxi Provincial Education Department (Nos. 23JP022, and 23JP014).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025  
A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15333, pp. 94–108, 2025.

[https://doi.org/10.1007/978-3-031-80136-5\\_7](https://doi.org/10.1007/978-3-031-80136-5_7)



precursors to colorectal cancer, the early identification of colon polyps is vital for the prevention and management of colorectal cancer.

Traditionally, the identification and analysis of colon polyps mainly rely on the experience of clinicians and manual segmentation, which not only requires significant time and effort but also is heavily influenced by subjective factors. As medical image analysis technology advances, automatic polyp segmentation has become a research hotspot that has attracted much attention. However, the low contrast of colonoscopic images, varying shapes of polyps, and the distinction between polyps and adjacent healthy tissue is notably vague, making automatic polyp segmentation an exceptionally challenging task [11, 29].

As deep learning continues to advance, scholars have proposed many automatic polyp segmentation models, which can generally be classified into two categories: Convolutional Neural Network (CNN)-based segmentation models [9, 10, 21, 25] and Transformer-based segmentation models [8, 14, 23, 31]. Although CNN-based segmentation models are effective for polyp segmentation tasks, they cannot fully consider the relationships and interdependencies between different regions in the image, lacking a comprehensive insight into the entire image. This leads to challenges when segmenting polyps that cover large areas or have uneven distribution. The Transformer-based model is adept at extracting effective global information, thereby enhancing the global context representation of networks. Nonetheless, due to only focusing on the global features, they may compromise the recognition of local details, resulting in a decrease of segmentation performance for tiny polyps and polyp boundaries.

Recently, the diffusion model [12] has received widespread attention due to its excellent feature learning ability in the reverse denoising process [1, 7]. Some scholars employed diffusion models in image segmentation tasks and achieved good segmentation results [5, 6]. For example, Baranchuk *et al.* [2] indicated that the U-shaped structure in the denoising diffusion model can effectively extract the semantic features of images, and initially employed diffusion models for semantic segmentation tasks. Wolleb *et al.* [26] were the first to utilize the diffusion model in medical image segmentation, leveraging stochastic noise in the model to generate a set of implicit segmentation masks, which effectively improves the segmentation performance. To intensify feature constraints in the diffusion process, MedSegDiff [28] integrates the noise segmentation mask into the encoding process and employs a feature frequency parser to suppress the high-frequency noise in the diffusion process, thus refining segmentation results. However, MedSegDiff is prone to produce incorrect masks due to limited global representation capabilities. To tackle this challenge, Wu *et al.* [27] enhanced the MedSegDiff by integrating Transformer to better capture the interaction between segmentation noise and semantic information, further improving the segmentation accuracy. In addition, Bozorgpour *et al.* [4] proposed DermoSegDiff with a novel boundary-aware loss, exhibiting impressive segmentation performance on skin lesion datasets.

In summary, diffusion model has shown great potential and application value in medical image segmentation tasks. However, there are still three main issues

in applying the diffusion model to polyp segmentation tasks: (1) The above methods cannot fully consider the complexity of polyps with different sizes when extracting conditional features, resulting in poor segmentation performance for various-sized and hidden polyps. (2) The above methods usually only extract conditional features from the original image or noise mask, ignoring the impact of feature diversity on segmentation performance. (3) The above methods fail to integrate features at different levels according to the diversity of semantic features, resulting in not fully capturing the crucial semantic information for polyp boundaries, thereby diminishing segmentation accuracy.

To address these issues, we propose a dynamic multi-scale conditional diffusion model for polyp segmentation, called PolypSegDiff. The principal contributions of our work are summarized as follows:

(1) We design the Dynamic Multi-scale Integration Network (DMIN) to fuse the features of the noise segmentation mask and the original image, and dynamically extract multi-scale conditional features, which not only effectively reduces the misleading of noise during feature fusion, but also significantly enhances the network’s capability to identify polyps with various shapes and sizes.

(2) We design the Hierarchical Feature Enhancement Module (HFEM) for multi-scale information fusion. HFEM effectively improves the semantic richness of conditional features by extracting information from different levels, and enables the denoising network to more accurately capture the spatial interrelations and semantic associations between polyps and the surrounding tissues.

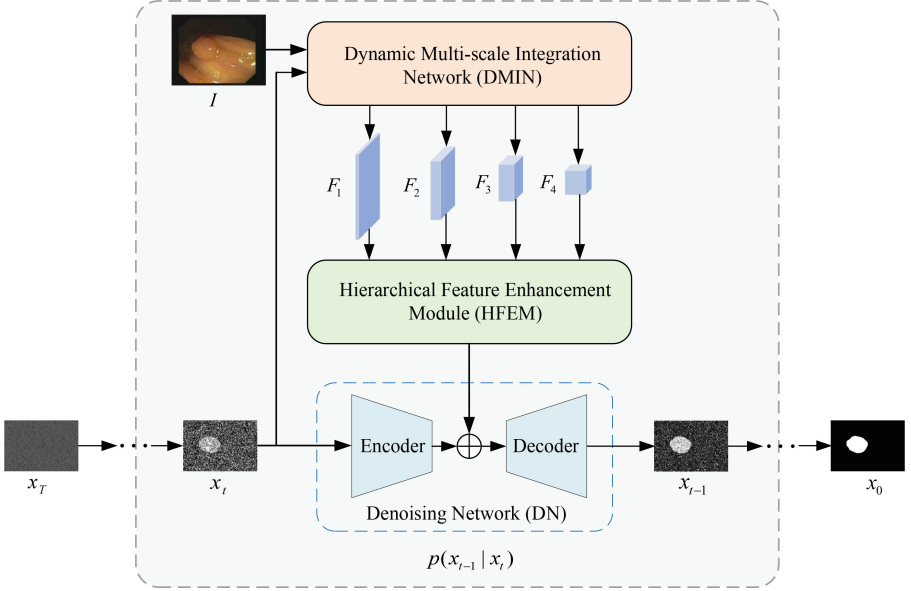
(3) Comprehensive experiments across five datasets for polyp segmentation indicate that PolypSegDiff is superior to several popular polyp segmentation methods in terms of accuracy. Notably, it achieves outstanding results, even on datasets containing more challenging and hidden polyps.

## 2 Method

Figure 1 illustrates the overall framework of PolypSegDiff. PolypSegDiff mainly includes three key components: Denoising Network (DN), Dynamic Multi-scale Integration Network (DMIN), and Hierarchical Feature Enhancement Module (HFEM). Specifically, PolypSegDiff utilizes the reverse process of the DN to iteratively obtain clear segmentation masks. Before this, the DMIN integrates the original image and the noisy segmentation mask from the current iteration to extract multi-scale features. After receiving the multi-scale features from DMIN, HFEM uses Detail Refinement Module (DRM) and Hierarchical Aggregation Module (HAM) to fuse features from different levels. The fused features are concatenated with the encoder output of the denoising network, and then input into the decoder of the denoising network for upsampling, allowing the denoising network to progressively generate clear segmentation results.

### 2.1 Denoising Network

PolypSegDiff regards the segmentation task as the reverse generative process based on diffusion model and generates predictions through the denoising net-



**Fig. 1.** The overall framework of PolypSegDiff, which mainly includes three important components: DMIN, HFEM and DN.

work. We implement the denoising network based on Denoising Diffusion Probabilistic Models (DDPM) [12]. The denoising network is divided into two main stages, which are the forward process and the reverse process. During the forward process, Gaussian noise is incrementally introduced to the original image  $x$  over a sequence of continuous steps. The forward process is represented as:

$$q(x_{1:T} | x_0) = \prod_{t=1}^T q(x_t | x_{t-1}), \quad (1)$$

where  $T$  represents the number of diffusion steps, and  $x_t$  represents the noise masks during the diffusion process. In each iteration, the addition of Gaussian noise follows:

$$q(x_t | x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}\right), \quad (2)$$

where  $\beta_t$  is a parameter that determines the schedule for introducing noise, and  $\mathbf{I}$  is the identity matrix of size  $n \times n$ . The forward process supports sampling at any step  $t$ :

$$q(x_t | x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t) \mathbf{I}\right), \quad (3)$$

$$\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{s=0}^t \alpha_s. \quad (4)$$

During the reverse diffusion process, the denoising network gradually restores the original features of the image  $x_t$  through multiple denoising iterations to generate the predicted mask  $\hat{x}_0$ . The reverse process can be expressed as:

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (5)$$

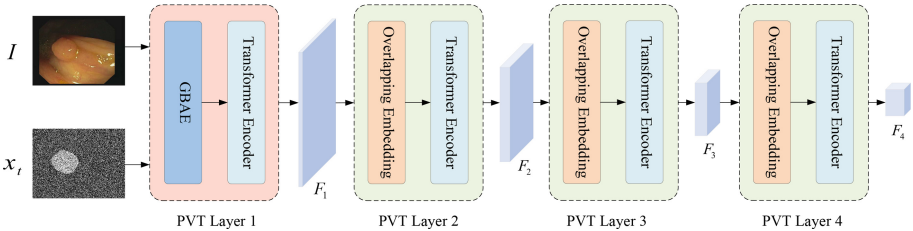
where  $\theta$  represents the reverse process parameter,  $\Sigma_\theta(x_t, t)$  is set to  $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ , and  $\mu_\theta(x_t, t)$  can be represented as:

$$\mu_\theta(x_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_{t-1}}x_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{x}_0. \quad (6)$$

We train the proposed PolypSegDiff by optimizing the loss function  $\mathcal{L}(\hat{x}_0, x_0)$ , which can be formulated as:

$$\mathcal{L}(\hat{x}_0, x_0) = \mathcal{L}_{\text{IoU}}^w(\hat{x}_0, x_0) + \mathcal{L}_{\text{BCE}}^w(\hat{x}_0, x_0), \quad (7)$$

where  $\mathcal{L}_{\text{IoU}}^w$  represents the weighted Intersection-over-Union loss,  $\mathcal{L}_{\text{BCE}}^w$  represents the weighted binary cross-entropy loss. The weighting coefficients are obtained by calculating the absolute difference between the predicted mask and the original mask. Specifically, first, we compute the average value of the local region of the input mask through an average pooling layer, then subtract it from the original mask and take the absolute value to obtain the importance weight for each position. Since the mask values of boundary and structural change regions usually differ significantly from their surrounding average values, this weight enhances the importance of these regions.

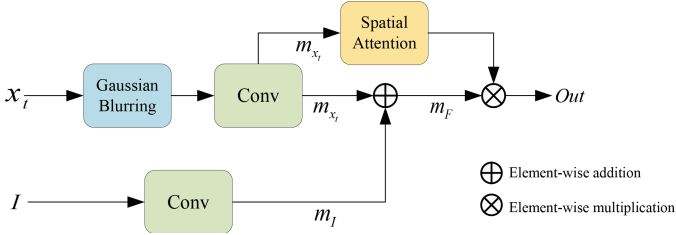


**Fig. 2.** The structure of DMIN. In the PVT layer 1, we employ GBAE to replace Overlapping Embedding for feature fusion and noise suppression.

## 2.2 DMIN

For the polyp segmentation task, the polyp regions are frequently hidden and difficult to distinguish from the background. Relying solely on a single original image  $I$  as the prior condition in the diffusion steps makes it challenging for the

network to learn effective semantic information. To effectively guide the denoising network, we design the DMIN module, which can extract rich multi-scale features and adaptively generate conditional features according to the reverse process of the denoising network. The main structure of DMIN is illustrated in Fig. 2. In DMIN, we cascade four PVT layers [24] to extract multi-scale features. Each PVT layer consists of an overlapping embedding and a Transformer encoder, which can extract multi-scale conditional features at different depths.



**Fig. 3.** The structure of GBAE. Especially, we introduce Gaussian blurring to reduce the impact of noise and use segmentation masks to emphasize the location and boundary of polyps.

In order to dynamically generate conditional features according to the denoising process, we design the Gaussian Blurring Attention Embedding (GBAE) and replace the overlapping embedding with GBAE in the first PVT layer. This mechanism aims to address the specific requirements of diffusion models in medical image segmentation applications. In diffusion model applications, noisy segmentation masks are used to guide the iterations of the model, but the noise in the masks may mislead the model. By introducing the Gaussian blur layer, we aim to mitigate the impact of noise, thereby making more effective use of the noisy masks to highlight the target areas.

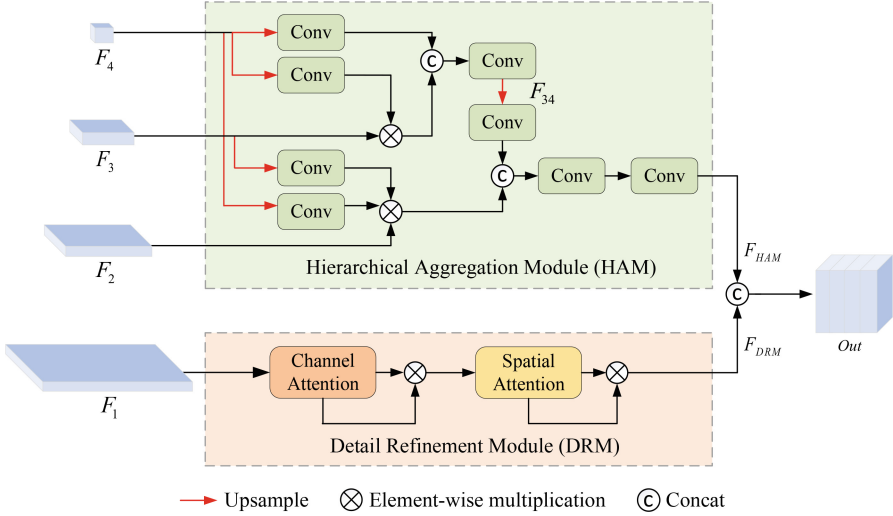
Figure 3 presents the detailed architecture of the GBAE. In the GBAE, we introduce the segmentation mask  $x_t$  and integrate it into the features of the original image to obtain adaptive conditional features, which can enhance the polyp areas. By combining GBAE and PVT layers, DMIN outputs dynamic multi-scale conditional features to guide network learning with the diffusion iteration process.

GBAE mainly includes three steps: first, in the reverse process, the segmentation mask  $x_t$  of the current step often contains a lot of noise. Directly using  $x_t$  to guide the features of the original image will diminish the precision of segmentation. Therefore, in GBAE, Gaussian blurring is employed to reduce the impact of noise in the segmentation mask  $x_t$ , thereby elevating the segmentation capability of the PolypSegDiff. Second, we conduct convolution operations on the segmentation mask  $x_t$  and the original image  $I$ , respectively, to obtain the feature maps  $m_{x_t}$  and  $m_I$ , then adding them to generate the fused feature map  $m_F$ . Finally, the output of the spatial attention applied on  $m_{x_t}$  is multiplied by  $m_F$ , to generate more precise conditional features.

GBAE can be formulated as:

$$Out = (Conv(GB(x_t)) \oplus Conv(I)) \otimes SA(Conv(GB(x_t))), \quad (8)$$

where  $\oplus$  represents the element-wise addition,  $\otimes$  represents the element-wise multiplication,  $GB$  is the Gaussian Blurring operation,  $SA$  is the Spatial Attention operation, and  $Conv$  is the Convolution operation.



**Fig. 4.** The structure of HFEM, which consists of HAM and DRM.

### 2.3 HFEM

To address the issues arising from the diverse sizes and shapes of polyps, along with the ambiguous boundaries between polyps and adjacent tissues in polyp segmentation tasks, we designed the HFEM to integrate multi-scale features and input the fused features into the denoising network to guide the reverse process.

After DMIN outputs the multi-scale features, it is necessary to effectively fuse them. Therefore, we design the HFEM to integrate multi-scale features and input the fused features into the Denoising Network to guide the reverse process. HFEM mainly consists of Detail Refinement Module (DRM) and Hierarchical Aggregation Module (HAM). HAM is used to gather semantic cues and localize the polyp by progressively aggregating multi-scale features. DRM aims to suppress noise and enhance the low-level feature representation of polyps, including texture, color, and edges. Finally, the features from HAM and DRM are fused through concatenation, and the output of HFEM is input into the denoising network as conditional features. The structure of HFEM is shown in Fig. 4.

Specifically, we first use DMIN to extract feature maps  $F_1$ ,  $F_2$ ,  $F_3$ , and  $F_4$  at four different layers. The low-level feature map  $F_1$  provides detailed information of the polyps, while the high-level feature maps  $F_2$ ,  $F_3$  and  $F_4$  provide the shape and boundary semantic features of the polyps.

To capture the hidden polyp details from different dimensions, we input  $F_1$  into DRM, which primarily comprises channel attention and spatial attention. We utilize these two mechanisms to enhance the precise extraction and analysis of polyp image features. The spatial attention focuses on critical spatial locations within the image, enabling the model to concentrate on lesion areas and thereby improving diagnostic accuracy. Concurrently, the channel attention assesses the contribution of each channel to the overall feature set, emphasizing the more significant feature channels. This allows the model to effectively capture the image’s colors and textures and to accurately detect polyps using specific channel information. By employing these mechanisms, the DRM not only accentuates important feature channels but also accurately pinpoints the locations of polyps, providing more detailed and precise information. This enhances the model’s capability to recognize and analyze polyp regions effectively. DRM can be represented as:

$$F_{DRM} = SA(CA(F_1) \otimes F_1) \otimes (CA(F_1) \otimes F_1). \quad (9)$$

For the HAM, it utilizes multi-layer cascading and convolution operations to deeply fuse features from different levels. We process and fuse these features through multiple convolutional layers, which assist the model in extracting and integrating features from various levels more effectively. This process enhances the richness and accuracy of feature representation. Specifically, we first fuse the high-level features  $F_2$ ,  $F_3$ , and  $F_4$  through multi-layer upsampling and cascading operations. Secondly, we upsample the high-level feature  $F_4$  to the same size as the feature maps  $F_3$  and  $F_2$ . Then, we conduct convolution operations on each feature map, and follow by element-wise multiplication and concatenation to perform feature fusion. Finally, we output the feature map  $F_{HAM}$  that contains rich high-level semantic features. HAM can be formulated as:

$$F_{34} = Conv(Concat(Conv(F_4 \otimes F_3, Conv(F_4))), \quad (10)$$

$$F_{HAM} = Conv(Conv(Concat(Conv(F_4) \otimes (F_3) \otimes F_2, Conv(F_{34}))))). \quad (11)$$

The final output of HFEM can be represented as:

$$Out = Concat(F_{HAM}, F_{DRM}). \quad (12)$$

## 3 Experiments

### 3.1 Datasets

To evaluate the effectiveness of the PolypSegDiff, we conduct experiments using five publicly available polyp datasets, including Kvasir-SEG [13], CVC-ClinicDB

[3], CVC-300 [22], ColonDB [20] and ETIS [18]. Following the previous method [9], the training dataset has a total of 1450 images, including 900 images from Kvasir-SEG and 550 images from CVC-ClinicDB. 100 images were randomly extracted from the training datasets to serve as the validation datasets. For testing, we evaluate the performance of PolypSegDiff on all five datasets.

### 3.2 Evaluation Metrics

We employ Dice and IoU, two widely used metrics, to quantitatively evaluate the performance of PolypSegDiff. In the experiments, we calculate the mean Dice (mDice) and mean IoU (mIoU) of all test samples. The higher these metric values, the more precise the segmentation results.

### 3.3 Implementation Details

PolypSegDiff is trained on a single NVIDIA A30 with GPU with 24GB memory. During training, the image size is adjusted to  $256 \times 256$ , the batch size is configured to 32, and the maximum number of epochs is set at 200. We employ the AdamW optimizer with a learning rate decay strategy, with an initial learning rate of 0.001.

### 3.4 Comparative Experiment

We compare the PolypSegDiff with current mainstream networks, including UNet [17], UNet++ [33], SFA [10], PraNet [9], MSEG [15], SANet [25], TGANet [21], APCNet [30], CFANet [32], CaraNet [16], and DermoSegDiff [4]. The comparative experimental results are shown in Table 1.

In Table 1, it can be seen that PolypSegDiff achieves the best results on the Kvasir-SEG, CVC-300, CVC-ColonDB, and ETIS, and achieves second-best results on the CVC-ClinicDB dataset. Especially on the two challenging datasets CVC-ColonDB and ETIS, PolypSegDiff achieves obvious performance advantages, mDice is 6.8% and 3.7% higher than the second-best method, respectively. It is 0.7% and 0.5% higher than the mDice of the second-best method on the Kvasir-SEG and CVC-300 datasets, respectively. On the CVC-ClinicDB dataset with second-best performance, the difference from the best method on mDice is only 0.2%. Similarly, in terms of the mIoU, PolypSegDiff also achieves the state-of-the-art results on four different polyp datasets. Notably, our method demonstrates significant advantages over the advanced diffusion-based method DermoSegDiff. PolypSegDiff significantly outperforms DermoSegDiff in both mDice and mIoU metrics. Therefore, PolypSegDiff has greater accuracy and generalization than the comparative methods, and can accurately segment the polyps. Especially for some challenging segmentation regions, PolypSegDiff can achieve more accurate segmentation than other comparative methods.

Figure 5 shows the visualization results of PolypSegDiff and eight state-of-the-art polyp segmentation methods on the polyp segmentation datasets. Based



**Table 1.** The comparative experiment on five polyp segmentation datasets. Blue represents the best results, and red represents the second-best results.

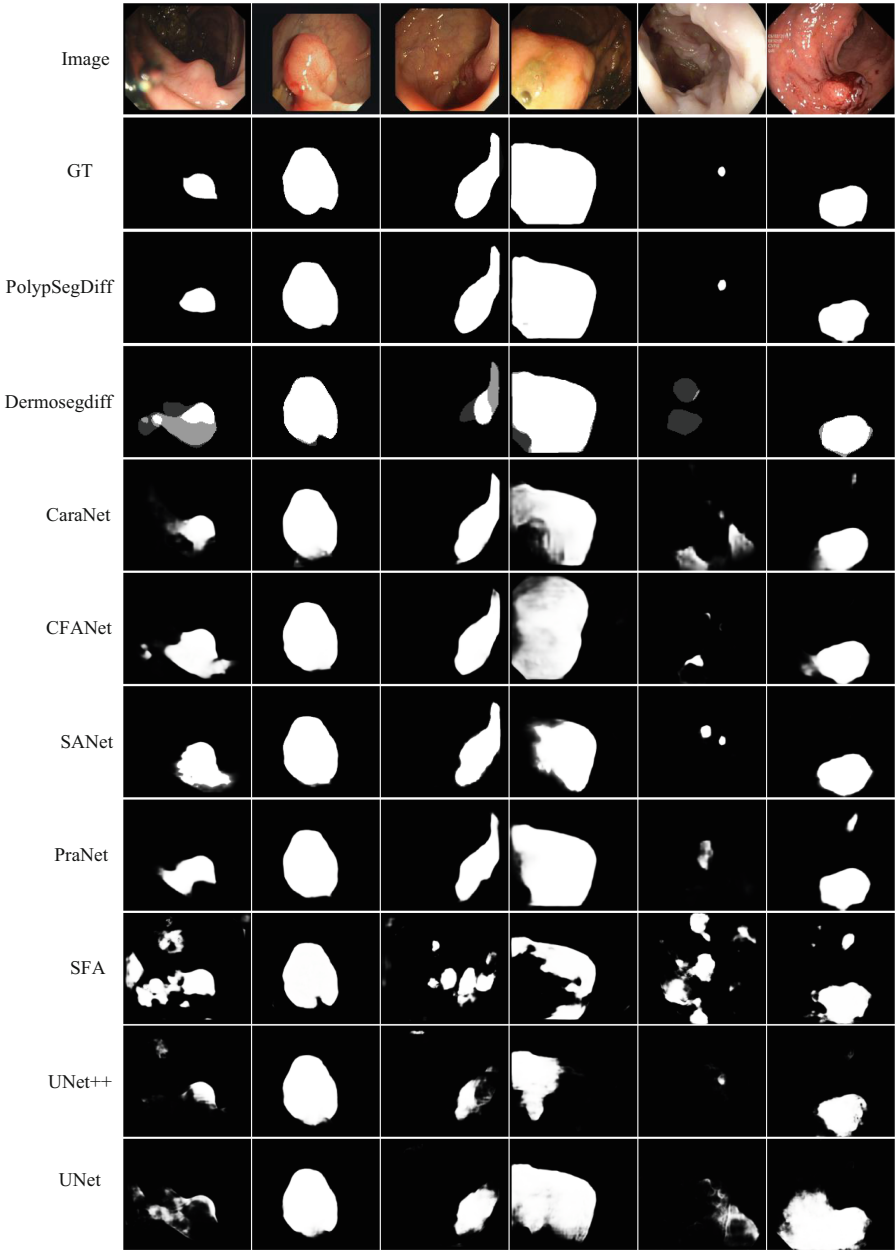
Method	Kvasir-SEG		CVC-ClinicDB		CVC-300		CVC-ColonDB		ETIS	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net [17]	0.818	0.746	0.823	0.755	0.710	0.627	0.512	0.444	0.398	0.335
U-Net++ [33]	0.821	0.743	0.794	0.729	0.707	0.624	0.483	0.410	0.401	0.344
SFA [10]	0.723	0.611	0.700	0.607	0.467	0.329	0.469	0.347	0.297	0.217
PraNet [9]	0.898	0.840	0.899	0.849	0.871	0.797	0.712	0.640	0.628	0.567
MSEG [15]	0.897	0.839	0.909	0.864	0.874	0.804	0.735	0.666	0.700	0.630
SANet [25]	0.904	0.847	0.916	0.859	0.888	0.815	0.753	0.670	<b>0.750</b>	0.654
TGANet [21]	0.894	0.839	0.907	0.855	0.886	0.819	0.707	0.633	0.653	0.578
CaraNet [16]	<b>0.918</b>	<b>0.865</b>	<b>0.936</b>	<b>0.887</b>	<b>0.903</b>	<b>0.838</b>	<b>0.773</b>	<b>0.689</b>	0.747	<b>0.672</b>
APCNet [30]	0.913	0.859	<b>0.934</b>	<b>0.886</b>	0.893	0.827	0.758	0.682	0.726	0.648
CFANet [32]	0.915	0.861	0.933	0.883	0.893	0.827	0.743	0.665	0.732	0.655
DermoSegDiff [4]	0.887	0.830	0.890	0.837	0.854	0.753	0.721	0.652	0.746	0.669
<b>PolypSegDiff</b>	<b>0.925</b>	<b>0.878</b>	<b>0.934</b>	<b>0.886</b>	<b>0.908</b>	<b>0.841</b>	<b>0.841</b>	<b>0.765</b>	<b>0.784</b>	<b>0.698</b>

on visual results, our model’s outputs most closely correspond to the ground truth. It is obvious that PolypSegDiff always shows robust segmentation capabilities for polyps with different sizes and types, outperforming other models in terms of adaptability and accuracy.

In addition, PolypSegDiff has superior perception of hidden polyps which are small in size and difficult to distinguish from surrounding tissues. Specifically, the fifth column of Fig. 5 presents a particularly challenging example. This polyp not only has very low contrast with the surrounding tissue but is also located in a shadow and is very small in area. Despite these challenges, our method is still able to make predictions that are very close to the actual situation, while avoiding the generation of artifacts and false-positive regions. Therefore, our method demonstrates excellent performance in handling difficult-to-recognize concealed polyps.

### 3.5 Ablation Study

The proposed PolypSegDiff has two important innovations: first, DMIN is designed to fuse the features of the noise segmentation mask and the original image, and dynamically extract multi-scale conditional features. Secondly, HFEM is designed to utilize the different characteristics of features at different levels to extract rich shape and boundary features of polyps as well as multi-scale information. To verify the effectiveness of DMIN and HFEM, we conduct ablation study on the CVC-ColonDB dataset between PolypSegDiff and Baseline based on PVT [24] using the same architecture. Specifically, in Baseline, we use PVT instead of DMIN, use layer-by-layer upsampling and concatenation instead

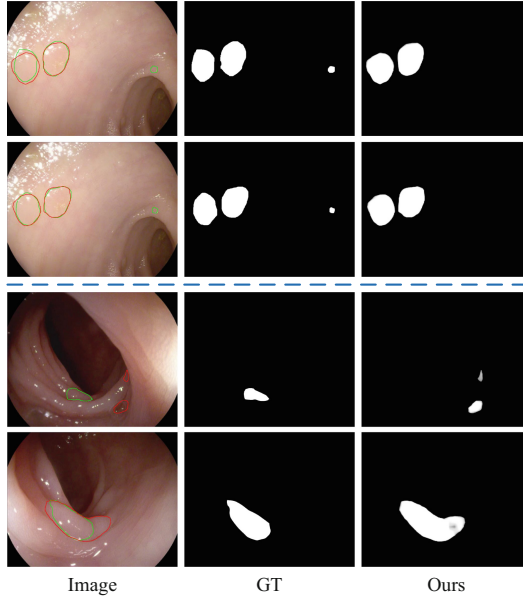


**Fig. 5.** The visualization results of PolypSegDiff and the comparative methods for polyps with diverse shapes. GT refers to the ground truth.

of HFEM. We gradually add DMIN and HFEM modules to Baseline, and calculate mDice and mIoU to prove the effectiveness of the innovations. The results of the ablation study are shown in Table 2. Experimental results show that both DMIN and HFEM can markedly enhance the accuracy of polyp segmentation tasks.

**Table 2.** Ablation Study on the CVC-ColonDB dataset.

DMIN	HFEM	mDice[%]	mIoU[%]
✗	✗	0.816	0.747
✓	✗	0.821	0.750
✗	✓	0.828	0.756
✓	✓	<b>0.841</b>	<b>0.765</b>



**Fig. 6.** Two failure cases of the proposed PolypSegDiff. The green outline represents the GT, and the red outline represents our segmentation results. (Color figure online)

## 4 Discussion

Despite achieving significant results, PolypSegDiff still has two limitations, which are shown in Fig. 6. First, in the first and second rows of Fig. 6, PolypSegDiff can accurately segment the two larger polyps on the left but overlooked the small

polyp on the right. While PolypSegDiff can identify small hidden polyps, it may still show bias when processing multiple polyps of significantly different sizes. Second, in the third and fourth rows of Fig. 6, irregular folds in the colon often intersect with and closely resemble polyps, potentially leading to misidentification. The high similarity between polyps and normal tissue, along with their diverse shapes, remains a key factor affecting segmentation accuracy. We aim to enhance PolypSegDiff to more effectively capture the structural details of polyps and more accurately distinguish them from normal tissue.

## 5 Conclusion

In this paper, we propose a dynamic multi-scale conditional diffusion model for polyp segmentation called PolypSegDiff. Initially, we design the DMIN, which effectively merges the noise segmentation masks with the original images and extracts multi-scale features, enhancing the perception of the network for polyp areas. Additionally, we design the HFEM for multi-scale feature fusion, guiding the denoising network to generate more accurate segmentation masks by producing conditional features that contain both the location and details of the polyps. Experimental results indicate that the proposed PolypSegDiff can effectively identify hidden polyps and precisely delineate the boundaries of polyps with various shapes.

## References





1. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. *Adv. Neural. Inf. Process. Syst.* **34**, 17981–17993 (2021)
2. Baranchuk, D., Rubachev, I., Voynov, A., Khruikov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126* (2021)
3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilarino, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **43**, 99–111 (2015)
4. Bozorgpour, A., Sadegheih, Y., Kazerouni, A., Azad, R., Merhof, D.: DermosegDiff: a boundary-aware segmentation diffusion model for skin lesion delineation. In: *Predictive Intelligence in Medicine*, pp. 146–158. Springer (2023)
5. Brempong, E.A., Kornblith, S., Chen, T., Parmar, N., Minderer, M., Norouzi, M.: Denoising pretraining for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4175–4186. IEEE (2022)
6. Chen, T., Li, L., Saxena, S., Hinton, G., Fleet, D.J.: A generalist framework for panoptic segmentation of images and videos. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 909–919. IEEE (2023)
7. Chen, X., Liu, Z., Xie, S., He, K.: Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404* (2024)

8. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-PVT: Polyp segmentation with pyramid vision transformers. arXiv preprint [arXiv:2108.06932](https://arxiv.org/abs/2108.06932) (2021)
9. Fan, D.-P., et al.: PraNet: parallel reverse attention network for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 263–273. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-59725-2\\_26](https://doi.org/10.1007/978-3-030-59725-2_26)
10. Fang, Y., Chen, C., Yuan, Y., Tong, K.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 302–310. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-32239-7\\_34](https://doi.org/10.1007/978-3-030-32239-7_34)
11. Guo, X., Yang, C., Liu, Y., Yuan, Y.: Learn to threshold: ThresholdNet with confidence-guided manifold mixup for polyp segmentation. *IEEE Trans. Med. Imaging* **40**(4), 1134–1146 (2020)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Adv. Neural. Inf. Process. Syst.* **33**, 6840–6851 (2020)
13. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: *MultiMedia Modeling*, pp. 451–462. Springer (2020)
14. Jha, D., Tomar, N.K., Sharma, V., Bagci, U.: TransNetR: transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing. In: *Medical Imaging with Deep Learning*, pp. 1372–1384. PMLR (2024)
15. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: MSeg: a composite dataset for multi-domain semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2879–2888. IEEE (2020)
16. Lou, A., Guan, S., Ko, H., Loew, M.H.: CaraNet: context axial reverse attention network for segmentation of small medical objects. In: *Medical Imaging 2022: Image Processing*, vol. 12032, pp. 81–92. SPIE (2022)
17. 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](https://doi.org/10.1007/978-3-319-24574-4_28)
18. Silva, J., Histance, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.* **9**, 283–293 (2014)
19. Sung, H., et al.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: Cancer J. Clin.* **71**(3), 209–249 (2021)
20. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med. Imaging* **35**(2), 630–644 (2015)
21. Tomar, N.K., Jha, D., Bagci, U., Ali, S.: TGANet: text-guided attention for improved polyp segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 151–160. Springer (2022)
22. Vázquez, D., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. *J. Healthc. Eng.* **2017**(1), 4037190 (2017)
23. Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., Song, S.: Stepwise feature fusion: local guides global. In: *Medical Image Computing and Computer Assisted Intervention MICCAI 2022*, pp. 110–120. Springer (2022)
24. Wang, W., et al.: PVT v2: improved baselines with pyramid vision transformer. *Comput. Vis. Media* **8**(3), 415–424 (2022)
25. Wei, J., et al.: Shallow attention network for polyp segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 699–708. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-87193-2\\_66](https://doi.org/10.1007/978-3-030-87193-2_66)

26. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: International Conference on Medical Imaging with Deep Learning, pp. 1336–1348. PMLR (2022)
27. Wu, J., Fu, R., Fang, H., Zhang, Y., Xu, Y.: MedSegDiff-V2: Diffusion based medical image segmentation with transformer. arXiv preprint [arXiv:2301.11798](https://arxiv.org/abs/2301.11798) (2023)
28. Wu, J., et al.: MedSegDiff: medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning, pp. 1623–1639. PMLR (2024)
29. Yin, Z., Liang, K., Ma, Z., Guo, J.: Duplex contextual relation network for polyp segmentation. In: IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2022)
30. Yue, G., Li, S., Cong, R., Zhou, T., Lei, B., Wang, T.: Attention-guided pyramid context network for polyp segmentation in colonoscopy images. *IEEE Trans. Instrum. Meas.* **72**, 1–13 (2023)
31. Zhang, Y., Liu, H., Hu, Q.: TransFuse: fusing transformers and CNNs for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 14–24. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-87193-2\\_2](https://doi.org/10.1007/978-3-030-87193-2_2)
32. Zhou, T., et al.: Cross-level feature aggregation network for polyp segmentation. *Pattern Recogn.* **140**, 109555 (2023)
33. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-00889-5\\_1](https://doi.org/10.1007/978-3-030-00889-5_1)



# Exploiting Text-Image Latent Spaces for the Description of Visual Concepts

Laines Schmalwasser<sup>1,2</sup> , Jakob Gawlikowski<sup>1</sup> , Joachim Denzler<sup>2</sup> ,  
and Julia Niebling<sup>1</sup> 

<sup>1</sup> Institute of Data Science, German Aerospace Center, 07745 Jena, Germany

<sup>2</sup> Computer Vision Group, Friedrich Schiller University Jena, 07743 Jena, Germany

`laines.schmalwasser@dlr.de`

**Abstract.** Concept Activation Vectors (CAVs) offer insights into neural network decision-making by linking human friendly concepts to the model's internal feature extraction process. However, when a new set of CAVs is discovered, they must still be translated into a human understandable description. For image-based neural networks, this is typically done by visualizing the most relevant images of a CAV, while the determination of the concept is left to humans. In this work, we introduce an approach to aid the interpretation of newly discovered concept sets by suggesting textual descriptions for each CAV. This is done by mapping the most relevant images representing a CAV into a text-image embedding where a joint description of these relevant images can be computed. We propose utilizing the most relevant receptive fields instead of full images encoded. We demonstrate the capabilities of this approach in multiple experiments with and without given CAV labels, showing that the proposed approach provides accurate descriptions for the CAVs and reduces the challenge of concept interpretation.

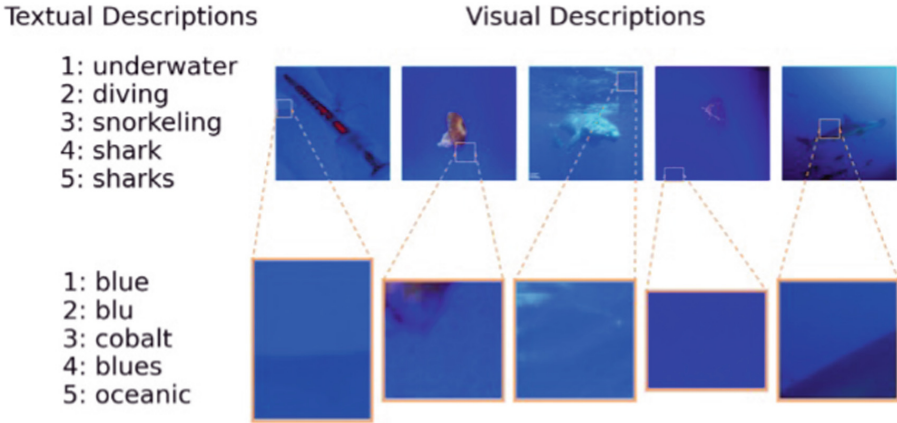
**Keywords:** XAI · Explainability · Concepts · Textual Description · Text-Image-Embeddings

## 1 Introduction

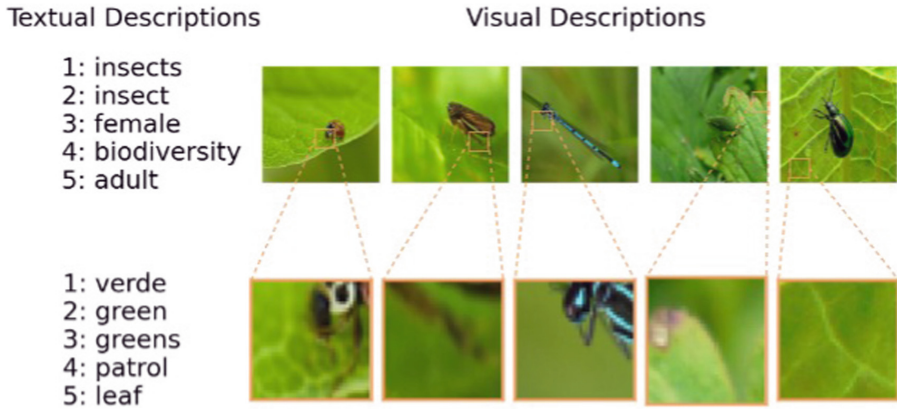
One major challenge of deep neural networks is their black-box nature which makes the interpretation of their behavior difficult. To mitigate this drawback, multiple approaches have been proposed to highlight relevant parts of the input data for a given prediction, for example, LIME [27], SHAP [20], GradCAM [28], LRP [2] and Feature Visualization [24]. Another idea is to explain the internal mechanism of a deep neural network in terms of concepts that are understandable and easy to communicate to humans [5, 15, 26]. One attempt to identify such concepts is with so-called Concept Activation Vectors (CAVs) [15]. A CAV is a

---

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/978-3-031-80136-5\\_8](https://doi.org/10.1007/978-3-031-80136-5_8).



(a) Top derived descriptions: *underwater* vs. *blue*



(b) Top derived descriptions: *insects* vs. *verde*

**Fig. 1.** Examples of two CAVs computed from the first residual block of a ResNet50, trained on Animals with Attributes 2 [31]. The first row of each subfigure shows the full representative images of the CAVs and the textual descriptions generated based on the full images. The second row shows the representative receptive fields for the same CAVs and the textual descriptions are derived from the receptive fields.

vector in the feature space of the activations of a specific network layer. It is designed to point to the direction of activations that are connected to a specific human understandable concept.

The idea behind CAVs is that a human defined concept that contributes to the model decisions has a representation in the model’s embedding space. For



example, the concept *stripe pattern* should have a corresponding representation when the model uses it in the decision-making process to predict a zebra.

In the literature, approaches have been suggested to find CAVs in a supervised and an unsupervised manner: While for the supervised approaches example images that contain the desired concepts are utilized [15, 21, 33], the unsupervised approaches use, for example, network bottlenecks to extract CAVs [32, 33].

We aim to describe the utilized concepts of a pretrained network without any assumptions about the concepts and without the need for example images for the concept. Hence, we focus on the description of an unsupervised discovered set of CAVs. As the discovered CAVs are given as vectors in the feature space, the encoded concepts need to be described for humans. A common way is to show images of a given dataset, which are most similar to the respective CAV in the hidden representation. However, this introduces the need for interpretation to derive a compact and communicable meaning from the given images.

To avoid the need for human interpretation, we propose to determine a ranking of textual descriptions for each concept. Depending on the CAV, the textual descriptions to be ranked, and the fine granularity of the text embedding, the highest ranked descriptions can be highly redundant. Therefore, we further derive a single common description based on the  $k$  highest ranked descriptions. Depending on the ranking, this common description can differ from the highest ranked description.

We build up on existing approaches to describe the information filtered by individual neurons in a textual way, for example, [23]. In this approach a neuron is described by generating a textual description for the relevant images of a neuron for which the neuron has the highest activation. The textual description is chosen as the best fitting one out of multiple candidates. In contrast to individual neurons, a major advantage of CAVs is that they represent vectors in the feature space and not only individual scalar neuron outputs. The total number of CAVs is usually significantly lower than the number of neurons in the corresponding layer.

The textual descriptions of the individual neurons in [23] are based on the full images that are relevant for the considered neuron. However, when the variety of images in a data set is not large enough, it is often not possible to separate highly correlated concepts, especially concepts of different degrees of abstraction, purely based on the full images. One example of the issue of highly correlated concepts are the concepts *insects* and *verde*, see Fig. 1a. An example of concepts of different degrees of abstraction are the concepts *underwater* and *blue*, as in many cases *underwater* is a specification of *blue*, see Fig. 1b. To address this limitation, we propose to use receptive fields instead of the full images for the generation of the textual descriptions. By replacing the full images with receptive fields, we can focus on the parts of the images, where an evaluated concept is most present. This reduces the noise that can affect the textual description of the concept.

In summary, the interpretation process of a neural network by ranking textual descriptions of human understandable concepts is represented by CAVs. Further,

we derive a single common textual description to decrease the redundancy. Our main contributions are:

- We enhance the automatic concept discovery in a trained model by interpreting the visual CAVs with textual descriptions.
- We derive a common concept description from the top- $k$  computed textual descriptions to reduce redundancy.
- We propose using receptive fields to derive the textual descriptions and introduce concept scores to measure the relevance of the receptive fields. By that the textual descriptions focus on the relevant parts of the images, e.g. only the parts of the image seen by the model up to that layer.

## 2 Related Work

**Concepts.** The idea that certain directions in a model’s latent representation align with human-understandable concepts was initially proposed by Kim et al. [15]. They propose to learn a hyperplane in the activation space of a neural network layer that separates images, which include the concept, from other images. The normal of the hyperplane in the direction of the images encoding the concept is the Concept Activation Vector (CAV). Since then, a lot of effort was put into the automatic discovery of such concepts activation vectors [9, 10, 22, 32, 35]. Interesting for our work is the novel concept discovery algorithm proposed by Yeh et al. [32], which combines interpretability with a new notion of *completeness* which measures how sufficient a set of CAVs is for the explanation of a model’s prediction behavior. They also introduce a method to rank the found CAVs by importance called ConceptSHAP which adapts Shapley values [1]. Shapley values assign importance to a feature by calculating its average contribution in all possible combinations. One drawback of approaches for automatic CAV discovery is that they rely on images as references for the explanation of a CAV.

**Network Dissection.** The idea of dissecting a network is to inspect the function of individual neurons in the network to get insights into the model. The first work about network dissection provided a method to quantify the interpretability of latent representations by comparing neuron activations with segmentation masks from a concept dataset [3]. This approach aligns individual neuron activations of a model with specific visual concepts given by the segmentation masks. One major limitation of this approach is, that the masks needed to be annotated by humans. Based on this, a segmentation model was proposed in [4] to annotate the masks for each concept. MILAN [12] extends the labeling of neurons to open-ended natural language descriptions: This approach generates descriptions of neurons by finding language strings that maximize the mutual information of the image regions where the neuron is active. To generate the language description, an image-to-text model is required, trained on a labeled data set. To avoid the need for labeled data, CLIP-Dissect [23] leverages the multimodal training of CLIP [25], a method that embeds image and text data to a joint feature space.

**Joint Text-Image Embeddings.** In recent years, there have been significant advancements in learning joint text-image embeddings [13, 19, 25, 34]. Text-image embeddings can be utilized to perform various tasks, such as zero-shot classification. Contrastive learning based approaches, such as CLIP [25], are trained to maximize the similarity between positive examples (e.g., images and matching image captions) and to minimize the similarity to negative examples (e.g., non-matching image-caption pairs). Approaches such as CLIP have shown good zero-shot image classification performance on multiple data sets by evaluating the similarity between the feature embeddings of the class labels and the images.

**Post-Hoc Concept-Bottleneck Models.** An alternative approach to generating post-hoc concept explanations is to first create a set of known CAVs and then find the subset of those CAVs that yield the best performance for a given model [21, 33]. Those approaches assume to have CAVs for all important concepts and then select the CAVs that can describe the essence of what was learned by the model. In our approach, the set of CAVs is discovered automatically by inspecting the model in more detail like in [32], and then designated by textual descriptions.

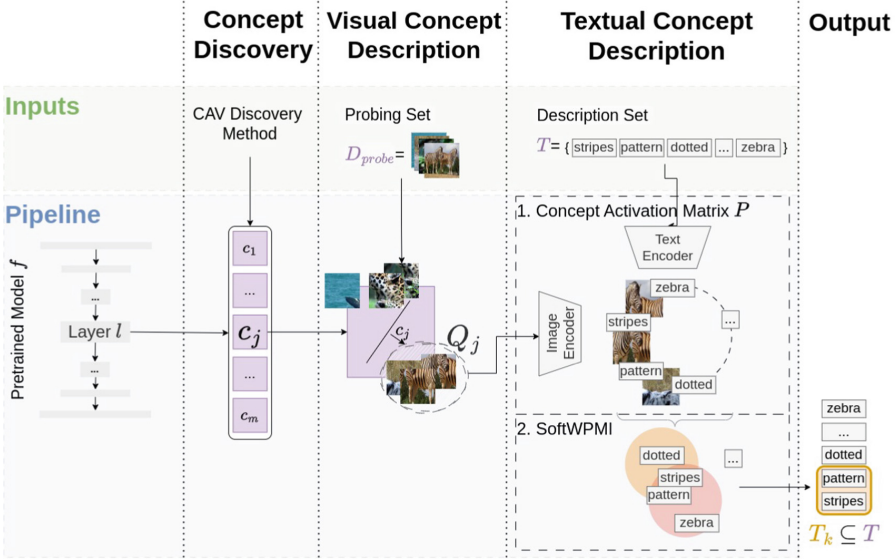
### 3 Method

We propose a method that derives textual descriptions for the concepts a neural network utilizes to solve an image classification task. The method consists of three steps, and each step represents a different level of concept description for a given neural network:

1. The **discovery of concepts** by concept activation vectors (CAVs), represented as directions in the feature space,
2. the **visual description** of the concepts (encoded by the CAVs) with representative images,
3. and the **textual description** of the concepts with words.

The steps are visualized in Fig. 2. In the following, the inputs, the three steps of the method, and the computed outputs are introduced in more detail.

**Inputs.** The method is based on a neural network trained on an image classification task,  $f$ , that maps input images to a  $K$ -dimensional output vector representing class probabilities. For a given layer  $l$ , for which concepts shall be extracted from the network, the network is decomposed into two functions  $h_l$  and  $\phi_l$ , such that  $f = h_l \circ \phi_l$ . Further, let  $\mathcal{D}_{probe} = \{x_1, \dots, x_n\}$  be a probing set, i.e., a set of  $n$  images that can be used for the visual description of the extracted CAVs. The textual descriptions of the concepts are based on a predefined and task dependent set of words  $T$ . For example,  $T$  can contain describing attributes [3], or the top 20.000 words of the English language [14].



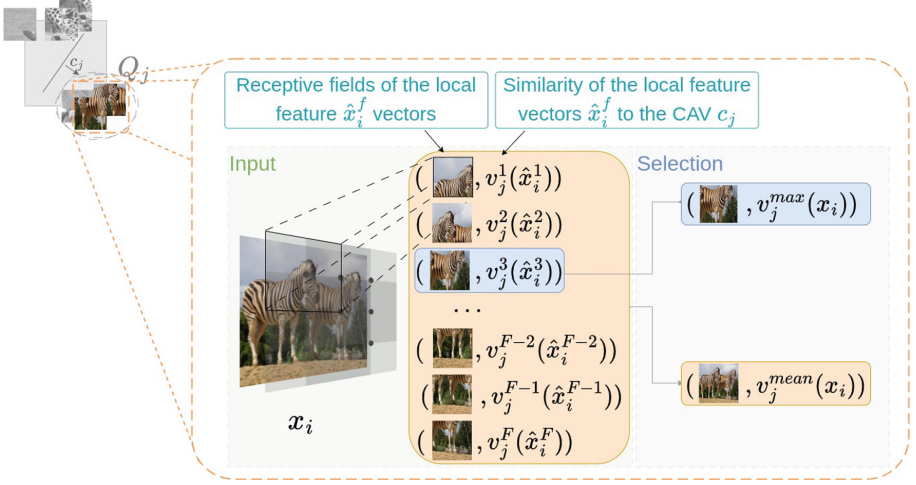
**Fig. 2.** Overview of our approach to describe the layer  $l$  of a pretrained model  $f$ . The *inputs* are a concept discovery method, a probing set  $D_{probe}$ , and a set of textual descriptions  $T$ . We apply *concept discovery* methods to find a set of CAVs, generate a set of *visual concept descriptions*  $Q_j$  for each CAV  $c_j$ , then *textual concept descriptions* and finally *output* the top- $k$  descriptions  $T_k \subseteq T$ .

**Concept Discovery.** We describe the embedding of layer  $l$  with Concept Activation Vectors (CAVs). A CAV is a vector that points in the direction of a concept learned by the model and is embedded in the feature space of the activations of layer  $l$ . The concepts learned at layer  $l$  are then represented by a set of  $m$  CAVs,  $C_l = \{c_{1,l}, \dots, c_{m,l}\}$ . We drop the index  $l$  in the following when considering only one specific layer. While the proposed method is independent of the underlying concept extraction approach, we follow the approach of Yeh et al. [32] to derive all concepts utilized for a given image classification task.

**Visual Concept Description.** For the visual description of a given CAV  $c_j$ , we follow the former work [32] to derive a set  $Q_j \subset D_{probe}$  of most relevant images from the probing set. This approach is illustrated in Fig. 3 and will be described in the following. The relevance of an image  $x_i \in D_{probe}$  is determined based on the similarity between the CAV  $c_j \in \mathbb{R}^k$  and its latent representation at layer  $l$ . In detail, consider the latent representation of an image  $x_i$  at layer  $l$ , which is

$$\phi_l(x_i) =: (\hat{x}_{i,l}^1, \dots, \hat{x}_{i,l}^F) \in \mathbb{R}^{F \times k}.$$

The vectors  $\hat{x}_{i,l}^1, \dots, \hat{x}_{i,l}^F$  are called *local feature vectors* of  $x_i$  and correspond to the activations of each channel of the convolutional neural network after layer  $l$ . We will omit the index  $l$  when the connection to the specific layer is clear.



**Fig. 3.** Selection of the visual representations for a given CAV  $c_j$ , compare with Fig. 2 column *Visual Concept Description*. The vector  $(v_j^1(\hat{x}_i^1), \dots, v_j^F(\hat{x}_i^F))$  represents the concept scores between each receptive field of  $x_i$  and the CAV  $c_j$ . While [23] select full images based on the mean score of all receptive fields, we also consider the receptive field with the highest concept score. Thus, we improve the visual input of the joint vision-text embedding by cropping  $x_i$  to the respective receptive field. This creates a more truthful and more detailed representation of the concepts learned in the hidden space.

For each local feature vector  $\hat{x}_i^f$  and each CAV  $c_j$  a *concept score*, which measures the similarity based on the scalar product, i.e.

$$v_j^f(\hat{x}_i^f) := \hat{x}_i^{fT} c_j.$$

This leads to a vector  $v_j(x_i) \in \mathbb{R}^F$  of  $F$  concept scores,

$$v_j(x_i) = (v_j^1(\hat{x}_i^1), \dots, v_j^F(\hat{x}_i^F)) \in \mathbb{R}^F. \quad (1)$$

Following [32], a larger concept score means a higher similarity of the corresponding receptive field of  $\hat{x}_i^f$  to the concept encoded by the CAV  $c_j$ .

While  $v_j(x_i)$  is a vector of similarities, the set of relevant images  $Q_j$  is chosen based on scalar values because they can be ordered. Former works such as [23] select full images of  $\mathcal{D}_{\text{probe}}$  for the set  $Q_j$ . To achieve this, they consider the average over the individual concept scores of the local feature vectors,

$$v_j^{\text{mean}}(x_i) = \frac{1}{F} \sum_{f=1}^F v_j^f(\hat{x}_i^f) \in \mathbb{R}. \quad (2)$$

As we are more interested in the most representative part of an image for a concept, we consider the maximum concept score of all local feature vectors:

$$v_j^{\max}(x_i) = \max_{f \in \{1, \dots, F\}} v_j^f(\hat{x}_i^f) \in \mathbb{R}. \quad (3)$$

Based on these two metrics, we introduce three different strategies to derive a set of most relevant images  $Q_j$  from  $\mathcal{D}_{\text{probe}}$ . Note that the subset  $Q_j$  can either contain the full image  $x_i$  or a receptive field associated with a local feature vector  $\hat{x}_i^f$ . We follow [22] and select the 100 most relevant images.

- $F_{\text{mean}}$  : Select the images with the highest  $v_j^{\text{mean}}(x_i)$ .
- $F_{\text{max}}$  : Consider those images with the highest  $v_j^{\max}(x_i)$  and choose the respective receptive fields where the maximum is reached.
- $F_{\text{mean} \rightarrow \text{max}}$  : Select images like  $F_{\text{mean}}$  but choose the receptive field with highest concept score  $v_j^f(\hat{x}_i^f)$ .

We search for the parts of the images with the highest presence of the concept encoded by the CAV. With  $F_{\text{mean}}$  we select the full images with the highest overall presence of the concept. As a result, the textual descriptions are calculated based on the full images. However, often the model can only see parts of the images at the layer where the CAVs were found. Due to this, and the fact that concepts may be more present in single parts of an image, we apply strategies to find the relevant receptive fields. Using  $F_{\text{max}}$  we select the receptive field of each image with the highest concept score. We propose  $F_{\text{mean} \rightarrow \text{max}}$  to combine the advantages of both strategies. This means that we find the images where the concept is highly present in the full image and reduce the noise introduced by other concepts by selecting the respective receptive field with the highest concept score.

**Textual Concept Description.** To derive a textual description for the visual descriptions collected in  $Q_j$ , we utilize joint text-image embeddings and corresponding image and text encoders  $E_{\mathcal{I}}$  and  $E_{\mathcal{T}}$  which map from the space of images,  $\mathcal{I}$ , and the space of texts,  $\mathcal{T}$ , respectively, to a joint feature space. This is, for example, provided by the CLIP model [25]. We compute a similarity matrix  $P$  based on the cosine similarity of the text and image embeddings of the textual descriptions set  $T = \{t_1, \dots, t_s\}$  and images in  $Q_j$ ,

$$P_{ij} = \frac{E_{\mathcal{I}}(x_i)^T E_{\mathcal{T}}(t_j)}{\|E_{\mathcal{I}}(x_i)\|_2 \|E_{\mathcal{T}}(t_j)\|_2}.$$

Intuitively, we want to find the textual descriptions that have a high similarity to all images in  $Q_j$ . To do this, we utilize the *Soft Weighted Pointwise Mutual Information* (SoftWPMI) [23], which indicates how well a word describes the mutual information of the representative images. SoftWPMI requires a weighting of the images in  $Q_j$ , which is determined by the concept scores. In particular, this vector  $q_j$  is calculated depending on the strategy to derive the set of most

relevant images  $Q_j$ :

$$q_j = \begin{cases} (v_j^{mean}(x_i))_{x_i \in Q_j} & \text{if } F_{mean} \\ (v_j^{max}(x_i))_{x_i \in Q_j} & \text{if } F_{max}, F_{mean \rightarrow max} \end{cases} \quad (4)$$

Finally, we find the subset  $T_k$  with the top- $k$  textual descriptions by:

$$T_k := \arg \max_{\hat{T} \subset T: |\hat{T}|=k} \sum_{t \in \hat{T}} \text{SoftWPMI}(t, q_j, P) \quad (5)$$

Note that, in practice,  $\text{SoftWPMI}(t, q_j, P)$  is computed for each  $t \in T$  separately, and finally, we take the top- $k$  textual descriptions. For the common textual description, we compute the weighted average of the top- $k$  descriptions in the feature space, with the weighting based on the SoftWPMI values. The common representation is then chosen as the textual description in  $T$  that is closest to this weighted average. Please note that we set all negative SoftWPMI values in  $\hat{T}$  to zero since we are only interested in positive similarities.

**Output.** The method returns the common description and the subset  $T_k$  from the human understandable textual descriptions set  $T$ , which are most similar to the concept represented by the CAV  $c_j$ .

## 4 Experiments

Our experimental procedure consists of three stages. First, we utilize CAVs with known concept labels to show that our approach is capable to yield meaningful textual explanations of CAVs. Second, we compare the different mappings  $F_{mean}$ ,  $F_{max}$ ,  $F_{mean \rightarrow max}$  for the generation of the set of best fitting textual descriptions. And finally, we consider a more complex scenario and explain a set of CAVs extracted from a model where we have no prior knowledge about the underlying concepts.

### 4.1 Explaining a Set of CAVs with Known Concept Labels

To be able to validate general idea of our approach, we follow Kim et al. [15] and design a set of CAVs where each CAV describes one class of a given data set. We achieve this by generating a set of CAVs after the last convolutional layer of a model and set the number of CAVs equal to the number of classes. It is important to note that the suggested strategy is closely related to the performance of the CLIP model. Hence, a bad classification performance of CLIP directly affects our approach in a negative way.

**Table 1.** Each row shows the Top-5 textual descriptions of a CAV computed with the proposed approach (ranked from left to right) and the derived common concept description. Each CAV is supposed to represent one class of the CIFAR10 dataset [17]. Imagenet is utilized [7] as  $\mathcal{D}_{\text{probe}}$  and google20k as the set of textual descriptions,  $T$ .

CAV-Label	Common Description	1	2	3	4	5
airplane	aircraft	aircraft	aviation	plane	airplanes	planes
automobile	vehicle	vehicle	vehicles	car	ambulance	automobile
bird	bird	avian	bird	birding	birds	juvenile
cat	cat	cat	kitts	kitty	kitten	katz
deer	deer	grazing	gnu	deer	female	wildlife
dog	dog	puppy	dog	canine	pundit	dug
frog	mating	mating	meal	head	emerging	frog
horse	horse	equine	horseback	horse	horses	equestrian
ship	sailing	sailing	yacht	sail	yachts	sailors
truck	trucks	truck	trailer	trucks	trailers	movers

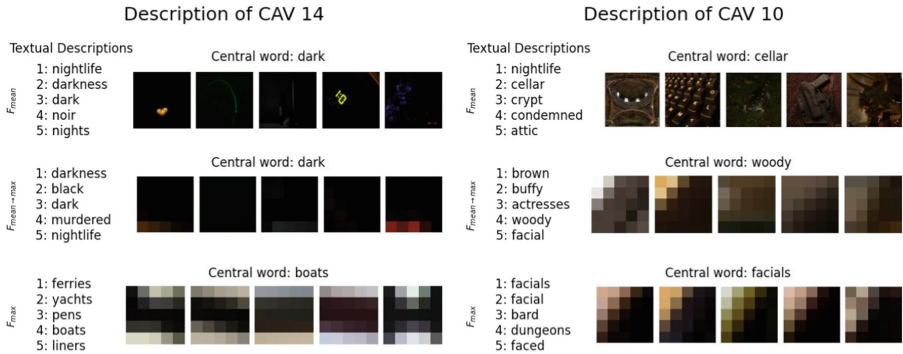
**Setup.** To make sure that the CLIP model itself performs well in this validation example, we use the datasets CIFAR10 [17] and MNIST [8] which have a zero-shot performance of 96.2% and 87.2%, with the vision encoder CLIP-ViT L/14 from CLIP [25]. For CIFAR10 we adapted a pre-trained ResNet50 [11] and finetuned it. The finetuned ResNet50 reaches an accuracy of 0.94. For MNIST we finetuned a simple ConvNet with 3 layers reaching an accuracy of 0.98. In this experiments we explain the the embedding after the last convolutional layer of the models (ResNet50 and ConvNet). For the set of textual descriptions we use google20k [14]. Details to the MNIST experiments can be found in the appendix.

**Results.** The results of this experiment for CIFAR10 are displayed in Table 1 (The table for MNIST can be found in the Appendix). The top-5 words, as well as the concept closest to the centroid for each class, are shown. Our approach is able to match each CAV which encodes a class as concept with fitting textual descriptions from the 20.000 textual suggestions given. The exception is the CAV encoding *Frog*. For MNIST our approach finds fitting textual descriptions for all classes except the CAV encoding “one” which is described by *makefile*.

## 4.2 Concept Discovery and Description

Compared to the class-wise concepts in the previous sections, automatically discovered CAVs usually describe more abstract concepts as colors and shapes. We utilize the approach of [32] to discover a set of CAVs automatically. The final set of CAVs is selected based on a hyper parameter search and the test accuracy of the classification task. The hyper parameter search includes the number of





(a) Most influential CAV for the class “cat” (b) Second most influential CAV for the class “cat”

**Fig. 4.** Comparison of the approaches to generate textual descriptions. Shown are the two most influential CAVs for the class “cat” after the first residual block of a ConvMixer [30]. The model was trained on *dark* cats and *light* dogs, a subset of the Cats vs. Dogs dataset [6]. The first approach uses the images with the highest mean activation for the CAV, the second takes the highest receptive fields of the images with the highest mean activation and the third takes the most activated receptive fields of all receptive fields over the whole probing data set. The probing dataset is the validation set from ImageNet [7] and the concept set is google20k [14]

concepts, the threshold value  $\beta$ , and scalars  $\lambda_1 > 0$  and  $\lambda_2 > 0$ . The parameters  $\lambda_1$  and  $\lambda_2$  are needed for the utilized concept discovery approach of [32]. They weight the similarity between the concepts and their most relevant images ( $\lambda_1$ ) and the pairwise dissimilarity between the concepts ( $\lambda_2$ ). Further, we calculated for each class the ConceptSHAP and explanation quality following [32]. The ConceptSHAP gives us an importance value for each CAV with respect to the class. The explanation quality serves as a measure how well a class is described by the set of CAVs discovered. In the following, we first compare the different approaches to select the relevant images, i.e., the receptive field-based approaches and the full image approaches.

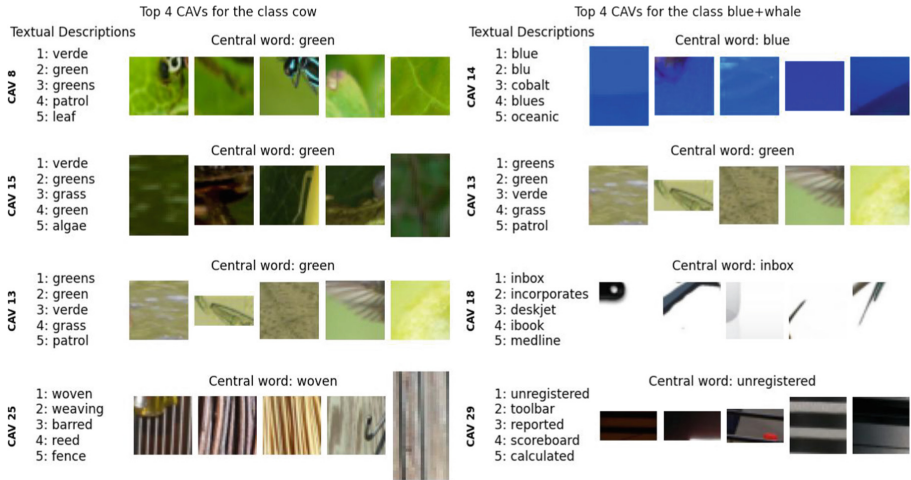
**Evaluation of Image Set Selection.** We consider concepts extracted from early layers, where concepts are assumed to be more abstract than in later layers. With this we can also evaluate the effect of  $F_{max}$  and  $F_{mean \rightarrow max}$  on highly correlated concepts and concepts of different degree of abstraction. We further introduce the abstract concept *dark* into the model by performing a classification of cat and dog images, where the training samples consist of dark cats and the bright dogs. We expect the trained model to mainly rely on those features due to the simplicity bias of neural networks [29].

**Setup.** We trained our model on a modified Cats vs. Dogs (CvD) dataset [6]. The Cats vs Dogs dataset was developed by Kaggle [6] and, following [16, 18], we split it by color, such that it consists of *dark* cats and *light* dogs. We call this dataset Dark Cats vs. Dogs (DCvD). In the following we refer to the original and the modified dataset as unbiased and biased dataset. Since all cats are *dark* and all dogs are *light* we make the assumption, that the color is a relevant concept for models trained on this dataset. To validate this we train a ConvMixer [30] with a depth of seven. The ConvMixer reaches an accuracy of 0.93 on the biased data and only an accuracy of 0.69 on the unbiased data (details in the appendix). This difference in accuracy indicates that the model learned to associate the color *black* with cats. We extract the set of CAVs after the first residual block of the model. The derived set consists of 20 CAVs and the classification based on the active and inactive CAVs yields an accuracy of 0.96 on the biased data. The hyper parameters used to learn the set are  $\lambda_1 = 0.2$ ,  $\lambda_2 = 0.2$  and  $\beta = 0.18$ . After we filter the CAVs where the dot product is over 0.95 we are left with 15 relevant CAVs. As the set of textual descriptions, google20k is used.

**Results.** Figure 4 shows the two most important CAV from left to right for the class cat. The CAVs are selected by the ConceptSHAP values. For each CAV we display the three approaches to select relevant images based on the concept scores. For each approach the textual descriptions and the top five images from the set of most relevant images are shown. It can be seen for CAV 14 that the approach  $F_{mean}$  returns as highest textual description *nightlife* and  $F_{max}$  returns *ferries* (See Fig. 4a). Only  $F_{mean \rightarrow max}$  returns a fitting highest textual description with *darkness*. Looking at the other descriptions we see that  $F_{mean}$  also yields similar textual descriptions in the top 5 descriptions. This results in the central word of  $F_{mean}$  and  $F_{mean \rightarrow max}$ , matching our expectations. For the CAV 10 we can see that all approaches return different textual descriptions (See Fig. 4b).  $F_{mean}$  returns *nightlife* and  $F_{max}$  returns *facials* which are both complex concepts. The approaches recognize different concepts which are relevant for the images. This is neither good nor bad. Only  $F_{mean \rightarrow max}$  returns a simple concept with *brown*.

**Animals with Attributes.** The objective of this experiment is to explore the performance of our approach for scenarios with increased complexity and to show its potential. The experiment is based on the Animals with Attributes2 dataset [31], which contains 37322 images from 50 different animals.

**Setup.** We finetuned a ResNet50 on the dataset AWA2 [31] that reaches a test accuracy of 0.9. The concept discovery method found a set of 30 CAVs after the first residual block. The found set of CAVs achieves an accuracy of 0.87 with the hyper parameter  $\lambda_1 = 3.1$ ,  $\lambda_2 = 3.1$  and  $\beta = 0.02$ . After filtering all duplicates 15 CAVs are left, describing the concepts learned by the first residual block.



(a) Description of the best represented class: "cow" (b) Description of the worst represented class: "blue whale"

**Fig. 5.** For each class the textual descriptions and the most activated receptive fields of the CAVs with the strongest influence are shown. The image set was selected by  $F_{mean \rightarrow max}$ . The set of CAVs describes the hidden representation after the first residual block of a ResNet50 finetuned on AwA2. The probing dataset is the validation set from ImageNet and the concept set is google20k.

**Results.** The results of this experiment can be seen in Fig. 5. Here, Fig. 5a shows the class which is best described by the set of CAVs and Fig. 5b shows the class which is worst described by the set of CAVs. Further, for each class the most influential CAVs ranked by ConceptSHAP are displayed. The descriptions are generated with the  $F_{mean \rightarrow max}$  approach. It can be observed that the model strongly connects the concept *green* with the class "cow" (See Fig. 5a). The class "blue whale" is connected to the concept *blue* (See Fig. 5b). When inspecting the descriptions of the CAVs 18 and 29 a mismatch becomes apparent. The descriptions for those CAVs seem to be hardly related and are not matching to the receptive fields.

## 5 Discussion

The experiments on the sets of CAVs with the known concept label show that the approach is capable of matching CAVs with the corresponding textual descriptions from a large set of general descriptions. This underlines that our approach is in general capable of identifying joint textual descriptions, even though the performance highly depends on the quality of the utilized joint text-image features space. For the experiment on CIFAR10, one can further see the redundancy in

the best-fitting descriptions which is successfully removed by selecting a common concept description (Table 1). Further, one can see the approach’s capabilities to detect biases in the training and/or probing images, e.g., the top five descriptions of the class *ship* are all related to sailing.

For the different approaches to select representative images for given CAVs, the ones using receptive fields help to correctly describe more abstract concepts that especially occur in earlier layers of a neural network (Fig. 4). Interestingly, 15 CAVs are detected as relevant, which is more than to separate the concepts of dark and bright. This can be explained by the fact, that dark and bright colors can also occur in the backgrounds of the images and hence the distinction purely based on color concepts is not feasible. However, the relevance of the *dark* concept shows that it is highly relevant to classify cats. The increased focus on abstract concepts when utilizing the receptive fields can also be explained by the nature of the CLIP model. CLIP was trained on images and corresponding captions, where specific colors (e.g., *green*) might be less relevant than the overall image description (e.g., *insect*). In Fig. 5b, the CAVs 18 and 29, which are relevant for the class “blue whale”, are examples where the approaches fail to generate matching textual descriptions. This can be attributed to limitations in the utilized CLIP model. For example, CAV 18 seems to show the concept *white* but the textual descriptions are *inbox*, *incorporate*, . . . This could be improved by applying a more fine-grained selection of the inputs for the joint text-image model or by utilizing other text-image feature spaces.

## 6 Conclusion

In this work, we proposed an approach to assist the interpretation of CAVs by suggesting textual descriptions and selecting common words for the individual CAVs. To improve the textual descriptions of CAVs found for earlier layers, we consider that for earlier layers of a model, the CAVs do not know the whole input and propose to use receptive fields for the generation of the textual descriptions. Through experiments on sets of CAVs where the underlying concepts are known, we showed that our method is capable of yielding meaningful descriptions for CAVs and that the usage of receptive fields improves the explanation quality for earlier layers. While this research already offers insights into the concept discovery process, further works on the computation of meaningful concepts as well as an exploration of other image-to-text projections are planned. The evaluation of the found textual descriptions regarding human understanding is also a topic for further research. To better understand the behaviour of the model, it would be interesting to extend the results of concept discovery methods with mismatched data. For the description of specific concepts, further insights into the capabilities of joint text-image feature spaces and the needed characteristics of probing sets are interesting for us, as well as the consideration of explicitly fine-tuning text-image embeddings to basic concepts.

**Acknowledgements.** We thank Niklas Penzel for preparing the Dark Cats vs. Dogs (DCvD) dataset and training the corresponding model.

## References

1. The Shapley Value: Essays in Honor of Lloyd S. Cambridge University Press, Shapley (1988)
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**(7), e0130140 (2015). <https://doi.org/10.1371/journal.pone.0130140>
3. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549 (2017)
4. Bau, D., Zhu, J.Y., Strobel, H., Lapedriza, A., Zhou, B., Torralba, A.: Understanding the role of individual units in a deep neural network. *Proc. Natl. Acad. Sci.* **117**(48), 30071–30078 (2020)
5. Chen, Z., Bei, Y., Rudin, C.: Concept whitening for interpretable image recognition. *Nat. Mach. Intell.* **2**(12), 772–782 (2020)
6. Cukierski, W.: Dogs vs. cats (2013). <https://kaggle.com/competitions/dogs-vs-cats>
7. 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 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
8. Deng, L.: The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Process. Mag.* **29**(6), 141–142 (2012)
9. Fel, T., et al.: CRAFT: concept recursive activation factorization for explainability. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721 (2023)
10. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. In: *Advances in Neural Information Processing Systems*, vol. 32 (2019)
11. 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)
12. Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., Andreas, J.: Natural language descriptions of deep visual features. In: *International Conference on Learning Representations* (2021)
13. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: *International Conference on Machine Learning*, pp. 4904–4916. PMLR (2021)
14. Kaufman, J.: google-10000-English: A list of the 10,000 most common English words. <https://github.com/first20hours/google-10000-english> (nd). Accessed 15 Mar 2024
15. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: *International Conference on Machine Learning*, pp. 2668–2677. PMLR (2018)

16. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: training deep neural networks with biased data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9012–9020 (2019)
17. Krizhevsky, A., Hinton, G., et al.: Learning Multiple Layers of Features from Tiny Images. ON, Canada, Toronto (2009)
18. Lakkaraju, H., Kamar, E., Caruana, R., Horvitz, E.: Discovering blind spots of predictive models: Representations and policies for guided exploration. arXiv preprint [arXiv:1610.09064](https://arxiv.org/abs/1610.09064) (2016)
19. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: A simple and performant baseline for vision and language. arXiv preprint [arXiv:1908.03557](https://arxiv.org/abs/1908.03557) (2019)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
21. Moayeri, M., Rezaei, K., Sanjabi, M., Feizi, S.: Text-to-concept (and back) via cross-model alignment. In: International Conference on Machine Learning, pp. 25037–25060. PMLR (2023)
22. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2022)
23. Oikarinen, T., Weng, T.W.: CLIP-dissect: automatic description of neuron representations in deep vision networks. In: The Eleventh International Conference on Learning Representations (2022)
24. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill (2017). <https://doi.org/10.23915/distill.00007>, <https://distill.pub/2017/feature-visualization>
25. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
26. Reimers, C., Runge, J., Denzler, J.: Determining the relevance of features for deep neural networks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12371, pp. 330–346. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-58574-7\\_20](https://doi.org/10.1007/978-3-030-58574-7_20)
27. 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)
28. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Why did you say that? arXiv preprint [arXiv:1611.07450](https://arxiv.org/abs/1611.07450) (2016)
29. Shah, H., Tamuly, K., Raghunathan, A., Jain, P., Netrapalli, P.: The pitfalls of simplicity bias in neural networks. Adv. Neural. Inf. Process. Syst. **33**, 9573–9585 (2020)
30. Trockman, A., Kolter, J.Z.: Patches are all you need? Transactions on Machine Learning Research (2023)
31. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE Trans. Pattern Anal. Mach. Intell. **41**(9), 2251–2265 (2018)
32. Yeh, C.K., Kim, B., Arik, S., Li, C.L., Pfister, T., Ravikumar, P.: On completeness-aware concept-based explanations in deep neural networks. Adv. Neural. Inf. Process. Syst. **33**, 20554–20565 (2020)
33. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2022)

34. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11975–11986 (2023)
35. Zhang, R., Madumal, P., Miller, T., Ehinger, K.A., Rubinstein, B.I.: Invertible concept-based explanations for CNN models with non-negative concept activation vectors. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11682–11690 (2021)



# Back to Supervision: Boosting Word Boundary Detection Through Frame Classification

Simone Carnemolla<sup>(✉)</sup>, Salvatore Calcagno, Simone Palazzo,  
and Daniela Giordano

Department of Electrical Electronic and Computer Engineering,  
University of Catania, Via Santa Sofia, 95123 Catania, Italy  
`simone.carnemolla@phd.unict.it`

**Abstract.** Speech segmentation at both word and phoneme levels is crucial for various speech processing tasks. It significantly aids in extracting meaningful units from an utterance, thus enabling the generation of discrete elements. In this work we propose a model-agnostic framework to perform word boundary detection in a supervised manner also employing a labels augmentation technique and an output-frame selection strategy. We trained and tested on the Buckeye dataset and only tested on TIMIT one, using state-of-the-art encoder models, including pre-trained solutions (Wav2Vec 2.0 and HuBERT), as well as convolutional and convolutional recurrent networks. Our method, with the HuBERT encoder, surpasses the performance of other state-of-the-art architectures, whether trained in supervised or self-supervised settings on the same datasets. Specifically, we achieved F-values of 0.8427 on the Buckeye dataset and 0.7436 on the TIMIT dataset, along with R-values of 0.8489 and 0.7807, respectively. These results establish a new state-of-the-art for both datasets. Beyond the immediate task, our approach offers a robust and efficient preprocessing method for future research in audio tokenization.

**Keywords:** Word Boundary Detection · Word Segmentation · Speech Processing

## 1 Introduction

Speech segmentation, from a psychological perspective, is the process by which our brain determines where a meaningful linguistic unit ends and the next begins in continuous speech [28].

In machine learning before and in deep learning nowadays, this capability is not easily achievable due to the dense information that the audio data conveys and the prosodic features that each speaker has. Furthermore, building specialized

---

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/978-3-031-80136-5\\_9](https://doi.org/10.1007/978-3-031-80136-5_9).



datasets, with high-quality recordings and well-annotated data, requires considerable effort. More specifically, in speech processing, word and phoneme boundary detection are intended both as a preprocessing phase for other downstream tasks such as speaker diarization, keyword spotting, or automatic speech recognition, and as a tool to extract semantically meaningful information from audio.

A correct boundary detection of words within a utterance would lead to a more accurate study of the prosody or emotional traits of speakers on a large scale, without necessary resorting to textual data alignment, that by their nature, lack important para-verbal information. Additionally, it would facilitate the discretization or tokenization of the elements that make up an utterance, a complex process given the variable length of speech units.

Due to the significant impact we believe this task may have on various downstream audio applications, we have opted for a supervised approach to maximize its performance, diverging from the self-supervised trend of the recent researches.

In this paper, we introduce a model-agnostic framework for word boundary detection. Our methodology integrates frame classification based on the BIO (begin, inside, outside) format with a label augmentation technique - to address the imbalance between *begin* and *inside/outside* frames - and a frame-selection strategy for post-processing. We trained various state-of-the-art models on the Buckeye dataset [30], a widely recognized benchmark for this task. To further validate and check the generalization capability of our method we also tested it on TIMIT dataset [15].

Our results indicate that our method set a new state-of-the-art on both the Buckeye and TIMIT datasets, achieving F-values of 0.8427/0.7436, and R-values of 0.8489/0.7807, respectively. The code is publicly available on Github<sup>1</sup>.

## 2 Related Work

Unlike phoneme boundary detection, which has a rich literature on supervised [12, 22, 24, 25], self-supervised [23, 34, 38], and unsupervised [3, 7, 11, 26] methods, the task of word boundary detection has been approached mainly from a self-supervised or unsupervised perspective.

Within supervised learning, most research focused its attention on probabilistic approaches and on the extraction of acoustic features making the preprocessing phase often long and complex. This is the case of Agarwal et al. [1] and Naganoor et al. [27]. The latter proposed a method that extracts rudimentary acoustic features and higher-order statistical features (HOS). The same work inspired Shezi et al. [33] for a word boundary detection task in IsiZulu language. In contrast, our approach focuses on utilizing raw audio data, sidestepping the need for extensive preprocessing steps.

Other methods such as [25] use word boundaries as a speech-text alignment system. However, our method takes a different route. We deliberately steer clear of relying on text label data, which sets our approach apart. This deliberate choice provides a significant advantage: our evaluation of performance is solely based on speech output. This independence from text data is a crucial feature

<sup>1</sup> <https://github.com/simonecarnemolla/Word-Segmenter>.

of our method, rendering it particularly advantageous in situations where text data is scarce or unavailable. This characteristic ensures the robustness and applicability of our approach across various speech processing tasks, even in challenging data environments.

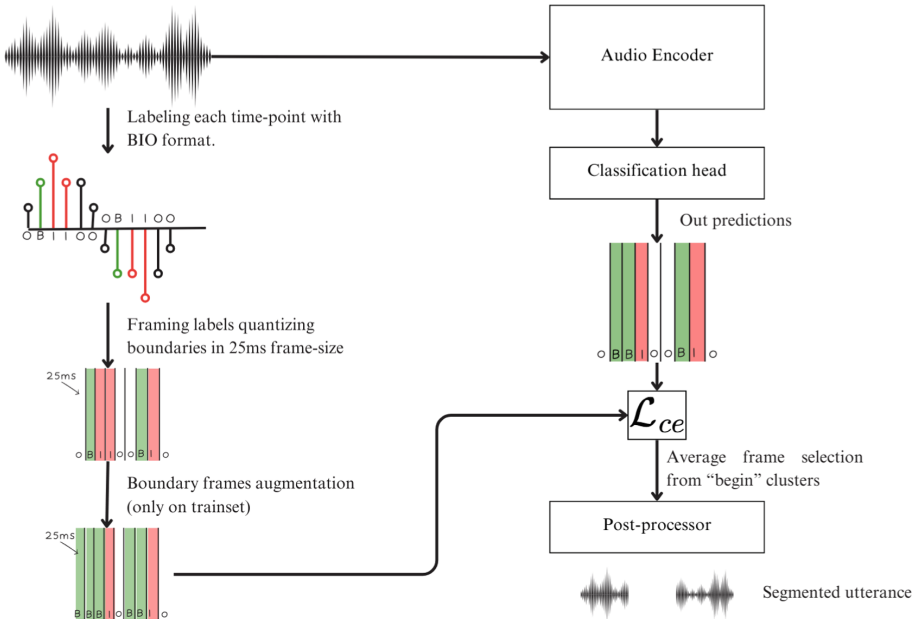


Fig. 1. Overview of our method

Moving to self-supervised and unsupervised fields, some of the main contributions come from the works of Kamper et al. [18–20]. In both [18,19], a Bayesian model that segments unlabeled speech and clusters the segments into hypothesized word groupings is proposed. In a subsequent work [20], an embedded segmental KMeans (ES-KMeans) model is proposed to solve the Bayesian model’s difficulty in scaling large speech corpora. Among the most interesting recent research advances are word boundary detection with vector quantized (VQ) neural networks [21], a segmental contrastive predictive coding (SCPC) approach [8,9], and the use of temporal gradients as pseudo-labels to find boundaries [14]. In the first work [21], the VQ neural networks - a vector-quantized variational autoencoder (VQ-VAE) [10] and a vector-quantized contrastive predictive coding (VQ-CPC) [5] - are trained in a self-supervised way, segmenting speech into discrete units, assigning blocks of contiguous feature vectors to the same code. Then, dynamic programming (DP) is used to merge frames and to optimize a quadratic error with a length penalty term to encourage fewer but longer segments. Bhati et al. [8,9] proposed a model that initially extracts frame-level representations and then identifies variable-length segments using

a differentiable boundary detector. Finally, Fuchs et al. [14] extracted temporal gradients and observed that gradients with low magnitude effectively identify far-from-boundary regions. Building on this observation, they proposed GradSeg, a method where frames with gradient magnitudes below a preset threshold are assigned a positive label (indicating far-from-boundary words). This approach outperformed other unsupervised methods. Furthermore, Fuchs et al. trained with a supervised approach to compare the impact of supervision on the results. We employed their supervised results as a state-of-the-art benchmark, as we surprisingly found no other recent methods to compare with. We also reproduced the experiments using our data distribution with their unsupervised method and their earlier work [13], reporting the scores. This was done not to compare their results with ours, but to validate the assumption that a supervised method can significantly enhance the applicability of word segmentation.

### 3 Method

#### 3.1 Overview

An overview of our method is shown in Fig. 1. Initially, the raw audio is labeled using a BIO format. Following framing, we apply a label augmentation technique to better handle the significant imbalance between beginning and inside/outside indices. An encoder architecture is then trained in a supervised setting for the frame classification task. Finally, a frame selection strategy is employed to post-process predictions and segment the input utterance. Each component of the method is described in detail in the following subsections.

#### 3.2 Problem Formulation

Let  $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,T})$  be a single utterance where  $x_t$  represents the amplitude of the signal at time  $t$ , with  $0 \leq t \leq T$ , considering a sampling frequency  $s$ . Let also denote  $\mathbf{y}_i = (y_{i,1}, y_{i,2}, \dots, y_{i,m})$  the corresponding sequence of framed labels. The generic value  $y_{i,j}$  refers to a frame within  $\mathbf{x}_i$ . All frames have the same fixed time-length (more details in Sect. 3.3). In particular,  $y_{i,j} \in \{0, 1, 2\}$  indicates if the corresponding frame in  $\mathbf{x}_i$  marks the start, is positioned inside, or lies outside a single word. Our objective is to accurately detect the word boundaries within the utterance  $\mathbf{x}_i$  by predicting the correct sequence of framed labels  $\mathbf{y}_i$ . Once the boundary frames have been identified, we aim to identify the exact time-point in the audio signal where words begin.

We adopted a multi-class cross-entropy as a loss function, defined as follows:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^m y_{i,j} \cdot \log(\hat{y}_{i,j}), \quad (1)$$

where  $N$  is the number of instances in the batch,  $y_{i,j}$  represents the label of the  $j$ -th frame of the  $i$ -th dataset element, and  $\hat{y}_{i,j}$  represents the model's prediction for the same frame.

### 3.3 Labels Processing

As showed in Fig. 1, we extracted all the utterances and the start-end boundaries of each word per utterance. We maintained the sample rate of the audio files at 16KHz without resampling. Subsequently, we created pre-labels with the same shape of the input waveforms, assigning to each time-point a value depending on if it was positioned inside or outside the boundaries or if it was a start index. This step was useful to investigate the distribution and the average duration of the words within the utterances. The final labels were obtained, framing the pre-labels along the temporal axis with a 25 ms frame duration. The total number of frames  $m$  in  $\mathbf{y}_i$  is given by:

$$m = \frac{T}{25 \cdot s} \quad (2)$$

where  $T$  is the number of time-points of the input sequence  $\mathbf{x}_i$  and  $s$  is its sampling frequency.

To address the significant imbalance of labels (i.e., begin, inside, and outside annotations) during training, we also considered the frames adjacent to the ground truth as *begin*. Specifically, we selected one frame to the left and one frame to the right of the actual start. We did not apply this augmentation during inference. We observed that this approach, combined with the frame selection strategy described in Sect. 3.5, considerably improve the final scores. More details about the interplay of labels augmentation and frame selection are described in Sect. 5.3.

### 3.4 Model Architecture

Given an input utterance  $\mathbf{x}_i$  as described in Sect. 3.2, the waveform passes through an audio encoder  $A_{enc}$ , resulting in a hidden representation  $\mathbf{z}_i \in \mathbb{R}^{n \times d}$ , where  $n$  is the number of frames produced by the encoder and  $d$  is the dimension of the hidden state.

$$\mathbf{z}_i = A_{enc}(\mathbf{x}_i) \quad (3)$$

Then, a linear layer  $A_{lp}$  projects  $\mathbf{z}_i$  into  $\mathbf{e}_i \in \mathbb{R}^{d \times m}$ , in order to match the dimensions of  $\mathbf{y}_i$ .

$$\mathbf{e}_i = A_{lp}(\mathbf{z}_i^\top) \quad (4)$$

Finally, a linear classification head  $A_{lc}$  is applied to obtain the predictions  $\mathbf{c}_i \in \mathbb{R}^{m \times p}$ , where  $p$  represents the output probabilities for each class.

$$\mathbf{c}_i = A_{lc}(\mathbf{e}_i^\top) \quad (5)$$

### 3.5 Post-processing

We extract word boundaries and post-process the audio input to return a sequence of variable-length segments, each corresponding to a word. Given the augmentation strategy outlined in Sect. 3.3, the model tends to over-segment during inference, resulting in clusters of predicted *begin* frames. To mitigate this

behaviour we select the average frame as predicted one discarding its neighbors. During the training stage no tolerance was applied, while at inference time, to check the correctness of the boundaries and compute the metrics, we consider a tolerance of 40 ms as [14]. Differently from this work we compare predicted boundaries with the time-point level ground truth instead of the frame quantized one, resulting in a more accurate evaluation. Testing on TIMIT dataset we applied a tolerance of 20 ms to align our scores with the ones got by [8].

**Table 1.** Architectures of employed models. We report convolutional layer details (channels, kernel size, stride), hidden size for recurrent model. For Wav2Vec, and HuBERT please refer to [6] [16].

	CNN	CRNN
# Convolutional channels	16, 32, 64, 128	16, 32, 64, 128
Kernel size	11, 3, 3, 3	11, 3, 3, 3
Stride	5, 2, 2, 2	5, 2, 2, 2
Hidden size	–	80, 40

## 4 Experimental Setup

### 4.1 Datasets

Buckeye [30] is a spontaneous speech corpus containing recordings of 40 talkers from central Ohio interviewed for about one hour. The group of people is stratified on age and gender and each recording is sampled at 16Khz and annotated to phoneme and word level. For our purpose we used the word annotations only. The whole corpus counts about 307,000 words.

TIMIT [15] is a speech corpus used for acoustic-phonetic studies, but it was also frequently employed for phoneme and word boundary detection tasks [8, 9, 23, 24]. It comprises recordings from 630 speakers, including 438 males and 138 females. Each speaker recorded ten utterances, resulting in a total of 6300 speech samples. These utterances are phonetically and lexically annotated, with indices marking the start and end of phonemes and words. For our study, we utilized only the test set, which includes 1680 samples.

On Buckeye, considering the length of each audio recording, we truncate them and split in train, validation and test with a similar strategy employed by [14]. We considered to use this way also to facilitate the comparison of results and the reproduction of the experiments.

### 4.2 Audio Pre-processing

The only manipulations applied on audio were standardization and padding. Since waveforms were with variable length, the standardization was done by

calculating the global weighted mean and global weighted standard deviation of the train set. The same values were then applied to the validation and test set. The padding was applied based on the longest waveform corresponding to a duration of approximately 9 s.

### 4.3 Encoders

Our framework is model-agnostic, meaning it can be applied regardless of the encoder architecture. However, in order to evaluate the method we chose several well-known architectures. Specifically, we selected a one-dimensional CNN and CRNN as from-scratch architectures, while we employed Hubert and Wave2Vec as pretrained models. The CNN takes raw audio as input and consists of four convolutional layers followed by batch normalization, ReLU activation, and max pooling, with the number of filters doubling at each layer. Its output is transposed and passed to a fully connected layer that produces a representation  $\mathbf{z} \in \mathbb{R}^{d \times m}$ , where  $d$  are the convolutional features and  $m$  the number of framed labels. Finally, the resulting vector is transposed again and passed to a linear classification layer. We chose the CRNN because it represents the state-of-the-art for various previous works in audio segmentation and classification [32, 35–37]. The network retains the same configurations as the CNN model for the convolutional layers, but without batch normalization. Additionally, it introduces two bidirectional Gated Recurrent Unit (B-GRU) layers with a similar configuration to [35, 36]. A final linear projection and a linear classification layer are applied. CNN and CRNN configurations are showed in Table 1.

We decided also to include the two main pretrained models at the state-of-the-art for several downstream tasks: HuBERT<sub>Large</sub> [16] and Wav2Vec2.0<sub>Base</sub> [6]. We kept both encoders frozen while fine-tuning the final linear layers, which were followed by layer normalization.

### 4.4 Training Procedure

All the models were trained on a NVIDIA RTX A6000. The average training time of our best fine-tuned model (i.e. HuBERT encoder) is around one hour. The inference time for the whole Buckeye test set is 26 s. We tuned learning rate and batch size hyperparameters with grid search on the validation set, choosing at the end  $10^{-3}$  as learning rate and 32 as batch size. We also applied an early stopping with a patience of 10 epochs if no improvement occurred on the best validation R-value. We did not use time error tolerance for the boundaries detection during the training phase. More details are available in the supplementary materials.

### 4.5 Metrics

The set of metrics, as defined in [2, 4, 29, 31], is composed by Precision, Recall, F-value, Over Segmentation (OS), and R-value.

**Precision** (PRC) and **Recall** (RCL) were employed as described by [31] and expressed in 6. In the equation  $N_{hit}$  represents the boundaries correctly detected, while  $N_{ref}$  stands for the total number of boundaries in the reference.

$$PRC = \frac{N_{hit}}{N_f}, RCL = \frac{N_{hit}}{N_{ref}} \quad (6)$$

**OS** [29] is the over segmentation rate and is given by the ratio of the total number of detected boundaries  $N_f$  over the total number of boundaries  $N_{ref}$  in the reference. Then the result is subtracted by one (Eq. 7).

$$OS = \frac{N_f}{N_{ref}} - 1 \quad (7)$$

**F-value** [2] is the harmonic average of PRC and RCL.

$$F\text{-value} = \frac{2 \cdot PRC \cdot RCL}{PRC + RCL} \quad (8)$$

**R-value** [31] is another composed metric derived from OS and Recall and is defined as a trade-off of these two metrics, being the balance between Recall and OS a suitable operating point for audio segmentation. R-value can be expressed as follow:

$$r_1 = \sqrt{(1 - RCL)^2 + OS^2}, r_2 = \frac{-OS + RCL - 1}{\sqrt{2}} \quad (9)$$

$$R\text{-value} = 1 - \frac{|r_1| + |r_2|}{2} \quad (10)$$

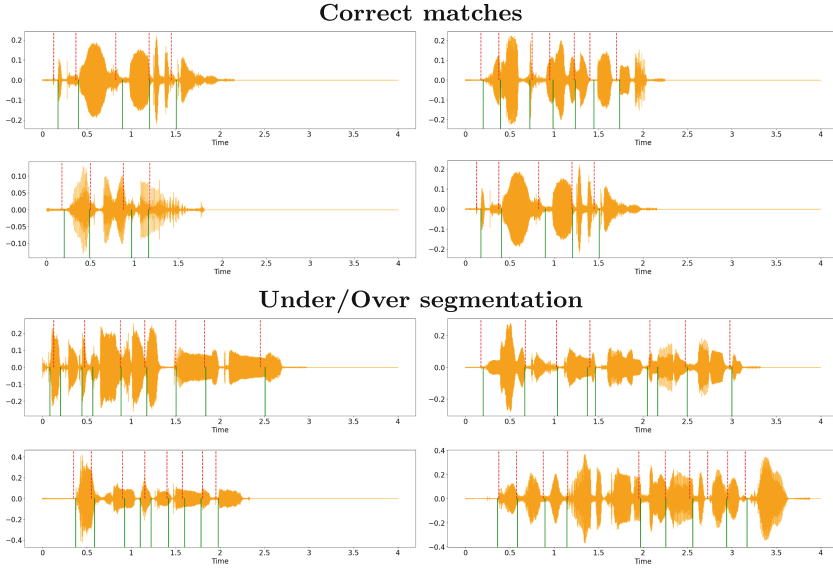
We decided to use R-value as main quality metric to evaluate the models, i.e. we saved the weights of models at the best validation R-value.

## 4.6 Experiments

To evaluate our method, we trained from scratch the CNN and CRNN architectures and fine-tuned HuBERT and Wav2Vec keeping both encoders frozen. For a comparative analysis we reproduced the training procedures of GradSeg [14] and DSegKNN [13]. All the architectures were tested on our Buckeye test set. To assess the generalization capability of our method, we also tested the above methods on the TIMIT dataset. Finally, to underline the need for label augmentation and frame selection, we assessed the performance of our method with and without the application of the two strategies.

## 5 Results

In the subsequent sections, we present the obtained results. Our method is benchmarked against a single supervised reference for Buckeye, while for TIMIT no supervised methods have been found. Finally, additional results without comparisons on NTIMIT [17] are reported in the supplementary materials. As discussed



**Fig. 2.** Segmentation comparison example between true (green solid line) and predicted (red dashed line) boundaries on Buckeye test set. The four pictures on top show a correct match of detected boundaries. The second group report three under segmentation scenarios and an over segmentation one. (Color figure online)

**Table 2.** Comparison of not-pretrained/pretrained supervised models and unsupervised ones on Buckeye test set. Tolerance was set to 40ms.

Unsupervised models					
Model	Precision	Recall	F-value	OS	R-value
DSegKNN [13]	0.3115	0.3226	0.3169	0.0355	0.4087
GradSeg [14]	0.4444	0.4356	0.4399	<b>-0.0197</b>	0.5251
Supervised models					
Model	Precision	Recall	F-value	OS	R-value
CNN	0.3842	0.3604	0.3708	-0.0575	0.4694
CRNN	0.4112	0.3711	0.3896	-0.0972	0.4923
GradSeg [14]	–	–	0.5960	–	–
Wav2Vec <sub>Base</sub> [6]	0.6556	0.4736	0.5494	-0.2766	0.6139
HuBERT <sub>Large</sub> [16]	<b>0.8999</b>	<b>0.7928</b>	<b>0.8427</b>	-0.1187	<b>0.8489</b>

in Sect. 2, the prevailing trend in recent research on word boundary detection involves self-supervised or unsupervised approaches making complex the possibility of comparison with those who want to tackle this task in a supervised way. On the other hand, a direct comparison with our supervised method may not be entirely fair. Nevertheless, we opted to include results from unsuper-



vised methods, not for direct comparison with our approach, but to underscore the potential impact of a supervised approach to word boundary detection in advancing applications within speech technology.

It is worth noting that certain supervised methods leverage word boundary detection for aligning speech to text transcriptions. For this task, metrics such as Word Error Rate (WER) are employed to assess the models and their efficacy. Given the disparate objectives of these metrics compared to ours, a direct correlation might be deemed unfair.

## 5.1 Models from Scratch

As shown in Table 2, the CNN model achieved lower results compared to its unsupervised counterpart. The CRNN model performed better than the CNN in most metrics except for the OS metric, yet it still lower than [14].

It is important to note that while the scores of both CNN and CRNN models are lower than those reported by [14], the latter utilized Wav2Vec pretrained as an encoder.

Table 3 presents the scores obtained on the TIMIT dataset [15]. As anticipated, both models demonstrated a general decline in performance compared to the results on the Buckeye dataset. Additionally, their scores are lower than those reported by [8]. However, it is important to consider that [8] was specifically trained on the TIMIT dataset.

## 5.2 Pretrained Models

HuBERT outperforms all models except in the over-segmentation metric, where GradSeg [14] in its unsupervised version achieved the value closest to zero. Additionally, we observed that both Wav2Vec and HuBERT exhibit higher Precision than Recall. This behavior is likely due to the frame selection strategy described in Sect. 4.4. Specifically, by selecting only the average boundary from the boundary clusters, we penalize instances where adjacent words (true positives) occur. Although we tested other frame selection strategies (see Sect. 5.3) to improve the recall metric, we ultimately favored a more precise model over a more sensitive one.

In the GradSeg article [14], the authors reported only the F-value metric for the supervised method, probably because it was not the main focus of their work. Nonetheless, this metric allowed us to compare our models with another benchmark, as we encountered difficulty finding recent supervised methods for comparison. Figure 2 shows some predictions done on Buckeye utterances by HuBERT model compared with the ground truth boundaries. The model tends to an under-segmentation, however the predicted boundaries are frequently close to the real ones.

Also on TIMIT dataset (Table 3) HuBERT outperforms the other models.

This section demonstrates that utilizing pretrained encoders for supervised training, along with other strategies like labels augmentation and output-frame

**Table 3.** Comparison of not-pretrained/pretrained supervised models and unsupervised one on TIMIT test set. Tolerance was set to 20 ms.

Unsupervised models					
Model	Precision	Recall	F-value	OS	R-value
SCPC [8]	0.2895	0.2302	0.2558	–	0.4003
Supervised models					
Model	Precision	Recall	F-value	OS	R-value
CNN	0.2490	0.1697	0.2008	–0.3169	0.3731
CRNN	0.2610	0.2247	0.2411	–0.1377	0.3794
Wav2Vec <sub>Base</sub> [6]	0.4433	0.3538	0.3930	–0.2005	0.5032
HuBERT <sub>Large</sub> [16]	<b>0.7566</b>	<b>0.7314</b>	<b>0.7436</b>	<b>–0.032</b>	<b>0.7807</b>

selection, can significantly enhance the quality of word segmentation. This improvement is evident not only on the specific dataset used to train the models, but also on other speech datasets (such as TIMIT), achieving even better results than unsupervised methods trained on them [8].

### 5.3 Effect of Labels Augmentation and Frame Selection

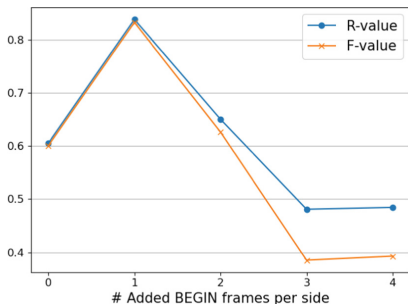
As shown in Table 4, applying only labels augmentation considerably worsens the scores. This is because increasing the number of boundaries enhances the model’s sensitivity, but it also leads to higher over-segmentation, significantly reducing precision and affecting both the F-value and R-value.

When the frame selection strategy is applied, we mitigate the over-segmentation issue, decreasing recall scores but improving overall performance. Different experiments were conducted with “begin” clusters during labels augmentation as reported in Fig. 3. Ultimately, we chose to label as “begin” one frame to the left and one frame to the right of the actual boundary because this setting yielded the best performance.

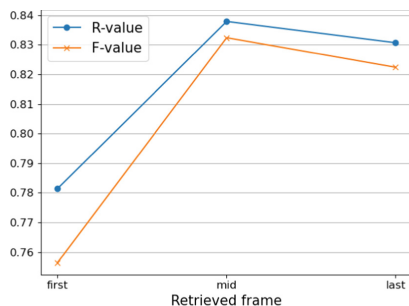
**Table 4.** Ablation study on labels augmentation and frame selection.

Model	Precision	Recall	F-value	OS	R-value
Wav2Vec <sub>Base</sub>	0.8805	0.0460	0.0874	–0.9475	0.3254
↔Label Augmentation	0.3013	0.6585	0.4131	1.1889	–0.1604
↔Frame selection	<b>0.6556</b>	<b>0.4736</b>	<b>0.5494</b>	<b>–0.2766</b>	<b>0.6139</b>
HuBERT <sub>Large</sub>	0.9302	0.4403	0.5971	–0.5327	0.6032
↔Label Augmentation	0.4840	0.9161	0.6332	0.8942	0.2049
↔Frame selection	<b>0.8999</b>	<b>0.7928</b>	<b>0.8427</b>	<b>–0.1187</b>	<b>0.8489</b>

On the frame selection side (Fig. 4), we tested three approaches: selecting the first frame, the last frame and the mid one. However, the best choice in terms of scores was to extract the mid-frame for each *begin* cluster.



**Fig. 3.** Comparison of R-value and F-value for the Buckeye validation set based on different window sizes for label augmentation. Each data point represents the number of frames labeled as *begin* to the left and right of the ground truth. The results are computed employing the HuBERT encoder.



**Fig. 4.** Comparison of R-value and F-value scores for the Buckeye validation set based on different frame selection strategies. The first approach retrieves the initial *begin* frame from the *begin* cluster, the second approach selects the middle frame, and the third approach picks the final frame. The results are computed employing the HuBERT encoder.

## 5.4 Discussion

As demonstrated in previous sections, in Table 2 and in Table 3, employing a supervised approach centered on frame classification significantly enhances the performance of word boundary detection (WBD). Notably, it’s not solely the choice of approach (frame classification) that influences the outcomes, but also the methodology we employ in handling labels imbalance during training and frame selection during inference as showed in Table 4 and discussed in Sect. 5.3, ensuring anyway that they don’t alter the inherent nature of the data and maintains a streamlined preprocessing pipeline. It’s crucial to acknowledge the significance we attribute to the WBD task. In contrast to the self-supervised methods outlined in Sect. 2, which strive to generate directly meaningful word-level audio latent representations, we interpret this task as an initial step to provide support and input for self-supervised models with discrete units, akin to tokens in text. In light of this intent, based on performance, we can state that the self-supervised models are not ready yet and probably this is not their goal. With this work, we also aim to stimulate the audio community to explore other supervised methods for word boundary detection. We strongly believe that this approach could significantly boost the performance of this task

and consequently enhance the development of recent audio application trends, such as speech-to-speech conversational models and real-time translators.

## 6 Conclusion

In this work we propose a robust and computationally light preprocessing approach for word boundary detection and evaluated its efficacy compared to other supervised and unsupervised methods, by using pre-trained and from scratch solutions. Our future work will leverage extracted words to build a tokenization-like method, thus enabling the variable-length discrete units to retain important para-verbal and prosodical features and paving the way to stronger self-supervised models and spoken dialogue systems.

**Acknowledgements.** Simone Carnemolla and Salvatore Calcagno acknowledge financial support from: PNRR MUR project PE0000013-FAIR.

## References

1. Agarwal, A., Jain, A., Prakash, N., Agrawal, S.: Word boundary detection in continuous speech based on suprasegmental features for Hindi language. In: 2nd International Conference on Signal Processing Systems (2010)
2. Ajmera, J., McCowan, I., Bourlard, H.: Robust speaker change detection. *IEEE Signal Process. Lett.* **11**(8), 649–651 (2004)
3. Almpantidis, G., Kotropoulos, C.: Phonemic segmentation using the generalised gamma distribution and small sample bayesian information criterion. *Speech Commun.* **50**(1), 38–55 (2008). <https://doi.org/10.1016/j.specom.2007.06.005>
4. Aversano, G., Esposito, A., Marinaro, M.: A new text-independent method for phoneme segmentation. In: *IEEE MWSCAS*. vol. 2 (2001)
5. Baevski, A., Schneider, S., Auli, M.: Vq-wav2vec: self-supervised learning of discrete speech representations. In: *IEEE ICLR* (2020)
6. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **33**, 12449–12460 (2020)
7. Bhati, S., Nayak, S., Murty, K.S.R.: Unsupervised Speech Signal to Symbol Transformation for Zero Resource Speech Applications. In: *Interspeech* (2017). <https://doi.org/10.21437/Interspeech.2017-1476>
8. Bhati, S., Villalba, J., Želasko, P., Moro-Velazquez, L., Dehak, N.: Unsupervised speech segmentation and variable rate representation learning using segmental contrastive predictive coding. *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 2002–2014 (2022)
9. Bhati, S., Villalba, J., Želasko, P., Moro-Velazquez, L., Dehak, N.: Segmental Contrastive Predictive Coding for Unsupervised Word Segmentation. In: *Interspeech* (2021). <https://doi.org/10.21437/Interspeech.2021-1874>
10. Chorowski, J., Weiss, R.J., Bengio, S., Van Den Oord, A.: Unsupervised speech representation learning using Wavenet autoencoders. *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**(12), 2041–2053 (2019)

11. Dusan, S., Rabiner, L.: On the relation between maximum spectral transition positions and phone boundaries. In: Interspeech (2006). <https://doi.org/10.21437/Interspeech.2006-230>
12. Franke, J., Mueller, M., Hamlaoui, F., Stueker, S., Waibel, A.: Phoneme boundary detection using deep bidirectional LSTMs. In: Speech Communication; 12. ITG Symposium (2016)
13. Fuchs, T., Hoshen, Y., Keshet, Y.: Unsupervised Word Segmentation using K Nearest Neighbors. In: Proceedings of Interspeech 2022, pp. 4646–4650 (2022). <https://doi.org/10.21437/Interspeech.2022-11474>
14. Fuchs, T.S., Hoshen, Y.: Unsupervised word segmentation using temporal gradient pseudo-labels. In: IEEE ICASSP (2023)
15. Garofolo, J.S.: TIMIT acoustic phonetic continuous speech corpus. Linguistic Data Consortium, 1993 (1993)
16. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: HuBERT: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 3451–3460 (2021)
17. Jankowski, C., Kalyanswamy, A., Basson, S., Spitz, J.: NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 109–112. IEEE (1990)
18. Kamper, H., Jansen, A., Goldwater, S.: Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(4), 669–679 (2016)
19. Kamper, H., Jansen, A., Goldwater, S.: A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Comput. Speech Lang.* **46**, 154–174 (2017)
20. Kamper, H., Livescu, K., Goldwater, S.: An embedded segmental k-means model for unsupervised segmentation and clustering of speech. In: IEEE ASRU (2017)
21. Kamper, H., van Niekerk, B.: Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks. In: Interspeech (2021). <https://doi.org/10.21437/Interspeech.2021-50>
22. Keshet, J., Shalev-Shwartz, S., Singer, Y., Chazan, D.: Phoneme alignment based on discriminative learning. In: Interspeech (2005). <https://doi.org/10.21437/Interspeech.2005-129>
23. Kreuk, F., Keshet, J., Adi, Y.: Self-supervised contrastive learning for unsupervised phoneme segmentation. In: Interspeech (2020). <https://doi.org/10.21437/Interspeech.2020-2398>
24. Kreuk, F., Sheena, Y., Keshet, J., Adi, Y.: Phoneme boundary detection using learnable segmental features. In: IEEE ICASSP (2020)
25. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal forced aligner: trainable text-speech alignment using Kaldi. In: Interspeech (2017). <https://doi.org/10.21437/Interspeech.2017-1386>
26. Michel, P., Rasanen, O., Thiollière, R., Dupoux, E.: Blind phoneme segmentation with temporal prediction errors. In: ACL Student Research Workshop, pp. 62–68 (2017)
27. Naganoor, V., Jagadish, A.K., Chemmangat, K.: Word boundary estimation for continuous speech using higher order statistical features. In: IEEE TENCON (2016)

28. Payne, B., Ng, S., Shantz, K., Federmeier, K.: Event-related brain potentials in multilingual language processing: The N's and P's, pp. 75–118. *Psychology of Learning and Motivation - Advances in Research and Theory*, Academic Press Inc., United States (2020). <https://doi.org/10.1016/bs.plm.2020.03.003>
29. Petek, B., Andersen, O., Dalsgaard, P.: On the robust automatic segmentation of spontaneous speech. In: *IEEE ICSLP*. vol. 2 (1996)
30. Pitt, M.A., Johnson, K., Hume, E., Kiesling, S., Raymond, W.: The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Commun.* **45**(1), 89–95 (2005)
31. Räsänen, O.J., Laine, U.K., Altsosaar, T.: An improved speech segmentation quality measure: the R-value. In: *Tenth Annual Conference of the International Speech Communication Association* (2009)
32. Salamon, J., MacConnell, D., Cartwright, M., Li, P., Bello, J.P.: Scaper: a library for soundscape synthesis and augmentation. In: *IEEE WASPAA* (2017)
33. Shezi, N., Reddy, S.: Word boundary estimation of isizulu continuous speech. In: *IEEE PICC*, pp. 1–6 (2020)
34. Strgar, L., Harwath, D.: Phoneme segmentation using self-supervised speech models. In: *IEEE SLT* (2023)
35. Venkatesh, S., et al.: Artificially synthesising data for audio classification and segmentation to improve speech and music detection in radio broadcast. In: *IEEE ICASSP* (2021)
36. Venkatesh, S., Moffat, D., Miranda, E.R.: Investigating the effects of training set synthesis for audio segmentation of radio broadcast. *Electronics* **10**(7), 827 (2021)
37. Venkatesh, S., Moffat, D., Miranda, E.R.: You only hear once: a YOLO-like algorithm for audio segmentation and sound event detection. *Appl. Sci.* **12**(7), 3293 (2022). <https://doi.org/10.3390/app12073293>
38. Wang, Y.H., Chung, C.T., Lee, H.Y.: Gate activation signal analysis for gated recurrent neural networks and its correlation with phoneme boundaries. In: *Interspeech* (2017). <https://doi.org/10.21437/Interspeech.2017-877>



# Semi-supervised Cross-Lingual Speech Recognition Exploiting Articulatory Features

Xinmei Su<sup>✉</sup>, Xiang Xie<sup>✉</sup>, Chenguang Hu<sup>✉</sup>, Shu Wu<sup>✉</sup>, and Jing Wang<sup>✉</sup>

School of Information and Electronics, Beijing Institute of Technology, Beijing, China  
suxinmei2022@126.com,  
{xiexiang,3220200551,3220230774,wangjing}@bit.edu.cn

**Abstract.** The state-of-the-art (SOTA) Automatic Speech Recognition (ASR) systems are mostly based on the data-driven methods. However, low-resource languages may lack data for training. Articulatory Features (AFs) describe the movements of the vocal organ which can be shared across languages. Thus, this paper investigates AFs-based semi-supervised techniques to share data between languages. First, the traditional acoustic features and the AFs are combined as front-end features to provide articulatory information for cross-lingual knowledge transfer. Then, the dropout-based lattice decoded are used as the pseudo-labels for the unsupervised data to address the problem of data deficiency. In addition, the Lattice-free Maximum Mutual Information (LF-MMI) objective is adopted to better adapt to small datasets. Experiments show that our system can obtain a relative improvement of 58.6% on Character Error Rate (CER) comparing to the baseline system. More specifically, the smaller the datasets are, the more obvious the advantages of our system can be.

**Keywords:** Automatic speech recognition · Semi-supervised · Articulatory features

## 1 Introduction

Multilingual or cross-lingual speech recognition has become one of the most important research directions in Automatic Speech Recognition (ASR), and received extensive attention since the 1990s. Since the emergence of the Hidden Markov Model (HMM), speech recognition has entered the data-driven era. But with more and more data, the model performance of the HMM itself has entered a bottleneck. With the rapid development of DNN, except the research on the neural networks of multilingual speech recognition such as, Time Delay Neural Network (TDNN) and the factored form of TDNN (TDNNF), current researches on multilingual speech recognition focuses more on training models driven by massive data [1, 2]. In recent years, end-to-end ASR uses multilingual tokens (words or subwords) and combines the tokens to achieve multilingual

speech recognition. The advantage of such system is that it gets rid of the limitations of different language pronunciation dictionaries [3]. Articulatory Features (AFs) based on articulatory attributes is also widely used in multilingual ASR. Multilingual speech recognition based on AFs generally uses a language with rich resources to train an articulatory attributes detector to extract AFs, and then combine AFs with traditional acoustic features [4].

Against the problem of domain mismatch between training data and test data in cross-lingual ASR, the Domain-adversarial training of Neural Networks (DANN) is proposed [5]. DANN reduces the difference between the target domain and the source domain through adversarial training, thereby improving the model’s general performance.

In this paper, we borrow the ideas from the prior work mentioned above to solve the problem of the cross-lingual ASR and propose several improvements to the previous ideas. The contributions of this paper are as follows: First, by combining articulatory features and traditional hand-crafted features (MFCCs) as front-end features, the system can take full advantage of phonemes, project phonemes of all languages to the same dimension and improve multilingual recognition performance. Second, making full use of unsupervised data and the dropout-based Lattice-free Maximum Mutual Information (LF-MMI) semi-supervised learning enables knowledge transfer from resource-rich languages to improve speech recognition performance in the target low-resource language domain. Finally, the system proposed has a relative Character Error Rate (CER) reduction of 58.6% compared with the baseline system.

## 2 Related Work

### 2.1 Articulatory Features for ASR

The advantage that articulatory attributes can be shared in multiple languages makes it widely used in the field of multilingual speech recognition [4]. The phoneme set of the International Phonetic Alphabet (IPA) can represent all sounds that humans can make through a set of phonetic symbol systems. Previous descriptions of articulatory attributes mostly focus on a single language or languages of the same language family [6, 7]. This paper uses a unified description of the articulatory attributes of Chinese, English, German, and French of different language families, which is adopted in our previous work, and can be extended to other languages as needed [8]. More details about the multilingual AFs defined by us can be seen in the work [8].

One-hot encoding is used to represent each phoneme. Especially, the same phonemes in different languages may have different encodings. For example, the articulatory attributes of n are “Alveolar, Nasal, Voiced, Nil, Nil, Nil”, “Alveolar, Nasal, Voiced, Nil, Nil, Nil”, “Nil, Nasal, Nil, Nil, Nil, Nil”, and “Alveolar, Nil, Nil, Nil, Nil, Nil” in Chinese, English, German, French. The one-hot encodings are “100000000, 000100, 100, 0001, 0001, 001”, “100000000, 000100, 100, 0001, 0001, 001”, “000000001, 000100, 001, 0001, 0001, 001”, “100000000, 000001, 001, 0001, 0001, 001”. If the basic modeling units are only phonemes, the phoneme



n is encoded samely in these four languages. However, the modeling units of articulatory attributes can distinguish the same phoneme in defferent languages, it is more fine-grained encoding comparing to the phoneme level modeling units. Although different languages have different sets of phonemes, when decomposed by articulatory attributes, it can be done at the articulatory level shared. This enables the articulatory attributes detector to obtain a feature space shared by multiple languages on the model.

## 2.2 Semi-supervised Learning for Multilingual ASR

The biggest difficulty faced by low-resource speech recognition is the lack of enough amount of supervised data for training [9]. In traditional Maximum Mutual Information (MMI)-based ASR system, the training of MMI depends on the lattice, while the lattice depends on the GMM-HMM acoustic model. In order to better improve the training of MMI, Dan Povey et al. proposed a training criterion of Lattice-free MMI (LF-MMI) [10]. In recent years, LF-MMI has gradually become one of the mainstream ASR models due to its training advantages in small datasets. Manohar et al. proposes a semi-supervised learning method based on LF-MMI which contains N-best lattices in the decoding path [11].

In DNN-based ASR systems, dropout is a common training strategy for improving system robustness. Monte Carlo refers to a class of algorithms that rely on repeated sampling to obtain a certain number of distributions, so the method of preserving dropout in inference is also called Monte Carlo dropout [12].

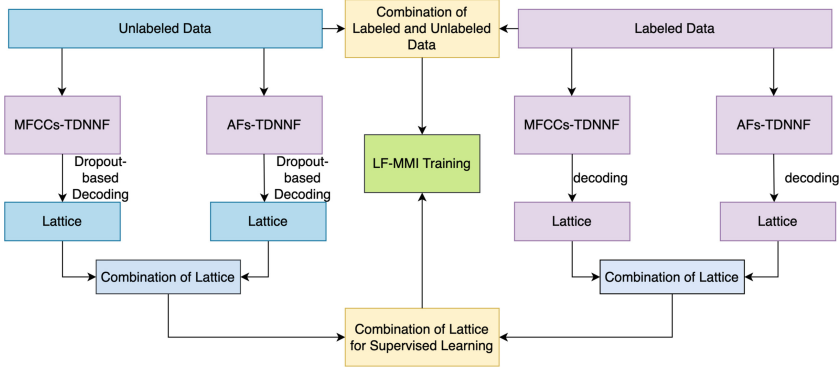
## 3 Method

The semi-supervised method proposed is shown in Fig. 1, and the training process is as follows:

First, train the DANN-based articulatory attributes detector on the source language domain, and train the TDNNF models based on MFCCs and AFs respectively, denoted as MFCCs-TDNNF and AFs-TDNNF, which are used as seed models of semi-supervised learning.

Second, use the DANN-based articulatory attributes detector to obtain the AFs of the unsupervised data, retain the dropout in AFs-TDNNF model on the decoding stage for N times, and merge it with the lattice decoded without dropout to get the N+1 dropout-based lattice. The lattice gained from AFs-TDNNF is denoted as AFs-dropout-lattice. Repeat the same operation on MFCCs-TDNNF and record the N+1 lattice as MFCCs-dropout-lattice.

Third, supervised and unsupervised data are both used as training data, and the obtained AFs-dropout-lattice and MFCCs-dropout-lattice are combined according to a certain weight distribution as all dropout-lattice for training. By migrating AFs through lattice, the paths of lattice are further enriched.



**Fig. 1.** The whole architecture of our proposed method.

Fourth, use the merged features as input, and use the merged dropout-lattice as the supervised lattice to carry out the molecular composition of LF-MMI for the subsequent training process.

### 3.1 DANN-Based AFs Detectors

The U-net structure diagram and the whole DANN-based AFs detector used is depicted in detail in our previous work [8]. The detector is applied to complete diverse tasks, and each task has a different loss function. The formula derivation of the weight adjustment method is described as follows [13]:

The likelihood probability of multi-classification task is defined in Eq. (1).

$$p(\mathbf{y}_1, \dots, \mathbf{y}_K | \mathbf{f}^{\mathbf{W}}(\mathbf{x})) = p(\mathbf{y}_1 | \mathbf{f}^{\mathbf{W}}(\mathbf{x})) \dots p(\mathbf{y}_K | \mathbf{f}^{\mathbf{W}}(\mathbf{x})), \quad (1)$$

where the function  $p(\mathbf{y} | \mathbf{f}^{\mathbf{W}}(\mathbf{x})) = \text{Softmax}(\mathbf{f}^{\mathbf{W}}(\mathbf{x}))$ .

Taking two tasks as an example, the loss function of multi-task is:

$$\begin{aligned} \mathcal{L}(\mathbf{W}, \sigma_1, \sigma_2) &= -\log p(\mathbf{y}_1, \mathbf{y}_2 = c | \mathbf{f}^{\mathbf{W}}(\mathbf{x})) \\ &= \text{Softmax}(\mathbf{y}_1 = c; \mathbf{f}^{\mathbf{W}}(\mathbf{x}), \sigma_1) \cdot \text{Softmax}(\mathbf{y}_2 = c; \mathbf{f}^{\mathbf{W}}(\mathbf{x}), \sigma_2) \\ &= \frac{1}{\sigma_1^2} \mathcal{L}_1(\mathbf{W}) + \frac{1}{\sigma_2^2} \mathcal{L}_2(\mathbf{W}) + \log \frac{\sum_{c'} \exp\left(\frac{1}{\sigma_2^2} f_{c'}^{\mathbf{W}}(\mathbf{x})\right)}{\left(\sum_{c'} \exp(f_{c'}^{\mathbf{W}}(\mathbf{x}))\right)^{\frac{1}{\sigma_1^2}}} \\ &\quad + \log \frac{\sum_{c'} \exp\left(\frac{1}{\sigma_2^2} f_{c'}^{\mathbf{W}}(\mathbf{x})\right)}{\left(\sum_{c'} \exp(f_{c'}^{\mathbf{W}}(\mathbf{x}))\right)^{\frac{1}{\sigma_2^2}}} \\ &\approx \frac{1}{\sigma_1^2} \mathcal{L}_1(\mathbf{W}) + \frac{1}{\sigma_2^2} \mathcal{L}_2(\mathbf{W}) + \log \sigma_1 + \log \sigma_2, \end{aligned} \quad (2)$$

where  $\sigma$  is a learnable weight factor between two tasks.

Therefore, the final loss function of the DANN-based AF-detector is:

$$L_{\text{ada}}(\sigma_1, \sigma_2, \sigma_i, \sigma_3) = \frac{1}{\sigma_1^2} L_{\text{phn}} + \frac{1}{\sigma_2^2} L_{\text{af-i}} + \frac{1}{\sigma_2^2} L_{\text{lid\_GRL}} \\ + \frac{1}{\sigma_3^2} L_{\text{spk\_GRL}} + \log(\sigma_1 \sigma_2 \sigma_i \sigma_3), \quad (3)$$

where  $L_{\text{phn}}$  is the loss function of the triphone classification,  $L_{\text{af-i}}$  is the loss function of the  $i$ th articulatory attributes classification. There are six types of articulatory attributes,  $L_{\text{spk\_GRL}}$ ,  $L_{\text{lid\_GRL}}$  are loss functions for language and speaker classification using gradient inversion and  $\sigma$  are the weight relation factor of the loss function. The loss function of all models is the Cross Entropy (CE) loss.

### 3.2 LF-MMI-Based Semi-Supervised Learning for Multilingual ASR

The basic task of ASR is to find the probability of the word sequence  $\mathbf{W}$  given the speech observation sequence  $\mathbf{O}$ , that is, to find the probability  $P(\mathbf{W} | \mathbf{O})$ . MMI is the direct maximization of  $P(\mathbf{W} | \mathbf{O})$ , namely:

$$\theta_{MMI} = \underset{\theta}{\operatorname{argmax}} P_{\theta}(\mathbf{W}_r | \mathbf{O}_r) \quad (4)$$

The loss function of MMI can be written as:

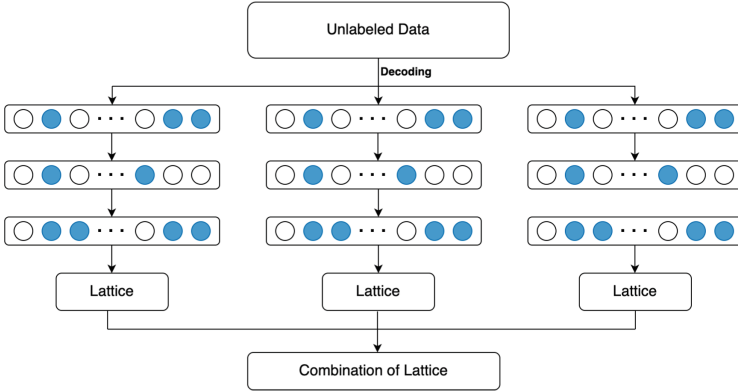
$$F_{MMI}(\theta) = \sum_{r=1}^R \log \frac{P_{\theta}(\mathbf{O}_r | \mathbf{W}_r) P(\mathbf{W}_r)}{\sum_{\hat{\mathbf{w}}} P_{\theta}(\mathbf{O}_r | \hat{\mathbf{W}}_r) P(\hat{\mathbf{W}})} \quad (5)$$

Based on the loss function of MMI, the semi-supervised loss function based on LF-MMI is defined as [11]:

$$\mathcal{L}_{MMI} = \max_{\theta} \sum_{u=1}^U \log \left( \sum_{\mathbf{w} \in \mathcal{G}_{\text{num}}^{(u)}} P(\mathbf{w} | \mathbf{X}^{(u)}, \theta) \right), \quad (6)$$

where  $\mathbf{X}^{(u)}$  represents the observation vector when the sentence  $u$  is given, and  $\mathcal{G}_{\text{num}}^{(u)}$  is the lattice result decoded by the supervised data under the supervised model.

The loss of LF-MMI has the following improvements: Firstly, when building a language model, if choosing characters or words as the basic token, the corpora will have hundreds of thousands or more characters. Thus, AFs are selected as the modeling token, considering the balance between modeling complexity and granularity. Secondly, in order to prevent over-fitting of the networks, LF-MMI introduces the CE function as a subtask for multi-task learning. Thirdly, the topological state of HMM is a single-state HMM.



**Fig. 2.** Schematic diagram of the semi-supervised speech recognition structure based on dropout and LF-MMI. Neurons with blue colors indicate the discarded neurons due to the existence of dropout. (Color figure online)

### 3.3 Dropout-Based Semi-supervised Learning for Multilingual ASR

For the reason that the amount of modeling units at the AFs-level is much smaller than at the word level, the multilingual ASR systems may be prone to over-fitting. The semi-supervised learning based on dropout is implemented in our system. Dropout is also retained during the inference stage, and the same data is decoded multiple times to obtain different lattices. In this way, the perturbation in the training and inference process can be increased to gain the uncertainty in the experiments, prevent the system from over-fitting, and increase the robustness of the ASR system. These lattices are combined to obtain more decoding space for the LF-MMI-based semi-supervised training, as shown in Fig. 2.

The semi-supervised LF-MMI loss function based on dropout is defined as [14]:

$$\mathcal{L}_{\text{MMI}} = \max_{\theta} \sum_{u=1}^U \log \left( \mathbb{E}_{\mathbf{W} \sim P(\mathbf{W} | \mathbf{X}^u, \mathbf{D}_s)} P(\mathbf{W} | \mathbf{X}^{(u)}, \theta) \right), \quad (7)$$

where  $\mathbf{X}^{(u)}$  represents the observation vector when a sentence  $u$  is given,  $\mathbf{D}$  denotes the supervised training data,  $\mathbf{W}$  is the result obtained for each sample.

For each specific sentence, N-best decoding is performed through a network with standard dropout. The lattice obtained by AFs and the lattice obtained by traditional acoustic features have different decoding spaces. By combining the AFs-based lattice with the MFCCs-based lattice, knowledge transfer can enrich the paths during inference. Consequently, a semi-supervised learning method based on articulatory attributes proposed in this paper is defined as:

$$\mathcal{L}_{\text{MMI}} = \max_{\theta} \sum_{u=1}^U \log \left( \sum_{\mathbf{W} \in \mathcal{G}_{\text{num-AFs-MFCCs}}^{(u)}} P(\mathbf{W} | \mathbf{X}^{(u)}, \theta) \right), \quad (8)$$

where  $\mathbf{X}^{(u)}$  represents the observation vector when the sentence  $u$  is given,  $\mathcal{G}_{\text{num-AFs-MFCCs}}^{(u)}$  is the molecular composition of LF-MMI for the final lattice that combines MFCCs-dropout-lattice and AFs-dropout-lattice, namely:

$$\mathcal{G}_{\text{num-AFs-MFCCs}}^{(u)} = \lambda \mathcal{G}_{\text{num-MFCCs-dropout}} \cup (1 - \lambda) \mathcal{G}_{\text{num-AFs-dropout}}, \quad (9)$$

where  $\mathcal{G}_{\text{num-MFCCs-dropout}}$  means MFCCs-dropout-lattice,  $\mathcal{G}_{\text{num-AFs-dropout}}$  represents AFs-dropout-lattice, and  $\cup$  represents the combination operation between lattices.

## 4 Experiments

### 4.1 Dataset

In this experiment, the source languages for training the AFs-based detector are English, German and French respectively and the target language is Chinese. Using these four languages can cover all phonemes, project the languages to the same dimension space and decompose all languages through AFs. Although Chinese is not a low-resource language, it is a relatively controllable language that is easier to analyze for subsequent experiments. For English, German, and French dataset, Librispeech and Multilingual Librispeech are adopted [15, 16]. In order to conduct semi-supervised learning, 15 h of speech is randomly selected from the Aishell (150 h) as supervised training, and the rest as unsupervised training [17]. There is no overlap between unsupervised and supervised data.

### 4.2 Experimental Details

The process of ASR is consistent with the standard Kaldi process. First, under the supervised data, the GMM-HMM model based on MFCCs and AFs is trained respectively, and the GMM-HMM model based on TDNNF and LF-MMI is trained as the seed model. The TDNNF has six hidden layers in total with 625 nodes in each hidden layer. In each layer, the dropout rate is set equally. For the context-sensitive decision tree required for training LF-MMI, this experiment only uses supervised data for decision tree training. The molecular FST is constructed with supervised and unsupervised data but has a higher weight for supervised data.

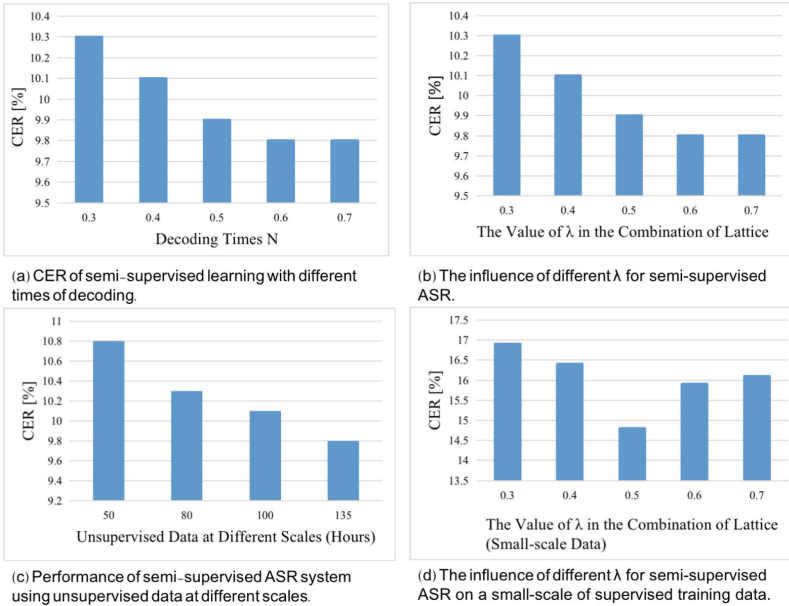
## 5 Results and Analysis

### 5.1 Dropout Settings

The experimental results tested on the baseline system are shown in Table 1. It shows that adding an appropriate dropout (dropout=0.1) can improve the generalization performance of the model and get a lower CER (CER=23.7%). When analyzing the influence of dropout in the test phase, the experimental

**Table 1.** Performance of different sizes of dropout on the LF-MMI-based ASR model.

	Dropout for Train Set	Dropout for Test Set	CER [%]	
LF-MMI	0.0	—	24.2	
		0.1	0.0	23.7
			0.1	26.3
			0.2	28.6
			0.5	40.5
	0.2	0.0	24.1	
		0.1	26.5	
		0.2	29.0	
		0.5	28.6	
	0.5	0.0	39.9	
		0.1	42.5	
		0.2	45.7	
		0.5	50.1	



**Fig. 3.** CER results with different settings for multilingual semi-supervised ASR.

results show that when the dropout used in the test is between 0.1 and 0.2, the performance will decrease slightly. However, when the dropout used in the test exceeds 0.5, the performance will drop significantly. This is because the more neurons the neural network randomly discards, the worse the predictive ability

**Table 2.** Results of AF-based multilingual ASR.

Systems	Weight of labeled and unlabeled data	Use of labeled data in Aishell (Hours)	CER[%]
MFCCs-baseline	–	5	34.9
MFCCs-baseline	–	15	23.7
MFCCs+AFs (Combined Lattice) † <sup>a</sup>	–	15	23.0
MFCCs+Semi-supervised	1.0:1.0	15	11.9
MFCCs+Semi-supervised	1.0:1.5	15	12.3
MFCCs+Semi-supervised	1.5:1.0	15	11.3
MFCCs+Semi-supervised	2.0:1.0	15	11.8
MFCCs+AFs+Semi-supervised	1.5:1.0	15	10.9
MFCCs+AFs+Semi-supervised (Dropout)	1.5:1.0	5	14.2
MFCCs+Semi-supervised (Dropout)	1.5:1.0	15	10.6
MFCCs+AFs+Semi-supervised (Dropout)	1.5:1.0	15	<b>9.8</b>

<sup>a</sup> † represents that the experimental results of the supervised module combine the lattice and decode it with sMBR.

<sup>b</sup> The “MFCCs-baseline” system represents the speech recognition model trained only with the MFCCs of the target language. The “AFs” listed are all DANN-AFs, and “+Semi-supervised” refers to the standard LF-MMI-based semi-supervised model, “+Semi-supervised (Dropout)” is a LF-MMI-based semi-supervised model based on Monte Carlo dropout

of the model is for the same dataset. The dropout size is set to 0.1 in the following experiments to balance between the CER result and the inference time.

## 5.2 Decoding Settings

In this Subsection, the number of decoding  $N$  is selected from 5 to 30 on the baseline system, and several experiments are carried out. The experimental results are shown in the Fig. 3 (a). As the number of decoding times  $N$  increases, the lattice merging and model training time will increase significantly. Although the CER of the system decreases slightly when  $N$  is set to 30 (CER=10.4%), comparing to the system performance when  $N$  equals to 20 (CER=10.5%), the training and decoding time is greatly increased when the number of decoding times is 30. In order to achieve a balance between experimental performance and training time, in the subsequent experiments, the number of decoding  $N$  is selected to 20.

## 5.3 Results of AF-Based Semi-Supervised Learning for Multilingual ASR

This Subsection studies the influence of  $\lambda$  in the Eq. (9). The experimental results of different  $\lambda$  values are shown in Fig. 3 (b). It can be seen from the figure that

**Table 3.** Results of semi-supervised ASR on unsupervised data at different scales.

Unsupervised Data Size	Systems	CER [%]	Relative CER decline [%]
–	Baseline	23.7	–
50 h	MFCCs+AFs+Semi-supervised (Dropout)	10.8	–
	MFCCs+Semi-supervised (Dropout)	11.9	9.2
	Standard Semi-supervised Learning	13.0	16.9
80 h	MFCCs+AFs+Semi-supervised (Dropout)	10.3	–
	MFCCs+Semi-supervised (Dropout)	11.3	8.8
	Standard Semi-supervised Learning	12.1	14.8
100 h	MFCCs+AFs+Semi-supervised (Dropout)	10.1	–
	MFCCs+Semi-supervised (Dropout)	11.0	8.2
	Standard Semi-supervised Learning	11.7	13.6
135 h	MFCCs+AFs+Semi-supervised (Dropout)	<b>9.8</b>	–
	MFCCs+Semi-supervised (Dropout)	10.6	7.5
	Standard Semi-supervised Learning	11.3	13.2
235 h (Aishell + Aidatatang)	MFCCs+Semi-supervised (Dropout)	<b>9.4</b>	–

the model has the lowest CER (9.8%) when  $\lambda$  equals 0.6 or 0.7. Accordingly, in the following experiments, the value of  $\lambda$  is set to 0.6 as the weight in the combination of MFCCs-dropout-lattice and AFs-dropout-lattice. Moreover, the results of AFs-based multilingual ASR systems are shown in Table 2. The following conclusions can be drawn from the experimental results:

First, compared with the baseline system trained only with the supervised data, when applying semi-supervised learning, all semi-supervised learning systems have more obvious decreases in CER than the baseline system (CER = 23.7%) and the supervised MFCCs+AFs (Combined Lattice) system (CER = 23.0%). When performing semi-supervised learning with the ratio of ‘1.5:1.0’ for supervised to unsupervised data, the model achieves the lowest CER (11.3%).

Second, compared with the standard LF-MMI semi-supervised learning method (CER = 11.3%), semi-supervised learning with dropout (CER = 10.6%) has a relative CER decrease of 6.2%, which shows that semi-supervised learning with dropout makes better use of unsupervised data.

Third, when using the AF-based semi-supervised learning method proposed, the best ASR performance (CER=9.8%) can be obtained. Compared with the MFCCs+semi-supervised (dropout) (CER=10.6%), which also uses the dropout semi-supervised learning method, there is a 7.5% relative CER drop, which obtains a 58.6% relative CER decline to the baseline system (CER=23.7%). This further shows that the semi-supervised learning method combined with articulatory attributes proposed can make the unsupervised data pass through DANNs-AFs, take full advantage of multilingual phoneme information and successfully transfer knowledge from data of rich-resource languages to low-resource languages.



#### 5.4 Unsupervised Data Scales Settings

In this Subsection, 15 h of speech in Aishell is still used as supervised training, and 100 h, 80 h, and 50 h are randomly selected from the 135 h of unsupervised data for unsupervised training. The model which achieves the best result in Table 2 is used for the remaining experiments. The experimental results are shown in Table 3 and the trend graph is shown in Fig. 3 (c). It can be seen from Table 3 that when there is more unsupervised data, the speech recognition system has better recognition performance. At the same time, when comparing the standard semi-supervised method with the semi-supervised method combined with AFs, the AF-based semi-supervised learning method gains more obvious improvements. It also shows that when the amount of data is smaller, the help of multilingual AFs is greater.

#### 5.5 Performance of Semi-Supervised Learning for Multilingual ASR on Small-Scale Supervised Data

In order to study the ASR performance of the proposed method under small-scale supervised data, 5 h of supervised training data is randomly selected from the 15 h of supervised data used above. Firstly, we explored the values of different weights in Eq. (9) when the supervised training data is reduced. The experimental results are shown in Fig. 3 (d). It can be seen that when the training data is only 5 h, assigning more weights to the AFs-lattice can make the model obtain lower CER for the reason that when the training data is smaller, more knowledge needs to be borrowed from the resource-rich languages. The CER results compared with the baseline system are shown in Table 2. The relative CER drop (59.3%) in the case of 5 h of supervised data in Aishell is more than the relative CER drop (58.6%) of 15 h of supervised data, which also shows that when the amount of training data is smaller, the improvement of our system is more obvious.

#### 5.6 Performance of AF-Based Semi-supervised Multilingual ASR System Across Datasets

In order to verify the scalability of the method proposed, 100 h of Chinese data (Aidatatang) is added for experiments. The 100 h speech corpus is randomly selected from the 139 h Aidatatang dataset. The experiments are implemented together with 135 h of Aishell unsupervised data and the results are shown in the bottom line of Table 3. After adding a certain size of out-domain unsupervised data, there is a 4.1% relative CER decrease comparing to the original experimental results, which shows the scalability of our method again.

## 6 Conclusions

In this paper, a semi-supervised learning method incorporating articulatory attributes is proposed. In order to solve the problem of lacking the linguistic knowledge in low-resource languages, this method combines Monte Carlo

dropout with articulatory attributes, retains dropout in the model inference stage and decodes unsupervised data multiple times to obtain the dropout-lattice, which enables semi-supervised learning to utilize the unsupervised data and transfer the linguistic knowledge from resource-rich to low-resource languages. Additionally, the MFCCs-based-lattice and the AFs-based lattice are combined for LF-MMI training of unsupervised data. Experiments show that semi-supervised learning using Monte Carlo dropout and articulatory attributes have relative declines of 58.6% and 7.5% compared to the baseline system. The experiments also show that the smaller the supervised data, the greater the enhancement is. The method is further verified on the out-domain dataset, which can also achieve a relative CER reduction of 4.1%. In the future, we will analyze the commonalities and characteristics between languages from an in-depth level, and expand the method of this paper to more application scenarios, such as robust speech recognition, speaker verification.

**Acknowledgments.** This work is supported by National Nature Science Foundation of China No.11590772, No.62071039, and Beijing Natural Science Foundation L22303.

## References

1. Waibel, A., et al.: Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust. Speech Sig. Process.* **37**(3), 328–339 (1989)
2. Povey, D., et al.: Semi-orthogonal low-rank matrix factorization for deep neural networks. *Interspeech* (2018)
3. Toshniwal, S., et al.: Multilingual speech recognition with a single end-to-end model. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2018)
4. Lee, C.-H., et al.: An overview on automatic speech attribute transcription (ASAT). Eighth Annual Conference of the International Speech Communication Association (2007)
5. Ganin, Y., et al.: Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17**(1), 2096–2030 (2016)
6. Qu, L., Weber, C., Lakomkin, E., Twiefel, J., Wermter, S.: Combining articulatory features with end-to-end learning in speech recognition. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) *ICANN 2018*. LNCS, vol. 11141, pp. 500–510. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01424-7\\_49](https://doi.org/10.1007/978-3-030-01424-7_49)
7. Qian, Y., Liu, J.: Articulatory feature based multilingual MLPs for low-resource speech recognition. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
8. Zhan, Q., et al.: Domain-adversarial based model with phonological knowledge for cross-lingual speech recognition. *Electronics* **10**(24), 3172 (2021)
9. Wessel, F., Ney, H.: Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Trans. Speech Audio Process.* **13**(1), 23–31 (2004)
10. Povey, D., et al.: Purely sequence-trained neural networks for ASR based on lattice-free MMI. In: *Interspeech* (2016)

11. Manohar, V., et al.: Semi-supervised training of acoustic models using lattice-free MMI. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2018)
12. Labach, A., Salehinejad, H., Valaee, S.: Survey of dropout methods for deep neural networks. arXiv preprint [arXiv:1904.13310](https://arxiv.org/abs/1904.13310) (2019)
13. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
14. Tong, S., et al.: Unbiased Semi-Supervised LF-MMI Training Using Dropout. In: INTERSPEECH (2019)
15. Panayotov, V., et al.: LibriSpeech: an ASR corpus based on public domain audio books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2015)
16. Pratap, V., et al.: MLS: A large-scale multilingual dataset for speech research. arXiv preprint [arXiv:2012.03411](https://arxiv.org/abs/2012.03411) (2020)
17. Bu, H., et al.: AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline. In: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). IEEE (2017)



# Collaborative Transformer Decoder Method for Uyghur Speech Recognition in-Vehicle Environment

Jiang Zhang, Liejun Wang<sup>(✉)</sup>, Yinfeng Yu, Miaomiao Xu,  
and Alimjan Mattursun

College of Computer Science and Technology, Xinjiang University,  
Urumqi, Xinjiang, China  
{zhangjiang,xmm,alim}@stu.xju.edu.cn, {wljxju,yuyinfeng}@xju.edu.cn

**Abstract.** In-vehicle automatic speech recognition plays a crucial role in the field of autonomous driving and in-car voice assistants, with one of the significant factors affecting recognition accuracy being noise interference in the vehicle environment. Most advanced automatic speech recognition methods are oriented towards training resource-rich languages and have some limitations for other small languages. This paper proposes a Collaborative Transformer Decoder Method (CTDM) for a Low-resource Uyghur Speech Recognition in-vehicle Environment. In the encoder, we adopt collaborative encoding to capture multi-scale detail information while focusing on global information. In the decoder part, we design a parallel decoding strategy for arbitrary sequences, breaking the traditional left-to-right or right-to-left decoding order in Transformer decoding methods to establish associations between different characters. Our CTDM enhances the model's ability to extract detailed information, reduces the model's dependence on large-scale training data, and weakens the impact of noise on the model. Experiments are conducted on the Uyghur General Speech 7, 8, 9, and 16 datasets combined with vehicle noise and human speech noise to simulate real vehicle speech environments. Experimental results indicate that in a vehicular noise environment with a signal-to-noise ratio (SNR) of 0, the average word error rates (WER) on datasets 7, 8, 9, and 16 decreased by 29.8%, 15.9%, 8.7%, and 5.1%, respectively. In a vocal noise environment with an SNR of 0, the WER on datasets 7, 8, 9, and 16 decreased by 11.0%, 35.5%, 32.9%, and 13.8%, respectively.

**Keywords:** Vehicle Environment · Uyghur Language · Speech Recognition · Collaborative Transformer

## 1 Introduction

In-vehicle automated speech recognition refers to a technology that recognizes speech within the interior environment of a vehicle. It finds significant applications in autonomous driving and in-car voice assistants [3, 8, 13, 26]. Current

methods of in-vehicle speech recognition mainly include cloud-based [12], embedded, and hybrid approaches. Cloud-based methods are susceptible to signal interference, embedded methods are unsuitable for deployment when model parameters are too large, and although hybrid methods demonstrate better performance, they are still affected by signal and model parameter issues. However, regardless of the approach, the accuracy of automated speech recognition is crucial. One of the key factors affecting recognition accuracy is noise interference in the vehicle environment [4].

Currently, most advanced methods are based on the Transformer architecture for end-to-end speech recognition, which is more reliant on large-scale training datasets and is mainly targeted at training resource-rich languages. Traditional Transformer-based decoders have two inputs: one from the encoder’s output and the other from the decoding of previous production. During training, a fixed lower triangular matrix masks text information to prevent information leakage [25], allowing the model to process previously generated only unidirectionally. While such methods achieve good training results with ample training data, in cases of limited training data or strong data noise, the assistance of contextual semantic information becomes crucial for accurate speech recognition.

To address the aforementioned challenges, we propose a Collaborative Transformer Decoder Method (CTDM) for a Low-resource Uyghur Speech Recognition in-vehicle Environment. In the encoder, we employ collaborative encoding to focus on multi-scale detailed and global information simultaneously, enhancing the model’s sensitivity to noise features. In the decoder, we design a parallel decoding strategy for arbitrary sequences, breaking the traditional left-to-right or right-to-left decoding order in Transformer-based methods, thus establishing correlations between different characters. Through this approach, the model learns internal language knowledge, which somewhat mitigates the impact of noise. Furthermore, multiple parallel decoding processes enhance the model’s utilization of training data, reducing its reliance on large-scale training datasets. Our contributions are outlined as follows:

1. In the encoder phase, we designed a multi-scale information extraction module that collaborates with the Conformer model, enhancing the model’s ability to extract detailed information and mitigating the impact of noise.
2. In the decoder, we introduced a collaborative decoding module with random masking, enabling the model to learn internal language knowledge better, further reducing the impact of noise, and enhancing the model’s utilization of training data.
3. In a signal-to-noise ratio of 0 noise environment, our CTDM method achieved outstanding performance.

## 2 Related Work

Traditional speech recognition methods rely on acoustic features and speech models to match speech signals with predefined patterns [21]. However, they struggle with complex speech signals and unstable speech variations and require

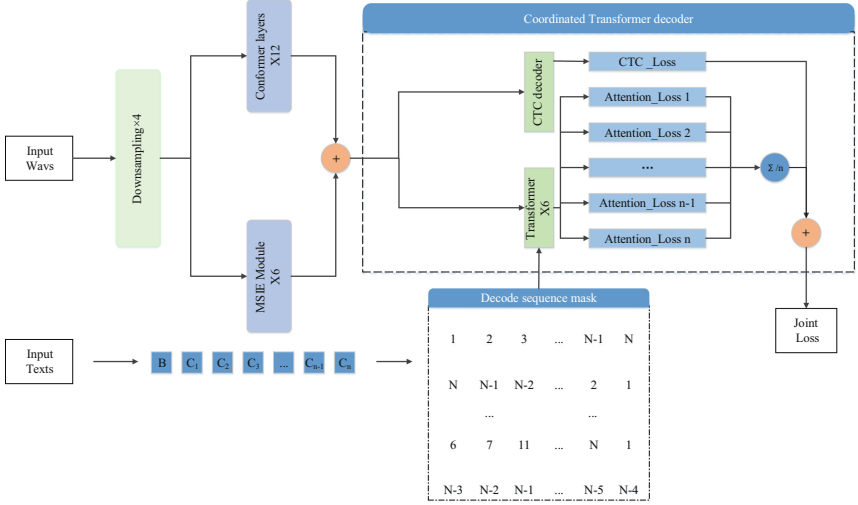
extensive manual feature engineering, limiting their generalization ability across different speech signals. With the advancement in computational capabilities, Hidden Markov Models (HMMs) [15] emerged as the mainstream approach for speech recognition. HMMs model speech signals and combine them with pronunciation dictionaries and language models for recognition [7]. However, they exhibit poor performance in recognizing long speech sequences and are sensitive to variations and noise in speech signals.

With the advancement of deep learning, Convolutional Neural Networks (CNN) [19] and Recurrent Neural Networks (RNN) [17] have been introduced into speech recognition tasks to improve feature extraction and sequence modeling capabilities. CNN excels at extracting local features and handling noise, and with Connectionist Temporal Classification (CTC) [10, 18], they address alignment issues in speech recognition, allowing models to learn output sequences directly from input speech without manual alignment. However, CTC assumes frame independence and suffers from label sparsity, resulting in poor handling of long sequence data and limited noise robustness. Additionally, CNN has limited receptive fields and struggles with effectively modeling long-term dependencies. Early RNN models enhanced sequence modeling capabilities, but traditional RNN models faced issues like vanishing and exploding gradients. Long Short-Term Memory Networks (LSTM) [14] addressed these problems, improving modeling capabilities for long sequences, albeit sequentially. Meanwhile, Transformer [9] models, leveraging self-attention mechanisms, enable parallel processing of sequence data, achieving significant performance improvements. Nevertheless, Transformer models still have limitations in handling local information. Scholars later combined CNN and Transformer models to tackle this issue, such as stacking CNN and attention mechanisms in Transformer models to capture local details while establishing long-term dependencies [2, 11, 27]. In recent years, end-to-end models [22, 29] have gained increasing attention in speech recognition. These models directly take speech signals as input and output corresponding text, eliminating traditional feature extraction and alignment steps and simplifying the system with promising results. However, their recognition accuracy may be inferior to traditional methods for complex speech signals and scenarios. Moreover, existing end-to-end models mostly rely on the Transformer architecture and can only capture forward semantic information. The emergence of bidirectional Transformers partially addresses this issue but still lacks perception capabilities between arbitrary characters.

In our proposed method, the encoder utilizes a multi-scale information extraction module in collaboration with the Conformer model. This module incorporates parallel multi-layer convolutional structures of different sizes to enhance the model's noise-handling capabilities and refine feature extraction. In the decoder part, correlations between arbitrary characters are established by employing a parallel decoding strategy for arbitrary sequences, allowing the model to learn internal language knowledge and somewhat mitigate the impact of noise. Semantic information between words is captured, enabling the model to perform well even with smaller datasets. Additionally, multiple parallel decoding

processes enhance the model’s utilization of training data and reduce its reliance on large-scale training datasets during training. The model adopts a unidirectional decoding approach to improve decoding efficiency during inference.

### 3 Methodology



**Fig. 1.** Overall Structure of the Collaborative Transformer Decoder Method

Our method, CTDM, follows an encoder-decoder architecture. The overall structure of our model is shown in Fig. 1.

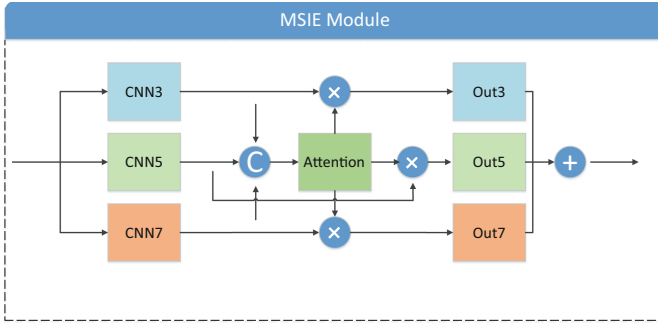
#### 3.1 Encoder

The encoder consists of a downsampling module, a Multi-scale Information Extraction (MSIE) Module, and a Conformer module.

Speech signals are sparse. The primary goal of the downsampling module is to reduce the data volume of the speech signal and simplify the model complexity. The input features  $x$  to the downsampling module have the shape of  $[B, T, F]$ , where  $B$  denotes the batch size,  $T$  represents the time length, and  $F$  is the feature dimension. In this experiment, we utilize 80-dimensional Fbank features. During the downsampling process, we first expand the dimensions of the features to  $[B, 1, T, F]$ . Subsequently, we apply two 2D convolutions with kernel sizes 3 and a stride of 2, each producing 256 channels. This operation downsamples both the speech length and feature dimension to one-fourth of their original sizes. Consequently, the shape of the feature maps becomes  $[B, 256, T/4, F/4]$ .

The feature maps are transformed to  $[B, T/4, 256 * F/4]$  following a reshape operation. Finally, after passing through a linear layer, the shape of the feature maps becomes  $[B, T/4, 256]$ . The downsampling process is represented by the (1).

$$x_d = \text{Linear}(\text{reshape}(\text{Conv2D}(\text{Expand}(x)))) \quad (1)$$



**Fig. 2.** MSIE module

The Multi-scale Information Extraction (MSIE) Module is designed to capture multi-scale features from speech data, refining feature extraction and improving the model's noise robustness. We employed six layers in our experiments, as illustrated in Fig. 2. The input  $x_d$  to MSIE is derived from the output of a downsampling module. It undergoes convolution with kernels of sizes 3, 5, and 7, resulting in three distinct scale features:  $x_{c3}$ ,  $x_{c5}$ , and  $x_{c7}$ . These feature maps are concatenated along the channel dimension and subsequently processed through convolutional layers to reduce their dimensionality to a single channel. Each resulting feature map is activated using the Sigmoid function to produce attention vectors. These attention vectors are then element-wise multiplied with  $x_{c3}$ ,  $x_{c5}$ , and  $x_{c7}$  respectively, and the resultant features are aggregated to obtain the final feature representation. The process is represented by the (2).

$$x_m = \sum_{i \in \{3,5,7\}} (\text{Sigmoid}(\text{Conv}(\text{Concat}(x_{ci}))) \odot x_{ci}) \quad (2)$$

where  $x_{ci}$  represents the feature map obtained from convolution with kernel size  $i$ ,  $\odot$  denotes the element-wise multiplication.

The Conformer module consists of 12 layers, primarily composed of the Multi-Head Attention (MHA) Mechanism, Feedforward Neural Network (FFN), Convolution, Layer Normalization (LN), and Skip Connections. The convolution and multi-head attention modules are situated between two Feed-Forward modules,



with a convolution kernel size of 15. We represent this using (3).

$$\begin{aligned}
 x_1 &= 1/2 * \text{FFN}(x_d) + x_d \\
 x_2 &= \text{MHA}(x_1) + x_1 \\
 x_3 &= \text{Conv}(x_2) + x_2 \\
 x_{con} &= \text{LN}(1/2 * \text{FFN}(x_3) + x_3)
 \end{aligned} \tag{3}$$

where  $x_{con}$  represents the final output of the conformer module. This output is then added to the output of the MSIE module,  $x_m$ , to serve as the final output of the encoder.

### 3.2 Decoder

The decoder comprises CTC, Transformer, and Attention\_Rescoring decoder. In the Transformer decoder, we employ a parallel collaborative decoding strategy.

The input  $x_E$  in the CTC decoder is the output of the encoder, obtained by adding the outputs of the MSIE and Conformer modules. CTC primarily utilizes a Linear layer to generate corresponding logits.

The Transformer decoder employs a 6-layer Transformer architecture, primarily composed of two MHA modules and one Multi-Layer Perceptron (MLP). The input to the first MHA in the decoder includes context information  $c$  and the mask  $m$ . The mask  $m$  is generated using Our Collaborative Transformer Decoder (CTD) to introduce random masking. The first MHA is represented by (4).

$$h_c = c + \text{MHA}(c, c, c, m) \tag{4}$$

The input to the second MHA module comprises the output  $h_c$  from the preceding MHA and the output  $x_E$  from the encoder.

$$h_i = h_c + \text{MHA}(h_c, x_E, x_E) \tag{5}$$

Finally,  $h_i$  undergoes MLP and linear layers to produce the logits of the transformer decoder.

$$\text{logits}_T = \text{Linear}(h_i) \tag{6}$$

This paper uses an alternative decoding method called Attention\_Rescoring. In this method, the decoding results from the CTC are fed into the attention model for re-decoding. This approach combines the advantages of the CTC decoder and attention decoder, achieving better performance by leveraging the strengths of each method. By utilizing the initial predictions from CTC as a guide, the attention model can refine these predictions, resulting in more accurate and reliable final outputs.

The typical Transformer decoding process starts with an initial point B, where both B and the encoder's output are fed into the network to generate the first character, C1, followed by sequential decoding of C2, C3, C4, and so on. The subsequent information is masked to prevent information leakage during training. In our CTD, after setting the initial point B, subsequent characters are

decoded randomly using different mask sequences. Figure 3 illustrates four randomly generated masked sequences produced by the Collaborative Transformer Decoder. It should be noted that we decode different sequences using various mask sequences while keeping the actual order of character labels unchanged. This random method is used only during training. During inference and testing, we decode using the conventional Transformer method.

Our experiments used 14 sets of masked sequences, resulting in 14 parallel decoders during training. Each decoder operates independently, reducing the risk of information leakage. Among the first seven masked sequences, one follows the traditional left-to-right decoding order, and the other is specifically designed for the Uyghur language. We observed that Uyghur often has verb-final sentences, with a significant distance between the subject and the verb, which can lead to the loss of dependency structures. We designed a special decoding sequence to address this issue, where the last four characters are decoded first. This means that after inputting the initial label B, decoding starts from the fourth-to-last character, then moves to the third-to-last, second-to-last, and finally the last character. Following this, the sequence is decoded from the first character onwards. This design brings the final verb to the beginning of the sentence, reducing the distance between the subject and the verb, thus alleviating the long-distance dependency problem. The remaining five masked sequences use random masking. The latter seven masked sequences are the reversed versions of the first seven.

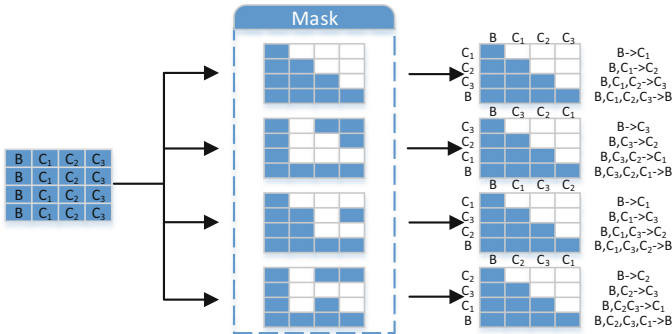
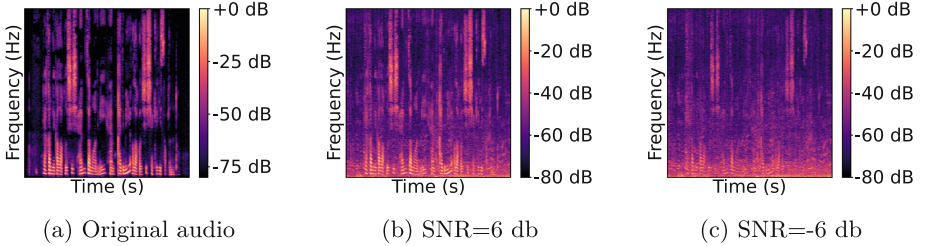


Fig. 3. Collaborative Transformer Decoder

## 4 Experiment

### 4.1 Datasets and Experimental Setup

In our experiments, we utilized four small-scale Uyghur language datasets: Mozilla Common Voice 7.0, 8.0, 9.0, and 16.1 [1], the total time for each dataset



**Fig. 4.** Spectrograms of audio files with different SNR

containing the training set, test set, validation set, and partially unavailable dirty data is 44, 64, 65, and 180 h, respectively. In this paper, we refer to them as Data 7, 8, 9, and 16 to represent these four datasets. For processing Uyghur text, we used the Byte-Pair Encoding (BPE) algorithm [24], which divides words into subwords based on frequency. In our experiments, we used BPE to split the vocabulary into 5,000 subwords. When processing speech data, we used the Librosa tool to extract 80-dimensional Fbank features as input for the network. These features accurately capture the spectral structure of the speech signal, thereby enhancing the model’s performance in speech recognition tasks. Additionally, we used noise data from the Thuyg20 [23] dataset, which includes two types of noise: vehicle noise and human voice noise. The paper employs Signal-to-Noise Ratio (SNR) to measure the relative strength between the signal and noise, as shown in Eq. 7. The SNR of zero indicates that signal and noise have equal intensity. We plotted the spectrogram of the original audio and those with different SNRs. As can be seen from Fig. 4, when the SNR is -6dB, there is a significant impact on the original data, with much of the detailed information lost. However, real-world SNR often fluctuates. To reflect this, we used Gaussian distribution when adding noise. That is, the added noise has an SNR with a mean of 0 and a variance of 1. This allows for controlling the randomness of the noise and simulating more realistic scenarios.

$$\text{SNR} = 10 \cdot \log_{10}\left(\frac{\text{SignalPower}}{\text{NoisePower}}\right) \quad (7)$$

Word Error Rate (WER) or Character Error Rate (CER) are commonly used evaluation metrics in speech recognition. For the Uyghur language, WER is generally adopted as the evaluation criterion. A lower WER indicates fewer errors, implying higher accuracy and better model performance.

In our experiments, our server’s CPU is an AMD Ryzen 5 5600, and it’s equipped with a graphics card with 24GB of VRAM, specifically an NVIDIA GeForce RTX 3090 Ti. Our hyperparameter settings are shown in the Table 1.

## 4.2 Comparison Experiments

To demonstrate the effectiveness of our CTDM, we conducted comparative experiments on Data 7, 8, 9, and 16 with added vehicle noise and human voice

**Table 1.** Experimental Parameter Settings.

Parameter Name	Parameter
Num_Conformer	12
Num_Decoder	6
Dff	2048
Epoch	140
Lr	0.0005
Num_Mel	80
Frame_Length	25
Frame_Shift	10
Spec_Aug	True
Accum_Grad	4

noise, comparing against existing 7 speech recognition methods. In the comparative experiments, Transformer [28] is a speech recognition method based on the Transformer architecture. Conformer [11] serves as our baseline, enhancing the encoder of the Transformer-based speech recognition architecture. Furthermore, our CTDM method employs a Collaborative Transformer Decoder in the decoding phase, allowing for flexible decoding strategies with learning of semantic information between arbitrary characters. Therefore, in our comparative analysis, we selected Conformer\_bi [20] and STBD [6], both capable of utilizing bidirectional context information to enhance speech recognition performance. Conformer\_bi utilizes a single decoder equipped with bidirectional context embedding for bidirectional decoding, whereas Squeeze\_bi employs a directional decoder comprising two distinct unidirectional decoders. Additionally, our CTDM method optimizes the encoder with a multi-scale information extraction module and collaborates with the Conformer model. Thus, this study compares Squeezeformer [16], Efficient Conformer V1 (E\_ConforV1) [5], and V2 (E\_ConforV2) [5]. Squeezeformer optimizes the encoder based on Conformer, significantly reducing model computational complexity and parameter count while maintaining high performance. E\_ConforV1 and E\_ConforV2 similarly optimize the encoder based on Conformer, substantially reducing computational complexity. V2 further improves upon V1 by reducing model parameters and computational complexity.

From the experimental data in Table 2 and Table 3, our CTDM performs excellently on datasets Data7, Data8, Data9, and Data16 when vehicle noise and human noise are added. A refers to the results of the Attention decoder, A\_r refers to the results of the Attention\_Rescoring decoder, and C\_g refers to CTC\_Greedy decoder. Compared to the baseline method Conformer, the Attention decoding WER in a vehicle noise environment decreased by 3.2%, 4.9%, 2.4%, and 0.8% on Data7, Data8, Data9, and Data16, respectively. In the human noise environment, the Attention decoding WER on these datasets decreased by 3.9%, 23.1%, 12%, and 2%, respectively. In the vehicle noise environment, the Atten-

**Table 2.** Comparison Experiment with Vehicle Noise

Vehicle SNR=0	Data 7			Data 8			Data 9			Data 16		
Model	A	A_r	C_g	A	A_r	C_g	A	A_r	C_g	A	A_r	C_g
Transformer	20.5	30.4	40.8	16.7	24.2	31.6	10.5	13.5	17.6	<b>9.9</b>	11.8	18.8
Conformer	21.8	50.2	58.3	10.0	23.1	30.9	10.0	14.1	21.2	10.8	11.7	17.5
Conformer_bi	<b>17.7</b>	18.9	25.4	9.2	18.4	29.8	9.2	14.3	18.6	10.5	7.6	12.0
STBD	22.8	31.1	41.6	17.8	19.2	22.3	15.7	20.3	30.1	14.0	6.6	10.3
Squeezeformer	24.4	31.2	40.8	16.6	19.0	26.6	15.3	45.2	54.5	13.5	8.7	14.1
E_ConforV1	115.2	98.9	99.1	114.3	98.7	99.8	11.1	15.3	19.8	10.0	7.6	9.3
E_ConforV2	111.4	99.1	99.8	17.7	23.5	29.7	11.0	16.4	22.0	<b>9.9</b>	6.5	<b>8.4</b>
CTDM(Ours)	18.6	<b>8.3</b>	<b>14.1</b>	<b>5.1</b>	<b>3.8</b>	<b>7.3</b>	<b>7.6</b>	<b>4.0</b>	<b>7.6</b>	10.0	<b>5.9</b>	8.7

**Table 3.** Comparison Experiment with Human Noise

Human SNR=0	Data 7			Data 8			Data 9			Data 16		
Model	A	A_r	C_g	A	A_r	C_g	A	A_r	C_g	A	A_r	C_g
Transformer	120.6	99.5	99.7	118.3	99.3	99.5	112.0	98.7	99.1	31.2	64.1	70.6
Conformer	77.9	84.0	86.4	40.1	72.2	76.7	35.1	66.4	70.5	18.4	35.1	43.4
Conformer_bi	110.9	99.1	99.9	107.5	99.0	99.1	115.0	98.9	99.6	17.9	17.0	23.2
STBD	<b>67.7</b>	69.2	74.9	<b>15.1</b>	31.6	41.0	23.6	44.1	53.6	19.6	24.0	31.7
Squeezeformer	72.6	90.1	91.6	25.0	67.5	74.1	24.0	54.5	63.9	20.8	24.8	32.9
E_ConforV1	131.5	99.8	99.7	115.9	99.4	99.9	113.1	99.4	99.7	18.8	19.7	26.5
E_ConforV2	105.6	98.9	99.4	114.8	99.1	99.8	28.6	52.0	59.9	17.9	18.4	25.0
CTDM(Ours)	74.0	<b>67.5</b>	<b>73.8</b>	17.0	<b>28.2</b>	<b>37.2</b>	<b>23.1</b>	<b>21.5</b>	<b>28.8</b>	<b>16.4</b>	<b>16.4</b>	<b>22.7</b>

tion\_Rescoring decoding WER on Data7, Data8, Data9, and Data16 decreased by 41.9%, 19.3%, 10.1%, and 5.8%, respectively. In the human noise environment, the Attention\_Rescoring decoding WER on these datasets decreased by 16.5%, 44%, 44.9%, and 18.7%, respectively. In the vehicle noise environment, the CTC\_Greedy decoding WER on Data7, Data8, Data9, and Data16 decreased by 44.2%, 23.6%, 13.6%, and 8.8%, respectively. In the human noise environment, the CTC\_Greedy decoding WER on these datasets decreased by 12.6%, 39.5%, 41.7%, and 20.7%, respectively. The above data shows that compared to the baseline method, CTDM shows significant improvements, especially on datasets Data7, Data8, and Data9. This is because our CTDM employs a parallel decoder with arbitrary sequences in the decoder, allowing for better learning of internal language knowledge. Additionally, during training, multiple parallel decoders enhance the model’s ability to utilize training data effectively, reducing dependence on large-scale training data. Hence, the improvements are particularly noticeable on the relatively smaller datasets Data7, Data8, and Data9.

From other comparative experiments, we can also observe that, in most cases, the model error rate decreases as data increases. Existing models are relatively

dependent on training data, performing well when the training data is relatively abundant. This is particularly evident in the Squeezeformer, E\_ConforV1, and E\_ConforV2 methods. As seen from the data in Table 2 and Table 3, these methods perform prominently on the relatively larger dataset Data16 but poorly on the smaller datasets Data7, Data8, and Data9. This is because, with smaller training data, models based on the Transformer architecture are less likely to converge. Additionally, the Squeezeformer, E\_ConforV1, and E\_ConforV2 methods are lightweight improvements based on the Conformer, making it difficult to extract effective features when the training data is limited. From the experimental data in Table 2, it can also be seen that in the vehicle noise environment, the recognition WER of the Transformer, Conformer, Conformer\_bi, and STBD methods decreases as the amount of training data increases. Among them, Conformer\_bi performs the best. This is because it combines convolutional neural networks' local feature extraction capabilities with the global feature extraction capabilities of attention mechanisms and uses bidirectional LSTM decoding during the decoding process. Although the STBD method uses bidirectional LSTM decoding, its encoder performs downsampling operations, losing some information. Therefore, it performs worse than Conformer\_bi when the dataset is smaller. However, when the training data is relatively abundant, its performance is comparable to that of Conformer\_bi.

From the experimental data in Table 3, it can be observed that in the Human Noise environment, most methods perform poorly on the small datasets Data7, Data8, and Data9, unlike Table 2. Conformer\_bi performs weaker than Conformer on these datasets. This is because human voice noise is quite similar to the content to be recognized. When the dataset is small, Conformer\_bi cannot effectively distinguish between noise and the target. As a result, Conformer\_bi is less sensitive to indistinguishable noise. With more abundant training data, Conformer\_bi can effectively learn valid features, leading to superior performance on Data16 compared to Conformer. In contrast, STBD, which undergoes lightweight processing in its encoder, performs better than Conformer\_bi on the relatively small datasets Data7, Data8, and Data9. This is because STBD filters out some noise during downsampling, which enhances its performance on these datasets. However, with more extensive training data, STBD's lightweight improvements inevitably result in the loss of some effective features, leading to weaker performance on Data16 compared to Conformer\_bi. It's notable that in Table 2, STBD performs worse than Conformer\_bi on Data7, Data8, and Data9, whereas in Table 3, it outperforms Conformer\_bi. This difference arises because human noise features are less distinct from recognition targets, making STBD more sensitive during downsampling, which improves its performance in Table 3. In contrast, vehicle noise features differ significantly from recognition targets, where Conformer\_bi's strong capability in detail handling enables better differentiation and recognition, thus outperforming STBD in Table 2.

From the experimental data in Table 2 and Table 3, it can be observed that the WER in Table 3 are generally higher than those in Table 2. This is because vehicle noise features differ significantly from recognition targets, resulting in

less impact on recognition methods and lower dependence on training data. In contrast, human noise features are less distinct from recognition targets, leading to a greater impact on recognition methods and, hence, higher dependence on training data.

### 4.3 Ablation Experiments

In the encoder part of our CTDM, we employed collaborative encoding by designing an MSIE module to collaborate with the Conformer. This module enables the model to focus on global information while capturing multi-scale detailed information, thereby enhancing the model’s sensitivity to noise features. In the decoder part, we improved upon the Transformer decoder by designing an arbitrary sequence parallel decoding strategy, CTD, which allows for correlations between different characters, facilitating better learning of intra-character semantic information. We conducted ablation studies on MSIE and CTD separately in vehicle and human noise environments on datasets Data7, Data8, and Data16, as shown in Table 4. Here, Av represents the average of the three types of decoding WER.

**Table 4.** Ablation Experiments with MSIE and CTD Modules

Noise	MSIE	CTD	Data 7				Data 8				Data 16			
			A	A_r	C_g	Av	A	A_r	C_g	Av	A	A_r	C_g	Av
Vehicle	✓	✓	18.6	8.3	14.1	13.7	5.1	3.8	7.3	5.4	10.0	5.9	8.7	8.2
	✓	✗	20.5	20.4	29.1	23.3	8.4	15.2	21.8	15.1	10.6	10.4	15.1	12.0
	✗	✓	19.7	13.6	19.9	17.7	7.5	10.2	17.4	11.7	10.0	9.5	12.7	10.7
	✗	✗	21.8	50.2	58.3	43.4	10.0	23.1	30.9	21.3	10.8	14.1	21.2	15.4
Human	✓	✓	74.0	67.5	73.8	71.8	17.0	28.2	37.2	27.5	16.4	16.4	22.7	18.5
	✓	✗	76.4	78.6	80.7	78.6	34.4	56.2	60.7	50.4	18.1	29.3	40.8	29.4
	✗	✓	75.2	70.8	76.8	74.3	25.8	48.9	54.5	43.1	17.2	22.6	30.4	23.4
	✗	✗	77.9	84.0	86.4	82.8	40.1	72.2	76.7	63.0	18.4	35.1	43.4	32.3

According to the experimental data in Table 4, we observe that using the CTD strategy on datasets Data7, Data8, and Data16 reduces the average WER under vehicle noise by 9.6%, 9.7%, and 3.8% respectively, and under human noise by 6.8%, 22.9%, and 10.9% respectively. These results demonstrate that the CTD decoding strategy, which correlates arbitrary characters and enhances the learning of inter-character semantic information, effectively mitigates the impact of vehicle and human noise on recognition results. Particularly noteworthy is the outstanding performance of the CTD strategy on the relatively smaller datasets Data7 and Data8, highlighting the effectiveness of CTD. Using the MSIE module on datasets Data7, Data8, and Data16 reduces the average WER under vehicle noise by 4.0%, 6.3%, and 2.5%, respectively, and under human noise by

2.5%, 15.6%, and 4.9% respectively. These findings indicate that the MSIE module enhances the capability to capture multi-scale detailed information, thereby better-distinguishing noise from target features and increasing the model’s sensitivity to noise features. Notably, the MSIE module performs exceptionally well on the relatively smaller dataset Data8, confirming its effectiveness. Combining both CTD and MSIE on datasets Data7, Data8, and Data16 reduces the average WER under vehicle noise by 29.7%, 15.9%, and 7.2%, respectively, and under human noise by 11.0%, 35.5%, and 13.8% respectively. These results highlight that the CTDM approach proposed in this paper effectively reduces recognition of WER in low-resource noise environments.

To demonstrate the robustness of our method under varying degrees of human noise, we conducted ablation studies in different levels of human noise environments, setting up control groups with varying SNRs. The experimental results are shown in Table 5.

**Table 5.** Ablation Experiment with Different SNR Levels with Human Noise

Model	SNR	Data 7				Data 8				Data 16			
		A	A_r	C_g	Av	A	A_r	C_g	Av	A	A_r	C_g	Av
base	-6	107.0	96.9	97.1	100.3	90.8	92.5	93.3	92.2	54.7	65.0	70.4	63.3
	-3	89.7	90.0	91.3	90.3	56.4	78.9	82.6	72.6	37.7	47.1	54.8	46.6
	0	77.9	84.0	86.4	82.8	40.1	72.2	76.7	63.0	18.4	35.1	43.4	32.3
	3	63.6	80.1	83.1	75.6	12.9	44.5	53.5	37.0	13.1	32.7	41.5	29.1
	6	61.6	75.0	78.7	71.8	10.3	34.6	44.6	29.8	11.9	23.8	31.8	22.5
ours	-6	98.9	98.7	99.5	99.0	50.1	70.1	75.7	65.3	45.1	52.3	58.9	52.1
	-3	86.2	84.6	87.5	86.1	19.1	30.7	39.1	29.6	30.0	43.5	51.9	41.8
	0	74.0	67.5	73.8	71.8	17.0	28.2	37.2	27.5	16.4	16.4	22.7	18.5
	3	41.5	27.0	36.8	35.1	6.6	8.2	13.5	9.4	11.4	9.4	13.5	11.4
	6	35.1	18.9	27.2	27.1	5.0	5.9	10.5	7.1	9.1	7.2	10.5	8.9

From the experimental data in Table 5, it is evident that when the SNR is 6, indicating minimal noise interference, the baseline method Conformer exhibits average WER of 71.8%, 29.8%, and 22.5% on datasets Data 7, Data 8, and Data 16, respectively. Our method, CTDM, in contrast, achieves an average WER of 27.1%, 7.1%, and 8.9% on the same datasets. Our CTDM model achieves a significantly lower WER on these datasets compared to the Conformer baseline. When the SNR is -6, indicating maximum noise interference, CTDM and Conformer perform poorly on Data 7, primarily due to strong noise interference. However, on Data 8 and Data 16, our method CTDM outperforms Conformer by a considerable margin, demonstrating that CTDM can converge on smaller datasets. This observation is further supported when the SNR is 0. At SNR 0, our method CTDM exhibits lower WER than Conformer on the relatively smaller dataset Data 7 and demonstrates excellent performance on Data 8 and Data 16.



#### 4.4 Cases Analysis

Conformer: رەقەملىك ئالاقلىشىش ھەم زامانىۋى ھەم قەدىمىي ئالاقلىشىش شەكلى ھېسابلىنىدۇ  
 Ours/Label: رەقەملىك ئالاقلىشىش ھەم زامانىۋى ھەم قەدىمىي ئالاقلىشىش شەكلى ھېسابلىنىدۇ

**Fig. 5.** Error Analysis

We listed the errors in speech recognition and analyzed them, as shown in Fig. 5. Our model produced correct results in the figure, aligning with the ground truth, whereas the Conformer model had two errors. Although both sentences convey the same meaning, the Conformer model made two mistakes: Error 1 is related to a suffix issue, where the words represent “modern” and “modernization.” Due to less effective context utilization, the Conformer model made a suffix error during recognition. Our model, which fully leverages contextual information, resolved the suffix error. Error 2 was caused by a highly unclear pronunciation in the original speech, especially in a noisy environment. The Conformer model failed to capture this detail, leading to misrecognition. Our Multi-scale Information Extraction (MSIE) module effectively compensated for this drawback, capturing the subtle information and thus achieving accurate recognition.

### 5 Conclusion

In-vehicle Uyghur speech recognition encounters significant challenges due to noise and other environmental disturbances. To address these challenges, this paper proposes a Collaborative Transformer model that introduces a Multi-Scale Information Extraction module at the encoder stage and optimizes decoding sequences at the decoder stage to coordinate multiple decoder sequences, thereby enhancing contextual information extraction.

Experiments were conducted on four datasets from Common Voice 7, 8, 9, and 16.1. The model’s performance was evaluated in vehicular and human noise environments with a signal-to-noise ratio (SNR) of 0. The average WER was calculated using three decoding methods (Attention, Attention\_Rescoring, and CTC\_Greedy). Compared to the baseline model, our model achieved the following improvements: In the vehicular noise environment, the average WER on datasets 7, 8, 9, and 16 decreased by 29.8%, 15.9%, 8.7%, and 5.1%, respectively. In the human noise environment, the average WER on datasets 7, 8, 9, and 16 decreased by 11.0%, 35.5%, 32.9%, and 13.8%, respectively. These results indicate that the Collaborative Transformer model substantially improves speech recognition performance in vehicle environments, enhancing the system’s accuracy and stability across various noisy conditions. However, we also recognize room for improvement in the model’s noise resistance. In the future, we will further enhance the model’s noise resistance capabilities.

**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. Mozilla common voice. <https://commonvoice.mozilla.org/zh-CN/datasets>
2. Andrusenko, A., Nasretidinov, R., Romanenko, A.: Uconv-conformer: high reduction of input sequence length for end-to-end speech recognition. In: 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
3. Banerjee, A., Maity, S.S., Banerjee, W., Saha, S., Bhattacharyya, T.: Facial and voice recognition based Ssecurity and safety system in car. In: 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), pp. 812–814. IEEE (2020)
4. Braun, S., Gamper, H.: Effect of noise suppression losses on speech distortion and ASR performance. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 996–1000. IEEE (2022)
5. Burchi, M., Vielzeuf, V.: Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 8–15. IEEE (2021)
6. Chen, X., Zhang, S., Song, D., Ouyang, P., Yin, S.: Transformer with bidirectional decoder for speech recognition. arXiv preprint [arXiv:2008.04481](https://arxiv.org/abs/2008.04481) (2020)
7. Cui, X., Gong, Y.: A study of variable-parameter gaussian mixture hidden markov modeling for noisy speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **15**(4), 1366–1376 (2007)
8. Deng, M., et al.: Using voice recognition to measure trust during interactions with automated vehicles. *Appl. Ergon.* **116**, 104184 (2024)
9. 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), pp. 5884–5888. IEEE (2018)
10. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 369–376 (2006)
11. Gulati, A., et al.: Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint [arXiv:2005.08100](https://arxiv.org/abs/2005.08100) (2020)
12. Homma, T., Obuchi, Y., Shima, K., Ikeshita, R., Kokubo, H., Matsumoto, T.: In-vehicle voice interface with improved utterance classification accuracy using off-the-shelf cloud speech recognizer. *IEICE Trans. Inf. Syst.* **101**(12), 3123–3137 (2018)
13. Ivanko, D., Ryumin, D., Axyonov, A., Kashevnik, A.: Speaker-dependent visual command recognition in vehicle cabin: methodology and evaluation. In: Karpov, A., Potapova, R. (eds.) *SPECOM 2021*. LNCS (LNAI), vol. 12997, pp. 291–302. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-87802-3\\_27](https://doi.org/10.1007/978-3-030-87802-3_27)

14. Jorge, J., et al.: LSTM-based one-pass decoder for low-latency streaming. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7814–7818. IEEE (2020)
15. Juang, B.H., Rabiner, L.R.: Hidden markov models for speech recognition. *Technometrics* **33**(3), 251–272 (1991)
16. Kim, S., et al.: Squeezeformer: an efficient transformer for automatic speech recognition. *Adv. Neural. Inf. Process. Syst.* **35**, 9361–9373 (2022)
17. 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)
18. Lee, J., Watanabe, S.: Intermediate loss regularization for CTC-based speech recognition. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6224–6228. IEEE (2021)
19. Liang, L., Zhang, Y., Zhang, S., Li, J., Plaza, A., Kang, X.: Fast hyperspectral image classification combining transformers and SimAM-based CNNs. *IEEE Trans. Geosci. Remote Sens.* **61**, 5522219 (2023)
20. Liao, L., et al.: A bidirectional context embedding transformer for automatic speech recognition. *Information* **13**(2), 69 (2022)
21. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* **77**(2), 257–286 (1989)
22. Rouhe, A., Grósz, T., Kurimo, M.: Principled comparisons for end-to-end speech recognition: attention vs hybrid at the 1000-hour scale. *IEEE/ACM Trans. Audio Speech Lang. Process.* **32**, 623–638 (2023)
23. Rouzi, A., Shi, Y., Zhiyong, Z., Dong, W., Hamdulla, A., Fang, Z.: THUYG-20: a free Uyghur speech database. *J. Tsinghua Univ. (Science and Technology)* **57**(2), 182–187 (2017)
24. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725 (2016)
25. Vaswani, A., et al.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, vol. 30 (2017)
26. Wang, S., Cao, J., Sun, K., Li, Q.: SIEVE: secure In-Vehicle automatic speech recognition systems. In: *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*, pp. 365–379. USENIX Association, San Sebastian (2020). <https://www.usenix.org/conference/raid2020/presentation/wang-shu>
27. Xu, M., Zhang, J., Xu, L., Silamu, W., Li, Y.: Collaborative encoding method for scene text recognition in low linguistic resources: the Uyghur language case study. *Appl. Sci.* **14**(5), 1707 (2024)
28. Zhang, B., 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)
29. Zhang, J., Wang, L., Yu, Y., Xu, M.: Nonlinear regularization decoding method for speech recognition. *Sensors* **24**(12), 3846 (2024)



# Audio-Visual Wake-up Word Spotting Under Noisy and Multi-person Scenarios

Cancan Li, Fei Su, and Juan Liu<sup>(✉)</sup>

Institute of Artificial Intelligence, School of Computer Science,  
Wuhan University, Wuhan 430072, China  
liujuan@whu.edu.cn

**Abstract.** The existing audio-visual wake-up word spotting (AVWWS) methods assume that the audio signal has been aligned with the lip movement video signal of a specific speaker in noisy environments, and are mainly applicable for scenarios with only a single speaker. However, in complex scenarios, there may be multiple people showing up in the video facing the camera simultaneously, and more than one person may be speaking at the same time. Wake-up word spotting in noisy and multi-person scenarios remains relatively under-explored. In this paper, we first propose a Wake-up Word Active Speaker Detection Model (WWASD) to recognize the face that is speaking the wake-up word. Based on the model, we propose two approaches, namely Two-stage detection and Three-stage detection, for audio-visual wake-up word spotting in noisy and multi-person scenarios. We compare the approaches from the perspectives of performance and computational complexity on MISP2021-AVWWS corpus. The best Two-stage detection approach, which contains WWASD and audio-visual wake-up word spotting model, achieves comparable performance against the systems with oracle visual speaker bounding boxes. Three-stage detection, which adds an audio-based single-modality wake-up word model as a front end greatly reduces the computational cost.

**Keywords:** Wake-up word spotting · Active speaker detection · Audio-visual modeling

## 1 Introduction

As the front end of voice assistant on smart terminal devices such as mobile phones, watches, headphones and speakers, Wake-up Word Spotting (WWS), also known as Keyword Spotting (KWS) or Wake-up Word Recognition, has achieved satisfactory performance in normal conditions. This task checks whether the text corresponding to the input voice is the same as the wake-up

---

Supported by the National Key Research and Development Program of China (1502-211171515).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025  
A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15333, pp. 170–184, 2025.  
[https://doi.org/10.1007/978-3-031-80136-5\\_12](https://doi.org/10.1007/978-3-031-80136-5_12)

word so that the back-end service can be switched from the sleep state to the working state.

However, in complex scenarios, the voice signal has strong noise, and the accuracy of wake-up word spotting in pure speech is greatly reduced. Lip movement video information is insensitive to audio noise, and can be used as a good supplement to the audio modality. Therefore, more and more works have chosen to construct multimodal wake-up word models based on joint audio and video modalities [1–6]. They achieved good performance in the WWS task. However, the active speaker information is given in most of the existing audio and video wake-up word spotting corpus [7]. The existing audio-visual wake-up word spotting methods assume that the audio signal has been aligned with the lip movement video signal of a specific speaker in noisy environments. That is to say, they assume that both the speech audio and the lip movement video signals come from the same speaker, lacking the ability to automatically identify the active speaker who speaks the wake-up word from multiple persons showing in the video. Hence, the aforementioned existing methods are mainly applicable for scenarios with only a single speaker. However, in complex scenarios, there may be multiple people showing up in the video facing the camera simultaneously, and more than one person may be speaking at the same time. In such a case, the recorded video may contain multiple human faces with lip movement, which makes the current audio-visual wake-up word spotting methods difficult to correctly match the wake-up word audio signal to the right speaker who said the word. It is necessary to first detect the active speaker in order to crop the correct video information and feed to the audio-visual wake-up word spotting model.

Our objective in this work is to solve the problem of multimodal wake-up word spotting in noisy and multi-person scenarios. We first develop a model that focuses on finding the face that is speaking the wake-up word. The task concerns active speaker detection (ASD) and we call the model Wake-up Word Active Speaker Detection Model (WWASD). There is limited work now designing a model that recognizes the face of active speaker speaking the wake-up word in multi-face and noisy scenarios. Based on this model, we further propose two audio-visual wake-up word spotting approaches in complex scenarios to achieve simultaneous recognition of active speakers and robust wake-up word spotting. We compare these solutions in detail in terms of performance and computational complexity, so that the solution can be selected according to the specific application needs.

## 2 Related Work

### 2.1 Multimodal Wake-up Word Spotting

Due to the robustness of visual information against audio noise, more and more researches are focusing on combining speech signals and speaker’s lip movement video to improve the recognition performance of wake-up word spotting [1–6], automatic speech recognition (ASR) [8–10], speech separation [11, 12], speaker verification [13] and so on. With the first Multimodal Information based Speech Processing Challenge (MISP Challenge 2021 [7]) being held, many multimodal

wake-up word spotting models have been designed. Yanguang Xu et al. proposed three models which are A-Transformer, A-Conformer, and AV Transformer as subsystems in [2], and conducted multimodal information fusion through voting at the decision-making level. A model using visual-assisted minimum variance distortionless response (MVDR) and cross-attention between audio-visual modalities is proposed in [3]. Haoxu Wang et al. improved their previous models [5,6] in [4], and proposed the Frame Level Cross Modal Attention (FLCMA) mechanism, which can help model audio-visual information at the frame level by synchronizing lip movements and speech signals. The proposed model achieves new state-of-the-art results on the far-field MISP datasets.

## 2.2 Active Speaker Detection

Active speaker detection (ASD) aims at determining which person or none is speaking in a video at each time window [14,15], and it serves as a frontend for multimodal speech recognition (ASR) [14,16,17]. Ruijie Tao et al. proposed MuSED to learn the denoising ability for low-quality noisy videos and it is fine-tuned with the AV-ASD task in [18]. Cross-modal contrastive learning and positional encoding in the attention modules of supervised ASD models is applied in [14]. In [19,20], Otavio Braga et al. proposed an attention layer to the ASR encoder that is able to soft-select the appropriate face video track, and they improved the previous work by presenting a single model that can be jointly trained with a multi-task loss in [21].

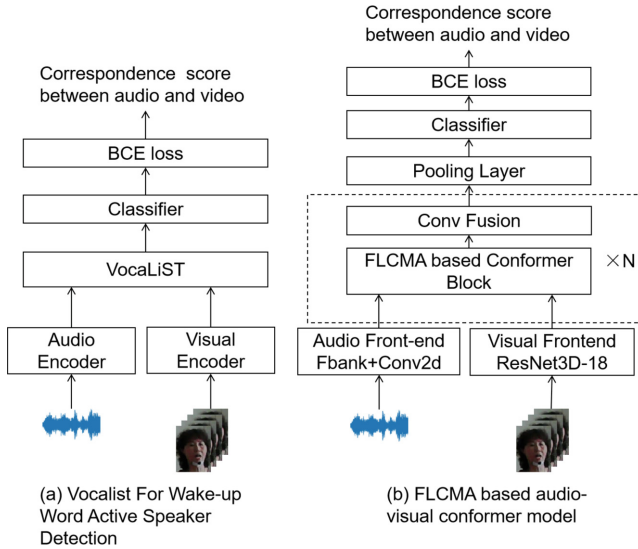
Lip synchronization model can also be used for ASD, due to that the speaker’s face is naturally the one with the highest correspondence between the audio and video signals [22]. The first deep learning based synchronization model SyncNet [22] proposed a two-stream ConvNet architecture that enables a joint embedding between the sound and the mouth images to be learnt from unlabelled data. A new learning strategy where the embeddings are learnt via a multi-way matching problem is used in [23]. Honglie Chen et al. compared a number of transformer-based architectural variants to model audio-visual synchronisation in [24]. Venkatesh S Kadandale et al. proposed a Transformer based audiovisual cross-modality model in [25], namely Vocalist, which outperformed several baseline models in the audiovisual synchronization task on the standard lip-reading speech benchmark dataset LRS2. At the same time, the special case of singing voice is also considered, and the model trained on the AV singing voice dataset Acappella also achieved state-of-the-art result.

In our work, we focus on detecting the active speaker with interactive intent and speaking wake-up words, so that the device can lock in the person with interactive intent in the video modality and constantly interact with the person later through audio-visual speech recognition or audio-visual speech extraction techniques.

## 3 Wake-Up Word Active Speaker Detection

To detect the active speaker, we compare two existing audio-video multimodal models and make them suitable for the task of wake-up word active speaker

detection, which are Vocalist [25] and FLCMA based audio-visual conformer model [4]. The frameworks of two models used for wake-up word active speaker detection are shown in Fig. 1.



**Fig. 1.** The frameworks of two models used for wake-up word active speaker detection.

### 3.1 Vocalist For Wake-up Word Active Speaker Detection

Vocalist [25] is originally designed for the audiovisual synchronization task. We extracted 64 frames of audio and video in this work instead of the 5 frames which is set in the paper [25].

Since the fps of the video is 25, 64 frames with 2.56s duration can cover the wake-up word length. Different from the Vocalist model, our active speaker detection model is designed for the wake-up word, and the input of 64 frames can directly compare the correspondence between video and audio at the sentence level, which is more convenient to compare the output score. We determine the active speaker in a multi-face scenario based on the fact that the face of the active speaker is the face with the highest correspondence between audio and video.

### 3.2 The Proposed WWASD Model

We use FLCMA based audio-visual Conformer as our final WWASD model, the same architecture as [4] which is originally used in AVWWS task. We re-design its training goal from detecting the wake-up word to the active speaker detection,

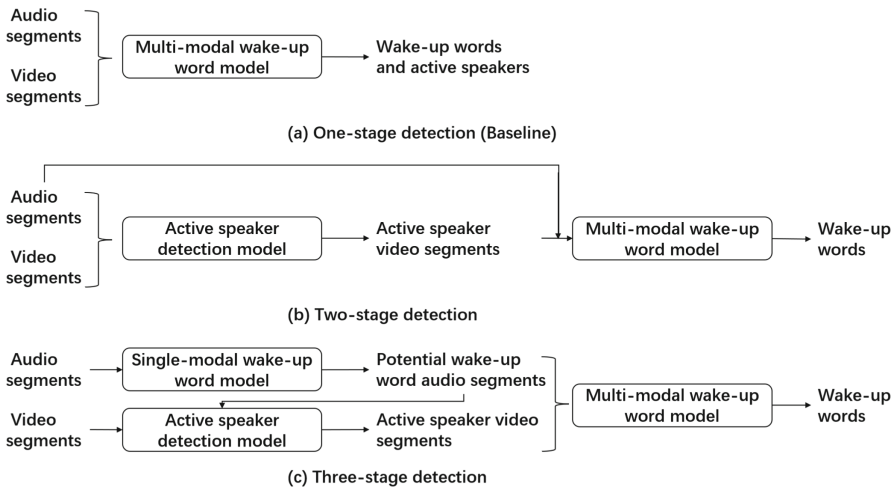
so that the output score represents the correspondence between audio and video, instead of the confidence score of whether is a wake-up word in AVWWS task.

The model obtains the complementary information between audio and video through the FLCMA based Audio-Visual Conformer Encoder for wake-up word spotting and obtains great performance, so we explore whether the encoder module can also obtain the synchronization information between audio and video if the model is applied to the wake-up word active speaker detection task.

6 self-attention blocks ( $N = 6$ ) and 256-dimensional hidden size were used in [4] for the FLCMA module based Conformer structure. In real application scenarios, there are often limitations in terms of computational complexity due to hardware reasons. Therefore, we test 6 self-attention blocks and 256-dimensional hidden size (WWASD-L), 4 self-attention blocks and 128-dimensional hidden size (WWASD-M) and 2 self-attention blocks and 64-dimensional hidden size (WWASD-S) respectively to decrease the computational cost.

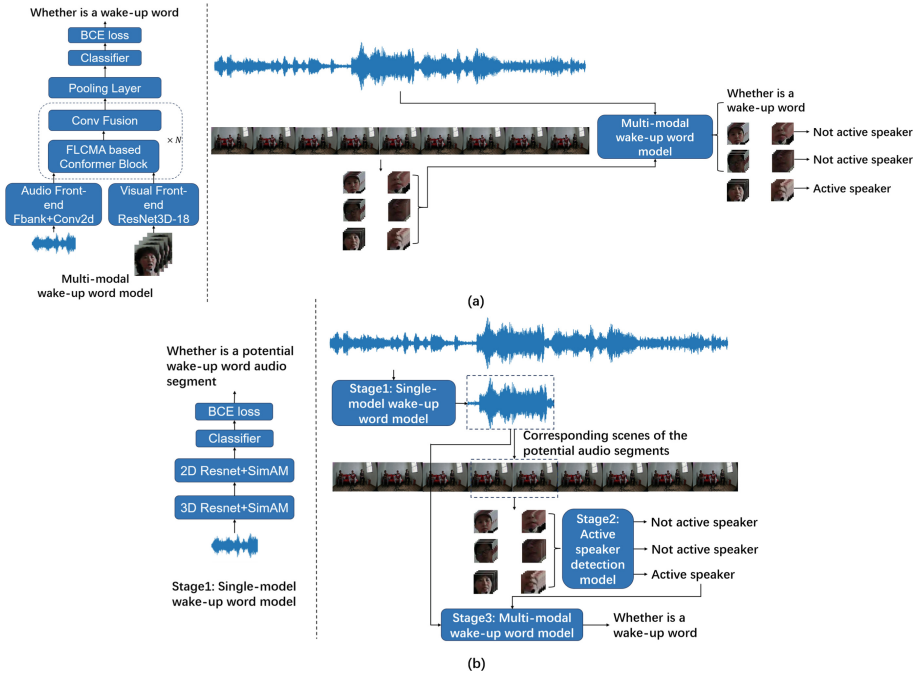
## 4 The Proposed Audio-Visual Wake-up Word Spotting Approaches

The overview diagram of wake-up word spotting approaches is shown in Fig. 2, and detailed picture is shown in Fig. 3 (Two-stage detection is not shown since Three-stage becomes Two-stage after removing the first stage).



**Fig. 2.** The overview diagram of compared approaches





**Fig. 3.** Detailed picture of compared approaches. (a) One-stage detection. (b) Three-stage detection.

#### 4.1 One-Stage Detection

The existing multimodal wake-up word spotting models [2–6] are trained in normal conditions that the audio signal has been aligned with the lip movement video signal of the speaker. In noisy and multi-person scenarios, they achieve lower recognition accuracy. In such scenarios, the multimodal wake-up word model detects the active speaker directly. The model first calculates all the scores of audio and corresponding videos of faces in a scene. If the highest score in the scene is greater than the threshold, it is judged as a wake-up word, and the person with the video corresponding to the highest score is selected as the active speaker. The behind reason is that a correctly aligned audio-visual signal pair could generate a higher score in judging the wake-up word. We choose the SOTA model in the MISP2021 wake-up word competition [4] for the task and the model serves as the baseline for audio-visual wake-up word spotting under noisy and multi-person scenarios.

#### 4.2 Two-Stage Detection

In order to increase the accuracy of Active Speaker Detection, we propose a specific Wake-up Word Active Speaker Detection Model (WWASD) to detect

the active speaker who is speaking the wake-up words and has interactive intent with the device. We use the model as the first stage, and the multimodal wake-up word spotting model as the second stage.

In this approach, every possible audio-lip pair in a scene is fed into the WWASD model to figure out the correct active speaker audio-video pair. The filtered video and corresponding audio in the scene are input into the back-end multimodal wake-up word spotting model, to detect whether it is a wake-up word.

### 4.3 Three-Stage Detection

On the basis of Two-stage detection, we add an audio-based single-modality wake-up word model as a front end. The small-scale audio wake-up word detection model is used to filter out most of the easy-to-identify non-wake-up word audio segments from the original signal and retain the potential wake-up word audio segments. This stage is characterized by small computational cost and a low false reject rate (FRR). The purpose of the stage is to reduce computational complexity, filtering out most of the simple samples with small computational resources, and provide segment-level audio. We choose the audio-only model in [6] as the single-modality wake-up word model, which has a unimodal architecture containing 3D-ResNet34, 2D-ResNet34 and simple attention module (SimAM).

In this approach, the single-modality wake-up word model filters a part of the audio clips first, and only inputs the scene corresponding to the remaining potential wake-up word fragments into the later two stages.

## 5 Experiments

### 5.1 Database and Evaluation Metrics

**Database:** In this work, the data is from Task-1 of the 1st Multimodal Information based Speech Processing Challenge (MISP 2021 [7, 26]). The database was collected in home TV scenes, with the wake-up word being "Xiao T Xiao T" and over 300 speakers. The accent of the dataset is Mandarin, and all data was collected in over 30 real rooms. If it contains a wake-up word, the sample will be considered positive, otherwise it will be considered negative. For each sample, a maximum of one wake-up word can be included. The subsets of the database is shown in Table 1. The training set and development set contain audio from three scenarios: far, middle, and near, while the evaluation set only contains far-field audio. The video includes mid-field and far-field. The mid-field video only includes the active speaker, while the far-field video includes everyone in the room.

We change the MISP2021 dataset for the wake-up word active speaker detection task. The dataset provides the facial box positions of the active speaker in the video, including the four coordinates of the four corners of the facial area. Therefore, multiple faces in the video can be compared based on this location. The Euclidean distance between the center coordinates of each face and the

**Table 1.** Subsets of the MISP2021-AVWWS corpus. P represents positive examples and N represents negative examples.

Dataset	P	N	Duration (h)
Training set	5K+	47K+	118.53
Development set	600+	2K+	3.39
Evaluation set	8057		2.87

labeled active speaker’s center coordinates can be calculated separately. The person with the smallest distance is the active speaker. The positive examples of this task include the videos of all active speakers and corresponding audios, while the negative examples include the videos of all non-active speakers and corresponding audios.

**Evaluation Metrics:** For evaluation metrics of wake-up word spotting, we use False Reject Rate (FRR), False Alarm Rate (FAR), and the score of WWS [7]. The FRR and FAR are defined as follows:

$$FRR = \frac{FN}{TP + FN}, FAR = \frac{FP}{TN + FP} \quad (1)$$

where  $TP$  represents the number of correctly identified positive samples,  $FN$  represents the number of positive samples identified as negative,  $TN$  represents the number of correctly identified negative samples, and  $FP$  represents the number of negative samples identified as positive. WWS is defined as:

$$WWS = FRR + FAR \quad (2)$$

We also calculate the area under the receiver operating characteristic curve (AUC) as a metric. It is necessary to consider the two situations of whether to include the active speaker when evaluating wake-up word spotting. For the case where the active speaker is not included, a positive sample is considered as a wake-up word. We call metric in the situation as metric(base). For the case of adding the active speaker, when the correct active speaker in the scene is detected and a word is a wake-up word, it is considered as a positive sample. We call metric in the situation as metric(+), WWS(+) and AUC(+).

For evaluation metrics of Active Speaker Detection, we consider a far-field video as a scene, and calculate the accuracy of judging the correct active speaker in the scene, which is the ratio between number of scenes where active speaker is chosen correctly and number of all scenes.

## 5.2 Training Details

**Preprocess:** We extract videos for each person according to the methods in [5]. We use the face detection model RetinaFace [27] to extract all facial images and

corresponding 5 facial coordinates of people in the video. Based on the sequential coordinates of the detected faces, the K-means algorithm is used to cluster the faces of the same person in a given video. Then, the center point and width of the lip region can be obtained through the coordinates of 5 facial coordinates to obtain the lip position and extract the lip video.

For inputs of Vocalist, the input dimension of the video is  $(3, 48, 96, t_v)$ , where  $t_v$  are set to 5 and 64, respectively. The input dimension of audio is  $(1, 80, t_a)$ . As the sampling rate of audio is 16kHz and the video is 25fps, the audio features with a length of  $t_v * 16000/25$  are first obtained. Then, the Mel spectrogram is obtained through an 80 Mel filter bank with a step size of 200 and a window size of 800.

For inputs of single-modality wake-up word model [5] and FLCMA Based Audio-visual Conformer model [4], we use the same preprocess strategy as in the papers. For the input of audio stream, we extract 80 dimensional FBank features, with a frame length of 25 ms and a frame shift of 10 ms. The time dimension is set to 256, resulting in an audio dimension of  $(256, 80)$ . Then, using a sliding window with a shape of  $(80, 80)$  and a step size of 4 to slice the features along the timeline to obtain the input with a shape of  $(T, H, W, C)$ , which is  $(64, 80, 80, 1)$ . For the input of the video stream, extract a lip region video with a resolution of  $112 \times 112$ , use 3 RGB channels, and sample it into 64 frames with dimensions of  $(64, 112, 112, 3)$ . In addition, each pixel value in the video is normalized to between  $[0, 1]$ .

**Data Augmentation:** For data augmentation, there is a 0.5 probability that audio can be enhanced using the following data augmentation methods: speed perturbation, volume perturbation, slight trimming, frequency masking, and time masking. And we perform offline noise/reverberation addition and beamforming on the audio, expanding this portion of the audio into the training set. There is a 0.5 probability that video can be enhanced using the following data augmentation methods: speed perturbation, frame-wise rotation, horizontal flip, frame-level cropping, and color jitters.

## 6 Results and Discussions

### 6.1 Wake-Up Word Active Speaker Detection

For active speaker detection, we first compare the scores of the Vocalist model trained with 5 and 64 frames on the evaluation set. After increasing the frames of input, the performance of the model is significantly improved, from 52% to 62%. Using 64 frames makes it more convenient for directly comparing the correspondence between wake-up word audio and each lip movement video at the sentence level, and resulting in better performance at determining the active speaker.

To figure out the problem that the accuracy of the model is only 62%, we explore whether the score is related to the positive or negative examples of wake-up words. We calculate the accuracy of the 64-frames Vocalist on the positive

wake-up words and non-wake-up words of Top N ( $N = 1, 2, 3$ ) on the evaluation set, that is, whether the active speaker is included in the first N highest confidence scores. The result is shown in Table 2.

**Table 2.** Accuracy of Vocalist and WWASD on positive and negative examples of wake-up words.

Wake-up words	TopN	Vocalist	WWASD
All examples	top3	96%	97%
	top2	90%	90%
	top1	62%	65%
Positive examples	top3	99%	99%
	top2	98%	98%
	top1	92%	94%
Negative examples	top3	96%	97%
	top2	88%	88%
	top1	55%	58%

Vocalist achieves 92% accuracy on positive wake-up words, while it only achieves accuracy of 55% on negative wake-up words. After investigating the negative case of wake-up word data of MISP2021, it is found that the active speakers of the negative examples provided by the challenge organizers are actually randomly labeled and do not correspond to the audio, which is consistent with the logic that for non-wake-up words, there is no need to detect the active speaker. Therefore, the model can be used to detect the active speaker who is speaking wake-up words if we only look at the positive cases where the labels are accurate.

Secondly, we also use the FLCMA Based Audio-visual Conformer model architecture to train the WWASD model to identify the active speaker, with the same training strategy of Vocalist model. The model outputs the correspondence score between video and audio. The accuracy of the WWASD model is shown in Table 2. All the WWASD model in this section refers to WWASD(L). Compared with the 92% accuracy of the Vocalist model, the model has improved to 94% on positive examples, which has better performance.

Based on the number of people in each video in the evaluation set, the accuracy of the two models on different number of people are tested separately, as shown in Table 3. It can be seen that the performance of WWASD will not decrease significantly due to the increase of the number of people, thus it is robust.

We compare the FLOPs (floating point operations) and the number of parameters of Vocalist and WWASD model, which are listed in Table 4. The number of parameters of the Vocalist model is 3.5 times that of WWASD, and the number

**Table 3.** Accuracy of active speaker detection methods under scenes with different number of persons.

Num of Speaker in the scene	Vocalist	WWASD
1	100%	100%
2	70%	71%
3	91%	92%
4	94%	98%
5	83%	67%
6	93%	92%
7	33%	67%
8	67%	73%

**Table 4.** FLOPs and the number of parameters of two wake-up word active speaker detection models.

Model	FLOPs(G)	Parameters(M)
Vocalist	87.5	70.7
WWASD	21.8	21.2

of FLOPs is 4 times that of WWASD. Our proposed WWASD model not only increases the accuracy, but also reduces the computational complexity.

## 6.2 Comparison of Wake-up Word Spotting Approaches

**Table 5.** Accuracy of the wake-up word spotting systems with three approaches. (L), (M) and (S) refers to using WWASD(L), WWASD(M) and WWASD(S) as the active speaker detection model, (0.5) refers to the threshold of the single-modality wake-up word model. Metrics(R) represents metrics when active speaker label is given as oracle.

Approach	Metrics(base)				Metrics(+)				Metrics(R)			
	FRR[%]	FAR[%]	WWS[%]	AUC[%]	FRR[%]	FAR[%]	WWS(+)[%]	AUC(+)[%]	FRR[%]	FAR[%]	WWS[%]	AUC[%]
One-stage detection	1.96	4.54	6.51	98.69	10.18	2.01	12.19	98.02	2.39	2.23	4.61	99.62
Two-stage detection(L)	3.43	2.04	<b>5.47</b>	<b>99.56</b>	8.28	1.79	<b>10.07</b>	99.26	same as above			
Two-stage detection(M)	3.37	2.30	5.68	99.49	8.71	1.79	10.50	99.22	same as above			
Two-stage detection(S)	3.86	2.04	5.90	99.54	8.95	1.79	10.74	<b>99.28</b>	same as above			
Three-stage detection(L)(0.5)	6.87	1.21	8.08	98.78	11.59	1.07	12.66	–	6.19	1.23	7.42	98.26

We analysed metrics of three wake-up word spotting approaches on the evaluation set. In the last column, Metrics(R), which means metrics when active speaker’s visual information is given, indicates the multimodal wake-up word model uses the correct active speaker directly. This column serves as a reference for ideal performance, with WWS of 4.61% and AUC of 99.62%. According

to Table 5, when traditional multimodal wake-up word spotting model is used in noisy and multi-person scenarios, which is One-stage detection, the score of WWS decreases from 4.61% to 6.51%, resulting in decreased recognition accuracy. Among the three approaches, Two-stage detection(L) achieves the best score, with WWS of 5.47%, close to 4.61% where all the active speakers are the right people. Two-stage detection(M) and Two-stage detection(S) all perform better than One-stage detection, while reducing computational cost at the same time. The comparison of computational cost is listed in Table 6. The performance of Three-stage detection(L) is definitely lower than the Two-stage detection(L), but it greatly reduces the computational complexity because of the single-modality wake-up word model in the first stage. We will analyze the impact of the threshold of the single-modality wake-up word model on the computation cost and performance in Fig. 4.

**Table 6.** FLOPs and the number of parameters of different models.

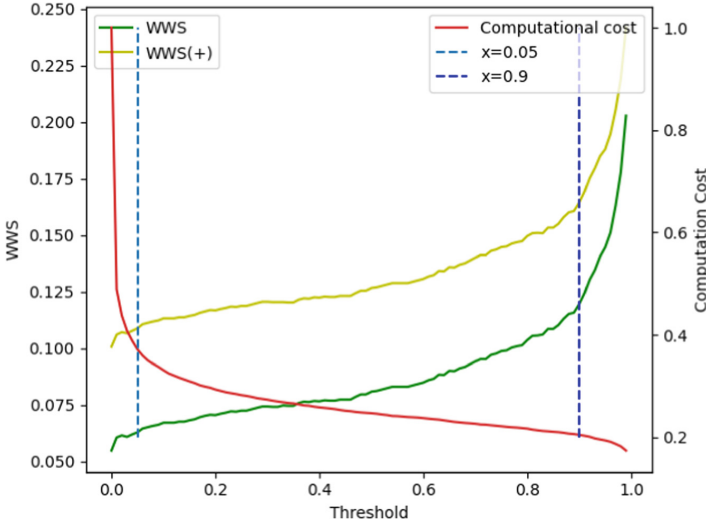
Model	FLOPs(G)	Parameters(M)
WWASD-L	21.78	21.19
WWASD-M	6.06	4.43
WWASD-S	1.84	0.89
Single-modality Wake-up Word Model [6]	4.86	11.10

We use WWASD model and Vocalist model for active speaker detection in Two-stage detection respectively, and test the score metrics on the evaluation set to examine that WWASD model is more suitable for wake-up word active speaker detection.

**Table 7.** Accuracy of the wake-up word spotting systems with the Two-stage detection strategy.

Active Speaker Detection Model	Metrics(base)				Metrics(+)				Metrics(R)			
	FRR[%]	FAR[%]	WWS[%]	AUC[%]	FRR[%]	FAR[%]	WWS(+)[%]	AUC(+)[%]	FRR[%]	FAR[%]	WWS[%]	AUC[%]
WWASD	3.43	2.04	<b>5.47</b>	99.56	8.28	1.79	<b>10.07</b>	<b>99.26</b>	2.39	2.23	4.61	99.62
Vocalist model	3.80	2.12	5.92	<b>99.58</b>	9.63	1.81	11.43	99.25				

As shown in Table 7, WWASD can identify the active speaker more accurately, resulting in improvements in most metrics compared to Vocalist. The WWS are 5.47% and 10.07% in metrics(base) and metrics(+), while WWS of Vocalist are 5.92% and 11.43% respectively. In Fig. 4, as the threshold of the single-modality wake-up word model for wake-up words increases, the WWS and WWS(+) score will slowly increase in the early stages and then rapidly increase thereafter. At the same time, the computational cost will experience a rapid decrease in the early stages and a slow decrease in the later stages. When



**Fig. 4.** The impact of different wake-up thresholds on WWS scores and computational costs.

**Table 8.** Accuracy of the wake-up word spotting systems with the Three-stage detection(L).

Threshold	Metrics(base)				Metrics(+)			Computational Cost [%]
	FRR[%]	FAR[%]	WWS[%]	AUC[%]	FRR[%]	FAR[%]	WWS(+)[%]	
0.05	4.66	1.62	6.28	99.18	9.44	1.43	10.87	37.16
0.5	6.87	1.21	8.08	98.78	11.59	1.07	12.66	24.66
0.9	11.04	0.90	11.94	98.54	15.63	0.79	16.43	20.43

the threshold is set from 0.05 to 0.5, Three-stage detection can significantly reduce computational complexity while maintaining good scores. Table 8 shows metrics of Three-stage detection(L) at different threshold in detail.

### 6.3 Conclusion

We propose a Wake-up Word Active Speaker Detection Model (WWASD) to recognize the face that is speaking the wake-up word. Based on the model, we propose two approaches, namely Two-stage detection and Three-stage detection strategies, for audio-visual wake-up word spotting in noisy and multi-person scenarios. We compare the approaches from the perspectives of performance and computational complexity, thus can select them according to actual application needs. The best approach Two-stage detection achieves comparable performance against the systems with oracle visual speaker bounding boxes while Three-stage detection greatly reduces the computational cost.



**Acknowledgements.** The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

## References

1. López-Espejo, I., Tan, Z.H., Hansen, J.H., Jensen, J.: Deep spoken keyword spotting: an overview. *IEEE Access* **10**, 4169–4199 (2021)
2. Xu, Y., et al.: Audio-visual wake word spotting system for MISP challenge 2021. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9246–9250. IEEE (2022)
3. Zhang, A., et al.: VE-KWS: visual modality enhanced end-to-end keyword spotting. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE (2023)
4. Wang, H., Cheng, M., Fu, Q., Li, M.: Robust wake word spotting with frame-level cross-modal attention based audio-visual conformer. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11556–11560. IEEE (2024)
5. Cheng, M., Wang, H., Wang, Y., Li, M.: The DKU audio-visual wake word spotting system for the 2021 MISP challenge. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9256–9260. IEEE (2022)
6. Wang, H., Cheng, M., Fu, Q., Li, M.: The DKU post-challenge audio-visual wake word spotting system for the 2021 MISP challenge: Deep analysis. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE (2023)
7. Chen, H., et al.: The first multimodal information based speech processing (MISP) challenge: data, tasks, baselines and results. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9266–9270. IEEE (2022)
8. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(12), 8717–8727 (2018)
9. Ma, P., Petridis, S., Pantic, M.: End-to-end audio-visual speech recognition with conformers. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7613–7617. IEEE (2021)
10. Shi, B., Hsu, W.N., Lakhota, K., Mohamed, A.: Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint [arXiv:2201.02184](https://arxiv.org/abs/2201.02184)* (2022)
11. Gao, R., Grauman, K.: VisualVoice: audio-visual speech separation with cross-modal consistency. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15490–15500. IEEE (2021)
12. Ephrat, A., et al.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *arXiv preprint [arXiv:1804.03619](https://arxiv.org/abs/1804.03619)* (2018)
13. Qian, Y., Chen, Z., Wang, S.: Audio-visual deep neural network for robust person verification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 1079–1092 (2021)
14. Wuerkaixi, A., Zhang, Y., Duan, Z., Zhang, C.: Rethinking audio-visual synchronization for active speaker detection. In: *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 01–06. IEEE (2022)
15. Kim, Y.J., et al.: Look who’s talking: active speaker detection in the wild. *arXiv preprint [arXiv:2108.07640](https://arxiv.org/abs/2108.07640)* (2021)

16. Peymanfard, J., Heydarian, S., Lashini, A., Zeinali, H., Mohammadi, M.R., Moza-yani, N.: A multi-purpose audio-visual corpus for multi-modal persian speech recognition: the Arman-AV dataset. *Expert Syst. Appl.* **238**, 121648 (2024)
17. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: deep speaker recognition. arXiv preprint [arXiv:1806.05622](https://arxiv.org/abs/1806.05622) (2018)
18. Tao, R., Qian, X., Das, R.K., Gao, X., Wang, J., Li, H.: Enhancing real-world active speaker detection with multi-modal extraction pre-training. arXiv preprint [arXiv:2404.00861](https://arxiv.org/abs/2404.00861) (2024)
19. Braga, O., Makino, T., Siohan, O., Liao, H.: End-to-end multi-person audio/visual automatic speech recognition. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6994–6998. IEEE (2020)
20. Braga, O., Siohan, O.: A closer look at audio-visual multi-person speech recognition and active speaker selection. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 6863–6867. IEEE (2021)
21. Braga, O., Siohan, O.: Best of both worlds: multi-task audio-visual automatic speech recognition and active speaker detection. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6047–6051. IEEE (2022)
22. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Chen, C.-S., Lu, J., Ma, K.-K. (eds.) *ACCV 2016. LNCS*, vol. 10117, pp. 251–263. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-54427-4\\_19](https://doi.org/10.1007/978-3-319-54427-4_19)
23. Chung, S.W., Chung, J.S., Kang, H.G.: Perfect match: improved cross-modal embeddings for audio-visual synchronisation. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3965–3969. IEEE (2019)
24. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Audio-visual synchronisation in the wild. arXiv preprint [arXiv:2112.04432](https://arxiv.org/abs/2112.04432) (2021)
25. Kadandale, V.S., Montesinos, J.F., Haro, G.: VocaLIST: an audio-visual synchronisation model for lips and voices. arXiv preprint [arXiv:2204.02090](https://arxiv.org/abs/2204.02090) (2022)
26. Zhou, et al.: Audio-visual wake word spotting in misp2021 challenge: dataset release and deep analysis. In: *Interspeech*, pp. 1111–1115 (2022)
27. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: RetinaFace: single-stage dense face localisation in the wild. arXiv preprint [arXiv:1905.00641](https://arxiv.org/abs/1905.00641) (2019)



# Missing Person Recognition Algorithms Based on Image Captioning and Visual Grounding

Ayeong Jeong<sup>1</sup>, Yeongju Woo<sup>2</sup>, Han-young Kim<sup>1</sup>, Gayun Suh<sup>1</sup>,  
Chae-yeon Heo<sup>1</sup>, Yeong-jun Cho<sup>2</sup>, and Hieyong Jeong<sup>2</sup>

<sup>1</sup> Department of Artificial Intelligence, Chonnam National University,  
77 Yongbong-ro, Buk-gu, Gwangju, 61186 Gwangju, Republic of Korea  
215001@jnu.ac.kr

<sup>2</sup> Department of Artificial Intelligence Convergence, Chonnam National University,  
77 Yongbong-ro, Buk-gu, Gwangju, 61186 Gwangju, Republic of Korea  
dndudwn31@jnu.ac.kr, {yj.cho,h.jeong}@chonnam.ac.kr

**Abstract.** Missing person searches are a critical societal challenge with significant implications for public safety and welfare. This study proposes two novel algorithms for efficient and rapid missing person detection based on video data. The first algorithm, *CaptionMP*, uses image captioning technology to generate descriptions of individuals' appearances in video footage, comparing these descriptions to missing person information. The second algorithm, *DINOMP*, employs visual grounding techniques to detect characteristics of missing persons within video streams directly via text prompts. Both algorithms were fine-tuned using the MALS dataset and demonstrated performance across diverse environmental conditions. Notably, they exhibited robust detection capabilities in low-light environments and with complex clothing patterns. The results showed that our proposed methods have considerable potential in the field of missing person detection, offering a solution to the limitations of traditional pedestrian attribute recognition (PAR) methods. This research is expected to substantially contribute to enhancing the practical applicability of intelligent CCTV systems in missing person searches.

**Keywords:** Missing person search · Image captioning · Visual Grounding

## 1 Introduction

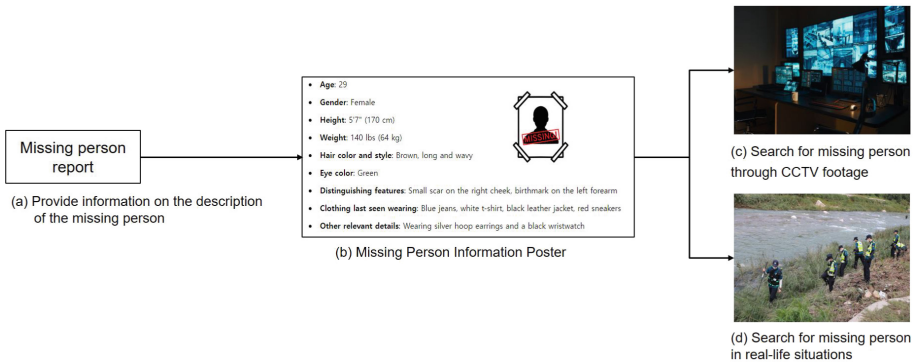
An individual may be classified as a missing persons when their whereabouts cannot be confirmed due to involuntary circumstances, such as accidents or dis-

A. Jeong, Y. Woo—Equal contributors.

---

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/978-3-031-80136-5\\_13](https://doi.org/10.1007/978-3-031-80136-5_13).

asters. Alternatively, they may disappear voluntarily through actions such as running away. In the Republic of Korea, 53,416 cases of missing adults aged 18 or older were reported to the National Police Agency in 2023, with 1,084 (2.05%) of these individuals discovered deceased. In cases of missing persons, the critical period of time which commonly is referred to as the “golden time” is 24 h. If a missing individual is not located within this crucial one-day period, the probability of successful recovery decreases significantly. The period is of particular importance, given that the majority of missing persons are vulnerable individuals, such as elderly individuals with dementia or young children.



**Fig. 1.** Standard processes in missing person searches

In general, missing person detection represents a critical societal challenge with profound implications for public safety and national well-being. In the real world, a missing person search involves the following processes, as shown in Fig. 1 (a) Receiving a missing person report, which may include information such as age, sex, and the individual’s most recent clothing. (b) Collecting and compiling descriptive information about the individual into a poster, which is then disseminated to search personal. (c) Utilizing this shared information, control centers analyze video footage, such as CCTV recordings, to identify potential matches. and (d) Law enforcement and rescue personnel, armed with the descriptive details, conduct on-site searches. However, this methodology often proves inadequate in urgent situations, as it relies heavily on search personnel’s ability to memorize and rapidly apply descriptive information across expansive search areas.

Recent advancements in deep learning technologies applied to intelligent CCTV systems offer promising solutions to these societal challenges. Progress in computer vision has enabled real-time detection, tracking, and analysis of individuals’ external characteristics within video streams [3, 4, 11, 15].

This study proposes algorithms for the swift and accurate identification of missing persons within vehicular black-box or CCTV footage, based on provided

descriptive information of the missing person, leveraging these artificial intelligence technologies.

## 2 Related Works

### 2.1 Pedestrian Attribute Recognition

Early iterations of Pedestrian Attribute Recognition (PAR) algorithms primarily relied on hand-crafted features for attribute identification. Layne et al. [6] proposed a method that combined color histograms and Scale-Invariant Feature Transform (SIFT) features to recognize attributes in pedestrian images. While these approaches demonstrated reasonable performance with limited datasets, they exhibited performance constraints in more complex environmental conditions.

The rapid advancement of Convolutional Neural Networks (CNNs) catalyzed the introduction of deep learning-based approaches in the PAR domain. Zhu et al. [19] introduced the DeepMAR (Deep Learning Multi-attribute Recognition) model, a CNN-based approach capable of simultaneously predicting multiple pedestrian attributes. This model achieved high accuracy by processing various image attributes in parallel. Furthermore, Sarfraz et al. [14] proposed an innovative method that integrated Spatial Transformer Networks (STNs) into a hybrid CNN-RNN (Recurrent Neural Network) model, enabling focused attribute recognition on salient image regions.

Recent developments in PAR have shifted towards leveraging video frames to maximize the utilization of temporal information, moving beyond the constraints of static image-based models. Zhu et al. [20] introduced a novel Visual-Text Fusion Transformer for video-based PAR, utilizing the CLIP [12] model. This approach enhances pedestrian attribute recognition in video sequences by learning the associations between visual and textual modalities. Experimental results demonstrated the superior performance of the proposed model compared to existing methods. However, this model is limited to binary classification of attributes and cannot recognize previously unseen classes.

### 2.2 Visual Understanding and Description

**Image Captioning.** Image captioning is a task that bridges the domains of computer vision and natural language processing, focusing on the generation of textual descriptions for visual content. This field primarily emphasizes the extraction of salient features from images and their subsequent translation into natural language.

Vinyals et al. [16] represents a seminal work in image captioning, integrating Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks to generate descriptive text for images. In this architecture, CNNs extract visual features, which are then utilized by LSTMs for text generation. Xu et al. [17] have made significant advances in the quality of captions through

the introduction of Visual Attention mechanisms that can generate more detailed and accurate descriptions by focusing on specific regions of the image in image captioning. Further improvements were achieved by Anderson et al. [1] with the implementation of Bottom-Up and Top-Down Attention mechanisms, markedly elevating the performance of image captioning systems. Cornia et al. [5] proposed an innovative Meshed-Memory network based on the Transformer architecture. This approach combines multi-head self-attention mechanisms with memory networks to optimize the interaction between visual and textual modalities, resulting in enhanced performance.

The evolution of image captioning has progressed from initial CNN-based models to more sophisticated architectures incorporating Visual Attention and Transformer structures. Contemporary research in this field is primarily focused on achieving higher accuracy and consistency in the extraction of key visual features and their translation into natural language descriptions.

**Visual Grounding.** While both Visual Grounding and Image Captioning address the relationship between images and text, they differ in their fundamental objectives. Image Captioning focuses on generating textual descriptions that encapsulate the overall content of an image, whereas Visual Grounding aims to identify specific objects or regions within an image based on given textual descriptions.

Early research in Visual Grounding primarily focused on modeling rudimentary associations between textual and visual modalities. Matuszek et al. [9] proposed a system that classified objects within an image according to given textual descriptions, utilizing these classifications for object recognition. Subsequent approaches involved segmenting images into multiple object proposal regions and evaluating their correspondence with textual descriptions. Ren et al. [13] advanced this concept by employing Faster R-CNN to generate object proposal regions within images and subsequently matching these regions with textual descriptions to identify corresponding objects. Recent developments in Visual Grounding have seen the application of Transformer-based models. Chen et al. [2] introduced the “UNITER” model, which leverages a Transformer architecture to model diverse interactions between textual and visual modalities. This model integrates textual and visual features, enabling more precise Visual Grounding.

Utilizing these Image Captioning and Visual Grounding techniques that address the relationship between textual and visual information, we propose two models that analyze the relationship between textual descriptions of missing persons’ appearances and video data, employing this analysis to facilitate the location of missing individuals.

### 3 Proposed Methods

This paper presents two different algorithms for the efficient detection of missing person from video inputs, such as CCTV, employing two distinct approaches: image capturing and visual grounding. The first methodology, *CaptionMP*, is

designed to detect missing persons using image captioning from video input to extract appearance features and compare them with the features of missing person information. The second methodology, *DINOMP*, uses a visual grounding technique that utilizes text prompts about the missing person’s appearance to detect individuals with that information in the video input.

### 3.1 *CaptionMP*

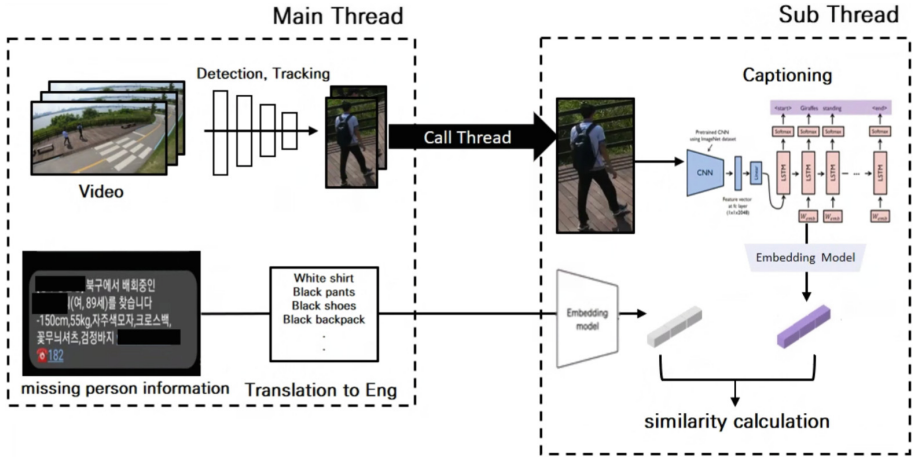


Fig. 2. An overall process for *CaptionMP*

Figure 2 illustrates the comprehensive framework of the Image Captioning-based missing person search system proposed in this study. The system operates on a dual-thread architecture:

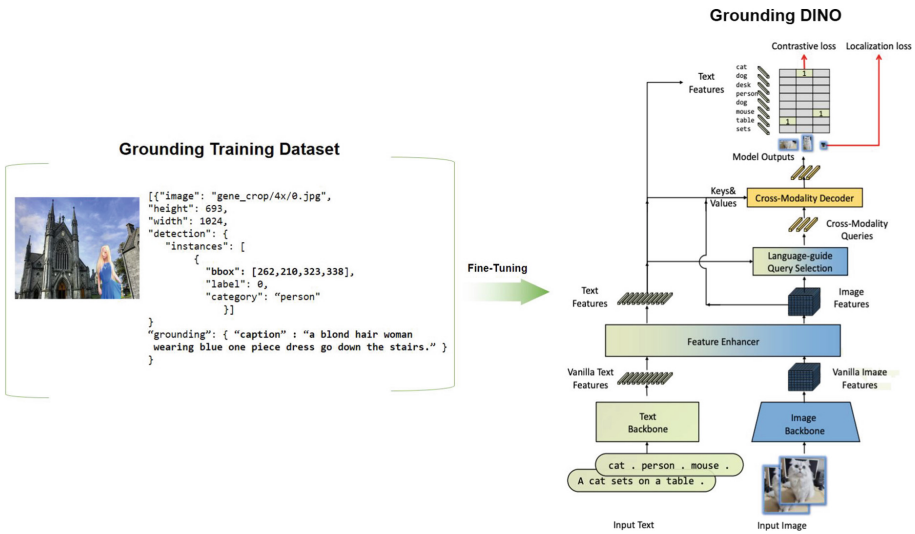
**Main Thread** : Employs object detection model and multi-object tracking model to detect and track human objects within the input video stream. When an object with a consistent ID is tracked for 90 consecutive frames, it is cropped and transmitted to the sub-thread. Concurrently, the textual input describing the missing person’s appearance is translated into English, standardized using regular grammar, and forwarded to the sub-thread.

**Sub Thread** : Utilizes an image captioning model to generate descriptive captions for the cropped human objects. These captions are then compared to the translated missing person information to assess similarity. The similarity is quantified using cosine similarity between the text embedding vectors. A similarity threshold of 0.7 or higher is used to classify an individual as potentially missing.

It is important to note that performing missing person detection on all pedestrians in CCTV and dashcam systems is computationally infeasible. To address this limitation, we implemented a task separation strategy:

CCTV systems perform only pedestrian detection and tracking, while the actual missing person determination is executed on high-performance systems such as central servers. This thread separation technique enables *CaptionMP* to be deployed across various hardware platforms, including IoT devices, by optimizing model efficiency. This approach significantly enhances the scalability and applicability of missing person detection systems.

### 3.2 *DINOMP*



**Fig. 3.** An overall process for *DINOMP*

Figure 3 illustrates the architecture and operational principles of the Visual Grounding model. Using visual grounding techniques for missing person detection enables accurate localization of the subject within the video frame using bounding boxes. This capability facilitates immediate and precise positional information during the search process.

However, when we executed inference using the pre-trained Visual Grounding model with text prompts describing the missing person’s appearance, we observed that the detection performance fell short of our expectations. This discrepancy arose because our objective to detect human subjects based on appearance descriptors differed from the pre-trained model’s tendency to detect objects corresponding to individual appearance attributes independently.



To address this limitation and enhance the model’s efficacy in detecting human subjects based on appearance characteristics, we conducted fine-tuning using the MALS dataset, as depicted in Fig. 3. This fine-tuning process was designed to optimize the model’s performance specifically for the task of missing person identification based on descriptive appearance features.

The implementation of visual grounding in missing person detection offers significant advantages in terms of localization and precise spatial information within video frames. This capability is crucial for enhancing the efficiency and accuracy of search operations. By fine-tuning the Visual Grounding model on domain-specific data, we aimed to bridge the gap between general object detection and the specialized task of human subject identification based on detailed appearance descriptors.

The integration of synthetic data generation techniques with the original MALS dataset allows us to overcome the initial limitations of the dataset while maintaining the rich descriptive information provided by the original captions. This approach not only addresses the lack of BBox annotations but also potentially increases the diversity of the training data, which may contribute to improved model generalization.

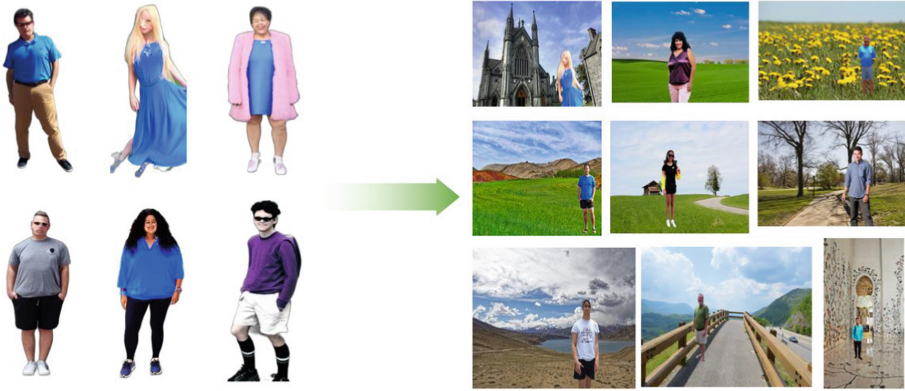
## 4 Experiment

### 4.1 Experimental Configuration

For the fine-tuning process, we used the MALS dataset [18], which comprises 1.5 million samples. The MALS dataset, generated through a diffusion model, maintains high-quality images while being free from privacy concerns. Additionally, MALS provides images with a wide range of variations, including background, viewpoint, occlusion, clothing, and body pose. Each image-text pair in MALS is annotated with appropriate attribute labels, making it effective for simultaneous training in attribute recognition and image-text matching. The fine-tuning processes for both *CaptionMP* and *DINOMP* used the following hyperparameters: a learning rate of  $2e-5$ , 10 epochs, a batch size of 40, and 5000 warm-up steps. We utilized the AdamW optimizer and the `get_linear_schedule_with_warmup` learning rate scheduler.

***CaptionMP***. In this model, we aimed to optimize the performance of an image caption generation model [10] for missing person detection. The image captioning model employed in *CaptionMP* underwent fine-tuning to enhance its specificity for missing person identification. The original model, which tends to generate global image descriptions, does not align with the purpose of this study. Therefore, to address this issue, we fine-tuned the model to create descriptions focused on specific objects using approximately 300,000 images from the MALS dataset.

This fine-tuning process aimed to produce captions that explicitly describe human appearance characteristics and attire, which are crucial for missing person identification.



**Fig. 4.** Data set preprocessing process for fine tuning

**DINOMP.** The data preprocessing methodology for fine-tuning is illustrated in Fig. 4. The MALS dataset we acquired lacks bounding-box (BBox) annotations, which are crucial for object detection and tracking tasks. To address this limitation, we implemented a comprehensive preprocessing pipeline:



















1. Segmentation: We performed semantic segmentation on each image to isolate human subjects and extract their positional information within the segmentation mask.
2. Background Removal: Utilizing the segmentation masks, we eliminated all non-human elements from the images, retaining only the human figures.
3. Image Synthesis: The isolated human figures were composited onto novel background images in randomly selected contextually appropriate locations.
4. BBox Generation and Labeling: We generated bounding boxes for the synthesized human figures and labeled them accordingly. The  $xywh$  coordinates of the BBoxes were specified relative to the entire image size and centered within the image. This approach aligns with our task and prioritizes accurately detecting personal attributes over precise localization.
5. Dataset Integration: The newly created BBox annotations were merged with the original caption labels from the MALS dataset.

This sophisticated preprocessing approach resulted in the creation of a novel augmented dataset specifically tailored for grounding-based learning. By synthesizing human figures onto diverse backgrounds and generating corresponding BBox annotations, we enhanced the dataset’s utility for our specific task of missing person detection. This augmented dataset provides the necessary spatial and descriptive information to effectively train the model in localizing human subjects based on textual appearance descriptions.

## 4.2 Experimental Results

To evaluate the performance of the models, we conducted real-time searches using a custom-designed test dataset. The custom-designed dataset was created

by recording 12 videos, each approximately 30 s long, under different environmental conditions. A total of five participants participated in the experiment. We set up the experimental environment to resemble real CCTV or car dashcam footage. Participants were captured walking from two different angles, with 3 to 4 participants appearing randomly in each video. To ensure the reliability of the test dataset, we collected footage in both bright and low-light conditions.

Input	Output			
<p><b>(a)</b></p>  <p><i>We are looking for a female, wearing a black striped sweater, gray pants, and brown boots.</i></p>	<b>Pars</b>			
	Target	Non_target #1	Non_target #2	Non_target #3
	 <p>LongSleeve UpperStride Trousers Age18-60 Female</p> <p><b>Similarity: 0.5334</b></p>	 <p>Hat LongSleeve Trousers Age18-60 Male</p> <p><b>Similarity: 0.2856</b></p>	 <p>Hat LongSleeve Trousers Age18-60 Female</p> <p><b>Similarity: 0.4993</b></p>	 <p>LongSleeve UpperLogo Trousers Age18-60 Female</p> <p><b>Similarity: 0.5410</b></p>
	<b>CaptionMP</b>			
	 <p>The woman is wearing a striped sweater, grey sweatpants, and tan slippers.</p> <p><b>Similarity: 0.7575</b></p>	 <p>The man is wearing a dark blue hoodie and black pants with black sandals.</p> <p><b>Similarity: 0.5078</b></p>	 <p>The person is wearing a green top, white pants, and a blue scarf, with black footwear.</p> <p><b>Similarity: 0.5467</b></p>	 <p>The woman is wearing a red hoodie and black pants.</p> <p><b>Similarity: 0.5367</b></p>
	<p><b>(b)</b></p>  <p><i>We are looking for a male, wearing black padding, black pants, black slippers.</i></p>	<b>Pars</b>		
Target		Non_target #1	Non_target #2	Non_target #3
 <p>LongSleeve Trousers Age18-60 Male</p> <p><b>Similarity: 0.4350</b></p>		 <p>ShortSleeve UpperLogo Trousers Age18-60 Female</p> <p><b>Similarity: 0.3624</b></p>	 <p>LongSleeve Trousers Age18-60 Female</p> <p><b>Similarity: 0.2758</b></p>	 <p>LongSleeve Trousers Age18-60 Female</p> <p><b>Similarity: 0.2758</b></p>
<b>CaptionMP</b>				
 <p>The person is wearing a black jacket, black pants, and black shoes with a black hat</p> <p><b>Similarity: 0.7700</b></p>		 <p>The person is wearing a light-colored top and dark pants.</p> <p><b>Similarity: 0.5664</b></p>	 <p>The person is wearing a dark long-sleeved top and light-colored pants with dark-colored footwear.</p> <p><b>Similarity: 0.5728</b></p>	 <p>The person is wearing a light-colored hoodie and matching pants with white sneakers.</p> <p><b>Similarity: 0.3829</b></p>

**Fig. 5.** A comparison of results between the existing PARS and the proposed *CaptionMP*

Figure 5 represents a qualitative comparison between the proposed *CaptionMP* and the existing PARS model [7] trained on the PA100K dataset [8]. Since PARS generates captions as words, we converted the label outputs

into sentence form and used the same methodology as the captioning approach to compare performance.

In Fig. 5(a), the search for a missing person took place in a bright environment where the individual wore clothing with complex patterns, such as stripes. The predefined keywords for the missing person’s appearance were: female, black striped sweater, gray pants, and brown boots. Although the existing PARS model generated some keyword captions, it failed to provide critical identification information such as color and produced nearly identical captions for most participants in the video. This limitation stems from the reliance of the PARS model on pre-trained attributes and binary classification. In contrast, the proposed *CaptionMP* demonstrated superior performance compared to the existing PARS model, achieving a high similarity score of 0.7575 for complex patterns like stripes, surpassing the results for solid-colored clothing. This model showed a better similarity for clothing with basic color combinations.

Figure 5(b) involves an experiment conducted in a low-light environment. The predefined keywords for the missing person’s appearance were: male, black padding, black pants, and black slippers. Unlike the previous experiment, this scenario was designed to be more challenging for feature extraction. Most of the captions generated by the PARS model lacked discrimination for the target, resulting in low similarity scores. Although the quality of captions by *CaptionMP* slightly decreased under low-light conditions, it maintained a similarity score above 0.7, verifying the model’s effectiveness in identifying the missing person. Our comparative analysis demonstrates that the fine-tuned model achieves a more stable object detection performance than the existing model. It effectively detects individuals in real-time, even in challenging scenarios such as low-light conditions or environments with multiple individuals with similar appearances.

The proposed *DINOMP* provides a method for detecting missing person using a fine-tuned Grounding DINO model, which differs from *CaptionMP* which employs caption generation and text comparison methods. Figure 6 shows the results of searching for missing persons using *DINOMP*, which identi-







	Scenario_1	Scenario_2	Scenario_3
<b>Input</b>	 <p>We are looking for a female, wearing a black striped sweater, gray pants, and brown boots.</p>	 <p>We are looking for a woman, wearing Red hoodie, black pants, white shoes.</p>	 <p>We are looking for a male, wearing black padding, black pants, black slippers</p>
<b>DINOMP</b>			

Fig. 6. Results of *DINOMP*

fies individuals matching the input prompt. In various scenarios, *DINOMP* directly detects distinctive clothing features within the video and marks them with bounding boxes, clearly identifying the missing person's location visually. The experimental results show that *DINOMP*, as *CaptionMP*, demonstrates robust detection performance based on descriptive features of the missing person's appearance in both bright and low-light conditions. These results suggest that our proposed algorithm can be effectively applied to missing person detection in diverse environmental conditions. This comprehensive evaluation highlights the potential of our proposed methodology to enhance the efficiency and accuracy of missing person search tasks.

## 5 Discussion and Conclusion

This research proposes two algorithms that can expeditiously and efficiently locate missing persons within the critical golden time, leveraging feature-based relationships between visual data and textual descriptions.

The proposed algorithms are designed to be user-friendly, requiring only textual descriptions of the missing individual's appearance and video imagery as inputs. This accessibility allows for utilization by users without specialized expertise.

The two-algorithms approach offers enhanced versatility, allowing users to select the most appropriate model based on specific operational requirements. This flexibility confers a competitive advantage in terms of utility. The systems can be integrated with intelligent camera networks to optimize the deployment of law enforcement personnel in search operations. Additionally, when incorporated into autonomous robotic systems, these models demonstrate the potential for rapid identification of missing persons even in congested environments.

However, it is important to note that the current study primarily utilizes descriptive features of the missing individual's appearance. Future research directions include the comprehensive analysis of diverse characteristics such as facial features, gait patterns, and body morphology, which are anticipated to enhance overall accuracy. Moreover, the practical commercialization of these systems necessitates model optimization for reduced computational demands, as well as the integration of advanced networking technologies and sophisticated database management systems.

In conclusion, this research lays the foundation for continuous advancement in the field of missing person identification. It is expected to make significant contributions towards enhancing the practical applicability of intelligent CCTV systems, with ongoing development guided by the aforementioned considerations.

**Acknowledgements.** This research was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea grant, funded by the Ministry of Education (NRF-2021R111A3055210), and partially by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government (MSIT).

## References

1. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)
2. Chen, Y.-C., et al.: UNITER: UNiversal image-TExt representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12375, pp. 104–120. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-58577-8\\_7](https://doi.org/10.1007/978-3-030-58577-8_7)
3. Cho, Y.J., Kim, S.A., Park, J.H., Lee, K., Yoon, K.J.: Joint person re-identification and camera network topology inference in multiple cameras. *Comput. Vis. Image Underst.* **180**, 34–46 (2019)
4. Cho, Y.J., Yoon, K.J.: Improving person re-identification via pose-aware multi-shot matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1354–1362 (2016)
5. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10578–10587 (2020)
6. Layne, R., Hospedales, T.M., Gong, S., Mary, Q.: Person re-identification by attributes. In: BMVC. vol. 2, pp. 8 (2012)
7. Li, D., Chen, X., Huang, K.: Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In: ACPR, pp. 111–115 (2015)
8. Liu, X., et al.: HydraPlus-Net: attentive deep features for pedestrian analysis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 350–359 (2017)
9. Matuszek, C., FitzGerald, N., Zettlemoyer, L., Bo, L., Fox, D.: A joint model of language and perception for grounded attribute learning. arXiv preprint [arXiv:1206.6423](https://arxiv.org/abs/1206.6423) (2012)
10. Mokady, R., Hertz, A., Bermano, A.H.: ClipCap: clip prefix for image captioning. arXiv preprint [arXiv:2111.09734](https://arxiv.org/abs/2111.09734) (2021)
11. Park, C., Kim, J., Chae, J., Lee, J.: A pilot study on intelligent surveillance system based on fashionpedia dataset. In: 2024 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 313–316. IEEE (2024)
12. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
14. Sarfraz, M.S., Schumann, A., Wang, Y., Stiefelhagen, R.: Deep view-sensitive pedestrian attribute inference in an end-to-end model. arXiv preprint [arXiv:1707.06089](https://arxiv.org/abs/1707.06089) (2017)
15. Ullah, R., et al.: A real-time framework for human face detection and recognition in CCTV images. *Math. Probl. Eng.* **2022**(1), 3276704 (2022)
16. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
17. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057. PMLR (2015)

18. Yang, S., Zhou, Y., Zheng, Z., Wang, Y., Zhu, L., Wu, Y.: Towards unified text-based person retrieval: a large-scale multi-attribute and language search benchmark. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 4492–4501 (2023)
19. Zhu, J., Liao, S., Lei, Z., Li, S.Z.: Multi-label convolutional neural network based pedestrian attribute classification. *Image Vis. Comput.* **58**, 224–229 (2017)
20. Zhu, J., Jin, J., Yang, Z., Wu, X., Wang, X.: Learning clip guided visual-text fusion transformer for video-based pedestrian attribute recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2626–2629 (2023)





# Contrastive and Restorative Pre-Training for Medical VQA

Vasudha Joshi<sup>1</sup>, Pabitra Mitra<sup>2</sup>, and Supratik Bose<sup>3</sup>

<sup>1</sup> Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India  
vasudhaj50@gmail.com

<sup>2</sup> Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India

<sup>3</sup> Varian Medical Systems Inc., San Ramon CA94582, CA, USA

**Abstract.** Medical Visual Question Answering (MedVQA) aims to develop models to answer clinically relevant questions on medical images. A major challenge in developing VQA for the Medical domain is the unavailability of large, well-annotated MedVQA datasets. Using transfer learning from ImageNet and finetuning on the MedVQA dataset is not helpful as visual concepts of ImageNet images and medical images differ. Therefore, this paper focuses on the problem of the lack of a large MedVQA dataset by employing a novel pre-training technique for the visual encoder. Our pre-training framework uses contrastive and restorative learning to learn fine-grained semantic representations from large, unlabelled medical images available online. We finetune our pre-trained visual encoder on the MedVQA dataset. Our experiments show that the combination of contrastive and restorative learning significantly improves the performance of MedVQA systems. We evaluate our model on three MedVQA datasets. The source code is available at <https://github.com/Vasudha27/CRP-for-MedVQA>.

**Keywords:** VQA · Medical Visual Question Answering · Self-supervised learning · Contrastive learning · Restorative learning

## 1 Introduction

Medical Visual Question Answering (MedVQA) is a domain-specific Visual Question Answering (VQA) system that answers questions related to the visual information present in the medical image. It can help doctors in diagnosing diseases, surgical planning, and triage. Besides doctors, it can benefit patients and medical students. Patients can use MedVQA to enhance their understanding of their radiology reports, thereby promoting health awareness. For medical students, it can enhance their learning experience. Therefore, it can contribute to the overall improvement of the healthcare system.

MedVQA systems leverage state-of-the-art computer vision and natural language processing (NLP) techniques. These areas use large deep-learning models. First, MedVQA systems require deep learning networks (CNN, Vision transformers) to extract visual features from the input medical image. Second, deep natural



language models (RNN, LSTM, etc.) are needed to extract semantic information from input questions. Third, a deep network captures the correlation between the input image features and question features. Finally, the correlation feature vector passes through a multilayer perceptron followed by a softmax layer to obtain the probability distribution on the set of answers.

Deep learning networks require large datasets for training. However, unlike VQA, no large-scale, well-annotated datasets are available in MedVQA. Visual encoders in computer vision tasks use Transfer learning to overcome the scarcity of data. They train on a large labelled imagenet dataset in a supervised manner, and the knowledge is transferred to tasks with small datasets. However, for MedVQA, finetuning models with an imagenet pre-trained visual encoder on the MedVQA dataset is not beneficial as the covariate shift between medical images and those in the imagenet hinders the effectiveness of transfer learning. To overcome this problem, MEVF [21] uses Convolutional Denoising Auto-Encoder (CDAE) [20] and Model-Agnostic Meta-Learning (MAML) [9] to pre-train the visual encoder. They manually construct a labelled dataset from the VQA-RAD [17] dataset to learn meta-weights using MAML. MTPT [10] pre-trains the visual encoder on two supervised tasks as a multi-task learning paradigm. They prepare a labelled dataset from the VQA-RAD [17] dataset. These two works cannot be generalized to other MedVQA datasets since pre-training tasks are based on the VQA-RAD dataset. CPRD [18] uses unlabelled radiology images from three body regions - brain, chest, and abdomen. They pre-train the visual encoder using contrastive learning [6]. Contrastive learning is highly effective at capturing discriminative high-level features. However, it is inefficient at learning fine-grained features.

In this work, we overcome the data scarcity problem by employing a pre-training framework that leverages Self-supervised learning to glean transferable knowledge from large, unlabelled medical image datasets available online. Self-supervised learning captures task-agnostic representations. We combine contrastive and restorative learning in the self-supervised pre-training framework. Contrastive learning aims to learn to differentiate between positive and negative samples. Restorative learning seeks to generate an original signal from the distorted signal. Contrastive learning excels in extracting high-level global features, while restorative learning performs well in extracting fine-grained features. MedVQA requires extracting high-level global features or fine-grained detailed features depending on the query posed. Our pre-training framework effectively equips the visual encoder to handle medical images' complex and diverse characteristics.

We summarize our contributions as follows: a) we propose a pre-training framework that uses unlabelled medical images to pre-train the visual encoder of the MedVQA model; b) we evaluate the performance of MedVQA model with the pre-trained visual encoder on three MedVQA datasets VQA-RAD [17], SLAKE [19] and VQA-Med-2019 [1] to demonstrate the efficiency of the proposed pre-training framework.

## 2 Related Work

**Medical Visual Question Answering.** Most of the existing MedVQA methods [2, 23, 25, 28, 31, 32, 36] use general domain VQA models. They use attention mechanisms such as SAN [33], BAN [16], and MFB [34] to combine visual and textual features. They use pre-trained CNNs such as VGG [29] and ResNet [14]. MEVF [21] and MTPT [10] propose a pre-training technique for MedVQA. MEVF [21] initializes visual feature encoder by MAML [9] and CDAE [20]. They manually construct an additional dataset using VQA-RAD [17] to train MAML. MTPT [10] treats visual feature encoder pre-training as a multi-task problem. One task is image classification and another is a binary task of identifying image-question compatibility. Both these methods require the preparation of an additional labeled dataset. MMQ [7] proposes multiple meta-model quantifying technique that increases meta-data by auto-annotation. It returns meta-models that have robust features for MedVQA. PubMedCLIP [8] model adopts CLIP [24] for MedVQA. They finetune CLIP [24] on Radiology Objects in COntext (ROCO) [22] dataset.

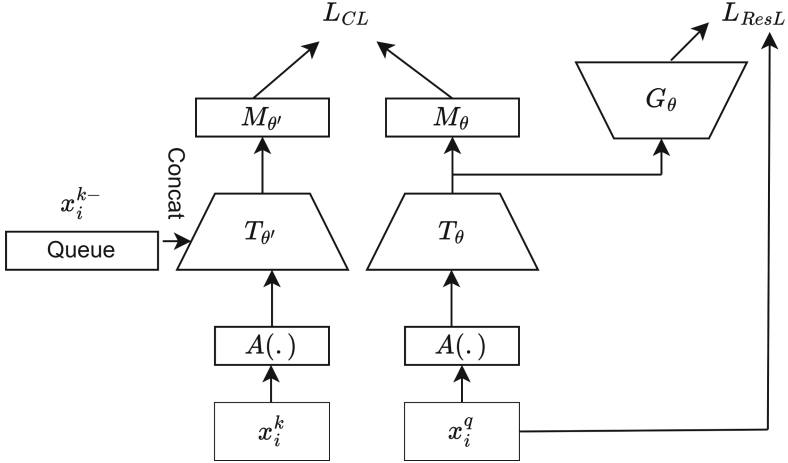
**Contrastive Learning.** Contrastive learning aims to learn an embedding space where representations of samples from the same class are close to one another and far apart from representations of samples from dissimilar classes. It excels in learning representations in self-supervised learning. SimCLR [5] generates distorted versions of each sample by applying a series of data augmentations. It maximizes the agreement between two representations of the same thing by minimizing the contrastive loss [5]. To avoid the computation of representations for negative samples in every batch, MoCo [13] uses a dictionary to save many negative samples. MoCo-v2 [12] combines the augmentation method of SimCLR with the dictionary of MoCo [13]. Hence, it takes the advantage of both techniques.

**Restorative Learning.** Restorative learning aims to reconstruct the distorted image. It is widely used in medical imaging. By reconstructing medical images, the model learns low-level features. Some of the works use only restorative learning [4, 37], while some combine restorative and adversarial learning [30]. [4] proposes a context restoration task. It swaps the two small patches of the image multiple times till its spatial information is changed. However, it preserves the intensity distribution during image distortion.

## 3 Proposed Approach

In this section, we describe the full pipeline of our approach. We first pre-train the visual encoder on medical images using contrastive and restorative pre-training framework (Sect. 3.1). After pre-training, the pre-trained visual encoder is used in the MedVQA framework to extract visual features from input medical images (Sect. 3.2).

### 3.1 Pre-training



**Fig. 1.** Contrastive and Restorative Pre-training framework (CRP) contains contrastive and restorative components. An input image  $x_i$  is distorted by applying a series of data augmentations  $A(\cdot)$ . Augmented images pass through contrastive and restorative components. The contrastive component has a visual encoder  $T_\theta$  and momentum encoder  $T_{\theta'}$ . The restorative component has visual encoder  $T_\theta$  and decoder  $G_\theta$ . Visual encoder  $T_\theta$  is trained using contrastive and restorative learning.

Figure 1 describes the pre-training framework used for the pre-training visual encoder. It contains two components:

**Contrastive Component.** It uses a contrastive learning method for pre-training the visual encoder represented by  $T_\theta$ . We use MoCo-v2 [12]. This component contains a twin visual encoder referred to as momentum encoder  $T_{\theta'}$ , a Dictionary, and two Multilayer perceptrons represented as  $M_\theta$  and  $M_{\theta'}$ . The dictionary in MoCo-v2 [12] is represented by a FIFO queue. It contains the encoded representations of image samples present in the previous mini-batch. The queue contains the representations generated by momentum encoder  $T_{\theta'}$ . Representations from the queue serve as negative samples  $x_j^{k-}$ . We randomly crop the input image  $x_i$  to produce two views  $x_i^q$  and  $x_i^k$ . We apply augmentation techniques  $A(\cdot)$  on  $x_i^q$  and  $x_i^k$ . We randomly crop and resize each image to  $224 \times 224$ . Augmentation techniques  $A(\cdot)$  include Gaussian blurring, random horizontal flipping, and color jittering. We pass the augmented views through visual encoder  $T_\theta$  and momentum encoder  $T_{\theta'}$  followed by  $M_\theta$  and  $M_{\theta'}$  respectively to generate the encoded representations  $\hat{x}_i^q = M_\theta(T_\theta(A(x_i^q)))$  and  $\hat{x}_i^{k+} = M_{\theta'}(T_{\theta'}(A(x_i^k)))$ . Since,  $\hat{x}_i^q$  and  $\hat{x}_i^{k+}$  are representations obtained from the different views of the

same input image  $x_i$ ,  $\hat{x}_i^q$  should be similar to  $\hat{x}_i^{k+}$  and dissimilar to  $x_j^{k-}$ . We use InfoNCE loss [13] to maximize the similarity between  $\hat{x}_i^q$  and  $\hat{x}_i^{k+}$ :

$$L_{CL} = -\log \frac{\exp(\hat{x}_i^q \cdot \hat{x}_i^{k+} / t)}{\exp(\hat{x}_i^q \cdot \hat{x}_i^{k+} / t) + \sum_{j=1}^N \exp(\hat{x}_i^q \cdot x_j^{k-} / t)} \quad (1)$$

where  $t$  represents the temperature,  $(\cdot)$  is dot product, and  $N$  is the size of the dictionary.  $T_\theta$  and  $M_\theta$  are updated through backpropagation.  $T_{\theta'}$  and  $M_{\theta'}$  are not differentiable due to large queue. Therefore, we update  $T_{\theta'}$  and  $M_{\theta'}$  using exponential moving average of  $T_\theta$  and  $M_\theta$  as

$$\theta' \leftarrow \alpha \theta' + (1 - \alpha) \theta \quad (2)$$

where  $\alpha$  is the momentum coefficient. This momentum base update is defined in [13].

**Restorative Component.** It uses a restorative learning method to enhance the visual representation of visual encoder  $T_\theta$  by using fine-grained image concepts. We use 2D U-Net [26] with  $T_\theta$  as base encoder and  $G_\theta$  as a decoder. We pass the augmented view  $x_i^q$  of  $x_i$  to  $T_\theta$ . The decoder  $G_\theta$  takes the representation obtained from the encoder to generate the original image. Restorative learning aims to minimize the distance between the restored image and the original image. We use mean square error as the restorative loss  $L_{ResL}$ .

$$L_{ResL} = MSE(x_i^q, y_i^q) \quad (3)$$

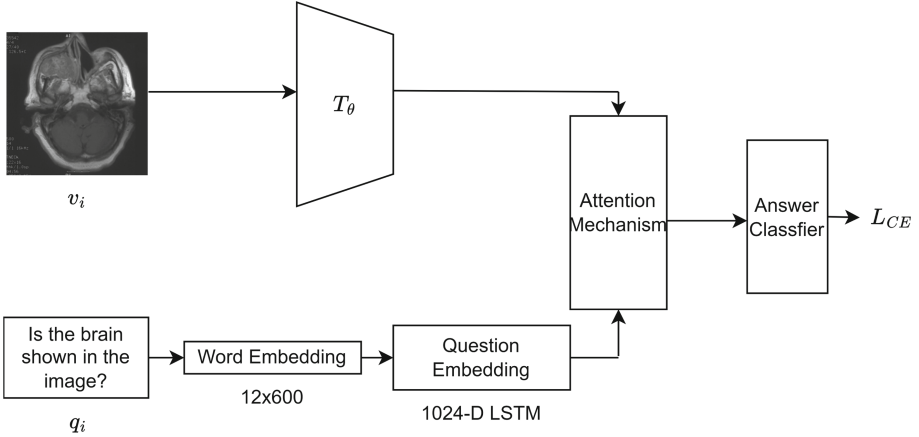
where  $y_i^q = G_\theta(T_\theta(A(x_i^q)))$  is the restored image.  $MSE()$  represents mean square error. As given in [12], during pretraining we optimize the two losses jointly. The final loss is:

$$L = \beta * L_{CL} + (1 - \beta) * L_{ResL} \quad (4)$$

where  $\beta$  signifies the importance of different losses.

### 3.2 Medical VQA Framework

Figure 2 describes the overall architecture of MedVQA. We use the pretrained visual encoder  $T_\theta$  to obtain the visual representation  $\hat{v}_i$  of input image  $v_i$ , i.e.  $\hat{v}_i = T_\theta(v_i)$ . For question features we use 1024-D LSTM. Each question is either trimmed or padded to 12 words depending on the length of the question. Question words are embedded to a 600-D vector by concatenating 300-D Glove word embeddings. The concatenated 600-D words embeddings pass through LSTM to produce question features  $\hat{q}_i$ . Both image embedding  $\hat{v}_i$  and question embedding  $\hat{q}_i$  are combined through attention mechanism like BAN [16] to produce a multimodal embedding  $z_{vq}$ . This multimodal embedding  $z_{vq}$  is passed through an answer classifier. Since we formulate MedVQA as a classification task, we use a Multilayer perceptron (MLP) followed by a softmax layer as an answer classifier. The multimodal embedding  $z_{vq}$  is passed through the answer classifier to obtain a probability distribution over a fixed set of candidate answers from the training set. To train the entire MedVQA model we use cross-entropy loss function  $L_{CE}$ .



**Fig. 2.** Medical VQA framework. Visual encoder  $T_\theta$  from the CRP framework is applied. BAN [16] attention mechanism combines the visual and textual features.

## 4 Experiments and Results

### 4.1 Datasets

**Pre-training Dataset.** For pre-training we use large-scale unlabelled dataset Radiology Objects in COntext (ROCO) [22]. It contains medical images belonging to several different modalities like MRI, CT, X-Ray, Angiography, PET, Mammograph, Ultrasound, and Fluoroscope. Besides medical images, this dataset also contains captions and keywords. However, we use only medical images for our pre-training task.

**MedVQA Datasets.** We evaluate the performance of our pre-training method on three MedVQA datasets: VQA-RAD [17], SLAKE [19], and VQA-Med-2019 [1]. SLAKE [19] dataset is a bilingual dataset. It contains 642 images and 14,028 question-answer pairs. Out of 14,028 question-answer pairs, there are only 7033 question-answer pairs in English. We use the 642 images and 7033 English question-answer pairs for our experiment. It has CT, MRI, and X-Ray images. They cover the head, neck, and chest organs. We follow the train-val-test split provided in the SLAKE dataset [19]. VQA-RAD [17] dataset has 315 images and a total of 3515 question-answer pairs. We follow the split given in [21]. It has abdominal CTs, chest X-Rays and head CTs or MRIs. VQA-Med-2019 [1] has 4200 medical images and 15292 question-answer pairs. The dataset is divided into train, validation, and test sets. The train set has 3200 medical images and 12792 question-answer pairs. The validation set has 500 images and 2000 question-answer pairs. The test set has 500 images and 500 question-answer pairs. It has questions about identifying plane, modality, organ, and abnormality. The dataset has 10 organs, 36 modalities, 16 planes, and around 1600 different abnormalities.

## 4.2 Implementation Details

**Pre-training Framework.** For contrastive component we use MoCo-v2 [12] with ResNet-50 [14] encoders. MLP heads  $M_\theta$  and  $M_{\theta'}$  have two layers of hidden dimension 2048 and ReLU activation unit. The size of the queue N is 65536. For the restorative component, we use 2D U-Net [26] architecture with ResNet-50 [14] as a backbone for  $T_\theta$  and a corresponding decoder network  $G_\theta$ . We train the different components of the Pretraining framework in a stepwise incremental manner similar to [11]. We first train only the contrastive component for 500 epochs. We optimize the contrastive component using InfoNCE loss given in Eq. 1, where temperature  $t$  is 0.07. After training the contrastive component for 500 epochs, we add the restorative component and train the entire Pre-training framework for 200 epochs using the joint loss function given in Eq. 4. We use an SGD optimizer with an initial learning rate of 0.03., weight decay of 0.0001, and momentum of 0.9 [12]. We use an early-stopping criterion on the validation set and save the image encoder weights with minimum validation loss. We use two Nvidia Tesla V100 GPUs and a mini-batch of size 128 for pre-training.

**MedVQA.** We initialize the image encoder  $T_\theta$  of MedVQA with the weights obtained in the pre-training stage. We freeze the image encoder and train the remaining MedVQA model end-to-end using the cross-entropy loss on MedVQA datasets. We use Adamax optimizer [3]. We vary the learning rate during training. We warm up the learning rate from 0.0025 to 0.01 for the initial four epochs. From the tenth epoch onwards we decrement the learning rate by 0.75 after every 18 epochs.

## 4.3 Comparison with State-of-the-Arts

We use accuracy to evaluate the performance of the MedVQA models [21]. We use our pre-trained visual encoder with BAN [16] attention mechanism without CR reasoning [35]. In Table 1, we compare our model with the general VQA baseline and existing works related to pre-training in MedVQA on the VQA-RAD, SLAKE, and CLEF datasets. We have not tested MTPT+BAN [10] for the SLAKE and VQA-Med-2019 datasets as the proposed pre-training task is designed specifically for the VQA-RAD dataset. We briefly describe these models in Sect. 2. For a fair comparison of the pre-training frameworks, we use 1024-D LSTM with GloVe word embeddings for extracting question features in all architectures. All the models are trained and tested at the same seed value. From Table 1, we see that our pre-training method based on combining contrastive and restorative learning learns more transferable features that give better results on varied datasets. Our proposed method outperforms MEVF+BAN [21], MTPT+BAN [10], PubMedCLIP+BAN [8], and MMQ+BAN [7]. MMQ+BAN [7] performs better on VQA-RAD dataset, while PubMedCLIP+BAN [8] works well on SLAKE dataset. Our CRP+BAN method outperforms MMQ+BAN [7] on VQA-RAD by 3.8%, and PubMedCLIP+BAN [8] on SLAKE dataset by 3.8%. We evaluate the performance of our method on VQA-Med-2019 dataset. We

**Table 1.** Comparison based on Test accuracies (%) on MedVQA datasets of SOTA models and baseline (BAN [16]). ‘fw.’ denotes framework. \* indicates our reimplementation using official codes. The best scores are written in **bold** while the second best are underlined.

Model	VQA-RAD [17]			SLAKE-EN [19]			VQA-Med-2019 [1]		
	Closed	Open	All	Closed	Open	All	Closed	Open	All
BAN fw.* [16]	66.5	27.4	51.0	67.5	47.3	55.2	60.9	34.6	38.0
MEVF+BAN [21]	75.1	43.9	62.6	—	—	—	—	—	—
MEVF+BAN* [21]	71.4	39.8	58.8	77.9	73.6	75.3	67.2	40.1	43.6
MTPT+BAN [10]	75.7	56.1	67.9	—	—	—	—	—	—
MTPT+BAN* [10]	71.9	48.0	62.3	—	—	—	—	—	—
PubMedCLIP+BAN+AE [8]	78.1	48.6	66.5	79.9	76.2	77.6	—	—	—
PubMedCLIP+BAN+AE* [8]	<u>75.7</u>	44.1	63.2	<u>80.8</u>	75.2	<u>77.4</u>	<u>78.1</u>	<u>47.0</u>	<u>51.0</u>
MMQ+BAN [7]	75.8	53.7	67.0	—	—	—	—	—	—
MMQ+BAN* [7]	74.6	<u>52.0</u>	<u>65.6</u>	79.8	<u>75.7</u>	77.3	75.0	41.7	46.0
CRP+BAN (ours)	<b>79.8</b>	<b>53.6</b>	<b>69.4</b>	<b>84.9</b>	<b>78.9</b>	<b>81.2</b>	<b>82.8</b>	<b>58.7</b>	<b>61.8</b>

outperform PubMedCLIP+BAN [8] and MMQ+BAN [7] by 10.8% and 15.8%, respectively. Although our method shows an improvement in all three datasets, its performance gain is more significant in the VQA-Med-2019 dataset, which contains more organs and abnormalities than the other two datasets. This shows that our pre-training framework learns generalized features that can be easily used on other datasets.

#### 4.4 Ablation Analysis

We perform an ablation study to evaluate the effectiveness of our pre-training framework. Table 2 summarizes the results. We use the same backbone architecture (BAN [16]) and text encoder (LSTM [15]) for all models in Table 2. These models differ only in the visual encoder. ResNet-50 (random init) uses a randomly initialized visual encoder, while ResNet-50 (ImageNet) uses a visual encoder that is pre-trained on ImageNet in a supervised manner. We train the visual encoder of ResNet-50 (Contrastive) using contrastive learning on the roco dataset [22] for different epochs. We obtain the best result on MedVQA datasets with a visual encoder trained for 600 epochs on the roco dataset. We observe that the visual encoder trained using contrastive and restorative learning outperforms the visual encoder trained only using contrastive learning. ResNet-50 (CRP) also significantly outperforms ResNet-50 (random init) and ResNet-50 (ImageNet).

**Table 2.** Comparison of different visual encoders (ResNet-50 [14]) based on Test accuracies (%) on MedVQA datasets. The best scores are written in **bold** while the second best are underlined.

Visual Encoder	VQA-RAD [17]			SLAKE-EN [19]			VQA-Med-2019 [1]		
	Closed	Open	All	Closed	Open	All	Closed	Open	All
ResNet-50 (random init)	67.6	28.5	52.1	69.2	47.8	56.2	59.4	34.4	37.6
ResNet-50 (ImageNet)	66.5	27.4	51.0	67.5	47.3	55.2	60.9	34.6	38.0
ResNet-50 (Contrastive)	<u>77.2</u>	<u>47.5</u>	<u>65.4</u>	<u>83.2</u>	<u>76.9</u>	<u>79.4</u>	<u>82.8</u>	<u>56.9</u>	<u>60.2</u>
ResNet-50 (CRP)	<b>79.8</b>	<b>53.6</b>	<b>69.4</b>	<b>84.9</b>	<b>78.9</b>	<b>81.2</b>	<b>82.8</b>	<b>58.7</b>	<b>61.8</b>

#### 4.5 Qualitative Analysis

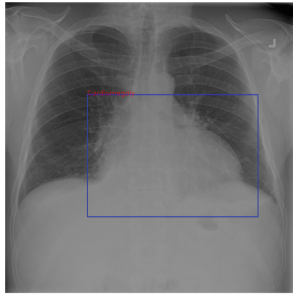
In this section, we show the visualizations of the Grad-CAM [27] heatmaps to see the region the model focuses on while answering the question. SLAKE dataset has additional visual annotations: segmentation mask and bounding box for object detection [19]. We compare the disease localization of our model with the segmentation mask or bounding box provided with the dataset. Figure 3 shows the visualizations of Grad-CAM heatmaps. The first column shows the original image, question, and answer from the SLAKE dataset. Image and Question are input to the MedVQA model. The Second column shows the bounding box or semantic segmentation mask provided with the image. The third column shows the activation maps of our visual encoder while predicting the answer.

The first row shows the Chest X-Ray. Figure 3b indicates the location of the disease with a bounding box. Figure 3c signifies the region of the image our model focuses on while detecting the disease. The highlighted region in Fig. 3c overlaps with the bounding box, marking the location of cardiomegaly in Fig. 3b. The second row shows a Brain MRI. Figure 3e is the segmentation mask of the brain edema region. The heatmap in Fig. 3f indicates that the highlighted region nearly overlaps with the brain edema region. The third row shows an Abdomen CT. Figure 3h shows the semantic segmentation of the Lung cancer. In Fig. 3i, the dark colour in the heatmap, indicating higher weights, cover the lung cancer region. From the Grad-CAM heatmaps, we see that our model accurately localizes the disease.

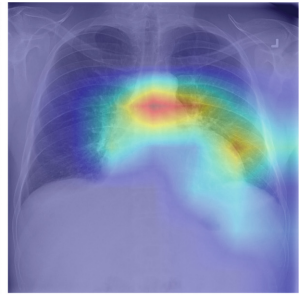




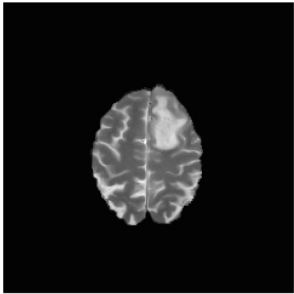
(a) Q: What diseases are included in the pictures?  
Answer: cardiomegaly



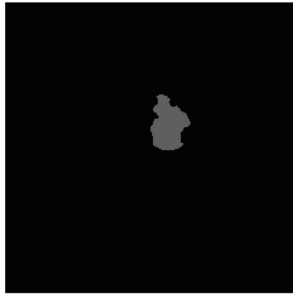
(b) Bounding Box highlighting the location of disease



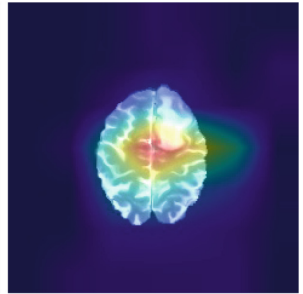
(c) Predicted: cardiomegaly ✓



(d) Q: Is the abnormality hyperdense or hypodense?  
Answer: hyperdense



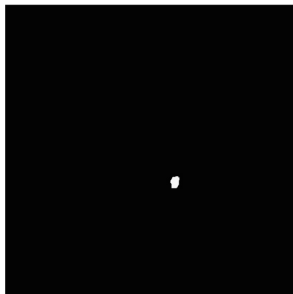
(e) Semantic segmentation mask of the brain edema



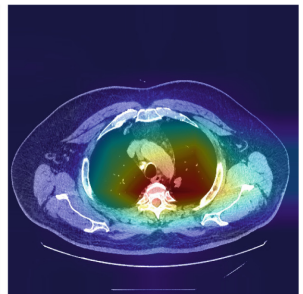
(f) Predicted: hyperdense ✓



(g) Q: What diseases are included in the picture?  
Answer: Lung Cancer



(h) Semantic segmentation mask of the Lung cancer



(i) Predicted: Lung Cancer ✓

**Fig. 3.** Grad-CAM heatmaps of our visual encoder. The left column shows the original image, the center column shows the bounding box or segmentation mask, and the right column shows the activation map. The dark colour in the activation map shows higher weights. The first, second, and third row shows X-Ray, MRI, and CT.

## 5 Conclusion

In this work, we overcome the problem of data scarcity. We combine contrastive and restorative learning to pre-train the visual encoder on a publically available large unlabelled dataset. We finetune the pre-trained visual encoder on three MedVQA datasets. We observe that our joint pre-training strategy outperforms present state-of-the-art methods on three datasets.

## References



1. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: overview of the medical visual question answering task at ImageCLEF 2019. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, 9-12 September 2019. CEUR Workshop Proceedings (2019)
2. Al-Sadi, A., Talafha, B., Al-Ayyoub, M., Jararweh, Y., Costen, F.: Just at ImageCLEF 2019 visual question answering in the medical domain. In: CLEF (working notes) (2019)
3. Ariff, N.A.M., Ismail, A.R.: Study of Adam and Adamax optimizers on Alexnet architecture for voice biometric authentication system. In: 17th International Conference on Ubiquitous Information Management and Communication, IMCOM 2023, Seoul, Republic of Korea, 3-5 January 2023, pp. 1–4 (2023)
4. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. *Med. Image Anal.* **58**, 101539 (2019)
5. 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)
6. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint [arXiv:2003.04297](https://arxiv.org/abs/2003.04297) (2020)
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](https://doi.org/10.1007/978-3-030-87240-3_7)
8. Eslami, S., Meinel, C., De Melo, G.: PubMedClip: how much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023, pp. 1181–1193 (2023)
9. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135. PMLR (2017)
10. Gong, H., Chen, G., Liu, S., Yu, Y., Li, G.: Cross-modal self-attention with multi-task pre-training for medical visual question answering. In: Proceedings of the 2021 International Conference on Multimedia Retrieval, pp. 456–460 (2021)
11. Guo, Z., Islam, N.U., Gotway, M.B., Liang, J.: Discriminative, restorative, and adversarial learning: Stepwise incremental pretraining. In: Kamnitsas, K., et al. (eds.) Domain Adaptation and Representation Transfer - 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, 22 September 2022, Proceedings, vol. 13542, pp. 66–76. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-16852-9\\_7](https://doi.org/10.1007/978-3-031-16852-9_7)

12. Haghighi, F., Taher, M.R.H., Gotway, M.B., Liang, J.: DiRA: discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20824–20834 (2022)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
14. 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 2016, Las Vegas, NV, USA, 27-30 June 2016, pp. 770–778 (2016)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
16. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
17. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. *Sci. data* **5**(1), 1–10 (2018)
18. Liu, B., Zhan, L.-M., Wu, X.-M.: Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 210–220. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-87196-3\\_20](https://doi.org/10.1007/978-3-030-87196-3_20)
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 (ISBI), pp. 1650–1654 (2021)
20. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6791, pp. 52–59. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-21735-7\\_7](https://doi.org/10.1007/978-3-642-21735-7_7)
21. 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](https://doi.org/10.1007/978-3-030-32251-9_57)
22. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in CContext (ROCO): a multimodal image dataset. In: Stoyanov, D., et al. (eds.) LABELS/CVII/STENT -2018. LNCS, vol. 11043, pp. 180–189. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01364-6\\_20](https://doi.org/10.1007/978-3-030-01364-6_20)
23. Peng, Y., Liu, F.: UMass at ImageCLEF medical visual question answering (MedVQA) 2018 task. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, 10-14 September 2018 (2018)
24. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
25. Ren, F., Zhou, Y.: CGMVQA: a new classification and generative model for medical visual question answering, pp. 50626–50636 (2020)
26. 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](https://doi.org/10.1007/978-3-319-24574-4_28)

27. 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)
28. Sharma, D., Purushotham, S., Reddy, C.K.: MedFuseNet: an attention-based multimodal deep learning model for visual question answering in the medical domain. *Sci. Rep.* **11**, 19826 (2021)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7-9 May 2015, Conference Track Proceedings (2015)
30. Tao, X., Li, Y., Zhou, W., Ma, K., Zheng, Y.: Revisiting rubik’s cube: Self-supervised learning with volume-wise transformation for 3d medical image segmentation. In: Martel, A.L., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part IV, LNCS, vol. 12264, pp. 238–248. Springer, Cham (2020). <https://doi.org/10.1007/978-3-030-59719-1-24>
31. Vu, M.H., Löfstedt, T., Nyholm, T., Sznitman, R.: A question-centric model for visual question answering in medical imaging, pp. 2856–2868 (2020)
32. Vu, M.H., Sznitman, R., Nyholm, T., Löfstedt, T.: Ensemble of streamlined bilinear visual question answering models for the ImageCLEF 2019 challenge in the medical domain. In: CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland, 9-12 Sept 2019 (2019)
33. 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)
34. 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)
35. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2345–2354 (2020)
36. Zhou, Y., Kang, X., Ren, F.: TUA1 at ImageCLEF 2019 VQA-med: a classification and generation model based on transfer learning. In: CLEF (working notes) (2019)
37. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. In: Medical image analysis, p. 101840 (2021)



# MRCI: Multi-range Context Interaction for Boundary Refinement in Image Segmentation

Yaqiang Wu<sup>1,2</sup> , Wanjun Lyu<sup>2</sup> , Xianchen Liang<sup>3</sup>, Qinghua Zheng<sup>1</sup> ,  
Jin Wei<sup>2</sup> , and Lianwen Jin<sup>4</sup> 

<sup>1</sup> Xi'an Jiaotong University, Xi'an, China  
qzheng@mail.xjtu.edu.cn

<sup>2</sup> Lenovo Research, Beijing, China  
{wuyqe,lvwj1,weijin4}@lenovo.com

<sup>3</sup> Beijing University of Posts and Telecommunications, Beijing, China  
liangxianchen@bupt.edu.cn

<sup>4</sup> South China University of Technology, Guangzhou, China  
eelwj@scut.edu.cn

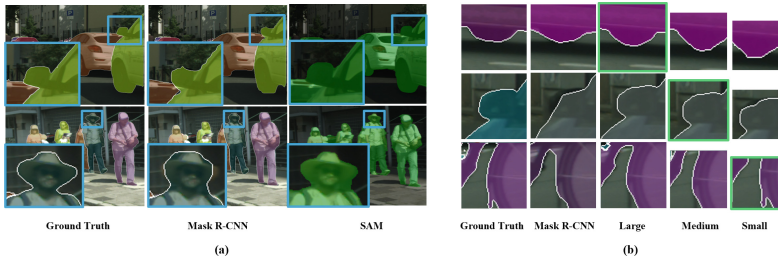
**Abstract.** In the era of foundational image segmentation models, there is a pressing need to leverage the outputs of these models and enhance the boundary accuracy of domain-specific segmentation results using lightweight post-processing techniques. Numerous existing boundary refinement approaches neglect the significance of incorporating diverse contextual scopes and global knowledge, resulting in restricted adaptability to different coarse segmentation errors. Moreover, the prevailing models are often lacking in lightweight design. To address these challenges, we propose a novel framework called Multi-Range Context Interaction (MRCI) that aims to refine the boundaries of predicted masks by incorporating comprehensive context knowledge while maintaining computational efficiency. Our approach utilizes a multi-range context-aware strategy to extract more informative local features and incorporates global knowledge prompts to guide the boundary refinement process. Experimental results on the widely used Cityscapes, ADE20K and satellite remote sensing dataset SpaceNet demonstrate the effectiveness of our approach, achieving top-tier Average Precision (AP) and mean IoU among the current state-of-the-art boundary refinement models while utilizing only 4M parameters. The source code will be available.

**Keywords:** Instance Segmentation · Semantic Segmentation ·  
Boundary Refinement · Post-processing

## 1 Introduction

Instance segmentation, which aims to assign a specific object category to each pixel in an image, has witnessed rapid advancements in recent years. A notable

example is Mask R-CNN [7], which proposes to combine Faster R-CNN [16] with a Fully Convolutional Network (FCN) to generate detection results and pixel-wise object masks simultaneously. SAM [10] represents another leap in computer vision. Built upon a vast dataset of labeled segmentations, SAM is promptable and processes the unique ability to transfer zero-shot to new image distributions and tasks. However, a disparity remains between the theoretical achievements of these algorithms and their efficacy in practical scenarios. A critical observation is that *most existing segmentation algorithms have a lower accuracy in predicting object boundaries compared to classifying internal regions*. This shortcoming is evident in Fig. 1 (a) where both Mask R-CNN and SAM struggle with accurate object boundary prediction.



**Fig. 1.** (a) Results of instance/semantic segmentation using Mask R-CNN and SAM. (b) A visual comparison of boundary refinement outcomes using different crop sizes. The most precise results are marked with green boxes for clarity. (Color figure online)

These observations underscore the significance of instance boundary prediction as a complementary task to instance segmentation. For large pre-trained models (e.g., SAM), there is a growing demand to enhance the accuracy of segmenting specific domain data through lightweight object boundary refinement methods. We assert several distinct advantages of post-processing boundary refinement: (1) It is model-agnostic, making it versatile across various architectures (2) It centers on refining segmentation boundaries, enabling detailed corrections and enhancements that are pivotal for high quality image segmentation. (3) Post-processing boundary refinement facilitates the incorporation of large models in segmentation tasks, eliminating the need for extensive pre-training.

Current state-of-the-art approaches for boundary refinement task can be broadly categorized into two groups: global-based and local-based approaches. Global-based approaches, such as [31], refine the boundary of an instance on a feature map or mask of the same size as the input image. They process inputs of high resolution, resulting in high computational costs and distraction from the boundary. Local-based approaches, such as [19], employ a “crop-and-refine” strategy, where the image is firstly cropped into small patches along the instance boundaries, and then the local boundary is refined. However, they are difficult to capture sufficient context knowledge for different segmentation instances with single-size patches. In Fig. 1 (b), we illustrate the effects of different patch sizes

on segmentation, small patches with high-resolution (HR) short-range context knowledge can be used to adapt small objects and preserve segmentation details, but the disadvantage is that limits the learned context dependencies to the cropped range. Conversely, large patches with low-resolution (LR) long-range context information can be used to adapt large stuff-regions and enhance low accuracy coarse predictions, but the details are not handled well.

To break through these limitations, we propose a novel post-processing framework, Multi-Range Context Interaction Network (MRCI), that effectively combines the strengths of HR short-range patches, LR long-range patches, and global knowledge to produce superior results for both foundational and non-foundational segmentation models.

Firstly, we propose a multi-range feature flow interaction module to enhance segmentation accuracy by refining object boundaries, particularly for challenging cases. Specifically, to extract multi-range features, we propose a four-stage feature extractor designed to comprehensively capture relevant information across different scales. Moreover, we incorporate an inter-branch interaction mechanism to facilitate the flow of multi-range features between neighboring branches. Secondly, we propose a global knowledge prompt that incorporates patch relative position and category information derived from segmentation models to enhance accuracy. MRCI employs a positional prompt of the input coordinates to assist convolution layers to represent high-frequency information. This is different from the positional encoding used in Transformer [20], where it serves to provide the discrete positions of tokens in a sequence as input that does not contain any notion of order. This positional prompt maps low-dimensional position information to a high-dimensional space, thus improving boundary accuracy. Moreover, by utilizing an encoding scheme for category mask prompts, we can distinguish various categories more effectively, thus enhancing boundary precision and shortening the training time by one-third.

In total, the main contributions of this work are summarized as follows:

(1) We propose a multi-range context interaction network (MRCI) as a post-processing framework to refine the boundaries of segmentation masks. MRCI utilizes a combination of low-resolution (LR) and high-resolution (HR) context knowledge in an interactive manner. This allows the network to capture multi-resolution local features, thereby enhancing the accuracy and robustness of boundary segmentation for various objects.

(2) To ensure accurate boundary localization and differentiate features across multiple instances, MRCI integrates global insights from both positional and category mask prompts. The positional prompt aids in enhancing boundary features, while the category mask prompt accelerates the convergence process.

(3) MRCI is a universal framework that consistently improves the performance of various segmentation models across multiple benchmarks. It provides a streamlined approach to integrate pre-trained models into segmentation tasks, allowing for improved results. Through extensive experiments conducted on widely used benchmarks, our method has achieved satisfactory results, thus



validating the effectiveness and generality of our approach for both instance segmentation and semantic segmentation tasks.

## 2 Related Work

### 2.1 Image Segmentation

Image segmentation is one of the most fundamental tasks in image processing and computer vision. Two prominent tasks within image segmentation are instance segmentation and semantic segmentation.

**Instance Segmentation.** Instance segmentation in computer vision involves two main approaches: top-down detection-based methods, and bottom-up segmentation-based methods. Top-down methods, like Mask R-CNN [7], use detection algorithms to find an instance’s bounding box and then generate segmentation maps. Bottom-up methods [2, 6], start with pixel-level segmentation and then differentiate instances using clustering or metric learning. [12] is both top-down and bottom-up which unify segmentation and detection tasks.

**Semantic Segmentation.** Semantic segmentation is long formulated as a pixel or mask classification problem. FCN [14] is the first model to employ a fully convolutional network for semantic segmentation. Following the proposal of Vision Transformer, several researchers [24, 29] have explored its potential for semantic segmentation research. Mask2Former [3] introduces masked attention for improved performance and faster convergence. SAM [10] is promptable and can transfer zero-shot to new image distributions and tasks. In comparison to SAM, SEEM [32] exhibits a broader scope in terms of both interaction and semantics levels, as it supports more prompt types and understanding semantics.

General segmentation of large models is currently a popular approach, but due to limitations in datasets and other factors, the accuracy of these models may not be sufficient for category or boundary. In contrast, our method offers a fine-tuning mechanism for segmentation outcomes within specific domains, utilizing only a set of such results for model training. The approach can be directly applied to the sophisticated post-processing of large models, resulting in a significant enhancement of its effect.

### 2.2 Boundary Refinement

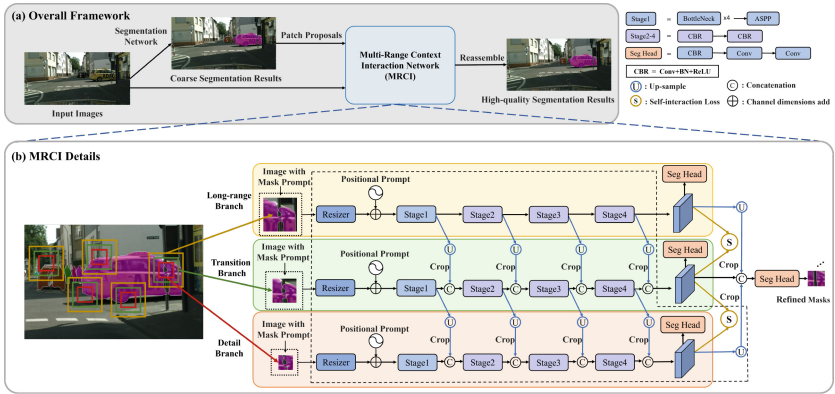
The boundary refinement work focuses on accurately perceiving and locating the segmentation boundary to improve the accuracy of the segmentation models. [27] offers a two-stage framework that integrates boundary refinement using fine-grained features. [26] introduces a model-independent real-time boundary processing method that maps from the boundary to the internal pixels, although it may incorrectly predict the internal areas. [11] treats refinement as a rendering task, using an MLP to adjust selected points on coarse segmentations. SharpCountour [31] uses a contour-based approach to refine the boundaries of



instances. BAProto [28] designs a boundary-aware prototype for boundary refinement through the similarity of the boundary representation and the specific prototype. Adopting the crop-and-refine strategy, [19] first cuts coarse segmentation masks along the boundary into small patches and then refines the boundary with [18] as the backbone.

However, these approaches often face challenges to achieve optimal segmentation accuracy or may be computationally intensive. In this paper, we present a lightweight and efficient strategy that tackles these limitations through a multi-range context-aware approach.

### 3 Method



**Fig. 2.** (a) The pipeline of our boundary refinement framework. (b) The proposed Multi-Range Context Interaction (MRCI) network comprises three branches, facilitating the integration of both short-range and long-range contextual knowledge. The virtual line is the network of the inference model, with the output of the detail branch being the final results.

#### 3.1 Boundary Patch Extraction

To enhance an instance mask generated by a segmentation model, we begin by extracting boundary patches for refinement. Inspired by the BPR [19], we employ a sliding window algorithm to meticulously select patches that align with the predicted boundaries of the instance. Following extraction, we refine our selection by applying a Non-Maximum Suppression (NMS) [15] technique, which helps to eliminate redundant patches and control overlap. By adjusting the NMS threshold, we can optimize the balance between processing speed and the precision of the mask. Alongside the image patches, we also precisely extract corresponding binary mask patches from the original instance mask.

### 3.2 Global Knowledge Prompt

The proposed framework takes into account global knowledge, such as category and positional prompt, from the coarse segmentation network to better represent the global features of the cropped patches. This combination of global knowledge and coarse segmentation masks helps capture the overall context of the image and facilitates the refinement of the boundaries of local contexts with different categories, sizes, and locations.

**Category Mask Prompt:** The input of our framework is obtained by stacking an image patch and a category mask encoding patch as follows:

$$I = P_{image} + P_{mask} \quad (1)$$

where  $P_{image}$  and  $P_{mask}$  represent image patches and coarse segmentation mask patches from object boundaries, respectively. The coarse segmentation mask has been specifically modified to represent different categories through an encoding scheme ranging from 0.2 to 1.0 for the object interior, 0.1 for the boundary and 0 for the background. By assigning these values to specific areas within the segmentation mask, we are able to create a highly detailed map of each object’s boundaries and internal structures.

**Positional Prompt:** We transform the low-dimensional positional information into a high-dimensional space by employing high-frequency functions to generate high-frequency data, such as the boundaries of objects. To inject positional prompt, we construct a vector with position information, including the location coordinates of the object  $(x_0, y_0, \dots, x_3, y_3)$  and patch  $(x_4, y_4, \dots, x_7, y_7)$ , and then encode it with  $\sin x$  and  $\cos x$  functions to a high-dimension space  $(x_{i0}, y_{i0}, x_{i1}, y_{i1}, \dots, x_{i7}, y_{i7}, \forall i \in [0 \dots 8])$ . Finally, we perform a linear transformation of the encoding result to obtain a positional prompt vector  $P_{position}$ , which is then added to the features generated by the resizer module.

### 3.3 Refinement Network Based on Multi-Range Context

As a post-processing framework, the proposed MRCI utilizes multi-range context-aware local features interactively to further enhance accuracy. The proposed framework enables a combination of short-range and long-range context knowledge, which improves the accuracy of segmentation boundary, as shown in Fig. 2. To construct multi-range context patches, we employ three scales, namely  $128 \times 128$ ,  $96 \times 96$ , and  $64 \times 64$ , to crop the images and corresponding segmentation masks with the same center located at the object boundary of the mask, as shown in Fig. 2 (b). This ensures that we have a comprehensive understanding of more aspects of the object’s shape and structure.

MRCI is composed of three branches, each of which shares the same structure to ensure sufficiently feature interactions. Each branch consists of a resizer module and four stages that work together to learn features and detect boundaries. The larger patch ( $128 \times 128$ )  $I_l$  is used as an input of the top branch to obtain long-range context knowledge. The smaller patch ( $64 \times 64$ )  $I_d$  is used as

an input of the bottom branch to obtain detailed knowledge. The middle patch ( $96 \times 96$ )  $I_t$  is taken as an input of the middle branch which serves as a transition between other two branches, ensuring smoothly transmission of features. Firstly, we enlarge the input by 1.5 times to learn large-scale features. Then in resizer module, we quickly reduce the features size of all the three branches to the size of detail branch input to persist a simple and fast backbone structure. Positional prompts  $P_{position}$  are added to obtain  $I'_l$ ,  $I'_t$  and  $I'_d$  respectively. The formulation is as follow:

$$I'_i = Resizer(Upsample(I_i)) + P_{position}^i, i \in \{l, t, d\} \quad (2)$$

where  $Resizer()$  means the resizing operation and  $P_{position}^i$  represents the input positional prompts of three branches.

To provide long-range context features to the detail branch, we add interaction between the neighboring branches after each stage of the network. The features are upsampled to the input size and then center-cropped before interaction. A segmentation head is cascaded to stage 4 of each branch. The formulations are as follows:

$$f_i^{(K)} = \begin{cases} ASPP(BottleNeck(I'_i)), & \text{if } K = 1, \\ CBR(f_i^{(K-1)}), & \text{if } K \in \{2, 3, 4\}. \end{cases} \quad (3)$$

$$S_i = SegHead(f_i^{(4)}) \quad (4)$$

the  $ASPP()$  operation refers to atrous spatial pyramid pooling, while  $CBR()$  operation represents ‘‘Convolution-Batch Normalization-ReLu’’, here we cascade two  $CBR()$  blocks for better feature extraction. In the transition and detail branches,  $f_i^{(K-1)}$  is obtained by concatenating features from neighboring branches and previous stages  $K - 1$ .  $S_i$  represents segmentation results from each branch.

Finally, the features produced by all three branches  $f_{final}$  are fused together and fed to a segmentation head, which generates the final refinement result. The formulations are as follows:

$$f_{final} = fusion(f_l^{(4)}, f_t^{(4)}, f_d^{(4)}) \quad (5)$$

$$S_{final} = SegHead(f_{final}) \quad (6)$$

where  $f_i^{(4)}$  denote features from stage 4 from each branch.

### 3.4 Training and Inference

To improve the robustness of the proposed model, we introduce random block noise along the boundary of the coarse mask during training. In terms of loss function, we calculate the Binary Cross Entropy loss, Dice loss, Active Boundary loss between the outputs of four SegHead modules and ground truth. In the implementation, we set the weights of the three branches to be equal.

To better convey long-range context knowledge, we calculate the self-interaction loss for the three features produced by the last stage. The self-interaction loss is calculated as follows:

$$\mathcal{L}_{SI} = MSE\left(f_t^{(4)}, f_l^{(4)}\right) + MSE\left(f_d^{(4)}, f_t^{(4)}\right) \quad (7)$$

where  $f_t^{(4)}, f_l^{(4)}, f_d^{(4)}$  represent the output features of the transition, long-range, and detail branch, respectively, in stage 4.  $MSE$  means Mean Squared Error loss function. The self-interaction loss not only enhances the stability and accelerates the convergence speed of training, but also improves the segmentation accuracy of the proposed model. The total loss is calculated as below:

$$\mathcal{L}_{total} = \lambda_{BCE} * \mathcal{L}_{BCE} + \lambda_{Dice} * \mathcal{L}_{Dice} + \lambda_{ABL} * \mathcal{L}_{ABL} + \lambda_{SI} * \mathcal{L}_{SI} \quad (8)$$

where  $\lambda_{BCE}, \lambda_{Dice}, \lambda_{ABL}, \lambda_{SI}$  denote the weight of  $\mathcal{L}_{BCE}, \mathcal{L}_{Dice}, \mathcal{L}_{ABL}, \mathcal{L}_{SI}$ , respectively, and we set 1, 1, 0.1, 0.05 in our implementation. After the training process, the output of the detail branch is almost identical to that of the fusion branch. Therefore, in the inference process, the result of the detail branch can be used as the final output to reduce computational cost.

## 4 Experiments

We illustrate that MRCI, a lightweight and efficient boundary refinement framework, generalizes well to both instance segmentation and semantic segmentation tasks on four widely used datasets. Furthermore, we provide extensive ablations to demonstrate the significance of MRCI's components.

### 4.1 Implementation Details

**Datasets.** We assess the effectiveness and transferability of our proposed approach on four public available datasets: Cityscapes [5], ADE20K [30], SpaceNet [1] and RWMD [23]. **Cityscapes** [5] is a real-world dataset that provides high-quality segmentation annotations of urban driving scenes. The dataset comprises 2,975 / 500 / 1,525 images for training/validation/testing, respectively. **ADE20K** [30] is a more challenging benchmark consisting of 20,210/2,000/3,000 images for training/validation/testing respectively. It encompasses 150 fine-grained semantic categories. **SpaceNet** [1] is hosted on Amazon Web Services (AWS) as a public dataset. The dataset encompasses diverse geographic regions of interest. For our experimental analysis, we have chosen to work with the AOI-1. This subset of the dataset consists of 4,858 images for training, 694 images for validation, and 1,388 images for testing.

**RWMD** [23] is a dataset of real-world mobile documents captured using mobile phones in natural environments. It consists of a total of 2,007 images, with 1,505 images designated for training and 502 images allocated for testing.

**Experimental Setup.** We utilize the distributed training PyTorch framework to train our model on 4 V100 GPUs for a duration of 3 d, with a batch size of 32, learning rate set at  $1e-5$  and weight decay at  $5e-4$ .

**Evaluation Metrics.** For the evaluation of instance segmentation results, we utilize the Average Precision (AP) and Boundary Average Precision (Boundary AP) metric. For the semantic segmentation task, we utilize the IoU score for each category and the mean IoU (mIoU) across all categories. Model size is quantified by the total number of parameters.

## 4.2 Comparisons with State-of-the-Arts

**Results on Instance Segmentation Task.** The aim of our framework is to acquire a general ability for the post-processing segmentation task, enabling us to refine the coarse segmentation results of various models directly and without requiring retraining or fine-tuning. We initially train a model using patches extracted from the coarse segmentation results produced by Mask R-CNN on Cityscapes. Subsequently, we utilize this model to refine predictions generated by other models (not limited to Mask R-CNN) on Cityscapes validation set. Specifically, we directly apply our proposed framework to the instance segmentation results of PointRend, SegFix, Mask2Former and OneFormer in Table 1. The results of Boundary  $AP_{val}$ ,  $AP_{val}$  and  $AP_{val50}$  are presented. As shown in Table 1, our proposed framework can significantly enhance the Boundary AP of SegFix and PointRend results by 3.6% (from 17.6 to 21.2) and 2.2% (from 20.6 to 22.8), respectively, while also improving the AP of SegFix and PointRend results by 1.8% (from 38.2 to 40.0) and 2.4% (from 37.9 to 40.3), respectively. In terms of the Transformer-based segmentation method Mask2Former and OneFormer, our approach also achieves 1.1% and 0.8% improvement in  $AP_{val}$  performance respectively, demonstrating its high compatible with these methods. Furthermore, our model surpasses other methods in terms of Boundary  $AP_{val}$  and  $AP_{val50}$  and achieve new state-of-the-art results on Cityscapes.

**Results on Semantic Segmentation Task.** We proceed to present the results of the semantic segmentation task on the Cityscapes and ADE20K validation datasets. MRCI only uses the positions of the patches relative to the whole image. As illustrated in Table 2, by employing the segmentation outputs from HRNet-W48 [22], SAM [10], and MetaPrompt-SD [21] as the training dataset, our method notably enhances the segmentation performance, yielding a notable 1.8% mIoU improvement over HRNet-W48, and attaining the highest mIoU of 87.8% over MetaPrompt-SD on Cityscapes.

Our results on ADE20K, as presented in Table 3, demonstrate that MRCI enhances the mIoU metric by 0.6% when applied to HRNet-W48, achieving a new state-of-the-art result of 62.8% mIoU based on OneFormer. This improvement underscores MRCI’s ability to complement strong baselines focused on improving segmentation boundary quality.

**Table 1.** SOTA results on Cityscapes validation set in terms of instance segmentation.

Method	Boundary AP <sub>val</sub>	AP <sub>val</sub>	AP <sub>val50</sub>
Mask R-CNN [7]	-	31.5	-
BMask R-CNN [4]	-	35.0	-
UPNet [25]	-	37.8	-
PANet [13]	-	41.4	-
Mask R-CNN [7]	15.1	36.4	59.2
+ SegFix [26]	17.6 (+2.5%)	38.2 (+1.8%)	63.4 (+4.2%)
+ Mask Transfiner [9]	18.0 (+2.9%)	37.9 (+1.5%)	64.1 (+4.9%)
+ <b>MRCI (Ours)</b>	20.9 (+5.8%)	39.8 (+3.4%)	<b>64.8 (+5.6%)</b>
+ SegFix + <b>MRCI (Ours)</b>	<b>21.2 (+6.1%)</b>	<b>40.0 (+3.6%)</b>	64.5 (+5.3%)
PointRend [11]	18.6	37.9	63.7
+ SegFix [26]	20.6 (+2.0%)	39.5 (+1.6%)	64.3 (+0.6%)
+ <b>MRCI (Ours)</b>	22.0 (+3.4%)	40.3 (+2.4%)	64.8 (+1.1%)
+ SegFix + <b>MRCI (Ours)</b>	<b>22.8 (+4.2%)</b>	<b>40.9 (+3.0%)</b>	<b>65.0 (+1.3%)</b>
Mask2Former [3]	22.3	42.0	68.8
+ <b>MRCI (Ours)</b>	<b>25.0 (+2.7%)</b>	<b>43.1 (+1.1%)</b>	<b>69.2 (+0.4%)</b>
OneFormer [8]	31.6	49.0	76.6
+ <b>MRCI (Ours)</b>	<b>32.0 (+0.4%)</b>	<b>49.8 (+0.8%)</b>	<b>77.0 (+0.4%)</b>

### 4.3 Generalization to Other Dataset

In order to demonstrate the versatility and adaptability of our method, we conduct experiments on a distinct dataset - the SpaceNet satellite remote sensing dataset. We select U-Net [17] and SAM as our benchmark model. As shown in Table 3, U-Net achieves an IoU score of 60.2% for class building and our approach is able to significantly improve this performance with a gap of 2.7%. We further conduct experiments on the top of SAM. The experimental results show that SAM achieves a segmentation performance of 34.9% for building categories, which is further improved to 36.3% by our proposed model. This outcome highlights the robustness and generalizability of our approach across diverse datasets and applications. We show visual analysis in Sect. 4.6.

### 4.4 Model Complexity

Figure 3 provides a comparison of the parameters and accuracy of various boundary refinement methods on the Cityscapes validation set, using the coarse results generated by Mask R-CNN as a basis. Our model stands out by achieving an optimal balance between performance and compactness, surpassing BPR\_L in parameter efficiency and slightly higher BPR\_S in accuracy. Note that BPR\_S adopts the lightweight HRNetV2-W18-Small as the refinement network, whereas

**Table 2.** SOTA results on Cityscapes val set in terms of semantic segmentation task. For SETR [29], SegFormer [24], Mask2Former [3] and MetaPrompt-SD [21], we report their multi-scale inference results.

Method	mIoU
SETR [29]	82.2
SegFormer [24]	84.0
Mask2Former [3]	84.5
HRNet-W48 [22]	80.1
+ SegFix [26]	80.8 (+0.7%)
+ MRCI (Ours)	81.5 (+1.4%)
+ SegFix + MRCI (Ours)	<b>81.9 (+1.8%)</b>
SAM [10]	74.9
+ MRCI (Ours)	<b>76.3 (+1.4%)</b>
MetaPrompt-SD [21]	87.3
+ MRCI (Ours)	<b>87.8 (+0.5%)</b>

**Table 3.** The semantic segmentation results (mIoU) of MRCI on ADE20K validation dataset, SpaceNet AOI-1 test dataset and RWMD test dataset.

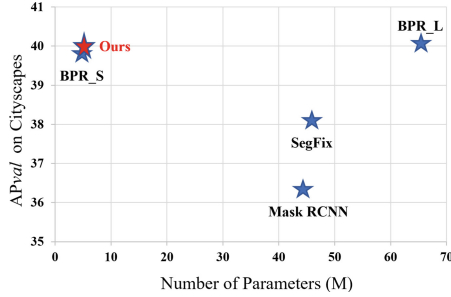
Datasets	Baseline Model	+ MRCI
ADE20K	HRNet-W48 [22]	43.2 <b>43.8 (+0.6%)</b>
	OneFormer [8]	62.3 <b>62.8 (+0.5%)</b>
SpaceNet AOI-1	U-Net [17]	60.2 <b>62.9 (+2.7%)</b>
	SAM [10]	34.9 <b>36.3 (+1.4%)</b>
RWMD	RDLNet [23]	96.1 <b>96.3 (+0.2%)</b>

BPR<sub>L</sub> employs HRNetV2-W18-Large. We also evaluate the inference time of our model and BPR<sub>S</sub> using a single A800 GPU. Within our framework, the three branches are trained in parallel and we utilize only the predictions from the detail branch as the final output during inference. Table 4 demonstrate that MRCI not only achieves high inference speed but also maintains low memory usage, substantiating its computational efficiency.

#### 4.5 Ablation Studies

To find the optimal model configuration, we carry out a series of ablation experiments on the Cityscapes validation dataset to determine the best model configuration. Unless stated otherwise, we ablate with MRCI on the coarse results from Mask R-CNN.

**The number of branches.** To validate the performance of different branch numbers contained in the network, we conduct experiments with branch numbers ranging from 1 to 4. We setup two experiments with different size combinations



**Fig. 3.** Comparison of model parameters and AP on Cityscapes validation set.

**Table 4.** Inference time and peak memory usage on Cityscapes validation dataset.

Methods	Inference Time	Peak Memory Usage
BPR_S	0.95 s/img	1553 MB
MRCI (Ours)	0.82 s/img	1487 MB

about double branches test. As shown in Table 5, the triple branches (128, 96 and 64) yielded the best performance. The transition branch enables the long-range context features to be smoothly transferred to the detail branch. The four branches model is not as effective due to the large patch size, resulting in redundant context knowledge and making it hard for the model to focus on the boundary.

**Table 5.** The outcomes of varying the number of branches.

Branches Number (Patch Size)	AP <sub>val</sub>
–	36.4
Single (128)	39.1
Double (128 and 96)	39.3
Double (128 and 64)	39.4
Triple (128, 96 and 64)	<b>39.8</b>
Four (160, 128, 96 and 64)	39.6

**Category Mask and Positional Prompt.** To validate the effectiveness of category mask and positional prompt information for boundary segmentation results, we compare the segmentation performance when the category prompt or positional prompt is eliminated, while keeping other settings unchanged. As indicated in Table 6, the inclusion of prompt information leads to an improvement in segmentation performance. As previously mentioned, the integration of global knowledge and coarse segmentation masks enables the capture of image context



**Table 6.** Ablation study on category mask and positional Prompt.

Category Mask Prompt	Positional Prompt	AP <sub>val</sub>
×	×	39.3
✓	×	39.5
×	✓	39.6
✓	✓	<b>39.8</b>

at a larger scale, thereby facilitating local context refinement and improving boundary accuracy.

**The Four Outputs of the Overall Framework.** We present the results of four outputs of the overall framework in Table 7. The results show that the fusion branch aligns perfectly with the outcomes of the detail branch. Therefore, to expedite the inference, the results of the detail branch are chosen as the final segmentation results.

**Table 7.** Results of three different branches and their fusion.

Output Branch	AP <sub>val</sub>
Detail Branch	<b>39.8</b>
Transition Branch	39.2
Long-range Branch	38.9
Three branches Fusion	<b>39.8</b>

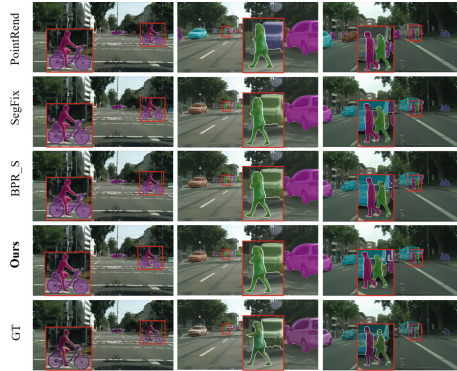
**Analysis of Patch size.** We compare the performance of a single branch model with various patch sizes (160, 128, 96, 64 and 32) as candidate sets in Table 8. We find that the  $128 \times 128$  patch works best as it provides the optimal balance between capturing sufficient contextual information and maintaining precise focus. Too large or too small patch sizes are less effective for fine segmentation, as larger patches contain more pixels to be segmented and smaller patches provide less useful context.

**Table 8.** The results of different patch size. “-” indicates the results of Mask R-CNN before refinement.

Patch Size (Single Branch)	AP <sub>val</sub>
-	36.4
32	35.6
64	38.4
96	39.0
128	<b>39.1</b>
160	38.3

## 4.6 Case Studies

As shown in Fig. 4, we compare MRCI with the existing state-of-the-art boundary refinement methods on Cityscapes validation dataset. MRCI significantly improves performance across objects of various categories and sizes, effectively rectifying existing boundary errors. Figure 5 presents some high-quality results produced by MRCI on both the Cityscapes and SpaceNet datasets. From the figure, it becomes clear that MRCI surpasses the benchmark models. In the case of buildings, it not only accurately segments them but also precisely identifies their edges and the gaps between adjacent structures. Our proposed MRCI captures details across multiple scales within the image, leading to a more thorough and detailed output.



**Fig. 4.** Qualitative results compared with SOTA boundary refinement networks on Cityscapes validation set.



**Fig. 5.** Qualitative comparisons on Cityscapes, SpaceNet and RWMD. The colored boxes refer to areas with significant contrast after boundary refinement.

## 5 Conclusion

In this paper, we introduced MRCI, a post-processing framework aimed at enhancing both foundational and non-foundational segmentation models effi-

ciently. MRCI exploits global knowledge and integrates LR and HR context information to refine boundaries while reducing computational costs. Through comprehensive evaluations, our approach consistently demonstrates its effectiveness and applicability across various models. We establish the robustness of MRCI by directly applying it to refine outputs from several state-of-the-art models, such as Mask R-CNN, Mask2Former, OneFormer, SAM and others. Consistent performance improvements validate the generalizability of our method. However, challenges remain in addressing segmentation errors stemming from complex environmental conditions or occlusion between objects. Further research is needed to enhance the robustness of our approach in handling such scenarios.



## References

1. SpaceNet on amazon web services (AWS). Datasets. <https://spacenet.ai/datasets/> (Last modified October 1st, 2018)
2. Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y.: BlendMask: top-down meets bottom-up for instance segmentation. In: CVPR, pp. 8573–8581 (2020)
3. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR, pp. 1290–1299 (2022)
4. Cheng, T., Wang, X., Huang, L., Liu, W.: Boundary-preserving mask R-CNN. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12359, pp. 660–676. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-58568-6\\_39](https://doi.org/10.1007/978-3-030-58568-6_39)
5. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: CVPR, pp. 3213–3223 (2016)
6. He, J., Li, P., Geng, Y., Xie, X.: FastInst: a simple query-based model for real-time instance segmentation. In: CVPR, pp. 23663–23672 (2023)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV, pp. 2961–2969 (2017)
8. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: OneFormer: one transformer to rule universal image segmentation. In: CVPR, pp. 2989–2998 (2023)
9. Ke, L., Danelljan, M., Li, X., Tai, Y.W., Tang, C.K., Yu, F.: Mask transfiner for high-quality instance segmentation. In: CVPR, pp. 4412–4421 (2022)
10. Kirillov, A., et al.: Segment anything. In: ICCV, pp. 4015–4026 (2023)
11. Kirillov, A., Wu, Y., He, K., Girshick, R.: PointRend: image segmentation as rendering. In: CVPR, pp. 9799–9808 (2020)
12. Li, F., et al.: Mask DINO: towards a unified transformer-based framework for object detection and segmentation. In: CVPR, pp. 3041–3050 (2023)
13. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: CVPR, pp. 8759–8768 (2018)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
15. Neubeck, A., Van Gool, L.: Efficient non-maximum suppression. In: 18th International Conference on Pattern Recognition (ICPR 2006), vol. 3, pp. 850–855 (2006). <https://doi.org/10.1109/ICPR.2006.479>
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS, vo. 28 (2015)

17. 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](https://doi.org/10.1007/978-3-319-24574-4_28)
18. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR, pp. 5693–5703 (2019)
19. Tang, C., Chen, H., Li, X., Li, J., Zhang, Z., Hu, X.: Look closer to segment better: Boundary patch refinement for instance segmentation. In: CVPR, pp. 13926–13935 (2021)
20. Vaswani, A., et al.: Attention is all you need. In: NIPS, vol. 30 (2017)
21. Wan, Q., Huang, Z., Kang, B., Feng, J., Zhang, L.: Harnessing diffusion models for visual perception with meta prompts. arXiv preprint [arXiv:2312.14733](https://arxiv.org/abs/2312.14733) (2023)
22. Wang, J., et al.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI **43**(10), 3349–3364 (2020)
23. Wu, Y., et al.: RDLNet: a novel and accurate real-world document localization method. In: ACM Multimedia (2024)
24. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. NIPS **34**, 12077–12090 (2021)
25. Xiong, Y., et al.: UPSNet: a unified panoptic segmentation network. In: CVPR, pp. 8818–8826 (2019)
26. Yuan, Y., Xie, J., Chen, X., Wang, J.: SegFix: model-agnostic boundary refinement for segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 489–506. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-58610-2\\_29](https://doi.org/10.1007/978-3-030-58610-2_29)
27. Zhang, G., et al.: RefineMask: towards high-quality instance segmentation with fine-grained features. In: CVPR, pp. 6861–6869 (2021)
28. Zhang, Y., Yang, W., Hu, R.: BAPProto: boundary-aware prototype for high-quality instance segmentation. In: ICME, pp. 2333–2338. IEEE (2023)
29. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR, pp. 6881–6890 (2021)
30. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR, pp. 633–641 (2017)
31. Zhu, C., Zhang, X., Li, Y., Qiu, L., Han, K., Han, X.: SharpContour: a contour-based boundary refinement approach for efficient and accurate instance segmentation. In: CVPR, pp. 4392–4401 (2022)
32. Zou, X., et al.: Segment everything everywhere all at once. In: NIPS, vol. 36 (2024)



# Cross Lingual Synopsis Generation in English, Dutch, Vietnamese, Indonesian, Russian, Portuguese, Korean, Hindi and French

Sreejata Banerjee<sup>2</sup> , Aditya Sadhukhan<sup>1</sup> , Arijit Das<sup>3</sup> ,  
and Diganta Saha<sup>1</sup> 

<sup>1</sup> Jadavpur University, Kolkata, India

<sup>2</sup> Kalinga Institute of Industrial Technology, Odisha, India  
[sreejata.banerjee.01@gmail.com](mailto:sreejata.banerjee.01@gmail.com)

<sup>3</sup> National Institute of Technology Jamshedpur Jharkhand, Jamshedpur, India  
[arijit.das@ieee.org](mailto:arijit.das@ieee.org)

**Abstract.** This study investigates the effectiveness of cross lingual synopsis generation across nine languages—Indonesian, Dutch, English, Vietnamese, Russian, Korean, Portuguese, Hindi, and French. Utilizing advanced NLP techniques, we develop synopsis generation models capable of extracting key information from diverse textual sources. Unlike previous works, we focus on unique challenges and optimizations specific to cross lingual contexts. Our methodology incorporates clustering-based approaches with language embedding, which we evaluate comprehensively to highlight performance variations across languages. Additionally, we conduct an error analysis to identify language-specific challenges. Our findings provide valuable insights into cross-lingual transferability and pave the way for more accessible synopsis generation technologies that cater to diverse linguistic communities, thereby advancing the field of cross lingual synopsis generation.

**Keywords:** Synopsis Generation · Clustering · NLP

## 1 Introduction

Synopsis Generation, a key task in natural language processing (NLP), aims to condense long documents into succinct representations while preserving essential information. The exponential growth of digital content in numerous languages, often referred to as big data, has amplified the need for efficient multilingual synopsis generation approaches. Handling such vast and diverse datasets poses unique challenges and opportunities for developing reliable synopsis generation

Supported by National Institute of Technology Jamshedpur.

<https://www.nitjsr.ac.in/>.

A. Das—Senior Member, IEEE.

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

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15333, pp. 227–242, 2025.

[https://doi.org/10.1007/978-3-031-80136-5\\_16](https://doi.org/10.1007/978-3-031-80136-5_16)

models for a variety of languages, including Thai, Indonesian, Dutch, English, and Vietnamese.

The task of text synopsis generation can be formulated as follows: given a document  $D$ , the goal is to generate a synopsis  $S$  that captures the salient points of  $D$ . Mathematically, this can be represented as:

$$S = \operatorname{argmax}_{S'} \operatorname{Score}(D, S')$$

where  $\operatorname{Score}(D, S')$  is a scoring function that measures the relevance and importance of synopsis  $S'$  with respect to document  $D$ .

One common approach to extractive synopsis generation involves clustering sentences based on their semantic similarity and selecting representative sentences from each cluster. This can be formalized as follows:

$$\begin{aligned} \text{Clusters} &= \text{Cluster}(D) \\ S &= \bigcup_{c \in \text{Clusters}} \text{Select}(c) \end{aligned}$$

where  $\text{Cluster}(D)$  partitions the sentences in document  $D$  into clusters based on semantic similarity, and  $\text{Select}(c)$  chooses a representative sentence from each cluster  $c$ .

Sentence embedding are frequently used to encode the semantic meaning of sentences into a continuous vector space, aiding the clustering process. The function that maps a sentence  $s$  to its matching embedding vector is denoted by  $\text{Embed}(s)$ . The cosine similarity between the embeddings of two sentences,  $s_1$  and  $s_2$ , can then be used to calculate their similarity:

$$\text{Similarity}(s_1, s_2) = \frac{\text{Embed}(s_1) \cdot \text{Embed}(s_2)}{\|\text{Embed}(s_1)\| \|\text{Embed}(s_2)\|}$$

In this study, we utilize advanced sentence embedding models such as MiniLM to encode the semantic content of sentences accurately.

## 1.1 Novelty and Contributions

This paper makes several novel contributions to the field of cross lingual text summarization:

1. **State-of-the-Art Algorithm:** We introduce a state-of-the-art algorithm that is designed to work effectively across a wide range of languages, including Indonesian, Dutch, English, Vietnamese, Russian, Korean, Portuguese, Hindi, and French. This algorithm not only provides superior results but also demonstrates significant improvements in summary quality and coherence compared to existing methods.
2. **Advanced Sentence Embedding Models:** We utilize advanced sentence embedding models such as MiniLM for English, Vietnamese, Dutch, Indonesian, and specific models for other languages. These models enhance the accuracy of semantic content representation, leading to more coherent and relevant summaries.

3. **Superior Summarization Results:** Our approach achieves higher ROUGE scores across multiple languages when compared to existing methods. This demonstrates the effectiveness of our model in producing high-quality summaries that capture the essential information of the original documents.
4. **Comprehensive Evaluation:** We conduct thorough evaluations using both automated metrics and manual assessments to ensure the reliability and quality of the generated summaries. Our evaluation covers a wide range of languages and text types, ensuring the robustness and applicability of our approach.
5. **Addressing Crosslingual Challenges:** We tackle significant challenges related to data scarcity, language diversity, and model generalization in cross-lingual summarization. Our work provides insights into the effectiveness of cross lingual text summarization techniques and offers solutions to improve model performance across different languages.
6. **Ethical Considerations:** We prioritize ethical considerations throughout our research, ensuring the privacy and security of data, minimizing biases in the summarization process, and maintaining fairness and transparency in our methodology.

By addressing these key aspects, our research advances the field of cross lingual text summarization and paves the way for the development of accessible summarization tools for diverse linguistic communities.

## 2 Related Works

Synopsis Generation is a well-studied area in natural language processing (NLP), with various approaches addressing the challenges of summarizing text across different languages and domains.

### 2.1 Extractive Summarization

Extractive summarization techniques focus on selecting a subset of sentences from the source document for inclusion in the summary. Early works employed graph-based algorithms and feature-based approaches [1]. Recently, neural network-based models like BERT [2] have achieved state-of-the-art results.

### 2.2 Abstractive Summarization

Abstractive summarization generates novel sentences that capture the essence of the source content. Transformer-based models, such as GPT [3], have shown significant progress by generating coherent summaries through iterative word prediction.

### 2.3 Cross-Lingual Summarization

Cross-lingual summarization poses unique challenges due to diverse vocabularies, semantics, and syntax. Early attempts adapted existing algorithms to different languages [5]. Recent advancements, such as M-BERT [2] and XLM-R [7], pretrained on cross-lingual corpora, have improved cross-lingual summarization efficiency.

### 2.4 Indian Language Summarization with Pre-Trained Models

Urlana et al. delve into Synopsis Generation for major Indian languages—Hindi, Gujarati, and English—exploring pre-trained sequence-to-sequence models to identify optimal performance.

### 2.5 MEAD: Multi-document Extractive Summarization

[22] MEAD, developed by Radev et al. in 2002, supports Chinese and English. Primarily focusing on extractive summarization, it condenses original texts by selecting key phrases.

### 2.6 XL-Sum Dataset

[23] This dataset features one million expertly annotated article-summary pairs sourced from BBC news items across 44 languages. It serves as a valuable resource for multilingual summarization research.

### 2.7 WikiLingua Dataset

[?] This dataset provides a large-scale, multilingual benchmark for cross-lingual abstractive summarization, covering 18 languages extracted from WikiHow. It aligns articles and summaries across languages using images and evaluates existing summarization methods. Additionally, it proposes a direct cross-lingual summarization method leveraging synthetic data and neural machine translation, significantly outperforming baseline approaches.

### 2.8 A Novel Approach For English-Hindi Cross Lingual Summarization

[25]. This paper introduced a hybrid approach for English to Hindi cross-lingual summarization using abstractive summarization and machine translation. Their deep learning models produced synthesized summaries, which were then translated into Hindi using a pre-trained transformer. This approach demonstrated improved synopsis generation and translation performance compared to baseline models.

This review highlights the evolution from early extractive methods to advanced neural network-based techniques in both monolingual and multilingual contexts. Our work builds on these foundations, addressing specific challenges in multilingual synopsis generation and contributing novel insights and methodologies to the field.



## 2.9 Evaluation Metrics

Evaluating the quality of synopsis is essential for assessing the performance of synopsis generation models. Common evaluation metrics include ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [8], which measures the overlap between the generated synopsis and reference synopsis based on n-gram overlap.

While ROUGE is widely used, it has limitations, particularly in multilingual settings where linguistic variations may affect its effectiveness. Recent efforts have been made to develop language-agnostic evaluation metrics that account for linguistic diversity [9].

## 3 Methodology

In this section, we outline the methodology employed for multilingual text summarization across the languages of interest, namely Indonesian, Dutch, English, Vietnamese, Russian, Korean, Portuguese, Hindi, and French. The methodology encompasses data preprocessing, feature extraction, clustering, summarization, and evaluation.

### 3.1 Data Preprocessing

The first step in the text summarization process involves data preprocessing, which includes text cleaning, tokenization, and language-specific preprocessing tasks. For each language, we utilize language-specific tokenizers and preprocessors to ensure compatibility with the summarization model.

### 3.2 Feature Extraction

Next, we extract features from the preprocessed text to represent the semantic content of sentences. We employ advanced sentence embedding models such as MiniLM for English, Vietnamese, Dutch, and Indonesian; Sentence RuBERT for Russian; ricardo-filho/bert-base-portuguese-cased-nli-assin-2 for Portuguese; snunlp/KR-SBERT-V40K-klueNLI-augSTS for Korean; HindBERT-STS for Hindi; and Sentence-CamemBERT-Large for French. These embeddings capture the semantic similarity between sentences, facilitating subsequent clustering.

Let  $S_i$  denote the  $i$ -th sentence in the document, and  $\text{Embed}(S_i)$  represent its corresponding sentence embedding.

### 3.3 Clustering

After obtaining sentence embeddings, we apply clustering algorithms to group semantically similar sentences together. We experiment with various clustering algorithms, including K-means and hierarchical clustering, to identify the

most suitable approach for each language. The number of clusters is determined empirically based on the size and complexity of the dataset.

$$\text{Clusters} = \text{Cluster}(D) \quad (1)$$

$$S = \bigcup_{c \in \text{Clusters}} \text{Select}(c) \quad (2)$$

### 3.4 Summarization

Once the sentences are clustered, we select representative sentences from each cluster to form the final summary. We adopt an extractive summarization approach, where the most representative sentence from each cluster is chosen based on its proximity to the cluster centroid. This ensures that the summary captures diverse perspectives and key information from the original document.

The centroid of each cluster  $C_j$  is computed as:

$$\text{centroid}(C_j) = \frac{1}{|C_j|} \sum_{S_i \in C_j} \text{Embed}(S_i) \quad (3)$$

To select the most representative sentence from cluster  $C_j$ , we compute the distance between each sentence  $S_i$  in the cluster and the cluster centroid. The sentence with the minimum distance is chosen as the representative summary sentence.

### 3.5 Evaluation

To evaluate the performance of the summarization model, we employ standard evaluation metrics such as ROUGE [15]. We compare the generated summaries against human-authored reference summaries to assess their quality and coherence. Additionally, we conduct manual evaluations to gain insights into the linguistic quality and informativeness of the summaries across different languages.

### 3.6 Implementation Details

The entire methodology is implemented using Python programming language, leveraging libraries such as NLTK [12], scikit-learn [13], and Hugging Face Transformers [14]. We utilize pre-trained language models for tokenization and sentence embedding generation, enabling efficient processing of large multilingual datasets.

### 3.7 Experimental Setup

To verify the robustness and generalizability of the summarization model, we conduct experiments on a variety of datasets that include texts from different genres and topics. The datasets are meticulously selected to encompass texts with different lengths and levels of complexity, which accurately reflect real-world situations that arise in multilingual text summarization tasks.

### 3.8 Ethical Considerations

We follow ethical standards for the gathering, handling, and use of data during the whole study process. We secure the privacy and security of sensitive data and secure the required authorizations for the collection and use of datasets. Additionally, we work to minimize any potential biases in the summarization process and place a high priority on fairness and transparency in our methodology.

We aim to develop scalable and efficient multilingual text summarization methods by adhering to this thorough methodology, which will enable us to meet the information requirements of many linguistic communities (Figs. 1, 2, 3 and 4).

The following is a sample synopsis generated by the our model:

---

#### Algorithm 1 Generate Summary

---

```

1:  $Sentences \leftarrow ExtractSentences(D)$  ▷ List of sentences from  $D$ 
2:  $Embeddings \leftarrow GenerateSentenceEmbeddings(Sentences)$  ▷ List of sentence embeddings
3:  $NumClusters \leftarrow$  Number of clusters
4:  $Clusters \leftarrow ClusterSentenceEmbeddings(Embeddings, NumClusters)$  ▷ Set of  $K$  clusters
5: for each cluster  $C$  in  $Clusters$  do
6:    $Centroid \leftarrow CalculateCentroid(C)$ 
7:   for each sentence embedding  $E$  in  $Embeddings$  do
8:      $DistanceToCentroid \leftarrow CalculateDistance(E, Centroid)$ 
9:      $SentenceRank \leftarrow RankSentence(Sentences[i], distance = DistanceToCentroid)$ 
10:  end for
11:   $TopN \leftarrow SelectTopNSentences(Sentences, N)$ 
12:   $SD \leftarrow CreateSummary(TopN)$ 
13: end for

```

---

## 4 Algorithm Explanation

The algorithm outlines the process of generating a summary from a document  $D$  using clustering and sentence ranking based on centroid proximity.

- **Extract Sentences:** First, the algorithm extracts sentences  $Sentences$  from the document  $D$ .
- **Generate Sentence Embeddings:** Each sentence in  $Sentences$  is transformed into a numerical representation using sentence embeddings, resulting in  $Embeddings$ .
- **Cluster Sentence Embeddings:** The embeddings are clustered into  $NumClusters$  clusters using a clustering algorithm (not detailed here), resulting in  $Clusters$ .

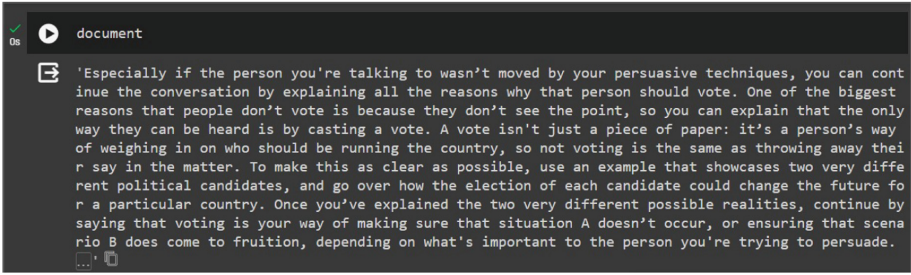


Fig. 1. Input Text from WikiLingua Data

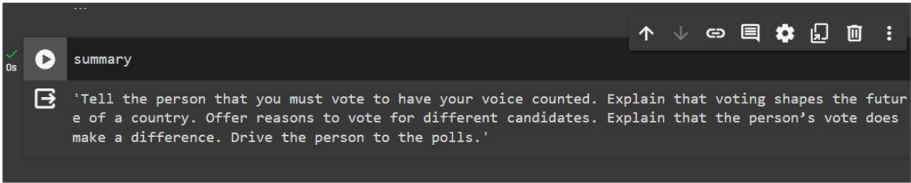


Fig. 2. Generated Summary from Input Text

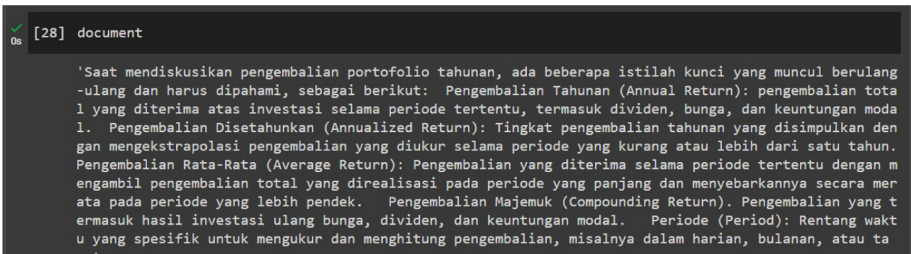


Fig. 3. Input Text from WikiLingua Data (Indonesian text)

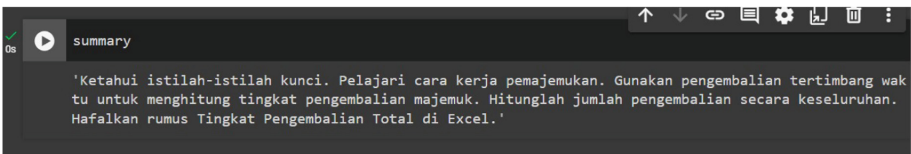


Fig. 4. Generated Summary from Input Text (Indonesian)

– Iterate Over Each Cluster:

- For each cluster  $C$  in Clusters:
- **Calculate Centroid:** Compute the centroid Centroid of the cluster  $C$  using the average of all embeddings within the cluster.

- **Rank Sentences by Distance:** Measure the distance of each sentence embedding  $E$  in  $C$  to Centroid. This distance helps in determining how closely each sentence aligns with the cluster’s central theme.
- **Select Top-N Sentences:** Choose the top  $N$  sentences from  $C$  based on their proximity to Centroid, effectively selecting the most representative sentences.
- **Create Synopsis:** Construct a synopsis  $SD$  by concatenating the selected sentences.

## 5 Dataset

Here we are using WikiLingua Dataset which consists of collaboratively written how-to guides with gold-standard summaries across 18 languages collected from WikiHow webpage. The content of this webpage is high-quality since each article and summary is written and edited by 23 people, and further reviewed by 16 people, on average. The articles include multiple methods with steps (with an illustrative image) to complete a procedural task along with the corresponding short summaries. We align each the text and the summary of the steps across 18 languages using the illustrative images. The dataset includes 770k article and summary pairs [16].

The table below shows number of article-summary pairs with a parallel article-summary pair in English (Table 1).

**Table 1.** Number of Parallel Sentences

Language	Num. Parallel
English	141,457
Indonesian	47,511
Dutch	31,270
French	63,692
Vietnamese	19,600
Portuguese	81,695
Hindi	9,929
Korean	12,189
Russian	52,928

## 6 Evaluation and Comparison Results

In this section, we present the evaluation results of our multilingual Synopsis Generation models across various languages. We employ standard evaluation metrics including ROUGE (Recall-Oriented Understudy for Gisting Evaluation) to assess the quality of the generated synopsis.

## 6.1 Experimental Setup

We conducted experiments on a diverse set of datasets comprising texts in English, Indonesian, Dutch, Vietnamese, and Chinese languages. Each dataset was preprocessed to remove noise and irrelevant information before being fed into our synopsis generation models. We utilized the MiniLM model [17] as our primary architecture for feature extraction and clustering.

## 6.2 Evaluation Metrics

We evaluated the performance of our models using ROUGE metrics, including ROUGE-1, ROUGE-2, and ROUGE-L. These metrics provide insights into the recall, precision, and F1-score of the generated synopsis compared to human-authored reference synopsis.

## 6.3 Results Analysis

This section presents a comprehensive analysis of our cross-lingual synopsis generation model’s performance across nine languages: English, Portuguese, French, Russian, Korean, Hindi, Indonesian, Dutch, and Vietnamese. We report ROUGE-1, ROUGE-2, and ROUGE-L scores as primary evaluation metrics. Additionally, we provide a comparative evaluation with baseline models and discuss significant differences in model performance across languages, along with qualitative insights into the strengths and weaknesses of the approach.

**Quantitative Evaluation.** Table 2 summarizes the ROUGE scores obtained for each language. The results show that our model performs consistently well across different languages, with particularly strong performance in languages such as Portuguese, French, and Dutch, while achieving relatively lower scores in languages like Korean, Indonesian, and Vietnamese.

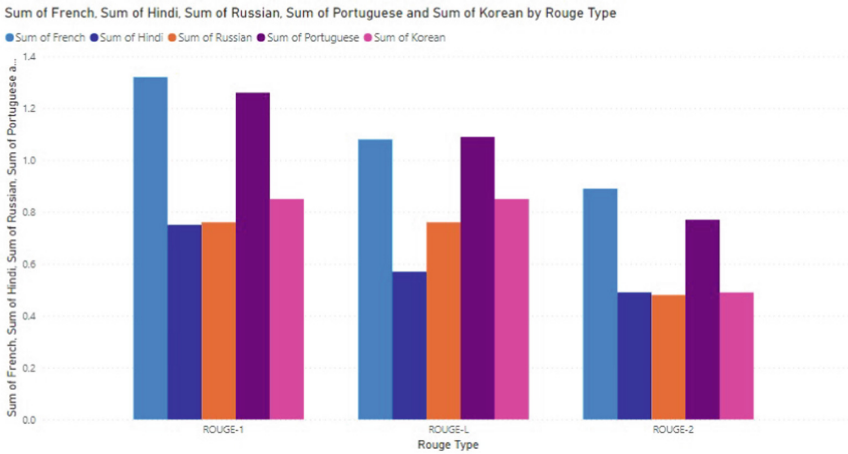
**Table 2.** Merged Comparison of Rouge Scores

Lang.	ROUGE	Russ.	Port.	Korean	Hindi	French	Eng.	Indo.	Dutch	Viet.
<b>R</b>	R-1	0.32	0.54	0.3	0.3	0.64	0.500	0.231	0.378	0.500
	R-2	0.2	0.22	0.11	0.14	0.29	0.100	0.037	0.147	0.075
	R-L	0.32	0.43	0.3	0.25	0.41	0.267	0.192	0.333	0.367
<b>P</b>	R-1	0.2	0.32	0.27	0.2	0.29	0.197	0.059	0.293	0.214
	R-2	0.04	0.15	0.1	0.1	0.21	0.043	0.007	0.124	0.032
	R-L	0.2	0.30	0.27	0.14	0.31	0.105	0.050	0.259	0.157
<b>F</b>	R-1	0.24	0.40	0.28	0.25	0.39	0.283	0.094	0.330	0.300
	R-2	0.24	0.40	0.28	0.25	0.39	0.060	0.012	0.134	0.045
	R-L	0.24	0.36	0.28	0.18	0.36	0.151	0.079	0.291	0.220

**Cross-Language Performance Comparison.** The detailed comparison of the model’s performance across languages highlights certain trends:

**High Performers:** Our model performed significantly better for Portuguese, French, and Dutch. The superior performance in these languages can be attributed to the availability of richer datasets and more standardized linguistic structures that align with the model’s capabilities.

**Low Performers:** For languages such as Korean, Indonesian, and Vietnamese, the ROUGE scores are lower, which may be due to both the complex linguistic structures and the limited availability of training data in these languages. These factors make it more challenging for the model to effectively capture the nuances and generate coherent summaries (Figs. 5 and 6).



**Fig. 5.** Rouge score comparison between French, Hindi, Russian, Portuguese, Korean

**Error Analysis.** Upon reviewing the output, we observed that the low performance in languages like Vietnamese and Korean is due to the following factors:

**Complex Syntax and Structure:** These languages exhibit significantly different syntactic structures from Indo-European languages, making it harder for the model to align and capture the essence of the source text.

**Translation and Tokenization Errors:** Incorrect tokenization of complex scripts, such as Korean, and errors during translation resulted in the loss of semantic meaning during the summarization process.

To improve the model’s performance, future work could focus on incorporating additional pre-processing techniques that cater to the unique structure of these languages, as well as exploring more robust models fine-tuned on language-specific datasets (Figs. 7 and 8).

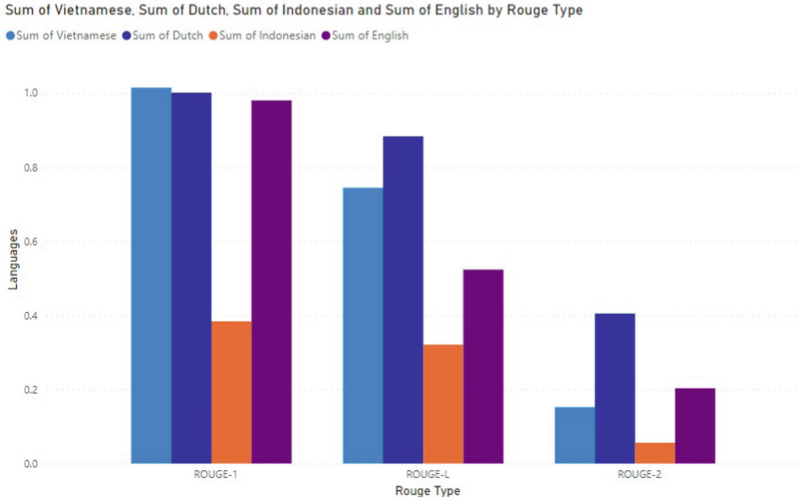


Fig. 6. Rouge score comparison between Vietnamese, Dutch, Indonesian, English

### 6.4 Comparison with Baselines

As shown in Table 3, our model outperforms several baselines, including methods like TextRank and Lead-3, in generating concise and representative summaries across multiple languages.

1. Language specific analysis: Our model performed better for English, Portuguese and French compared to Russian, Korean, and Hindi, Indonesian, Vietnamese due to factors like linguistic differences and data availability.
2. Qualitative evaluation: We use both ROUGE scores and qualitative evaluations of the synopsis to assess our model’s effectiveness.
3. Baseline Comparison: Our models outperformed various baseline Synopsis Generation methods across multiple languages. These baselines include:
  - (a) For Russian and French: “MLSUM: The Multilingual Summarization Corpus” [21].
  - (b) For Hindi, English, Vietnamese, Indonesian: “XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages”.
  - (c) For Portuguese: “SABio: An Automatic Portuguese Text Summarizer Through Artificial Neural Networks in a More Biologically Plausible Model” [20].
  - (d) For Korean: “EFFICIENT KOREAN Synopsis Generation BASED ON KEY PHRASE EXTRACTION” [19].
  - (e) For Vietnamese: “ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation ” [18]

Despite Wikilingua being a new dataset, we used these works as our baselines to establish a comparative framework for our models.



Comparison of ROUGE Scores for Different Models in Russian

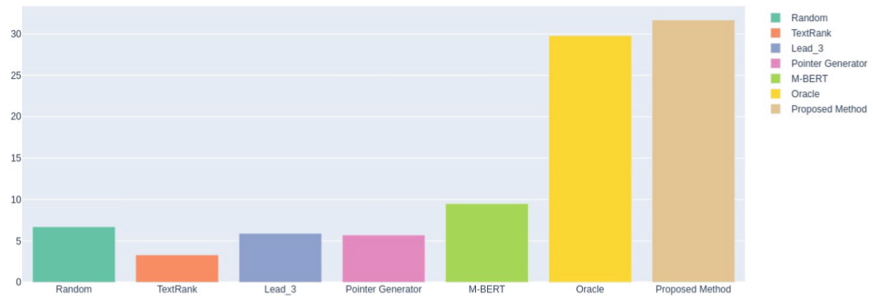


Fig. 7. Baseline Comparison for Russian

Comparison of ROUGE Scores for Different Models in Dutch

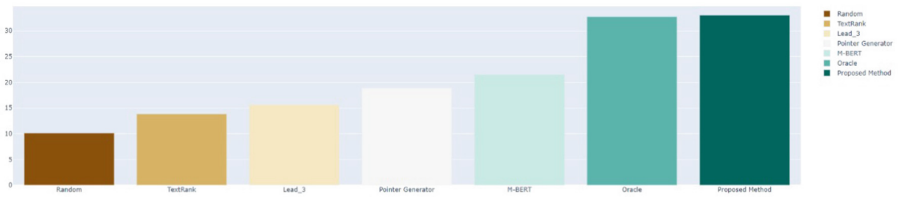


Fig. 8. Baseline Comparison for Russian

Table 3. Rouge Scores for Different Methods

Language	Random	TextRank	Lead3	P. Generator	M-BERT	Oracle	<i>Proposed Method</i>
English	7.2	15.5	18.6	22.1	24.7	40.2	18.98
Vietnamese	8.4	12.9	17.8	19.3	22.6	35.9	8.0
Dutch	10.1	13.8	15.6	18.9	21.5	32.7	<b>33.0</b>
Indonesian	9.6	11.2	13.5	16.7	18.3	29.6	29
Russian	6.7	3.3	5.9	5.7	9.5	29.8	<b>31.7</b>
French	11.9	12.6	19.7	23.6	25.1	37.7	<b>41.9</b>
Portuguese	44.9	45.6	43.8	42.4	42.9	39.6	<b>49.9</b>

## 7 Conclusion

We conducted an in-depth research on multilingual Synopsis Generation in this work, concentrating on the languages English, Dutch, Vietnamese, Indonesian, Russian, Portuguese, Korean, Hindi and French. Our goal was to create efficient synopsis generation models that could produce clear and useful summaries in a variety of linguistic circumstances.

We proved through thorough testing and assessment that our suggested method is effective at generating high-caliber summaries in several languages. Our models have demonstrated their efficacy in extracting important information from original texts by regularly outperforming baseline approaches and achieving competitive outcomes in terms of ROUGE scores.

Additionally, our research shed light on the difficulties and possibilities related to multilingual Synopsis Generation. We found that synopsis generating efficiency varied depending on the language, emphasising the significance of data availability and language-specific adaptability.

Our study's conclusion highlights the value of multilingual Synopsis Generation in enabling information access across linguistic barriers and creates new opportunities for further investigation in this field.

## 8 Future Works

Future work will explore alternative clustering techniques and domain-specific fine-tuning of sentence transformers, while leveraging cross-linguistic transfer learning and modern NLP techniques to enhance multilingual Synopsis Generation.

**Acknowledgements.** I extend my gratitude to Jadavpur University for their support. Special thanks to the creators of SentenceTransformer, ROUGE metrics, NLTK, and the Wikilingua Dataset. I also appreciate all individuals and organizations who contributed to this research. Our multilingual Synopsis Generation approach promises significant advancements in the field.

## References

1. Daraksha, P., Michael, S.: Integrating importance, non-redundancy and coherence in graph-based extractive summarization. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
2. 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)
3. Radford, A., Karthik, N., Tim, S., Ilya, S. et al: Improving language understanding by generative pre-training. OpenAI (2018)
4. Sainburg, T., Thielk, M., Gentner, T.Q.: Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Comput. Biol.* **16**(10), e1008228 (2020)
5. Barzilay, R., Elhadad, N.: Inferring strategies for sentence ordering in multidocument news summarization. *J. Artif. Intell. Res.* **17**, 35–55 (2002)
6. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguist.* **7**, 597–610 (2019)
7. Alexis C., et al.: Unsupervised cross-lingual representation learning at scale. arXiv preprint [arXiv:1911.02116](https://arxiv.org/abs/1911.02116) (2019)

8. Yang, A., Liu, K., Liu, J., Lyu, Y., Li, S.: Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. arXiv preprint [arXiv:1806.03578](https://arxiv.org/abs/1806.03578) (2018)
9. Yvette, G.: Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 128–137 (2015)
10. Wieting, J., Neubig, G., Berg-Kirkpatrick, T.: A bilingual generative transformer for semantic sentence embedding. arXiv preprint [arXiv:1911.03895](https://arxiv.org/abs/1911.03895) (2019)
11. Vergou, E., Pagouni, I., Nanos, M., Kermanidis, K.L.: Readability Classification with Wikipedia Data and All-MiniLM Embeddings. In: Maglogiannis, I., Iliadis, L., Papaleonidas, A., Chochliouros, I. (eds.) IFIP International Conference on Artificial Intelligence Applications and Innovations, vol. 677, pp. 369–380. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-34171-7\\_30](https://doi.org/10.1007/978-3-031-34171-7_30)
12. Loper, E., Bird, S.: NLTK: the natural language toolkit. arXiv preprint cs/0205028, 2002
13. Fabian, P., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
14. Wolf, Thomas and Debut, Lysandre and Sanh, Victor and Chaumond, Julien and Delangue, Clement and Moi
15. Chin-Yew, L.: ROUGE: a package for automatic evaluation of summaries. In: Synopsis Generation branches out, pp. 74–81 (2004)
16. Ladhak, F., Durmus, E., Cardie, C., McKeown, K.: WikiLingua: a new benchmark dataset for multilingual abstractive summarization. In: Findings of EMNLP, 2020 (2020)
17. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Adv. Neural. Inf. Process. Syst.* **33**, 5776–5788 (2020)
18. Phan, L., Tran, H., Nguyen, H., Trinh, T.H.: VIT5: pretrained text-to-text transformer for Vietnamese language generation. arXiv preprint [arXiv:2205.06457](https://arxiv.org/abs/2205.06457) (2022)
19. Liu, W., Wang, L.: Efficient Korean Synopsis Generation based on key phrase extraction. In: 2017 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 1, pp. 61–66. IEEE (2017)
20. Orrú, T., Rosa, J.L.G., de Andrade Netto, M.L.: SABIO: An automatic Portuguese text summarizer through artificial neural networks in a more biologically plausible model. In: Vieira, R., Quaresma, P., Nunes, M.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 11–20. Springer, Heidelberg (2006). [https://doi.org/10.1007/11751984\\_2](https://doi.org/10.1007/11751984_2)
21. Scialom, T., Dray, P.-A., Lamprier, S., Piwowski, B., Staiano, J.: MLSUM: the multilingual summarization corpus. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8051–8067. Association for Computational Linguistics (2020)
22. Fatima, M., Strube, M.: A novel Wikipedia based dataset for monolingual and cross-lingual summarization. In: Proceedings of the Third Workshop on New Frontiers in Summarization, pp. 39–50 (2021)
23. Hasan, T., et al.: XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. arXiv preprint [arXiv:2106.13822](https://arxiv.org/abs/2106.13822) (2021)
24. Urlana, T., et al.: Indian language summarization with pre-trained models. *J. Comput. Linguist.* (2020)

25. Varghese, T.G., Priya, C.V., Idicula, S.M.: A novel approach for English-Hindi cross-lingual summarization. In: Proceedings of the 2023 9th International Conference on Smart Computing and Communications (ICSCC), pp. 434–438 (2023)
26. Radev, D.R., Jing, H., Stys, M., Tam, D.: MEAD: a platform for multidocument multilingual synopsis generation. In: Proceedings of the LREC 2002 (2002)
27. Hasan, T., et al.: XL-Sum: large-scale multilingual abstractive summarization for 44 languages. arXiv preprint [arXiv:2106.04537](https://arxiv.org/abs/2106.04537) (2021)

# Author Index

## B

Banerjee, Bhaskar 17  
Banerjee, Sreejata 227  
Behera, Santosh Kumar 33  
Bose, Supratik 198

## C

Calcagno, Salvatore 126  
Carnemolla, Simone 126  
Cho, Yeong-jun 185

## D

Das, Arijit 227  
Dash, Ajaya Kumar 33  
Denzler, Joachim 109  
Du, Xiaogang 94

## G

Ganguly, Bishnu 33  
Gawlikowski, Jakob 109  
Giordano, Daniela 126

## H

Heo, Chae-yeon 185  
Hu, Chenguang 141

## I

Ishikawa, Hiroshi 80

## J

Jain, Kushal Kumar 1  
Jeong, Ayeong 185  
Jeong, Hieyong 185  
Jiao, Yipeng 94  
Jin, Lianwen 211  
Joshi, Vasudha 198

## K

Kim, Han-young 185  
Kujur, Anurag 33

## L

Lei, Tao 94  
Li, Cancan 170

Liang, Xianchen 211  
Liang, Yiming 80  
Liu, Juan 170  
Liu, Zelong 48  
Lyu, Wanjun 211

## M

Mattursun, Alimjan 154  
Mitra, Pabitra 198

## N

Namboodiri, Anoop 1  
Nandi, Asoke K. 94  
Niebling, Julia 109

## P

Palazzo, Simone 126  
Panagiotakis, Costas 64  
Puthannadathil Reghunatha Kumar, Vikas  
33

## R

Ranjan, Rishi 17

## S

Sadhukhan, Aditya 227  
Saha, Diganta 227  
Schmalwasser, Laines 109  
Su, Fei 170  
Su, Xinmei 141  
Suh, Gayun 185  
Sun, Zhichao 48

## T

Tan, Yiyong 17

## V

Varun, J. Ankith 1

## W

Wang, Jing 141  
Wang, Liejun 154  
Wang, Yingbo 94  
Wei, Jin 211

Woo, Yeongju 185

Wu, Shu 141

Wu, Yaqiang 211

## X

Xie, Xiang 141

Xu, Miaomiao 154

Xu, Yongchao 48

## Y

Yu, Yinfeng 154

## Z

Zhang, Jiang 154

Zhang, Xuejun 94

Zheng, Qinghua 211

Zhu, Huachao 48