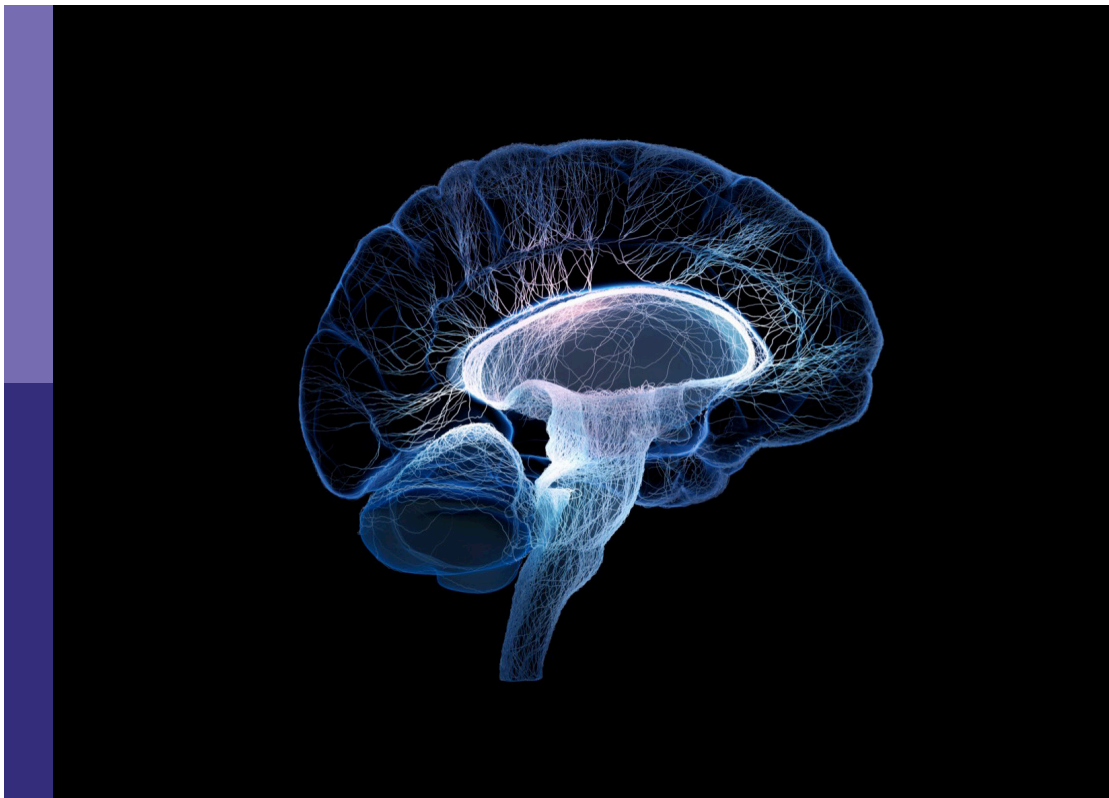
frontiers | Research Topics

Advances in computer vision: from deep learning models to practical applications

Edited by
Hancheng Zhu, Lu Tang, Rui Yao and Yanqiu Huang

Published in
Frontiers in Neuroscience



ISSN 1664-8714
ISBN 978-2-8325-6373-1
DOI 10.3389/978-2-8325-6373-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Advances in computer vision: from deep learning models to practical applications

Topic editors

Hancheng Zhu — China University of Mining and Technology, China
Lu Tang — Xuzhou Medical University, China
Rui Yao — China University of Mining and Technology, China
Yanqiu Huang — University of Twente, Netherlands

Citation

Zhu, H., Tang, L., Yao, R., Huang, Y., eds. (2025). *Advances in computer vision: from deep learning models to practical applications*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-6373-1

Table of contents

frontiers | Frontiers in Neuroscience

TYPE Editorial
PUBLISHED 09 May 2025
DOI 10.3389/fnins.2025.1615276

OPEN ACCESS

EDITED AND REVIEWED BY
Yi Bao,
Stevens Institute of Technology, United States

*CORRESPONDENCE
Lu Tang
✉ xztanglu@xzhmu.edu.cn

RECEIVED 21 April 2025
ACCEPTED 25 April 2025
PUBLISHED 09 May 2025

CITATION
Zhu H, Yao R and Tang L (2025) Editorial: Advances in computer vision: from deep learning models to practical applications. *Front. Neurosci.* 19:1615276. doi: 10.3389/fnins.2025.1615276

Editorial: Advances in computer vision: from deep learning models to practical applications

Hancheng Zhu[1], Rui Yao[1] and Lu Tang[2]*

[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China, [2]Xuzhou Medical University, Xuzhou, China

KEYWORDS

computer vision, deep learning models, practical applications, lightweight network, medical image analysis and processing system

Editorial on the Research Topic
Advances in computer vision: from deep learning models to practical applications

Computer vision has emerged as one of the most transformative areas of artificial intelligence, with deep learning models driving unprecedented advancements in both theoretical understanding and practical applications. Over the past decade, the rapid development of deep learning techniques has enabled machines to perform tasks such as image recognition, object detection, and video analysis with remarkable accuracy and efficiency. However, as the field continues to evolve, there is a growing need to bridge the gap between theoretical models and real-world applications to ensure that these technologies are powerful but also practical, efficient, and scalable. This Research Topic, "*Advances in computer vision: from deep learning models to practical applications*," is dedicated to exploring the latest innovations in computer vision that are addressing these challenges and pushing the boundaries of what is achievable.

The articles in this Research Topic represent a diverse range of research directions and applications, reflecting the interdisciplinary nature of computer vision. From efficient single-image super-resolution techniques to lightweight network architectures for traffic sign recognition, and from medical image processing to action recognition in autonomous systems, the contributions highlight the versatility, and potential of computer vision technologies. Below, we provide a brief overview of the accepted articles, emphasizing their key contributions and practical implications.

Efficient and lightweight deep learning models for real-world applications

One of the central themes of this Research Topic is the development of efficient and lightweight deep learning models that can operate effectively in resource-constrained environments. An et al. presented a lightweight network architecture based on an enhanced LeNet-5 model for traffic sign recognition. By optimizing the network structure

and reducing the number of parameters, they achieved state-of-the-art performance on standard benchmarks, making their solution suitable for deployment in real-world autonomous driving systems.

Qu and Ke proposed an asymmetric large kernel distillation network for single image super-resolution, which leveraged asymmetric kernels to achieve high computational efficiency while maintaining superior performance in image restoration. Their approach demonstrated the importance of balancing model complexity and practical applicability, particularly in scenarios where computational resources are limited.

Xie et al. proposed an extremely lightweight pathological myopia instance segmentation method (SMLS-YOLO) that combined attention mechanisms with efficient network design to achieve real-time performance. Their approach was particularly valuable for applications in ophthalmology, where rapid and accurate segmentation is critical for diagnosing and monitoring conditions such as pathological myopia. The integration of attention mechanisms into lightweight models highlighted the importance of optimizing both computational efficiency and accuracy to ensure that these technologies can be deployed in real-world settings.

Deep learning in medical image processing

Medical image processing is another area where deep learning has shown tremendous potential. Li L. et al. explored the use of Swin Transformer-based automatic delineation of the hippocampus in MRI scans for hippocampus-sparing whole-brain radiotherapy. Their work showcased the effectiveness of transformer-based architectures in medical image segmentation, providing a more accurate and automated approach to treatment planning.

Tian and Zhang presented a GAN-guided nuance perceptual attention network (G2NPAN) for multimodal medical fusion image quality assessment. Their work combined generative adversarial networks (GANs) with attention mechanisms to evaluate the quality of fused medical images, ensuring that the outputs were both visually appealing and diagnostically useful. This approach highlighted the importance of integrating advanced deep learning techniques with practical applications in healthcare, where image quality and interpretability are critical.

Zhou et al. focused on lymph node segmentation in lung cancer diagnosis, introducing a dual-stream feature-fusion attention U-Net (DFA-UNet). By incorporating attention mechanisms into the U-Net architecture, they achieved improved segmentation accuracy and computational efficiency, which are essential for clinical applications where time and resource constraints are significant.

Gu et al. proposed a motion-sensitive network for action recognition in autonomous systems, leveraging insights from motion perception to improve decision-making in real-time control scenarios. Their work underscored the importance of designing models that can efficiently process dynamic visual inputs, which have direct applications in robotics and autonomous vehicles.

Practical applications and beyond

The practical applications of computer vision are vast and varied, ranging from culture to education to security systems. Ramesh et al. presented a hybrid manifold smoothing and label propagation technique for handwritten Kannada character recognition, demonstrating how advanced deep learning methods can be adapted for tasks involving handwritten text. Their work had implications for document analysis, OCR systems, and cultural heritage preservation, where handwritten text recognition remains a challenging yet important task.

Li X. et al. proposed a parameter-dense three-dimensional convolution residual network for classroom teaching applications. By incorporating dense 3D convolutions, they demonstrated improved performance in handling complex, multidimensional data.

The work by Wang on attention-enhanced computation in multimedia affective computing explored how attention mechanisms can be used to evaluate and analyze visual perception, particularly in the context of affective computing. By integrating attention-based models, Wang demonstrated how visual perception can be quantified and optimized for applications such as emotion recognition and human-computer interaction.

Ma et al. proposed a study on face anti-spoofing based on pseudo-negative feature generation, addressing the critical issue of ensuring security and reliability in facial recognition systems. By generating pseudo-negative features to enhance robustness against spoofing attacks, their work contributed to the development of more secure and trustworthy biometric systems. This is particularly relevant in practical applications where facial recognition is increasingly used for authentication and access control.

This Research Topic, "*Advances in computer vision: from deep learning models to practical applications*," reflects the current state of the field and its trajectory toward solving real-world problems. Collectively, the articles demonstrate how deep learning models can be optimized for efficiency, scalability, and practicality while maintaining high performance in diverse applications. From lightweight architectures for traffic sign recognition and super-resolution imaging to advanced attention mechanisms for medical image segmentation and action recognition, these contributions highlight the interdisciplinary nature of computer vision and its potential to revolutionize diverse domains.

This Research Topic of articles inspires further research and collaboration between researchers, practitioners, and industry. The integration of deep learning with practical applications not only enhances the functionality of computer vision systems but also ensures their relevance and usability in addressing real-world challenges. As the field continues to grow, we anticipate even more exciting developments that will bridge the gap between theoretical

models and practical deployment, paving the way for smarter, more efficient, and more reliable vision technologies.

Finally, we would like to extend our gratitude to all the authors for their insightful contributions and to the reviewers for their valuable feedback. This Research Topic will serve as a valuable resource for researchers and practitioners alike, fostering innovation and advancing the field of computer vision toward practical and impactful applications.

Author contributions

HZ: Writing – original draft. RY: Writing – review & editing. LT: Writing – review & editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 25 March 2024
DOI 10.3389/fnins.2024.1370024

OPEN ACCESS

EDITED BY
Hancheng Zhu,
China University of Mining and Technology, China

REVIEWED BY
Lei Zhang,
Nanjing Normal University, China
Rashid Khan,
Shenzhen Technology University, China

*CORRESPONDENCE
Jialiang Gu
✉ gujliang@mail2.sysu.edu.cn

RECEIVED 13 January 2024
ACCEPTED 04 March 2024
PUBLISHED 25 March 2024

CITATION
Gu J, Yi Y and Li Q (2024) Motion sensitive network for action recognition in control and decision-making of autonomous systems. *Front. Neurosci.* 18:1370024. doi: 10.3389/fnins.2024.1370024

Motion sensitive network for action recognition in control and decision-making of autonomous systems

Jialiang Gu*, Yang Yi and Qiang Li

Computer Science and Engineering, Sun Yat-sen University, Guangdong, China

Spatial-temporal modeling is crucial for action recognition in videos within the field of artificial intelligence. However, robustly extracting motion information remains a primary challenge due to temporal deformations of appearances and variations in motion frequencies between different actions. In order to address these issues, we propose an innovative and effective method called the Motion Sensitive Network (MSN), incorporating the theories of artificial neural networks and key concepts of autonomous system control and decision-making. Specifically, we employ an approach known as Spatial-Temporal Pyramid Motion Extraction (STP-ME) module, adjusting convolution kernel sizes and time intervals synchronously to gather motion information at different temporal scales, aligning with the learning and prediction characteristics of artificial neural networks. Additionally, we introduce a new module called Variable Scale Motion Excitation (DS-ME), utilizing a differential model to capture motion information in resonance with the flexibility of autonomous system control. Particularly, we employ a multi-scale deformable convolutional network to alter the motion scale of the target object before computing temporal differences across consecutive frames, providing theoretical support for the flexibility of autonomous systems. Temporal modeling is a crucial step in understanding environmental changes and actions within autonomous systems, and MSN, by integrating the advantages of Artificial Neural Networks (ANN) in this task, provides an effective framework for the future utilization of artificial neural networks in autonomous systems. We evaluate our proposed method on three challenging action recognition datasets (Kinetics-400, Something-Something V1, and Something-Something V2). The results indicate an improvement in accuracy ranging from 1.1% to 2.2% on the test set. When compared with state-of-the-art (SOTA) methods, the proposed approach achieves a maximum performance of 89.90%. In ablation experiments, the performance gain of this module also shows an increase ranging from 2% to 5.3%. The introduced Motion Sensitive Network (MSN) demonstrates significant potential in various challenging scenarios, providing an initial exploration into integrating artificial neural networks into the domain of autonomous systems.

KEYWORDS

deep learning, action recognition, computer vision, visual perception, motion information, spatial-temporal feature, practical application

1 Introduction

With the rapid development of computer vision technology, action recognition in videos (Sun et al., 2022) has become a crucial challenge, finding applications in areas such as autonomous driving and virtual reality. In this context, video action recognition is not just an academic research field but a key component for addressing real-world problems and enhancing the intelligence of AI systems. Recently, action recognition methods based on convolutional neural networks (CNNs) have gained significant attention. Among them, 3D convolutional networks are renowned for directly extracting spatiotemporal features from videos, but they suffer from high computational costs, limiting their efficiency for human action recognition. On the other hand, 2D convolutional networks (Yao et al., 2022), especially two-stream networks, extract motion information by capturing multimodal cues. However, fusing multimodal information still poses challenges, and the pre-computation of optical flow is computationally expensive (Alayrac et al., 2022; Islam et al., 2023). In recent years, successful approaches have emerged by extracting motion features from RGB using embeddable modules within 2D convolutional networks, achieving satisfactory performance at a lower cost (Wu et al., 2020). Although these modules capture some motion features, they may overlook spatial scale variations over time and inconsistent action frequencies across different actions. This motivates us to propose a novel approach aimed at handling spatiotemporal features in video action recognition more comprehensively and efficiently.

This paper explores the utilization of artificial neural networks (ANNs) in the context of spatial and temporal modeling, contributing to the theoretical foundations and practical applications of ANNs in autonomous system control and decision-making. However, applications in the field of ubiquitous Human Activity Recognition (HAR) have been relatively limited. To address the issue of information loss during channel compression, researchers have proposed a multi-frequency channel attention framework based on Discrete Cosine Transform (DCT) to better compress channels and utilize other frequency components (Xu et al., 2023). On the other hand, Federated Learning (FL) shows potential in HAR tasks, but the non-IID nature of sensor data poses challenges for traditional FL methods. To tackle this, researchers have introduced the ProtoHAR framework, which leverages global prototypes and personalized training to address representation and classifier issues in heterogeneous FL environments (Cheng et al., 2023). Additionally, wearable sensor-based HAR has gained significant attention, where the phenomenon of channel folding in existing methods impairs model generalization. Researchers have proposed a channel equalization method to balance feature representation by reactivating suppressed channels (Huang et al., 2022). These studies provide important references and guidance for the development and practical applications in the HAR field. In the realm of video-based action recognition (Zheng et al., 2022a), complexities arise from the need to handle intricate data distributions and extract both spatial and temporal information concurrently. Distinguishing diverse action classes, addressing scale changes, and accommodating inconsistent action frequency require sophisticated spatial and temporal modeling (Cob-Parro et al., 2024). For instance, discerning actions like "Running" from "Walking" involves not only recognizing visual tempo differences but also understanding spatial scale variations. Similarly, "Brush Teeth" and "Apply Eye Makeup" have great differences in spatial scale despite sharing high similarities in the temporal dimension (Kulsoom et al., 2022). Learning the intention of human action from such data in videos poses a great challenge (Zheng et al., 2024). In certain scenarios, fine-grained recognition of actions becomes crucial, requiring more detailed spatial and temporal modeling. Learning the intent behind human behavior from such data in videos poses a significant challenge. Additionally, there are substantial challenges in the fusion of multimodal information, especially when it involves additional modalities such as optical flow. Existing methods face difficulties in effectively integrating different modalities, and the computational cost of pre-computing modalities like optical flow remains a bottleneck. Similarly, modeling actions in long-term videos often encounters challenges related to memory. Models may struggle to capture the evolution of actions over long time spans and maintain consistent understanding throughout the entire sequence. Imbalance and scarcity of samples across different action categories in the dataset present another problem, as the models may exhibit bias when learning minority class actions, thereby affecting overall performance. In some application scenarios, real-time requirements for action recognition models are high. For example, in autonomous driving systems, achieving high accuracy while ensuring fast inference speed to adapt to real-time environments is crucial (Lin and Xu, 2023). Meanwhile, action recognition models are susceptible to adversarial attacks, where subtle perturbations to the model inputs can lead to misclassification. Improving model adversarial robustness and resilience remains a challenge (Chen et al., 2019).

In this paper, we propose a new approach called Motion Sensitive Network (MSN) that addresses the challenge of efficiently recognizing complex actions with varying spatial scales and visual tempos. To achieve this, we introduce two new modules: the Temporal Spatial Pyramid Motion Extraction (STP-ME) module and the Deformable Scales Motion Excitation (DS-ME) module. The STP-ME module extracts implicit motion information by taking consecutive frames as input and using feature difference to focus on the position and tempo of the action occurring between frames. This information is incorporated into the single RGB frame (Liu et al., 2021), allowing for better alignment of the temporal and spatial dimensions at different scales. The DS-ME module addresses irregular deformation of the action subject in space and long-range feature alignment issues. It uses multiscale deformable convolutions to model the complete action region (He and Tang, 2023), allowing for more accurate representation of different motion splits. Additionally, to address numerical problems with negative values, we use the absolute value of the feature. Overall, our framework can be broken down into three steps: extracting effective motion information in the early stage, giving higher weight to motion features in the later stage, and doing numerical processing to avoid harmful results during processing (Luo, 2023). Our proposed MSN method effectively handles the challenges of action recognition, improving on existing 2D and 3D CNN-based methods. By leveraging ANNs in spatial and temporal modeling, this work contributes to enhancing the theoretical foundations and

practical applications of ANNs in autonomous system control and decision-making.

The contributions of this paper can be summarized in the following three aspects:

(1) The paper introduces a novel approach known as the Motion Sensitive Network (MSN) for action recognition. This method is characterized by its simplicity and effectiveness in accurately estimating scale variations, thereby enhancing overall network performance in action recognition tasks.
(2) The paper proposes a unique Time-Space Pyramid Motion Extraction (STP-ME) module. This module leverages a pyramid structure to extract multi-scale temporal features, thereby fortifying the model's robustness across diverse action scenarios. The STP-ME module is designed to address challenges associated with scale variations and capture motion information across different time scales.
(3) The paper introduces the Variable Scale Motion Excitation (DS-ME) module as an innovative solution to challenges posed by unique and irregular motion patterns in dynamic scenes. This module utilizes deformable scale convolutions to adaptively modify the motion scale of target objects before computing temporal differences on consecutive frames. This approach aims to enhance the model's ability to handle objects with varying scales during motion.

The organizational structure of this paper is as follows: The introduction (Section 1) sets the stage by presenting the background, significance, and motivation for the research, highlighting challenges in existing action recognition methods, and outlining the contributions of the proposed Motion Sensitive Network (MSN). Section 2, "Relevant Work," conducts a comprehensive review of existing literature, emphasizing prior research on motion sensitivity in action recognition and identifying gaps in current approaches. The third section, "Method," provides a detailed exposition of the MSN architecture, elucidating its design principles and showcasing its motion-sensitive modules. Moving on to Section 4, "Experiment," the paper delves into the experimental setup, detailing the datasets used, metrics employed for performance assessment, and the methodology for training MSN, while Section 5, "Discussion," critically analyzes experimental results. This section interprets findings, assesses MSN's effectiveness in addressing motion sensitivity, and discusses potential applications and limitations. Finally, in Section 6, "Conclusion," the paper synthesizes key discoveries, underscores the contributions made by MSN, discusses broader implications for the field of action recognition, and proposes avenues for future research. This organized structure guides readers through a coherent narrative (Han et al., 2022), facilitating a comprehensive understanding of the research from problem introduction to proposed solution, experimental validation, discussion, and ultimate conclusion.

2 Related work

The realm of action recognition within computer vision has undergone significant exploration (Zhang et al., 2022; Dai et al., 2023; Wu et al., 2023), with convolutional neural networks (CNNs) at the forefront of innovation (Xu et al., 2022). Two major categories, two-stream CNNs and 3D CNNs, have shaped the landscape. Next, we will delve into the theoretical foundations and practical applications of artificial neural networks in the field of autonomous system control and decision-making.

Simonyan and Zisserman (2014) proposed a multi-stream network for action recognition, consisting of two separate branches: a temporal convolutional network and a spatial convolutional network. Both branches have the same architecture, with the temporal stream learning motion features from stacked optical flows and the spatial stream extracting spatial features from still images (Wang et al., 2020). The two streams are then fused to obtain the final classification result. However, this approach has some drawbacks. Firstly, the computational cost is relatively high, particularly due to the complexity of optical flow computation. The stacking of optical flows may result in expensive computational overhead, especially when dealing with long video sequences or high frame-rate videos. Secondly, the method's reliance on optical flow makes it sensitive to video noise and motion blur, impacting the reliability of accurately extracting motion features. Additionally, the dependence on optical flow introduces sensitivity to video noise and motion blur, affecting the reliability of accurately extracting motion features. Moreover, the challenge of modal fusion is also a concern, as effective fusion requires careful design to ensure that features extracted from both streams collaborate without interference. Lastly, the method may have limitations in modeling spatiotemporal relationships, especially in complex motion scenarios, such as non-rigid motion or rapidly changing movements. This may result in constraints on the comprehensive capture of complex spatiotemporal dynamic relationships. Wang et al. (2016) proposed a Temporal Segment Network (TSN) based on the two-stream CNN, which utilizes a sparse time sampling strategy to randomly extract video fragments after time-domain segmentation. TSN addresses the insufficient modeling ability of long-range temporal structure in two-stream CNNs. However, this approach may have some potential limitations. Sparse temporal sampling strategies may result in the loss of crucial temporal information during the model training process, especially for modeling long-duration actions, which may not be adequately captured. Furthermore, this randomness in sampling may hinder the model's ability to effectively capture critical temporal patterns for specific types of actions, thereby impacting its performance. Building on TSN, Zhou et al. (2018) attempted to extract connections between video frames of different scales by convolving video frames of different lengths, performing multi-scale feature fusion, and obtaining behavior recognition results. However, applying convolution to video frames of different lengths may increase the computational complexity of the model in handling information at different scales, thereby impacting the training and inference efficiency of the model. He et al. (2019) proposed a local and global module to hierarchically model temporal information based on action category granularity, while Li et al. (2020) proposed motion excitation and multiple temporal aggregation modules to encode short- and long-range motion effectively and efficiently, integrated into standard ResNet blocks for temporal modeling. Wang et al. (2021) focused on capturing multi-scale temporal information

for efficient action recognition, presenting a video-level motion modeling framework with a proposed temporal difference module for capturing short- and long-term temporal structure. However, these methods may share some potential common drawbacks. Firstly, approaches such as local and global modules based on action category granularity, hierarchical networks from coarse to fine, motion excitation, multiple temporal aggregation modules, video-level motion modeling frameworks, and temporal offset modules may require more complex network structures and additional parameters to achieve layered modeling of temporal information. This may lead to increased computational complexity, heightened training difficulty, and an increased demand for hardware resources. Secondly, these methods might necessitate carefully designed hyperparameters and model structures to adapt to different time scales and action categories. In practical applications, this could require extensive parameter tuning and model optimization, raising the method's usage threshold and operational difficulty. Additionally, these methods may encounter memory issues when dealing with long temporal video sequences in temporal modeling. The model might struggle to effectively capture the evolution of actions over extended time ranges and maintain consistent understanding throughout the entire sequence. When handling long temporal videos, these methods might need additional mechanisms to ensure the model's effectiveness and stability.

Another type of method attempts to learn spatio-temporal features directly from RGB frames using 3D CNNs. The 3D convolutional network for action recognition was introduced by Yang et al. (2019), which uses a 3D convolution kernel to perform 3D convolution on the input and directly extracts spatio-temporal features along the spatial and temporal dimensions of the video. Tran et al. (2018) constructed a C3D network framework using 3D convolution and 3D pooling operations. Carreira and Zisserman (2017) combined a two-stream network and a 3D CNN to propose an I3D network framework based on the inception-V1 model, using RGB and optical flow as inputs. Diba et al. (2018) and others improved the I3D by using different scales of convolution to build the TTL layer and using 3D-DenseNet as the basic network to build the T3D network framework. Qiu et al. (2019) and others proposed a P3D network, which uses 133 convolution and 311 convolutions instead of 333 convolutions to greatly reduce the amount of computation. Nevertheless, directly processing videos using 3D convolutional networks may result in a larger number of parameters along the temporal dimension, increasing the risk of overfitting. Tran D proposed a similar structure called R(2+1)D. Our proposed method is inspired by TDN and TEA with short- and long-range temporal modeling, taking several continuous frames as input. Our work differs from previous works in that we employ a strategy for long- and short-range temporal modeling to better extract motion information. Although our approach shares similarities with these works, we focus on addressing the problem of spatio-temporal inconsistency (Zheng et al., 2022b). In addition, in the field of autonomous driving, integrating MSN into autonomous systems offers potential advantages for enhancing the environment perception and decision support of the vehicle system. By performing real-time analysis of video and sensor data, MSN can perceive the surrounding environment, accurately recognize the movements of other vehicles, pedestrians, and obstacles, thereby providing autonomous vehicles with richer environmental information. This enables vehicles to more accurately predict the behavior of other traffic participants, thereby improving overall driving safety. However, this application also faces some challenges, especially in terms of real-time requirements, particularly in autonomous driving scenarios that require immediate decision-making. Accurate and efficient action recognition is crucial for rapidly changing traffic environments, making it imperative to address the reduction of algorithm inference time. By incorporating short- and long-range temporal modeling (Wu et al., 2021), our approach aims to enhance the efficiency of action recognition methodologies, showcasing the potential of artificial neural networks in the complex landscape of autonomous system control and decision-making.

3 Method

The overall flowchart of the algorithm in this article is shown in Figure 1:

3.1 MSN sports sensitive network

The MSN is a video-level framework that learns action models using entire video information. To improve efficiency, we follow the TSNTSN framework with a sparse and holistic sampling strategy for each video. Our main contribution is to fully consider the scale changes in the space-time dimension when obtaining implicit action information through feature difference and inject this action information into the network in two ways: element-wise addition of the implicit action information extracted by the STP-ME module to the keyframe-wise information extracted by the backbone, and embedding the DS-ME module into the CNN block to increase the processing weight of motion features adaptively. Its structural diagram is shown in Figure 2.

In first stage, each video V is divided into T segments of equal duration without overlapping. We randomly sample 5 frames $I^i = I^i_{k-2}, I^i_{k-1}, I^i_k, I^i_{k+1}, I^i_{k+2}$ from each segment. We select the third frame in as the keyframe and totally obtain T key frames $I_k = I^1_k, ..., I^T_k$. These keyframes are separate fed into a 2D CNN to extract keyframe-wise features $F = [F_1, F_2, ..., F_T]$. Besides we applied STP-ME module to extract motion information from the whole 5 frames and supplied it to the original keyframe process pipeline, so as to increase the amount of effective information input and improve the feature's representation power. Specifically, we fuse the keyframe-wise feature and implicit motion 1 information using the following Equation 1:

$$F' = F_i + S(I_i) \tag{1}$$

Where F' denotes the fused feature for segment i, F_i is the keyframe-wise feature, S denotes our STP-ME module, and it extracts implicit motion information from adjacent frames I_i.

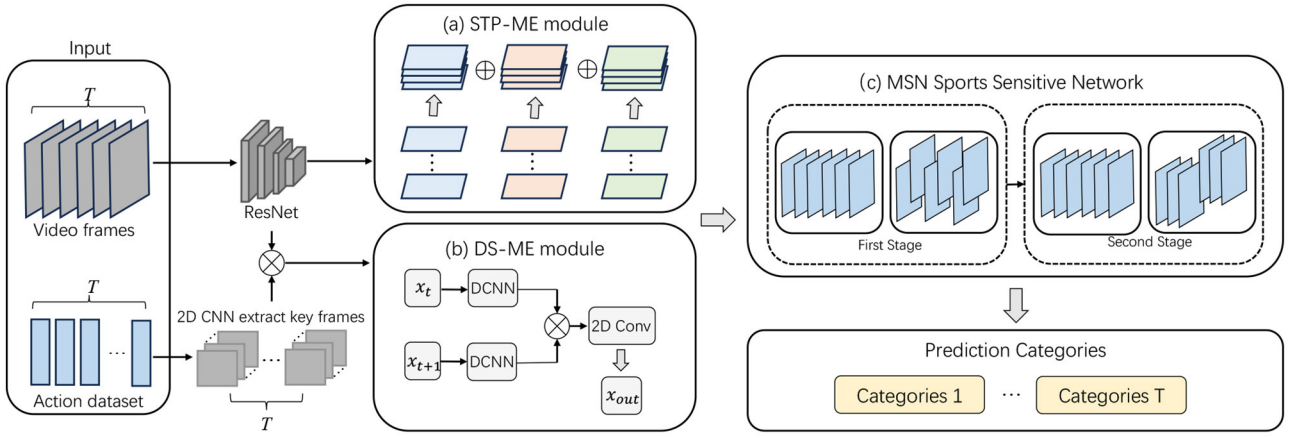


FIGURE 1
Overall algorithm flowchart. The input section comprises video frames and an action label dataset, where video frames capture the spatiotemporal information of actions, and the action label dataset is used to supervise the model learning process. Subsequently, video frames undergo processing through ResNet and 2D CNN to extract advanced features and key frames, providing a robust foundation for subsequent modules. The STP-ME module addresses the consistency of motion information across different temporal scales using a pyramid structure. This structure enables the module to adaptively handle motion information at different temporal scales, enhancing the model's robustness to scale variations. By applying the pyramid structure to consecutive frames, the STP-ME module focuses on motion information at different temporal scales, including modeling the position and motion rhythm between frames at each pyramid level. The DS-ME module introduces deformable scale convolutions, adaptingively modifying the motion scale of target objects to address unique and irregular motion patterns in dynamic scenes. Adopting multi-scale deformable convolutions covers a broader range of motion, allowing the DS-ME module to intricately model the entire action region and improve modeling accuracy for objects at different scales, particularly when dealing with spatiotemporal inconsistency challenges. Within the overall network architecture, the MSN Sports Sensitive Network's two stages process different levels of feature representations to comprehensively express action information in the video. Finally, the model aggregates multi-category predictions at each time step through Temporal Aggregation, producing the ultimate action classification results.

In the second stage, we embed the DS-ME module into the CNN block and calculate the channel weight by multiscale cross-segmentation difference. In this way, we could distinguish some feature channels that contain different scales of motion information and enhance these channels to make our net-work pay more attention to the motion. We establish the channel enhance process as follows (Equation 2):

$$F^{'} = F + D(F) \odot F \tag{2}$$

Where D represents our DS-ME module, F is the origin features and F' is the enhanced features. In the current implementation, we only consider adjacent segment-level information for channel weight calculation in each DS-ME module, Details will be described in the following subsections.

3.2 STP-ME module

In a video, the action is reflected in the change of pixel value between adjacent frames. We argue that modest variances across adjacent frames respond well to the nature of the action. Many previous works sample a single frame from a segment which extracts appearance information instead of the motion information contained in each segment. To tackle this problem, we propose the STP-ME module shown in Figure 3.

In STP-ME module, we selected 5 frames in a segment and extracted implicit motion information by feature difference. Furthermore, the time interval often shows a positive correlation with the variance of spatial scale. In specific, as the time interval increases, the spatial scale also increases. Therefore, we aligned the temporal dimension with the spatial dimension from the perspective of scales and extracts implicit motion information from adjacent frames by three steps. Then, make each step corresponds to a different temporal spatial scale.

(1) In the first step, we set the time interval is 1 frame. For each sampled frame I_i, we extract several feature differences and then stack them along channel dimension (Equations 3–6):

$$F_{12} = conv1(I_2) - conv1(I_1) \tag{3}$$

$$F_{23} = conv1(I_3) - conv1(I_2) \tag{4}$$

$$F_{34} = conv1(I_4) - conv1(I_3) \tag{5}$$

$$F_{45} = conv1(I_5) - conv1(I_4) \tag{6}$$

Where F_{ij} is feature difference between I_i and I_j, $conv1$ is a convolution layer.

(2) At the second step, we set the time interval is 2 frames. We select 3 feature map contains I1, I3, I5 to extract the mid step feature and stack it (Equations 7, 8).

$$F_{13} = conv2(I_3) - conv2(conv1(I_1)) \tag{7}$$

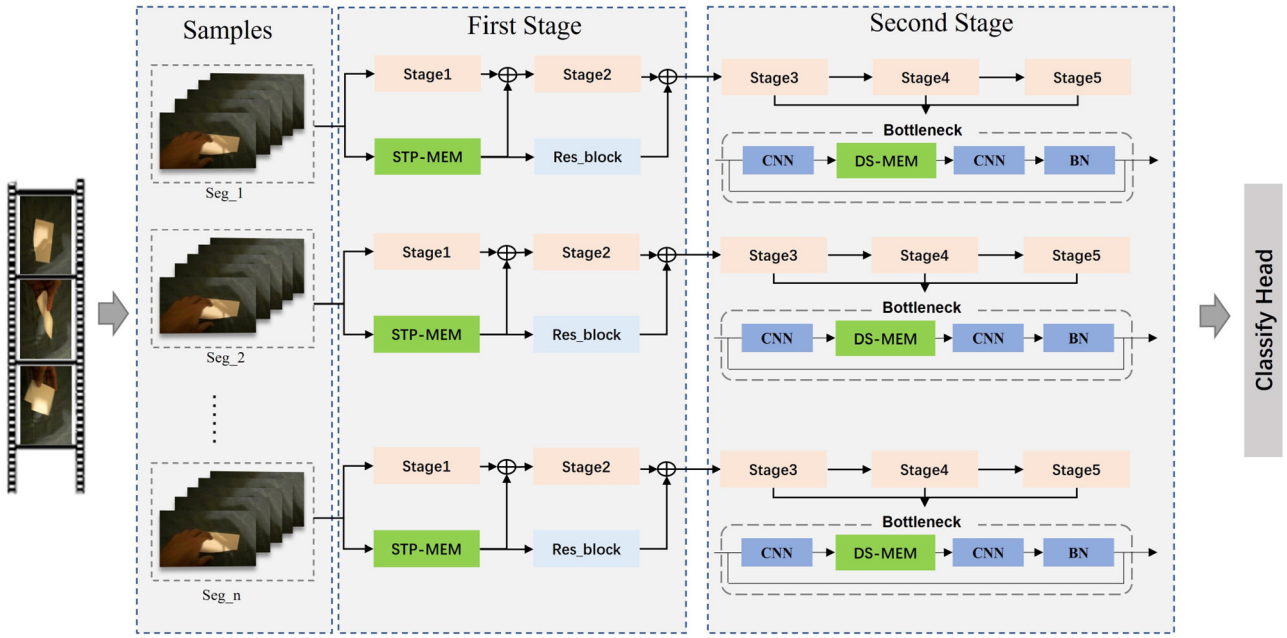


FIGURE 2
Structure diagram of MSN Sports Sensitive Network. First, input samples containing video frames and corresponding action labels are provided to supervise the model learning process. Proceeding to the First Stage, it is subdivided into Stage1, Stage2, STP-MEM, and Res_block. In Stage1 and Stage2 of the First Stage, advanced feature representations are extracted from the input samples. The STP-MEM module in the First Stage enhances the representation of motion information, adaptively handling motion information at different temporal scales through a pyramid structure, thereby improving the model's robustness to scale variations. Meanwhile, the Res_block strengthens feature propagation through residual connections, helping alleviate the vanishing gradient problem and making the model easier to train. Moving on to the Second Stage, which includes Stage3, Stage4, Stage5, Bottleneck, CNN, DS-MEM, and BN. In each stage of the Second Stage, the model further processes features obtained from the First Stage, gradually forming more abstract and high-level representations. The Bottleneck structure is employed for dimensionality reduction and increased network depth to extract more expressive features. CNN and DS-MEM modules in this stage introduce deformable scale convolutions to better model irregular and unique motion patterns, enhancing modeling accuracy for objects at different scales. BN normalizes features, accelerating convergence, and improving the training stability of the model. Finally, classification is performed through the Classify Head, providing predictions for action categories. The entire flowchart integrates these key components organically, forming a motion-sensitive action recognition network with powerful modeling capabilities for complex motion scenes.

$$F_{35} = conv2(I_5) - conv2(conv1(I_3)) \tag{8}$$

(3) At the third step, we set the time interval is 4 frames. We select 2 feature maps in last step [f1; f5] to extract the final step feature (Equation 9).

$$F_{15} = cons3(conv2(conv1(I_5))) - conv3\left(conv2\left(conv1\left(I_1\right)\right)\right) \tag{9}$$

(4) Finally, we realize the consistency of each dimension by up-sampling f_u the above features, and fuse them by elementwise addition (Equation 10).

$$F = concat(F_{12}, F_{12}, F_{12}, F_{12}) + f_u\left(concat\left(F_{12}, F_{12}\right)\right) + f_u\left(F_{15}\right) \tag{10}$$

The implicit motion information F is fused with the keyframe features, so that the original frame-level representation is aware of motion pattern and able to better describe a segment.

3.3 DS-ME module

The STP-ME module provides a powerful representation for capturing spatial-temporal features, including local motion information within a segment. However, it is essential to leverage this motion information in the second stage to enhance action recognition. While the channel attention strategy has been shown to improve the importance of certain types of information, for action recognition, we need to consider more details. We observe that a complete action is comprised of different scales of motion split and irregular deformations of the action subject in space. To address these issues, we propose the DS-ME module, which employs a multiscale convolution kernel to capture different scales of motion splits and achieve more accurate channel attention calculation. In addition, to smooth the irregular deformation of the action subject, we devise a deformable CNN architecture, as illustrated in Figure 4.

The proposed DS-ME module operates as follows. Firstly, we compress the feature dimension by a ratio of r and split the feature segmentation in the temporal dimension as follows (Equation 11):

$$[X_1, X_2, \ldots, X_t] = f_{split}(Conv(F_{in})) \tag{11}$$

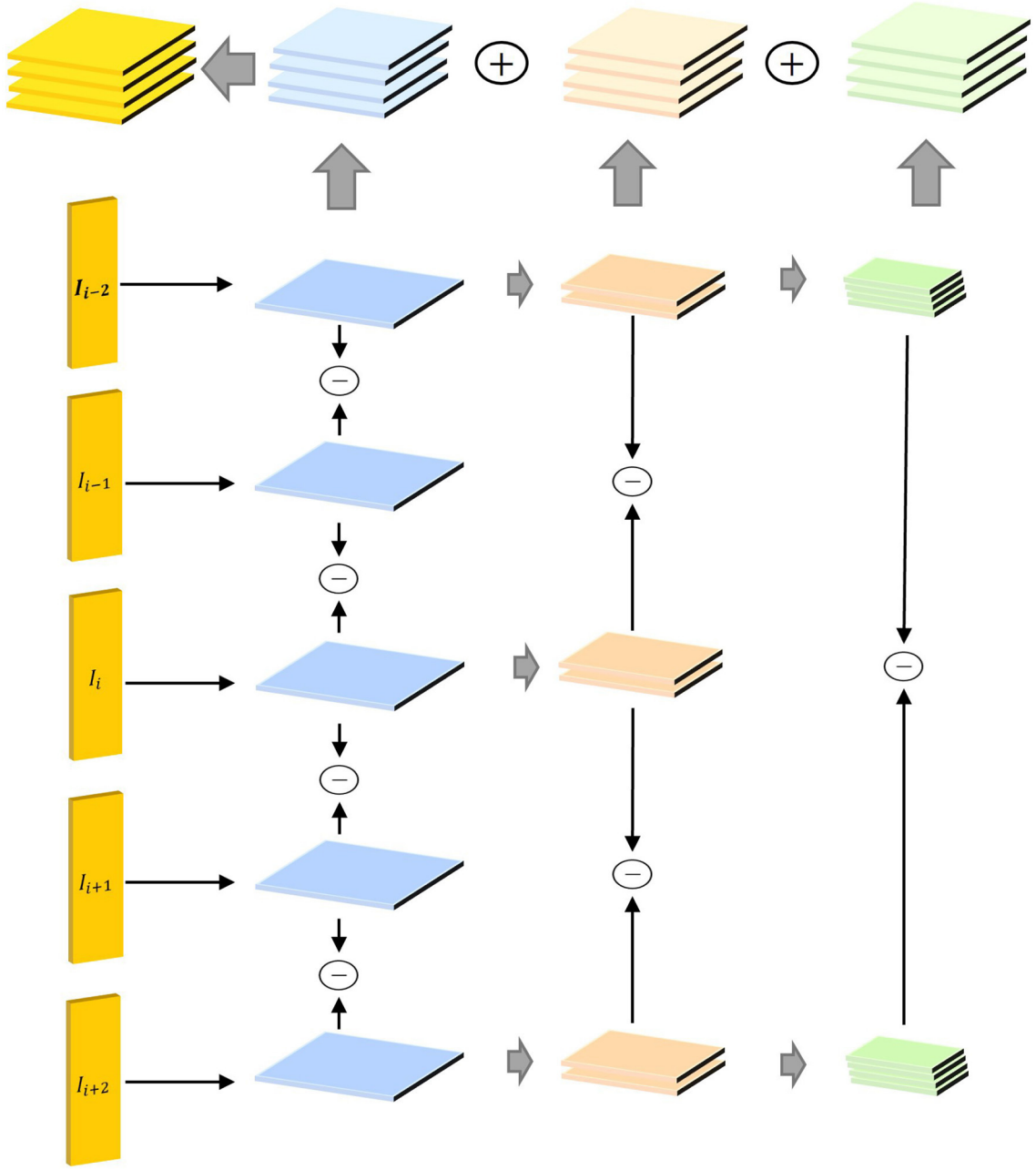


FIGURE 3
Structure diagram of STP-ME module. The components I_{i-2}, I_{i-1}, I_i, I_{i+1}, and I_{i+2} represent frames at the current time step and two preceding and two succeeding time steps, respectively. These frames are introduced as inputs to the STP-ME module, capturing action information in the video at different time steps. Through these inputs, the STP-ME module aims to address the consistency of motion information across various time scales. The primary task of the STP-ME module is to adaptively process motion information at different time scales through a pyramid structure. By applying the pyramid structure between consecutive frames, the module can focus on motion information at different time scales, including modeling the position and motion rhythm between frames. This design enables the STP-ME module to better capture motion information at different time scales, enhancing the overall model's robustness to scale variations.

Where $[X_1, X_2, ..., X_t]$ is a set of split features in the temporal dimension with a size of T, *Conv* is the channel-wise convolution, and F_{in} is the input feature.

Next, these split features undergo three different scale Deformable CNN (DCNN) operations, namely: (1) a 1 × 1 deformable CNN, (2) a 3 × 3 deformable CNN, and (3) a 5 × 5 deformable CNN. This operation is computed as follows (Equations 12–14):

$$X_t^1 = DCNN_1(X_t) \tag{12}$$

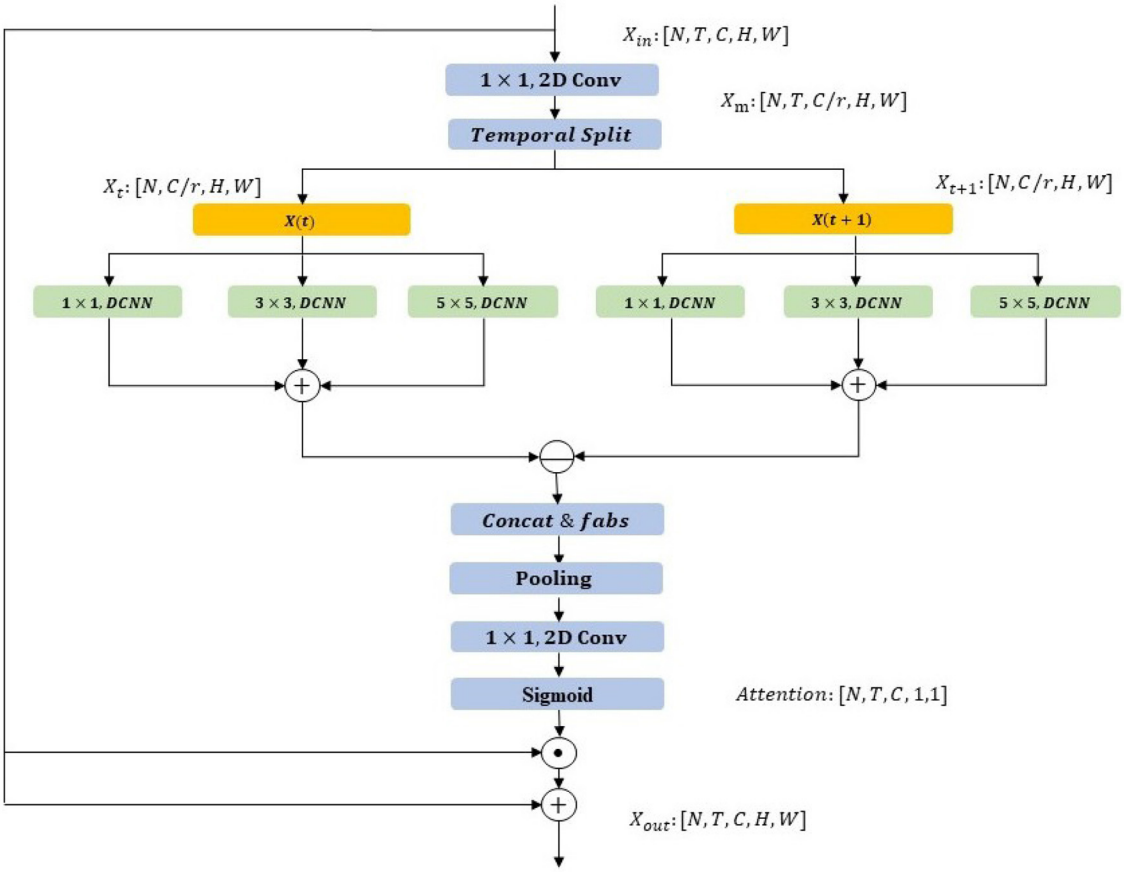


FIGURE 4
Structure diagram of DS-ME module. The module takes an input tensor [N, T, C, H, W], where N is batch size, T is the number of time steps, and C is the channel dimension. Initial processing involves 1 × 1 and 2D convolutions to adjust channels and extract spatial features, forming a robust foundation for subsequent spatiotemporal modeling. The Temporal Split operation separates the temporal dimension into X(t) and X(t+1), introducing temporal dynamics. These undergo independent 1x1 convolution, deformable convolution (DCNN) with 3x3 kernel, DCNN, and 5 × 5 convolution to intricately model spatiotemporal information, particularly addressing irregular motion patterns. Further operations, including Concat&fabs, pooling, 1 × 1 convolution, 2D convolution, and Sigmoid, fuse and process features, achieving nuanced spatiotemporal modeling. This enhances adaptability to diverse scales and irregular motions, ultimately improving action recognition. The output is a tensor [N, T, C, H, W], representing motion excitation distribution per time step.

$$X_t^2 = DCNN_2\left(X_t\right) \tag{13}$$

$$X_t^3 = DCNN_3\left(X_t\right) \tag{14}$$

Where X_t^1, X_t^2, X_t^3 are the deformable features from X_t. After that, we could fused X_t^1, X_t^2, X_t^3 and calculate feature difference between consecutive segments as follows (Equation 15):

$$X_{diff} = (X_{t+1}^1 + X_{t+1}^2 + X_{t+1}^3) - (X_t^1 + X_t^2 + X_t^3) \tag{15}$$

where X_{diff} is the segment-wise feature difference. To avoid the loss of information caused by negative numbers after subtraction, we add an additional absolute value operation f_{abs} and then perform the maximum value pooling operation $f_{pooling}$ as follows (Equation 16):

$$W_{\mathrm{raw}} = f_{\mathrm{pooling}}\left(\left|X_{\mathrm{di}}\frac{f}{f'}\right|\right) \tag{16}$$

where W_{raw} is the raw weight. To obtain the channel attention weight, we upgrade the channel dimension with a 1x1 convolution $conv$ and activate it using the sigmoid function W_{raw} as follows (Equation 17):

$$W = F_{sig}(conv\left(W_{raw}\right)) \tag{17}$$

Finally, we enhance the video-level representation through a channel attention operation and combine it with the original feature map via a residual connection.

The MSN framework is based on sparse sampling of TSN and operates on a sequence of frames uniformly distributed over the entire video. The framework employs a two-stage motion modeling mechanism that focuses on capturing motion information at different space-time scales. The STP-ME module is inserted in the early stages for fine and low-level motion extraction, while the DS-ME module is used in the latter stages to further strengthen the role of action information in the network. We use a ResNet backbone for the MSN instantiation. Similar to V4D, we use the first two stages of ResNet (also known as the early stage) for implicit motion

information extraction within each segment using the STP-ME module. The latter three stages of ResNet (also known as the later stage) are embedded with the DS-ME module for channel attention by capturing different scales of motion splits across segments. To fuse motion information with spatial information in the early stage, we add residual connections between the STP-ME module and the main network for Stage 1 and Stage 2. To enhance the action feature, we embed the DS-ME module to the CNN block and add a channel attention mechanism in each residual block of Stages 3-5.

The pseudocode of the algorithm in this paper is shown in Algorithm 1:

Input: Training data $\mathcal{D}$ from Kinetics-400, Something-Something V1, Something-Something V2 datasets
Initialize: MSN model parameters Θ randomly
Set learning rate η, batch size B
for each training epoch **do**
  for each mini-batch $\mathcal{B}$ in $\mathcal{D}$ **do**
    Sample video clips $\mathcal{C}$ from $\mathcal{B}$
    Extract spatial features X_s and temporal features X_t from $\mathcal{C}$
    Compute motion stream features X_m using optical flow or other motion extraction methods
    Generate spatiotemporal proposals using STP-ME: $\mathcal{P}_{stp} = STP\_ME(X_s, X_t)$
    Generate discriminative spatiotemporal proposals using DS-ME: $\mathcal{P}_{ds} = DS\_ME(X_s, X_t)$
    Fuse spatiotemporal proposals using MSN: $X_{fuse} = MSN(\mathcal{P}_{stp}, \mathcal{P}_{ds}, X_m)$
    Perform action recognition using the fused features: $Y_{pred} = Action\_Recognition(X_{fuse})$
    Compute loss $\mathcal{L}$ using ground truth labels Y_{gt}
    Update MSN parameters using backpropagation: $\Theta = \Theta - \eta \frac{\partial \mathcal{L}}{\partial \Theta}$
  end for
  Evaluate model on validation set for metrics: Accuracy, Precision, Recall
end for

Algorithm 1. MSN training process.

4 Experiment

The experimental flow chart of this article is shown in Figure 5:

4.1 Lab environment

- Hardware environment:

  This experiment utilized a high-performance computing server that offers excellent computational and storage capabilities, providing robust support for research on motion-sensitive network action recognition. The server is equipped with an Intel Xeon E5-2690 v4 @ 2.60GHz CPU, a high-performance multi-core processor that delivers substantial computational power suitable for deep learning tasks. With 512GB of RAM, the server ensures abundant memory resources for model training and data processing, contributing to enhanced experimental efficiency. Additionally, the server is outfitted with 8 Nvidia Tesla P100 16GB GPUs, renowned for their outstanding performance in deep learning tasks, significantly accelerating both model training and inference processes.
- Software Environment:

  In this research, we have chosen Python as the primary programming language and PyTorch as the deep learning framework to explore effective methods for motion-sensitive network model. Leveraging the powerful capabilities of deep learning, our objective is to enhance both the performance and efficiency of the model. Taking full advantage of the convenience and flexibility of Python, we rapidly constructed the model. PyTorch, as our preferred deep learning framework, provides us with a rich set of tools and algorithm libraries, significantly streamlining the process of model development and training. With PyTorch's dynamic computation graph mechanism and built-in automatic differentiation functionality, we can more easily build, optimize, and fine-tune the model to achieve superior results in action recognition.

4.2 Experimental data

- Kinetics-400 Dataset

  The Something-Something V1 dataset is a video dataset focused on action recognition, renowned for capturing various common actions and object interactions in daily life. The dataset comprises thousands of video clips, with an average duration of around 3 seconds, covering a diverse range of action categories such as stirring, wiping, twisting, and rubbing. Through meticulous annotation, each video clip is explicitly labeled with the ongoing action and involved objects, providing reliable ground truth labels. To collect this diverse data, the dataset's creation leveraged online communities, inviting participants to upload short video clips of themselves performing various actions. This collection method makes the dataset more representative of real-world daily actions, increasing the diversity and complexity of the data. Given the inclusion of many subtle and complex actions, along with diverse interactions between objects and actions, the Something-Something V1 dataset poses a challenge in action recognition tasks. This dataset not only serves as a rich resource for researchers to understand human daily activities but also provides robust support for evaluating model performance in handling fine-grained and multi-category interaction tasks.
- Something-Something V1 Dataset

  The Something-Something V1 dataset stands out as a comprehensive video dataset meticulously crafted for advancing the field of action recognition research. Its

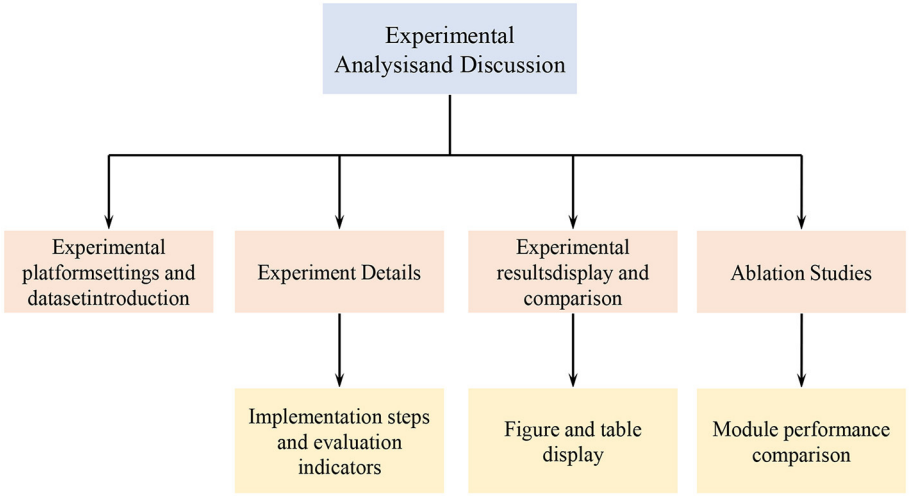


FIGURE 5
Experimental flow chart.

distinguishing feature lies in its ability to capture a diverse range of everyday actions and the interactions between individuals and objects, offering valuable insights into human daily activities. With a multitude of action categories, including stirring, wiping, twisting, and rubbing, the dataset encompasses thousands of short video segments, each lasting around 3 seconds. These segments vividly portray a rich variety of actions performed by individuals, contributing to the dataset's diversity. What sets Something-Something V1 apart is its detailed annotation process. Each video segment undergoes careful labeling, providing explicit information about the ongoing action and the objects involved. This meticulous annotation serves as robust ground truth data, essential for training and evaluating action recognition models. The dataset's creation involved a unique approach, leveraging online communities to encourage participants to contribute short video clips featuring diverse actions. This methodology ensures that the dataset captures a more realistic representation of daily activities, adding an extra layer of complexity and authenticity. One of the dataset's notable challenges lies in its inclusion of subtle and complex actions, coupled with diverse interactions between objects and actions. This complexity poses a significant challenge for models aiming to accurately recognize and categorize these nuanced action scenarios.

- Something-Something V2 Dataset

  The Something-Something V2 dataset builds upon the foundation laid by its predecessor, Something-Something V1, and stands as a significant contribution to the realm of action recognition research. Designed to deepen our understanding of human actions, this dataset introduces new challenges and complexities. Something-Something V2 features a diverse array of common actions performed in everyday scenarios, spanning activities such as stirring, wiping, pouring, and more. The dataset comprises a substantial number of video clips, each lasting approximately 3 seconds, offering a rich collection of short segments capturing various actions and interactions. Annotations play a crucial role in Something-Something V2, with meticulous labeling of each video segment specifying the action and involved objects. This detailed annotation serves as invaluable ground truth data for the training and evaluation of action recognition models. What sets Something-Something V2 apart is its introduction of additional challenges, making it more intricate than its predecessor. Notably, the dataset includes actions performed with hands only, pushing the boundaries of action recognition tasks and introducing a new layer of complexity. Intentionally incorporating challenging scenarios, such as ambiguous or subtle actions, Something-Something V2 serves as a benchmark dataset for evaluating the robustness and adaptability of action recognition models.

4.3 Experimental comparison and analysis

In this section, we present the experimental results of our MSN framework. Firstly, we describe the evaluation datasets and implementation details. Next, we compare our MSN with state-of-the-art methods. Then, we perform ablation studies to verify the effectiveness of the proposed modules. Finally, we show some visualization results to further analyze our MSN.

In our experiments, we use ResNet50 as the backbone to implement our MSN based on TSN framework, and sample T = 8 or T = 16 frames from each video. For training, each video frame is resized to have the shorter side in [256; 320], and a crop of 224 × 224 is randomly cropped. The total training epoch is set to 100 in the Kinetics dataset and 60 in the Something-Something dataset. We adopt a multi-step learning rate adjustment strategy, where it would be divided by a factor of 10 in each step. In different experiments, our batch size was set to a fixed value of 32. For testing, the shorter side of each video is resized to 256. We implement two

TABLE 1 Comparisons with state-of-the-art approaches on the Something-something v1&v2 test set.

SSV1(Zhou et al., 2018)				
Method	Backbone	frames	Top 1	Top 5
TSN-RGB	BNInception	8	19.50%	-
S3D	Inception	64	48.20%	78.70%
TSM	ResNet50	8+16	49.70%	78.50%
TEINET	ResNet50	8+16	52.50%	-
TANet	ResNet50	8+16	50.60%	-
TEA	ResNet50	16	51.90%	80.30%
TAM	bLResNet50	16	48.40%	-
I3D	ResNet50	32	41.60%	72.20%
TDN	ResNet50	8	52.30%	80.60%
TDN	ResNet50	16	53.90%	82.10%
MSN	ResNet50	8	53.00%	81.50%
MSN	ResNet50	16	54.10%	82.30%
SSV2 (Materzynska et al., 2020)				
Method	Backbone	frames	Top 1	Top 5
TAM	bLResNet50	16*2	61.70%	88.10%
TSM	ResNet50	16*6	63.40%	88.50%
TEINET	ResNet50	8+16	65.50%	89.80%
GST	ResNet50	16	62.60%	87.90%
STM	bLResNet50	16*30	64.20%	89.80%
SmallBigNet	ResNet50	8+16	63.30%	88.80%
TDN	ResNet50	8	64.00%	88.80%
TDN	ResNet50	16	65.30%	89.50%
MSN	ResNet50	8	63.90%	89.20%
MSN	ResNet50	16	65.50%	89.90%

kinds of testing schemes: the 1-clip and center-crop, where only a center crop of 224 × 224 from a single clip is used for evaluation, and the 10-clip and 3-crop, where three crops of 256 × 256 and 10 clips are used for testing. The first testing scheme is with high efficiency, while the second one is for improving accuracy with a denser prediction scheme.

We compare our model with state-of-the-art methods including I3D, TAM, GST, SmallBigNet, TEA, and TDN on two benchmarks: Something-Something and Kinetics-400. We report the details used by each method and use the 1 clip and center crop testing scheme for Something-Something and 10 clips and 3 crops for testing on the Kinetics-400 dataset.

Results on something-something. As expected, sampling more frames can further improve accuracy but also increases the FLOPs. We report the performance of both 8-frame MSN and 16-frame MSN. Table 1 shows the comparison results for the proposed MSN on the Something-Something test set, the visualization is shown in Figure 6. Using a ResNet-50 backbone, MSN achieves 53.0% and 54.1% with 8/16 frames, respectively, which are 2.2% and 0.2% better than TEA and TDN, respectively. On the Something-Something v2 dataset, a similar improvement is observed as in SSV1 datasets, especially on 16 frames, which achieved the highest results.

Results on kinetics. On Kinetics-400, we compare our MSN with other state-of-the-art methods. We note that these are comparisons of systems which can differ in many aspects. Nevertheless, our method surpasses all existing RGB or RGB+flowbased methods by a good margin. Without using optical flow and without any bells and whistles, Table 2 shows our model achieved the best performance of 77.1%, the visualization is shown in Figure 7.

We present the results of our experiments to verify the effectiveness of the proposed STP-ME and DS-ME modules, using ResNet50 as the backbone and evaluating the model's accuracy on the something-to-something v1 dataset.

Study on the effect of STP-ME module and DS-ME module. To investigate the impact of the STP-ME and DS-ME modules, we conducted a comparative study and evaluated four different combinations, as summarized in Table 3, the visualization is shown in Figure 8. First, we established a baseline network without any of these modules, which achieved an accuracy of 46.6%. Then, we separately added the STP-ME module and the DS-ME module to the early layers of the network. As the number of STP-ME modules increased, the accuracy improved, achieving 48.8% and 51.8%, respectively. Similarly, the DS-ME module improved the baseline accuracy by 2.3%, achieving an accuracy of 48.8%. Finally, we included all usable modules in our final model, which achieved the best performance of 52.3% and 53.0% on the something-to-something v1 dataset.

In addition, we compared our STP-ME module with similar modules from other works, including S-TDM proposed in TDN and the super image proposed in StNet (as shown in Table 4). From the results, we found that the super-image module could increase the top-1 accuracy by 2%, and S-TDM could increase it by 4.9%. However, our STP-MEM achieved the maximum performance gain of 5.3%.

In our study on the STP-ME module, we found that the fusion operation of different scale features is a crucial step. Therefore, we compared different fusion operations of the STP-ME module, including (1) channel concatenation, (2) element-wise addition, and (3) element-wise average. As shown in Table 5, the element-wise addition achieved the best accuracy of 53.0%, while the element-wise average and channel concatenation obtained top-1 accuracies of 52.3% and 52.1%, respectively. We note that the action information captured by different scale operators is complementary, and therefore the performance of the feature can be maximized when only element-wise addition is used.

Furthermore, we conducted a study on the DS-ME module, where we made several improvements to the deficiencies present in the ME modules of the previous TEA. We tested these improvements one by one, including four networks: (1) using ME modules, (2) using multi-scale ME modules, (3) using DCNN ME modules, and (4) using DS-ME module. As shown in Table 6, these improved modules provided performance improvements of 0.2%, 0.4%, and 0.5%, respectively, the visualization is shown in Figure 9.

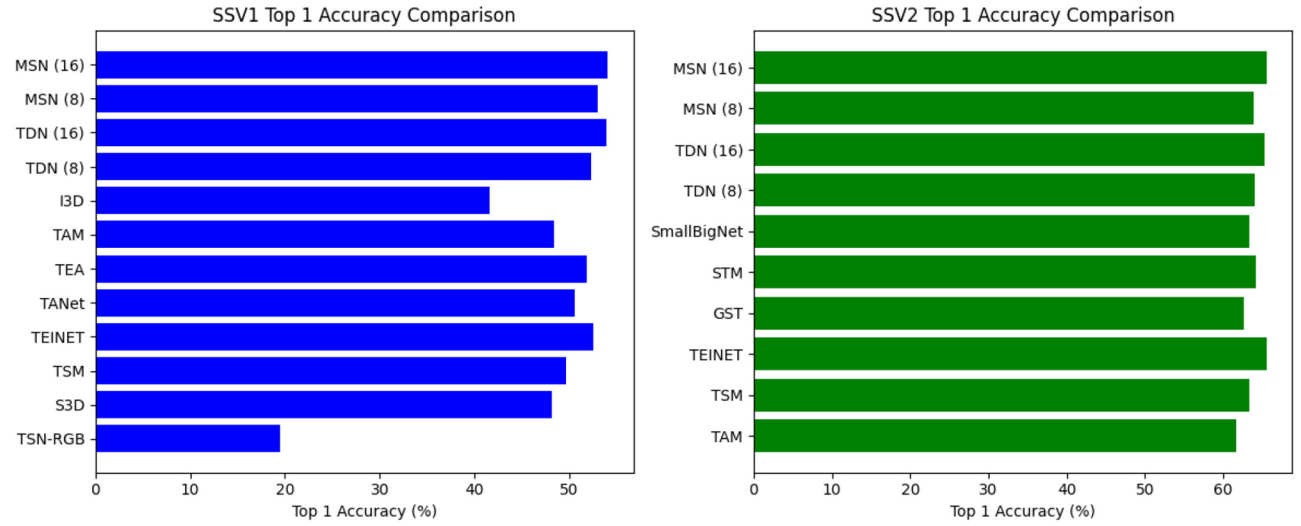


FIGURE 6
Comparison visualization with state-of-the-art methods on Something-something v1&v2 test set.

TABLE 2 **Comparisons with state-of-the-art approaches on the Kinetics-400 test set.**

Kinetics-400(Carreira et al., 2018)					
Method	Backbone	frames	GFLOPs	Top 1	Top 5
ARTNet	R18	16	23.5	69.20%	88.30%
R(2+1)D	R34	16	152	74.30%	91.40%
I3D	Inception	64	108	71.10%	89.30%
S3D-G	Inception	64	71.4	74.70%	93.40%
TSN	Inception	25	16	72.50%	90.20%
TEA	R50	16	70	76.10%	92.50%
SlowOnly	R50	8	41.9	74.90%	91.50%
SlowFast	R50	4+32	36.1	75.60%	92.10%
SlowFast	R50	8+32	65.7	77.00%	92.60%
NL I3D	R50	32	N/A	74.90%	91.60%
NL I3D	R50	128	282	76.50%	92.60%
GloRe	R50	8	28.9	75.10%	N/A
TDN	R50	8	36	76.60%	92.80%
SmallBigNet	R50	8	57	76.30%	92.50%
TSM	R50	16	65	74.70%	N/A
MSN	R50	8	36.2	77.10%	93.10%

Figure 10 shows the performance metrics of various models, including Xing Z et al., Ahn D et al., Chen T et al., Liu Y et al., Wu L et al., Xu B et al., and "Ours," evaluated across three distinct datasets: Kinetics-400, Something-Something V1, and Something-Something V2. Notably, our model, labeled as "Ours," consistently outshines the others across all datasets, boasting the highest accuracy, precision, recall, and AUC-ROC values. Then, Figure 11 provides a comprehensive overview of various models, including Xing Z et al., Ahn D et al., Chen T et al., Liu Y et al., Wu L et al., Xu B et al., and "Ours," assessed across three different datasets: Kinetics-400, Something-Something V1, and Something-Something V2. The models' performance is evaluated based on three key parameters: the number of parameters (in millions), inference time (in milliseconds), and training time (in seconds). Notably, our model, labeled as "Ours," stands out with the lowest number of parameters, efficient inference times, and remarkably short training durations across all datasets. Specifically, on the Kinetics-400 dataset, "Ours" exhibits a competitive parameter

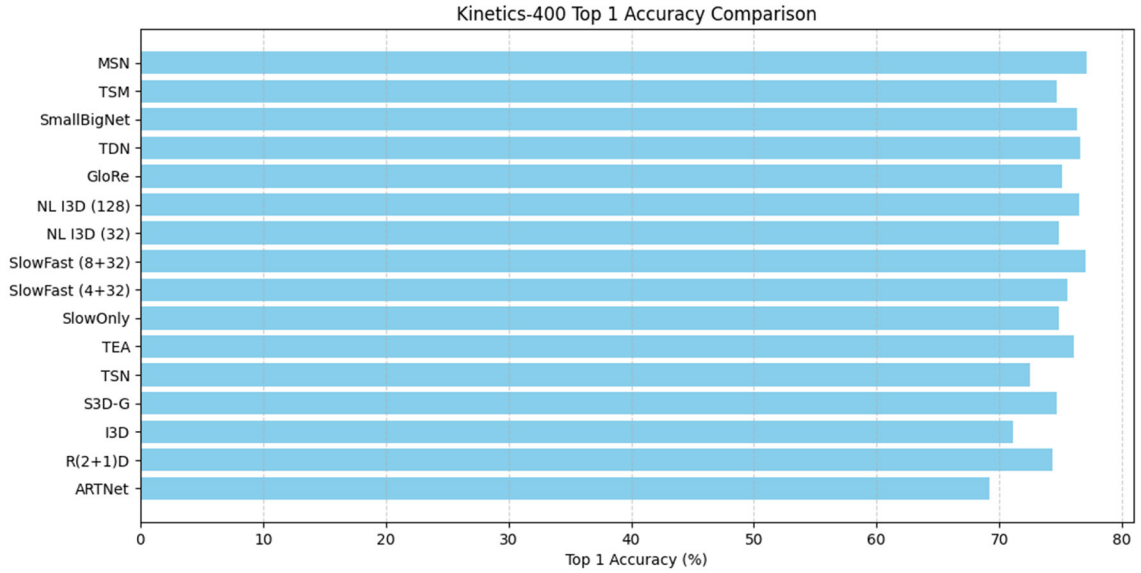


FIGURE 7
Comparative visualization of Kinetics-400 test system and state-of-the-art methods.

TABLE 3 Evaluation of four different combinations.

STP-ME module		DS-ME module			Top 1
stage1	stage2	stage3	stage4	stage5	
					46.60%
✓					48.80%
✓	✓				51.80%
✓	✓	✓	✓		52.30%
		✓	✓	✓	48.90%
✓	✓	✓	✓	✓	53.00%

count (227.64 M), efficient inference time (182.46 ms), and notably quick training time (87.62 s). This trend continues across the Something-Something V1 and V2 datasets, reinforcing the efficiency of our model in terms of model complexity, real-time inference, and training speed compared to other evaluated models.

We visualize the class activation maps with Grad-CAM++ (Chattopadhay et al., 2018) and results are shown in Figure 12. Specifically, we used 8 frames as input and only visualized the activation maps in the center frames. The visualization results clearly demonstrate that the baseline method with only temporal convolutions cannot effectively focus on motion-salient regions, while our proposed MSN with the STP-ME module and DS-ME module for motion modeling is able to more accurately localize action-relevant regions.

5 Discussion

The experimental results demonstrate the effectiveness of the introduced STP-ME and DS-ME modules, marking a significant advance in the field of spatiotemporal modeling for action recognition. The research focuses on the theoretical foundations and practical applications of artificial neural networks (ANN) in autonomous system control and decision-making, and our experimental results bring valuable insights. The quantitative evaluation of MSN against state-of-the-art methods on Kinetics-400 and Something-Something datasets reveals compelling results. Achieving an accuracy of 77.1% on Kinetics-400, MSN outperforms existing RGB or RGB+flow-based methods by a significant margin. This demonstrates not only the theoretical effectiveness of the proposed method but its practical superiority in large-scale action recognition benchmarks. The experiments involving different frame sampling rates (8-frame MSN and 16-frame MSN) showcase the scalability of MSN in handling varied input scenarios. While using more frames generally improves accuracy, MSN maintains competitive performance even with a reduced frame sampling rate. This scalability is crucial for applications where computational resources are limited. Ablation studies offer detailed insights into the impact of module additions. The step-wise improvement in accuracy with the introduction of the STP-ME and DS-ME modules provides a clear understanding of their individual contributions. This data-driven analysis substantiates the claim that these modules are not merely additions but essential components for enhancing action recognition performance. The comparative analysis of fusion operations within the STP-ME module provides nuanced information on the best strategy for integrating multiscale features. The superior performance of element-wise addition in achieving an accuracy of 53.0% underscores its effectiveness in preserving and maximizing valuable information across different temporal scales. When compared with similar modules from previous works, such as S-TDM and super image modules, the STP-ME module exhibits the highest performance gain of 5.3%. This data-driven comparison quantifies the advancements achieved by MSN in capturing intricate motion information, setting it apart as a

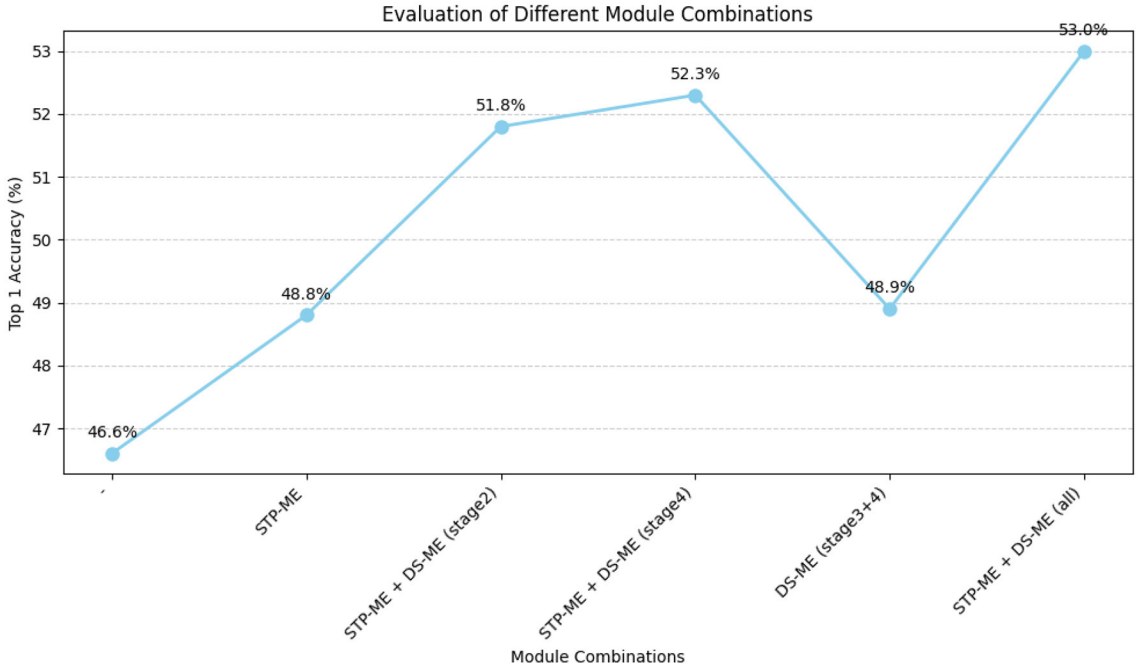


FIGURE 8
Evaluation comparison visualization of four different combinations.

TABLE 4 Performance comparison of STP-ME module and other modules.

Fusion mode	GFLOPs	Top 1
Concatation		52.30%
Element-wise average		52.10%
Element-wise addition		53.00%

TABLE 5 Comparison of ablation experiments of STP-ME modules.

Similar module	Model	dataset	Top 1 (Increase)
Super-Image	St-Net	Kintics-600	2%
S-TDM	TDN	SSV1	4.90%
STP-MEM	MSN	SSV1	5.30%

TABLE 6 DS-ME module improvement testing.

Fusion mode	GFLOPs	Top 1
ME module		48.40%
Multi-scale ME module		48.80%
DCNN-ME		48.70%
DS-MEM		49.00%

leading method for spatial-temporal modeling. Beyond accuracy, the evaluation of MSN's efficiency in resource utilization is critical. The balance achieved between accuracy and computational efficiency, particularly with the sparse sampling strategy and two-stage motion modeling mechanism, positions MSN as a practical solution for real-world applications where both accuracy and efficiency are paramount. MSN's consistent performance across diverse datasets, such as Kinetics-400 and Something-Something, highlights its ability to generalize well to various action recognition scenarios. This generalization is a key characteristic, indicating the adaptability and versatility of MSN in handling different types of actions, scales, and temporal variations. This adaptability aligns with the requirements of autonomous systems, which often encounter diverse action types, scales, and temporal variations. In summary, the theoretical implications of MSN in enhancing spatial-temporal modeling, coupled with its practical performance and efficiency, resonate with the goals of advancing artificial neural networks within the realm of autonomous system control and decision-making. The identification of specific scenarios where MSN excels opens avenues for future optimizations, ensuring its robustness and applicability in specialized application domains within autonomous systems.

6 Conclusion

In this paper, we present a novel network architecture for action recognition, called MSN. Our approach is both simple and effective, and involves leveraging multiple temporal rates in actions using the temporal pyramid module, which captures motion information at different scales by adjusting the size of the convolution kernel and time interval simultaneously. Additionally, we introduce a new motion excitation module that employs a multi-scale deformable CNN to adjust the motion scale of the target object, which is often non-uniform and irregular. We evaluate our method on four challenging datasets, namely Something-Something V1, Something-Something V2

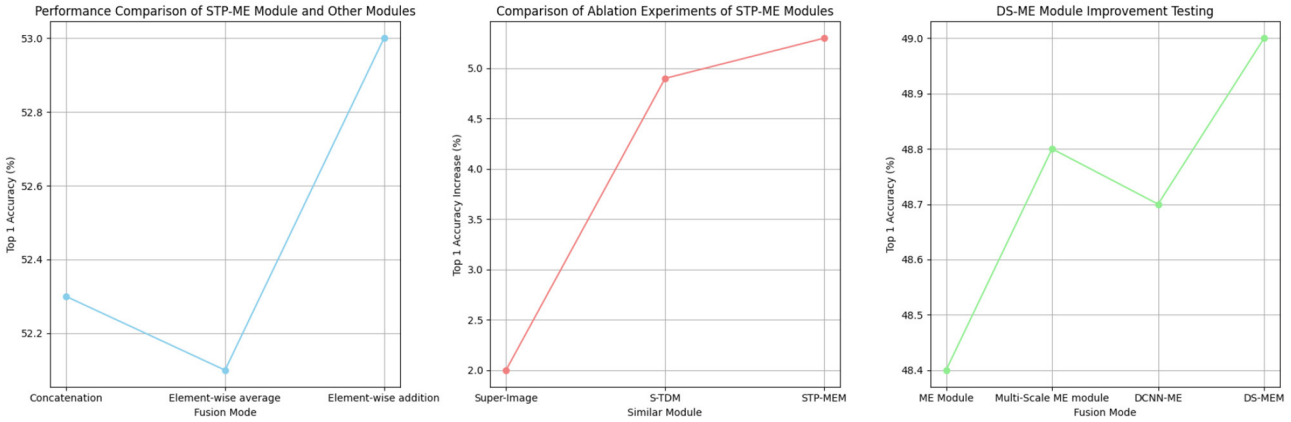


FIGURE 9
Comparative visualization of module performance and ablation experiments.

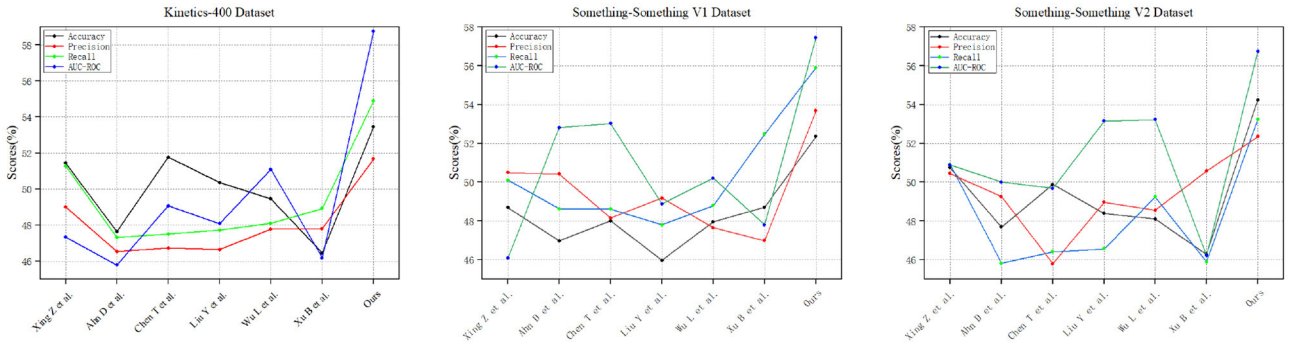


FIGURE 10
Comparison of experimental indicators between this method and other methods on three data sets.

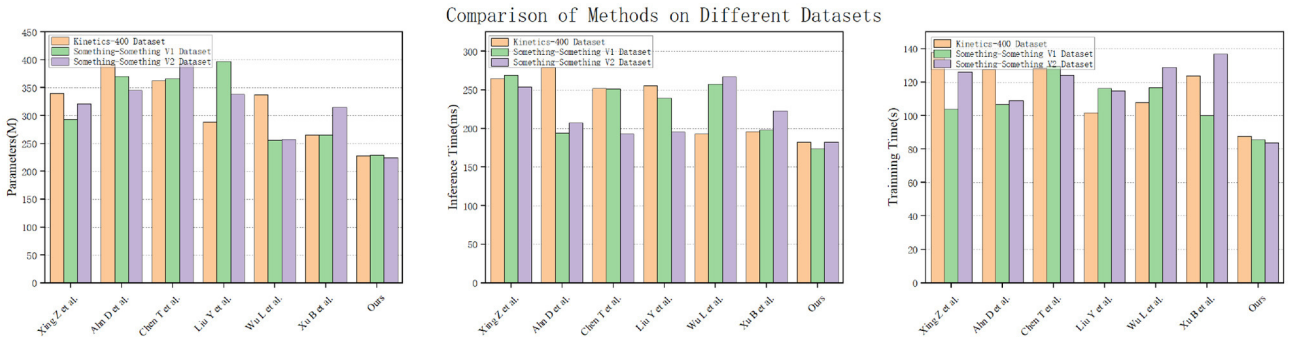


FIGURE 11
Experimental comparison of parameters, inference time and training time of this method with other methods on three datasets.

and Kinetics-400, and compare our results to those of other state-of-the-art (SOTA) approaches. The results demonstrate that MSN performs exceptionally well in a variety of challenging scenarios. The theoretical foundation of MSN is in line with the continuously evolving landscape of spatiotemporal modeling, resonating with the broader discussions about the integration of Artificial Neural Networks (ANN) in autonomous system control and decision-making. Its adaptability across datasets and scenarios, coupled with efficiency, positions MSN as a promising tool that not only advances action recognition but also contributes to the theoretical and practical foundations of ANN in the autonomous systems domain. While MSN demonstrates commendable performance in action recognition, it is essential to acknowledge its computational cost, interpretability challenges,

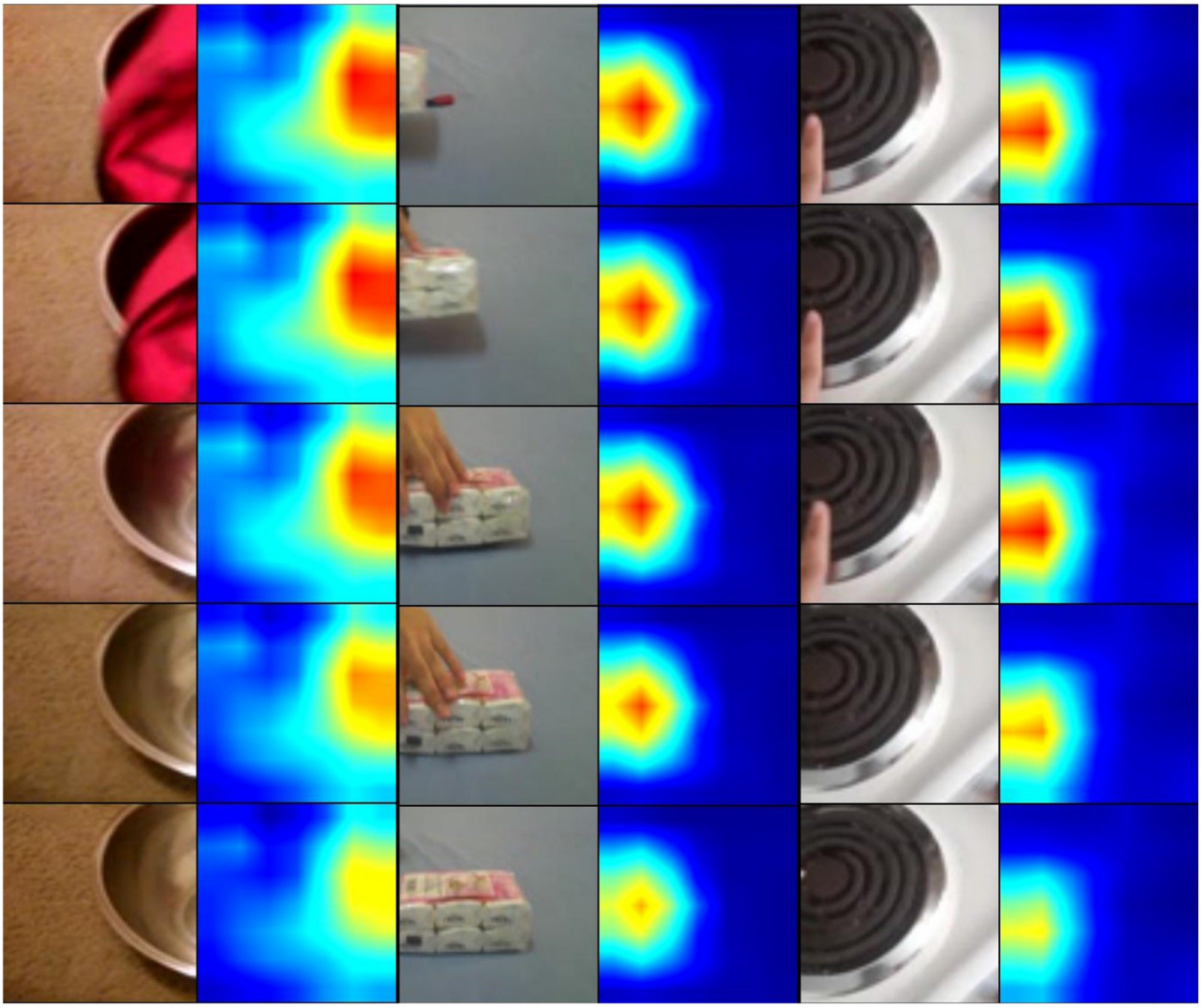


FIGURE 12
Visualization of activation maps with Grad-CAM++.

and the need for further extension to new environments. Future developments should prioritize enhancing the interpretability of MSN, achieving real-time adaptability, exploring transfer learning in diverse environments, delving into human interaction understanding, and seamlessly integrating MSN into autonomous systems. We can design network structures with enhanced interpretability, introduce attention mechanisms, or employ visualization techniques to illustrate the model's key steps and rationale during decision-making. Additionally, domain adaptation, transfer strategy design, improving model robustness, and incorporating online learning mechanisms are also indispensable aspects to consider. These steps will pave the way for establishing a more robust, transparent, and versatile network, aligning with the ongoing developments in the field of artificial neural networks within autonomous system control and decision-making.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JG: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Writing – original draft. YY: Data curation, Methodology, Project administration, Resources, Software, Visualization, Writing – review & editing. QL: Conceptualization, Data curation, Funding acquisition, Project administration, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The funding for this work was partly supported by the National Natural Science Foundation of China under Grant No. 61672546, and the Guangzhou Science and Technology Project under Grant No. 201707010127.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al. (2022). Flamingo: a visual language model for few-shot learning. *Adv. Neur. Inf. Proc. Syst.* 35, 23716–23736. doi: 10.48550/arXiv.2204.14198

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., and Zisserman, A. (2018). A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

Carreira, J., and Zisserman, A. (2017). "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308. doi: 10.1109/CVPR.2017.502

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE), 839–847. doi: 10.1109/WACV.2018.00097

Chen, F., Luo, Z., Xu, Y., and Ke, D. (2019). Complementary fusion of multi-features and multi-modalities in sentiment analysis. *arXiv preprint arXiv:1904.08138*.

Cheng, D., Zhang, L., Bu, C., Wang, X., Wu, H., and Song, A. (2023). Protohar: Prototype guided personalized federated learning for human activity recognition. *IEEE J. Biomed. Health Inform.* 27, 3900–3911. doi: 10.1109/JBHI.2023.3275438

Cob-Parro, A. C., Losada-Gutiérrez, C., Marrón-Romera, M., Gardel-Vicente, A., and Bravo-Mu noz, I. (2024). A new framework for deep learning video based human action recognition on the edge. *Expert Syst. Applic.* 238:122220. doi: 10.1016/j.eswa.2023.122220

Dai, W., Mou, C., Wu, J., and Ye, X. (2023). "Diabetic retinopathy detection with enhanced vision transformers: the twins-pcpvt solution," in *2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI)* (IEEE), 403–407. doi: 10.1109/ICETCI57876.2023.10176810

Diba, A., Fayyaz, M., Sharma, V., Hossein Karami, A., Mahdi Arzani, M., Yousefzadeh, R., et al. (2018). "Temporal 3D convnets using temporal transition layer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1117–1121.

Han, Z., Lu, Y., Li, Y., Wu, R., and Huang, Z. (2022). Strategy to combine two functional components: efficient nano material development for iodine immobilization. *Chemosphere* 309:136477. doi: 10.1016/j.chemosphere.2022.136477

He, D., Zhou, Z., Gan, C., Li, F., Liu, X., Li, Y., et al. (2019). STNET: local and global spatial-temporal modeling for action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 8401–8408. doi: 10.1609/aaai.v33i01.33018401

He, S., and Tang, Z. (2023). Fabrication and control of porous structures via layer-by-layer assembly on pah/paa polyelectrolyte coatings. Shuyue He and Ziyu Tang. Fabrication and Control of Porous Structures Via Layer-By-Layer Assembly on PAH/PAA Polyelectrolyte Coatings. *Biomed. J. Sci. Tech. Res.* 51:8165. doi: 10.26717/BJSTR.2023.51.008166

Huang, W., Zhang, L., Wu, H., Min, F., and Song, A. (2022). Channel-equalization-har: a light-weight convolutional neural network for wearable sensor based human activity recognition. *IEEE Trans. Mobile Comput.* 22, 5064–5077. doi: 10.1109/TMC.2022.3174816

Islam, M. M., Nooruddin, S., Karray, F., and Muhammad, G. (2023). Multi-level feature fusion for multimodal human activity recognition in internet of healthcare things. *Inf. Fusion* 94, 17–31. doi: 10.1016/j.inffus.2023.01.015

Kulsoom, F., Narejo, S., Mehmood, Z., Chaudhry, H. N., Butt, A., and Bashir, A. K. (2022). A review of machine learning-based human activity recognition for diverse applications. *Neur. Comput. Applic.* 34, 18289–18324. doi: 10.1007/s00521-022-07665-9

Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., and Wang, L. (2020). "Tea: temporal excitation and aggregation for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 909–918. doi: 10.1109/CVPR42600.2020.00099

Lin, Z., and Xu, F. (2023). "Simulation of robot automatic control model based on artificial intelligence algorithm," in *2023 2nd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)* (IEEE), 535–539. doi: 10.1109/AIARS59518.2023.00113

Liu, K., He, S., Li, L., Liu, Y., Huang, Z., Liu, T., et al. (2021). Spectroscopically clean au nanoparticles for catalytic decomposition of hydrogen peroxide. *Sci. Rep.* 11:9709. doi: 10.1038/s41598-021-89235-y

Luo, Z. (2023). "Knowledge-guided aspect-based summarization," in *2023 International Conference on Communications, Computing and Artificial Intelligence (CCCAI)* (IEEE), 17–22. doi: 10.1109/CCCAI59026.2023.00012

Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., and Darrell, T. (2020). "Something-else: Compositional action recognition with spatial-temporal interaction networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1049–1059. doi: 10.1109/CVPR42600.2020.00113

Qiu, Z., Yao, T., Ngo, C.-W., Tian, X., and Mei, T. (2019). "Learning spatio-temporal representation with local and global diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12056–12065. doi: 10.1109/CVPR.2019.01233

Simonyan, K., and Zisserman, A. (2014). "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 27.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., and Liu, J. (2022). Human action recognition from various data modalities: a review. *IEEE Trans. Patt. Analy. Mach. Intell.* 45, 3200–3225. doi: 10.1109/TPAMI.2022.3183112

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6450–6459. doi: 10.1109/CVPR.2018.00675

Wang, D., Wang, T., and Florescu, I. (2020). Is image encoding beneficial for deep learning in finance? *IEEE Internet Things J.* 9, 5617–5628. doi: 10.1109/JIOT.2020.3030492

Wang, L., Tong, Z., Ji, B., and Wu, G. (2021). "TDN: temporal difference networks for efficient action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1895–1904. doi: 10.1109/CVPR46437.2021.00193

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., et al. (2016). "Temporal segment networks: towards good practices for deep action recognition," in *European Conference on Computer Vision* (Springer), 20–36. doi: 10.1007/978-3-319-46484-8_2

Wu, J., Hobbs, J., and Hovakimyan, N. (2023). "Hallucination improves the performance of unsupervised visual representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16132–16143. doi: 10.1109/ICCV51070.2023.01478

Wu, M., Pan, S., Zhou, C., Chang, X., and Zhu, X. (2020). "Unsupervised domain adaptive graph convolutional networks," in *Proceedings of the Web Conference 2020*, 1457–1467. doi: 10.1145/3366423.3380219

Wu, R., Han, Z., Chen, H., Cao, G., Shen, T., Cheng, X., et al. (2021). Magnesium-functionalized ferro metal-carbon nanocomposite (MG-FEMEC) for efficient uranium extraction from natural seawater. *ACS EST Water* 1, 980–990. doi: 10.1021/acsestwater.0c00262

Xu, K., Ye, F., Zhong, Q., and Xie, D. (2022). "Topology-aware convolutional neural network for efficient skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2866–2874. doi: 10.1609/aaai.v36i3.20191

Xu, S., Zhang, L., Tang, Y., Han, C., Wu, H., and Song, A. (2023). Channel attention for sensor-based activity recognition: embedding features into all frequencies in DCT domain. *IEEE Trans. Knowl. Data Eng.* 45, 12497–12512. doi: 10.1109/TKDE.2023.3277839

Yang, H., Yuan, C., Li, B., Du, Y., Xing, J., Hu, W., et al. (2019). Asymmetric 3D convolutional neural networks for action recognition. *Patt. Recogn.* 85, 1–12. doi: 10.1016/j.patcog.2018.07.028

Yao, K., Liang, J., Liang, J., Li, M., and Cao, F. (2022). Multi-view graph convolutional networks with attention mechanism. *Artif. Intell.* 307:103708. doi: 10.1016/j.artint.2022.103708

Zhang, M., Xie, K., Zhang, Y.-H., Wen, C., and He, J.-B. (2022). Fine segmentation on faces with masks based on a multistep iterative segmentation algorithm. *IEEE Access* 10, 75742–75753. doi: 10.1109/ACCESS.2022.3192026

Zheng, J., Li, W., Hong, J., Petersson, L., and Barnes, N. (2022a). "Towards open-set object detection and discovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3961–3970. doi: 10.1109/CVPRW56347.2022.00441

Zheng, J., Yao, Y., Han, B., Wang, D., and Liu, T. (2024). "Enhancing contrastive learning for ordinal regression via ordinal content preserved data augmentation," in *The Twelfth International Conference on Learning Representations*.

Zheng, Y., Qi, Y., Tang, Z., Hanke, F., and Podkolzin, S. G. (2022b). Kinetics and reaction mechanisms of acetic acid hydrodeoxygenation over pt and pt-mo catalysts. *ACS Sustain. Chem. Eng.* 10, 5212–5224. doi: 10.1021/acssuschemeng.2c00179

Zhou, B., Andonian, A., Oliva, A., and Torralba, A. (2018). "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 803–818. doi: 10.1007/978-3-030-01246-5_49

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 12 April 2024
DOI 10.3389/fnins.2024.1362567

OPEN ACCESS

EDITED BY
Hancheng Zhu,
China University of Mining and Technology, China

REVIEWED BY
Thangairulappan Kathirvalavakumar, Virudhunagar Hindu Nadars' Senthikumara Nadar College (Autonomous), India
Li Yan,
China University of Mining and Technology, China

*CORRESPONDENCE
J. Shreyas
✉ shreyas.j@manipal.edu

†These authors have contributed equally to this work

RECEIVED 28 December 2023
ACCEPTED 20 March 2024
PUBLISHED 12 April 2024

CITATION
Ramesh G, Shreyas J, Balaji JM, Sharma GN, Gururaj HL, Srinidhi NN, Askar SS and Abouhawwash M (2024) Hybrid manifold smoothing and label propagation technique for Kannada handwritten character recognition. *Front. Neurosci.* 18:1362567. doi: 10.3389/fnins.2024.1362567

Hybrid manifold smoothing and label propagation technique for Kannada handwritten character recognition

G. Ramesh[1], J. Shreyas[2]*, J. Manoj Balaji[3], Ganesh N. Sharma[3†], H. L. Gururaj[2], N. N. Srinidhi[4], S. S. Askar[5] and Mohamed Abouhawwash[6]

[1]Department of AIML-Artificial Intelligence & Machine Learning, Alva's Institute of Engineering and Technology, Mangalore, Karnataka, India, [2]Department of Information Technology, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, Karnataka, India, [3]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bengaluru, Karnataka, India, [4]Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, Karnataka, India, [5]Department of Statistics and Operations Research, College of Science, King Saud University, Riyadh, Saudi Arabia, [6]Department of Mathematics, Faculty of Science, Mansoura University, Mansoura, Egypt

Handwritten character recognition is one of the classical problems in the field of image classification. Supervised learning techniques using deep learning models are highly effective in their application to handwritten character recognition. However, they require a large dataset of labeled samples to achieve good accuracies. Recent supervised learning techniques for Kannada handwritten character recognition have state of the art accuracy and perform well over a large range of input variations. In this work, a framework is proposed for the Kannada language that incorporates techniques from semi-supervised learning. The framework uses features extracted from a convolutional neural network backbone and uses regularization to improve the trained features and label propagation to classify previously unseen characters. The episodic learning framework is used to validate the framework. Twenty-four classes are used for pre-training, 12 classes are used for testing and 11 classes are used for validation. Fine-tuning is tested using one example per unseen class and five examples per unseen class. Through experimentation the components of the network are implemented in Python using the Pytorch library. It is shown that the accuracy obtained 99.13% make this framework competitive with the currently available supervised learning counterparts, despite the large reduction in the number of labeled samples available for the novel classes.

KEYWORDS

computer vision, convolutional neural networks, handwritten character recognition, machine learning, manifold smoothing, label propagation

1 Introduction

The challenge of converting manuscripts and printed documents into digital formats has been the focus of computer vision research (Nasir et al., 2021; Gowda and Kanchana, 2022). Recent advances have blurred the interface between physical copies of text and their digital counterparts. Large scale scanning of thousands of historical documents has been performed. Enabling visually impaired individuals to read signboards and paper, and faster processing of checks. Legislative bodies have benefited from the ease of digitizing legal documents, allowing for seamless transfer, signing, and searching. The field

of handwritten character analysis has strive to make effective algorithms to achieve various goals, such as the classification of handwritten characters, the classification of the writers of different manuscripts, generating text matching the handwriting of a writer, and so on (Dhiaf et al., 2023). Prior to supervised deep neural networks, handcrafted methods were used for handwritten character recognition, which often required several different steps such as binarization of images, rescaling and rotating the images, performing statistical aggregations on different parts of the images, etc. This required fine-tuning a large number of parameters to obtain accurate results and could not generalize well to variations in the input images (Aradhya et al., 2010; Ramesh et al., 2020).

Supervised learning using deep neural networks has allowed most of the explicit tasks to be replaced by a single neural network model that, by virtue of back-propagation, is able to learn the weights required for effective extraction of features from the images that are used for classification. By providing a large training set that includes diverse samples of each character, the neural network is rendered more robust in its accuracy in classifying a larger range of handwriting samples. However, the creation of a large labeled training set of images is laborious, and certain character classes have few real-world samples. By utilizing already pre-trained models to predict the new classes, sample efficiency is improved. The difficulty in obtaining such a dataset for Kannada handwritten characters is compounded by the large number of possible graphemes in the Kannada script, stemming from the use of combinations of base characters to form digraphs. Semi-supervised learning techniques, which exploit the use of a large unlabeled dataset to improve the robustness and accuracy of a model trained on a small labeled training set, have been successfully used to achieve this goal. The scenario of novel classes' incorporation is modeled with the episodic learning approach (Nichol et al., 2018). Recent works in few shot learning make use of this framework to mimic meta-learning tasks (Gidaris et al., 2019). Improved generalization of the neural network is achieved through the use of data augmentation where sample images are rotated in four different orientations, increasing the number of training samples the network is trained on (Zhou et al., 2004). The use of label propagation allows the incorporation of new classes into the classification framework with very few extra training samples (Alsuhibany and Alnoshan, 2021). The handwritten CAPTCHA image then asks visitors to choose the joints between Arabic letters. In the latter approach, a novel generator of Arabic handwritten CAPTCHA pictures is devised; once the image is formed, the user is required to input the letters depicted in the image (Weldegebriel et al., 2019). Although both have showed encouraging outcomes, this experimental study compares both in terms of security and usability for mobile device applications.

The enormous success of supervised neural network-based machine learning approaches can be ascribed to the minimal amount of manual parameter adjustment needed as well as the models' flexibility to learn efficient feature representations that work for a variety of inputs. However, supervised neural network models need well-curated, sizable, labeled datasets to obtain strong generalization capabilities and robustness. This makes it feasible for the models to accurately learn the various potential variations they might experience. Due to the bias introduced by unbalanced datasets, these models may favor predicting the classes that were represented more frequently in the training set, which would lead to subpar performance when identifying previously undiscovered classes of characters. Being one of the acknowledged regional languages in India, Kannada also serves as the province of Karnataka's official language of communication (Ramesh et al., 2019b). The literature and artistic diversity of the language makes it a priceless repository of information and culture. Many of these regional languages need the power of technology to retain the language directed at them (Thippeswamy and Chandrakala, 2020; Parikshith et al., 2021). The preservation of the language's scripture is greatly aided by advances in digitization, which also give the language a significant edge in terms of reaching a wider audience given the pervasiveness of internet access around the world. Building precise pattern recognition models is also a difficult task due to the absence of readily accessible annotated data relevant to the local languages. The suggested study addresses the issue of "Recognizing Kannada Handwritten Characters in a Few-Shot Learning viewpoint" by utilizing a strong, cutting-edge technique that offers best-in-class accuracy and consistent outcomes. There are 47 basic characters in the Kannada alphabet. Main contribution of this paper are as fallows, we introduce a Manifold Smoothing and Label Propagation-based Approach for Offline Handwritten Kannada Character Recognition. In particular, our contributions are outlined as follows: The goal of this work is to combine a few techniques in order to create an offline Kannada handwritten character classifier that can be trained to retain high accuracies on classes with as few as one or five samples. This allows for the rapid incorporation of classes with minimal extra samples required.

- A novel classes incorporation is modeled with the episodic learning approach.
- Improved generalization of the neural network is achieved through the use of data augmentation.
- The label propagation allows the incorporation of new classes into the classification framework with very few extra training samples.

2 Related work

Weldegebriel et al. (2019) presented by the Handwritten Ethiopian Character Recognition (HECR) dataset was used to prepare a model, and the HECR dataset for images with more than one shading pen RGB was considered. This framework employs a half breed model comprised of two super classifiers: CNN and eXtreme Gradient Boosting (XGBoost). CNN-XGBoost characterization error rate brings about HECR dataset 0.1612%. This proposed work got an accuracy of 99.84% in the CNN-XGBoost strategy. Sahlol et al. (2020) proposed a hybrid ML approach that uses area binary whale improvement calculation to choose the most suitable highlights for the recognition of handwritten Arabic characters. This strategy utilized the CENPARMI dataset and This strategy results show away from of the proposed approach as far as memory footprint, recognition accuracy, and processor time than those without the features of the proposed technique. This proposed BWOA-NRS approach beats any remaining works in both execution and time utilization got an accuracy of 96% in 1.91 s time. Cilia et al. (2018) has considered various univariate measures to create an feature ranking and

proposed a greedy search approach for picking the element subset ready to maximize the characterization results. One of the best and broadly set of features in handwriting recognition and we have utilized these features for considering to tests of three genuine word information bases. Karthik and Srikanta Murthy (2018) presented by the recognition of isolated handwritten characters of Kannada proposed a new method based on deep belief network with DAG features. The recognition accuracy for consonants and vowels to achieve an accuracy of 97.04% using deep belief network.

Weng and Xia (2019) proposed technique using Convolutional neural network has been approved by previous work with the results of existing strategies, utilized for optical character recognition. In this strategy, First, build a Shui character dataset for applying a Convolutional neural network to manually written character recognition, at that point during the proposed of the CNN, analyzed the consequences of various parameters so that proposed the parameter tuning suggestions and accuracy is around 93.3%. Guha et al. (2019) presented by CNN has been a well-known way to deal with remove features from the image data. in this work, we consider as different cnn models freely accessible Devanagari characters and numerals datasets. This method uses a Kaggle Devanagari character dataset, UCI character dataset, CVPR ISI Devanagari dataset, and CMATERdb 3.2.1 dataset. Using the DevNet, the recognition accuracies obtained on UCI DCD, CVPR ISI Devanagari character dataset, CMATERdb 3.2.1, and Kaggle Devanagari character dataset have obtained an accuracy of 99.54, 99.63, 98.70, and 97.29%, respectively. Khan et al. (2019) proposed technique presents a efficient handwriting identification framework which joins Scale Invariant Feature Transform (SIFT) and RootSIFT descriptors in a bunch of Gaussian mixture models (GMM). This proposed system using six different public datasets are IAM dataset obtained accuracy of 97.85%, IFN/ENIT dataset obtained an accuracy of 97.28%, AHTID/MW dataset obtained an accuracy of 95.60%, CVL dataset obtained an accuracy of 99.03%, Firemaker dataset obtained an accuracy of 97.98%, and ICDAR2011 dataset obtained an accuracy of 100.0%.

Sahare and Dhok (2018) proposed robust algorithms for character segmentation and recognition are introduced for multilingual Indian document images of Latin and Devanagari contents. Perceiving the input character utilizing the KNN classifier technique, as it has characteristically zero preparing time. This strategy got the highest segmentation and recognition rates of 98.86% is acquired on an exclusive information base of Latin content and the Proposed recognition algorithm shows the most best accuracy of 99.84% on the Chars74k numerals data set. Zheng et al. (2019) proposed strategy separate a novel component from pooling layers, called uprooting highlights, and join them with the features coming about because of max-pooling to catch the structural deformations for text recognition tasks. This strategy utilizes three content datasets, MNIST, HASY, and Chars74K-textual style, and contrasted the proposed technique and CNN based models and best in class models. Mhiri et al. (2018) work depends on deep CNN and it doesn't need explicit segmentation of characters for the recognition of manually written words. Proposed strategy presentation forward and in reverse ways or robust representation. This proposed approach use IAM and RIEMS information base and this methodology achieve a word error rate of 8.83% on the IAM information base and 6.22% on the RIEMS dataset.

Sueiras et al. (2018) proposed a technique framework for recognizing offline handwritten words and use of another neural architecture design that consolidates a deep cnn with an encoder-decoder, called sequence to sequence. This proposed technique utilizes two handwritten databases are IAM and RIMES datasets and these datasets acquire a word error rate in the test set of 12.7% in IAM and 6.6% in RIMES datasets. Katiyar and Mehfuz (2016) proposed presents hybrid feature extraction and GA based feature selection for off-line handwritten character recognition by utilizing adaptive MLPNN classifier. The proposed technique has been performed utilizing the standard database of Center of Excellence for Document Analysis and Recognition for the English alphabet. It is obvious from the outcomes that the proposed strategy beats the other state of art techniques with an accuracy of 91.56 and 87.49% individually for capital alphabet and little alphabet in order (Singh et al., 2020). The proposed technique contains six non-Indic-contents and eight Indic contents specifically, Persian, Roman, Thai, Chinese, Japanese, Arabic, Chinese, Japanese Assamese, Bangla, Devanagari, Gurmukhi, Tamil, Telugu, Kannada, and Malayalam. This strategy conversation about the classification tools, pre-processing steps, include feature extraction, and approach's utilized, and different online handwriting recognition methods advancement have been carried out. Ramesh et al. (2019a) demonstrate the use of Convolutional Networks in generating extremely accurate handwritten character classifiers. They assembled the vowels and consonants freely and utilized 400 images for each character for preparing the CNN. They have claimed the accuracy of 98.7%.

3 System architecture

The proposed method's architecture is based on the episodic framework for few-shot learning shown in Figure 1. The dataset consists of images of handwritten characters in Kannada with 400 examples, written by multiple writers each for 47 classes with a size of 84 x 84 for each image. The episodic framework is utilized to evaluate the architecture in a few-shot environment.

3.1 Experimental steps

The experiment is carried out with the following steps:

- **Collection of the dataset:** The dataset consists of 47 classes representing each base character of the Kannada abugida. Each class consists of 400 samples obtained from different writers. 50% of the dataset is used for pretraining (24 classes), 2% is used for finetuning (12 classes), and 25% is used for the validation set (11 classes).
- **Preprocessing the images:** The images are rescaled to 84 x 84px using the Python Image Library (PIL) library. Bilinear interpolation is used to achieve this. The images are converted to RGB format.
- **Training the handwritten character classifier:** Two different convolutional networks are used, the Conv4 network and the

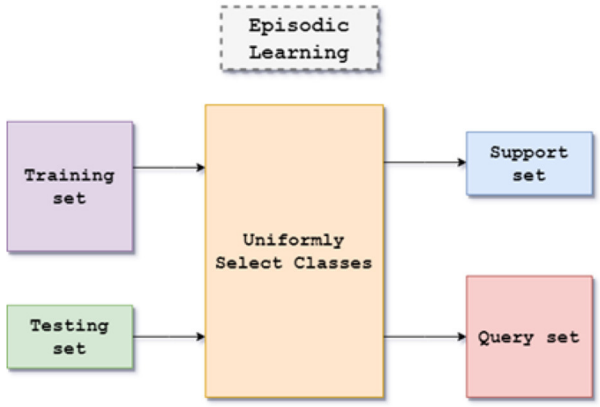


FIGURE 1
Flow diagram of episodic framework.

Resnet-12 network. The training consists of the pretraining phase where the network is trained on the base set. The next phase is the finetuning phase, where the network is trained in an episodic fashion on the unseen classes.

- **Analyzing the result:** The accuracy and loss of the two different networks are plotted and compared. Training and Validation accuracy are plotted for the pretraining phase (seen characters), while Test and Validation accuracy are plotted for the finetuning phase. 1-shot and 5-shot finetuning are performed (one example per class and five examples per class, respectively).

3.2 Episodic framework

The episodic framework was introduced by Nasir et al. (2021). It provides a simulation for training a meta-learning model for few-shot classification tasks. In the episodic framework is a large labeled dataset C_{train} is present. The goal is to train the classifier on a previously unexplored set of classes C_{test}, where there are only a few labeled samples available. To create a support set S and query set Q for each episode, a small subset of N classes from the C_{train}, each task has N classes that need to be classified in N way K shot learning, which has K available labeled samples. In contrast to the query set Q's different examples from the same N classes, the support set S's K examples from each of the N classes. In this work, N = 5 classes are chosen, and the size of the query set is 15 examples per class. The five classes are chosen uniformly over the union of sets (C_{train}) U (C_{test}) and sample accordingly. A transductive setting is used due to the small size of K in the support set. The entire query set Q can be used for predicting labels rather than predicting each example independently. This helps alleviate the bias caused by the small number of samples while improving generalization.

4 Proposed approach

The proposed work uses the combination of manifold smoothing and label propagation to solve the considered problem statement. For better generalization, Manifold Smoothing is used to regularize the features extracted for better generalization, while Label Propagation allows few-shot inference on unseen classes.

4.1 Manifold smoothing with metric learning

In order to make the decision boundaries of the hidden layer of the model more smooth, resulting in better robustness and generalization, a layer to smoothen the extracted features is used (Lee et al., 1995). Given the feature vectors $z_i \in R^m$ (R^m is the set of m-dimensional real number vectors) which are extracted using the Convolutional Neural Network layers, a smoothing function is applied to obtain the smoothed feature vectors $\tilde{z}_i$, which are forwarded to the fully connected layer for classification. This smoothing process consists of using a Gaussian similarity function using the L2 norm as a measure of the similarity/dissimilarity of the different features. $d_{ij}^2 = \|z_i - z_j\|_2^2$ where d_{ij}^2 is the distance between feature vectors z_i and z_j and $\|z_i - z_j\|_2^2$ is the square of the L2 norm between the feature vectors, for pairs of features z_i, z_j and

A similarity matrix is constructed using Equation 1:

$$A_{ij} = e^{\frac{-d^2_{ij}}{\sigma^2}} \tag{1}$$

where A_{ij} is the element of the similarity matrix A, d_{ij}^2 is the distance between feature vectors z_i and z_j and $\sigma^2 = \mathrm{Var}(d_{ij}^2)$ is the variance of d_{ij}^2.

The similarity matrix A is normalized using the Laplacian in order to ensure convergence:

$$L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{2}$$

where L is the Laplacian similarity matrix computed using normalizing matrix D defined as Equation 3.

$$D_{ii} = \sum\nolimits_j A_{ij} \tag{3}$$

Power iteration is used to successively increase the weights of the closest features while reducing the weights of the features that are not too close to each other. This is similar to the power iteration needed in label propagation, and the propagator matrix P is thus obtained by:

$$P = (I - \alpha L)^{-1} \tag{4}$$

where P is the propagator matrix, I is the identity matrix, α is the smoothing factor and L is the Laplacian obtained using Equation (2). The new feature vectors are calculated as Equation 5:

$$\tilde{z}_i = \sum\nolimits_j P_{ij} z_j \tag{5}$$

where P is the matrix calculated in Equation (4), z_j is the input feature vector and $\tilde{z}_i$ is the smoothed feature vector.

This is similar to a weighted sum of neighbors, resulting in a reduction in the noise present in each feature vector.

4.2 Label propagation

The prediction of labels for the query set Q using label propagation is obtained using the similarity matrix that is equivalent to the one used in the manifold smoothing step. Given the query set Q, the equation for the label matrix Y is given by:

$$Y = \frac{Y_S}{0} \tag{6}$$

where Y is the label matrix,

- The matrix Y_S of size (nk×n) corresponds to the support set S. In each row of Y_S, the column corresponding to the correct label is 1, ($Y_{ij} = 1$) if $y_i = j$. The rest of the elements are 0.
- The matrix 0 is a matrix of 0s of size (t×n) and corresponds to the query set Q. n is the number of classes, k is the number of samples per class in S, and t is the number of samples in Q.

Label propagation iteratively determines the unknown labels for the union set S ∪ Q (Ramesh et al., 2020):

$$F_{t+1} = \alpha L F_t + (1 - \alpha) Y \tag{7}$$

where L is the normalized similarity matrix calculated in Equation (2), F_t is the label propagation after t iterations, Y is the label matrix defined in Equation (6) and α is the smoothing factor between 0 and 1. The sequence F_t converges to

$$F^* = (I - \alpha L)^{-1} Y \tag{8}$$

where F^* is the matrix obtained on convergence of Equation (7) as $t \to \infty$. The different features are clustered in a similar fashion to graph spectral clustering (Equation 8).

4.3 Feature extraction using convolutional neural networks

The features are extracted from the input images using convolutional neural network layers (CNNs). Two CNN feature extractors are used in the experiments to determine the one with greater efficacy.

- The first feature extractor is a standard CNN Model with four layers Each layer consists of a convolution (kernel of size 3 × 3), as mentioned in Table 1 followed by Max-Pooling which reduces the size of the image progressively in each layer. The window of the Max-Pool layer is (2 × 2). The ReLU (Rectified Linear Unit) is used as the activation function which zeroes negative values.

The second is a Resnet Model with 12 layers (Karthik and Srikanta Murthy, 2018). This model is deeper, and each block has an identity shortcut path that helps prevent the vanishing gradient problem that is exacerbated as the number of layers increases. This increased depth improves the feature representation of the model, resulting in greater accuracy.

TABLE 1 Layer of Conv4 network.

Layer name	Output shape	Next layer
Input layer	(84, 84, 3)	Conv0
Conv0	(42, 42, 64)	Conv2
Conv2	(10, 10, 64)	Conv3
Conv3	(5, 5, 64)	AvgPool
AvgPool	(64)	Output

TABLE 2 Layer of RestNet12 network.

Layer name	Output shape	Next layer
Input layer	(84, 84, 3)	Block0
Block0	(26, 26, 64)	Block1
Block1	(9, 9, 128)	Block2
Block2	(3, 3, 256)	Block3
Block3	(512)	Output

```
Input: Batch of input images
rotated_input_batch = rotate(input_batch, 0,90,180,270) Rotating the images;
z = backbone_network(rotated_input_batch) z is the feature representation;
A = new matrix(size = z.len × z.len) Manifold Smoothing;
for <z_i in z> do
    for z_j in z do
        eIfi==j A[i][j] = 0;
        A[i][j] = exp(-(L2Norm(z_i,z_j))^2/Var(L2Norm(z_i,z_j));
    end for
end for
A = laplacian(A) Normalizing the matrix;
I = new matrix(size = A.size, type = Identity);
P = matrix_invert(I α × A) Smoothing factor α taken as 0.9;
z_smooth = P×z;
predicted_label = fully_connected_classifier(z_smooth, z_smooth.labels) #C1;
predicted_rotation = fully_connected_classifier(z_smooth, rot_labels) #C2;
Output: Predicted labels of the input images;
```

Algorithm 1. Pretraining algorithm.

As mentioned in Table 2 each block has 3 convolutional layers, a shortcut connection between the first and the third layer and a Max-Pool layer (of window (3 × 3)). The shortcut connection adds the output of the first layer and third layer before passing it to the activation function (ReLU again).

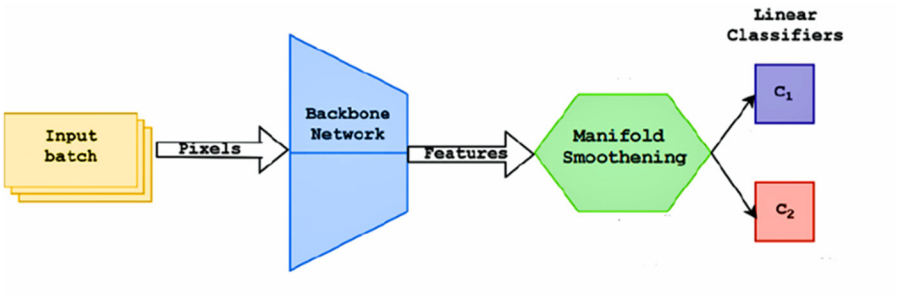


FIGURE 2
Flow diagram of the pretraining process.

4.4 Pretraining process

The pretraining process is similar to a supervised training schedule. The training set C_{train}, It contains classes that have a large number of labeled examples. The objective of the pretraining phase is to learn a good feature representation of the images, which can later be fine-tuned to classify unseen classes. Input batches of size 128 are used to improve the efficiency of batch normalization (He et al., 2016), reducing overfitting and improving the smoothness of gradients. Each image is rotated four times for the self-supervision loss (Dhiaf et al., 2023). Stochastic Gradient Descent is used to train the network. The pretraining process is defined in Algorithm 1. Two fully connected classifiers are trained as shown in Figure 2, which use the features extracted by the CNN backbone networks and regularized using the manifold smoothing process.

- The first classifier C1 is trained to predict the class labels of the input images. A standard cross entropy loss for classification is used to train this classifier.

  The loss function is given by Equation (9):

$$L_{C1}(x_i, y_i;\, W_l, \theta) = -lnp((y_i|\tilde{z}_i), W_l) \tag{9}$$

- The second classifier C2 is utilized to provide a self-supervision type learning signal, where the rotation angle of each input image (after being rotated by $0^\circ, 90^\circ, 180^\circ, 270^\circ$), is predicted. This helps improve the learning signal and provides a certain degree of rotation invariance to the model.

  The loss function is given by:

$$L_{C2}(x_i, y_i;\, W_\gamma, \theta) = -lnp(r_i|\tilde{z}_t), W_\gamma) \tag{10}$$

  where W_γ is the fully connected layer with softmax activation representing C_r and r_i is the prediction of the rotation angle.

  The overall loss to be minimized is given by:

$$argmin \sum_{i=1}^{128} \sum_{j=1}^{4} L_{C1}(x_i, y_i;\, W_l, \theta) + L_{C2}(x_i, y_i;\, W_\gamma, \theta) \tag{11}$$

  where L_{C1} $(x_i,y_i;W_l,\theta)$ is defined in Equation (10), L_{C2} $(x_i,y_i;W_l,\theta)$ is defined in Equation (11) and argmin optimizes the arguments to minimize the sum.

```
Input: Episode of input images
z = backbone_network(rotated_input_batch) # z is
  the feature representation;
A = new matrix(size = z.len × z.len) # Manifold
  Smoothing;
for <z_i in z> do
    for z_j in z do
        if i==j then
            A[i][j] = 0;
        else
            A[i][j] =
             exp(-(L2Norm(z_i,z_j))^2/Var(L2Norm(z_i,z_j)));
        end if
    end for
end for
A = laplacian(A);
I = new matrix(size = A.size, type = Identity);
P = matrix_invent(I α × A) Smoothing factor α
  taken as 0.9;
z_smooth = P×z;
lp = label_propagation(z_smooth.support_set,
  z_smooth.query_set, P);
predicted_unseen =
  fully_connected_classifier(lp, lp.labels) #Label
  propagation predicted_all =
  fully_connected_classifier(z_smooth,
  z_smooth.labels) ;
Output: Predicted labels of the input images;
```

Algorithm 2. Finetuning algorithm.

4.5 Finetuning process

The finetuning process is performed after the model has been trained on the training set C_{train}. Here, the objective is learning to recognize the unseen classes (part of the test set C_{test}. The label propagation method is used to find the labels of the unseen classes. Each epoch in finetuning consists of generating an episode calculating the loss obtained and using backpropagation to adjust

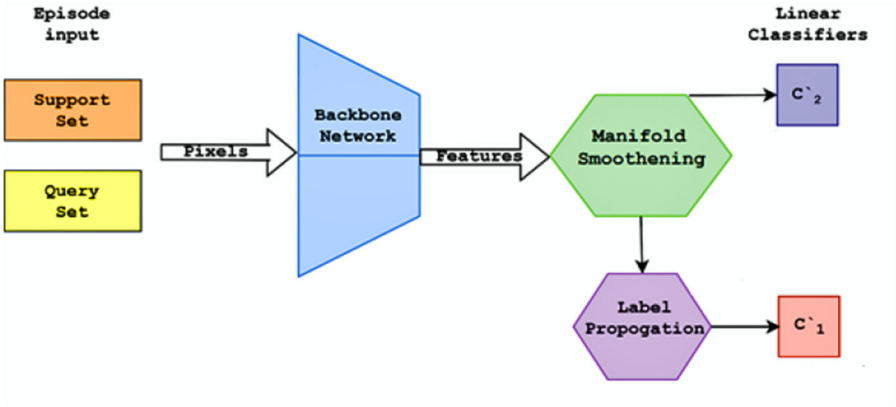


FIGURE 3
Flow diagram of the finetuning process.

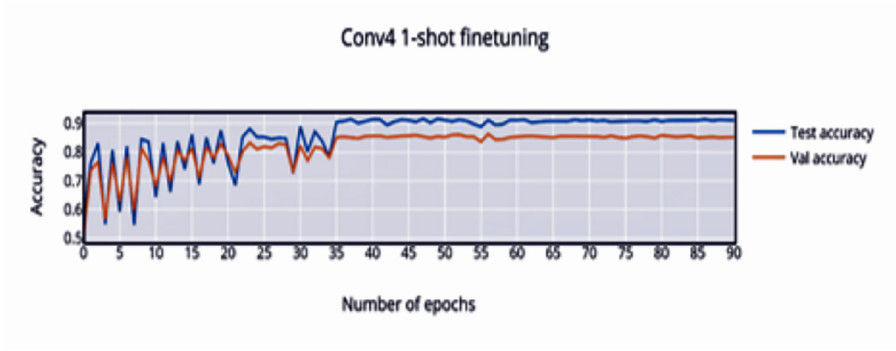


FIGURE 4
1-shot finetuning accuracy vs. number of epochs.

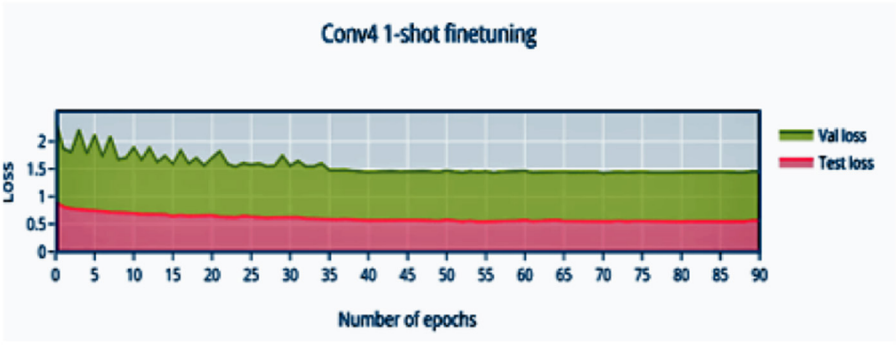


FIGURE 5
1-shot finetuning loss vs. number of epochs.

the weights accordingly. The finetuning process is defined in Algorithm 2, Two linear classifiers are once again used as shown in Figure 3.

1. The classifier C‘$_1$ utilizes label propagation to compute the probabilities of the classes in the query set. The logits are converted to class probabilities using the SoftMax function.

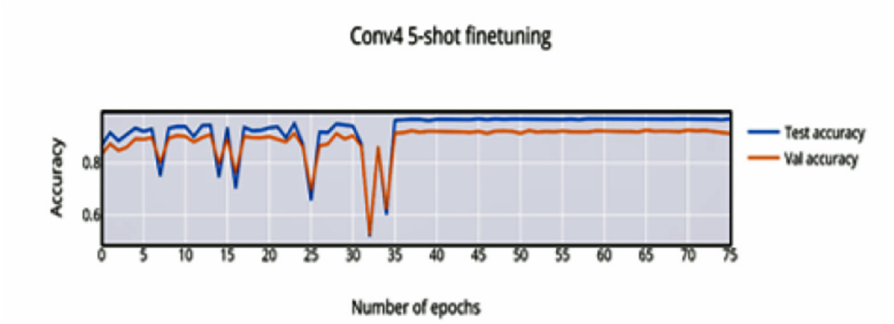


FIGURE 6
5-shot finetuning accuracy vs. number of epochs.

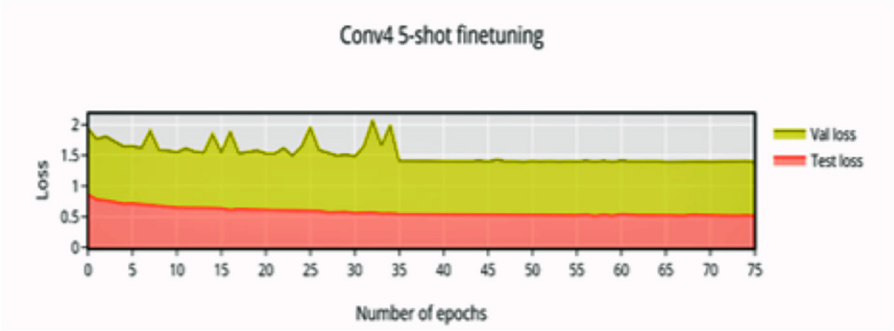


FIGURE 7
5-shot finetuning loss vs. number of epochs.

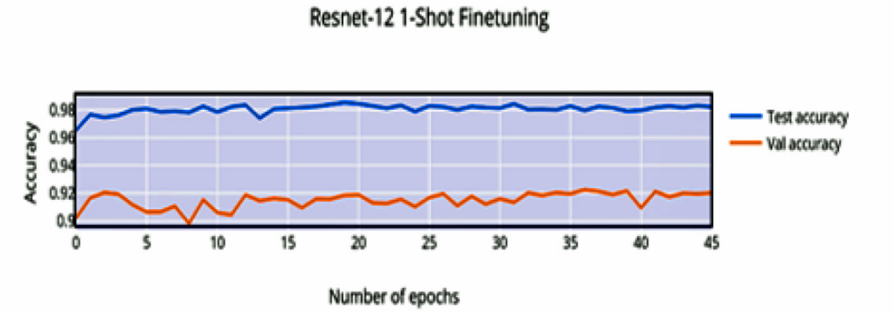


FIGURE 8
1-shot finetuning accuracy vs. number of epochs.

The loss function is given by Equation (12):

$$L_{C'1}(x_i, y_i; \theta) = -lnp(y_i|(\tilde{z}_l), \tilde{Z}, Y_S) \quad (12)$$

where x_i is the input image, y_i is the label of the input image, θ is the CNN feature extractor and -ln $p(y_i|\, z_i, \tilde{Z}, Y_S)$ is the cross-entropy loss defined on predictions using label propagation (Y_S) defined in Section V.

2. Since the label propagation loss tends to favor mixing of features, impacting the discriminativeness of the feature representation, a second classifier C'_2 is trained with the standard cross entropy loss on the union S∪Q. This helps in preserving the discriminativeness of the feature representation.

The loss function is given by

$$L_{C'2}(x_i, y_i;\, W_1, \theta) = -lnp(y_i|\tilde{z}_l, W_1) \quad (13)$$

The overall loss to be minimized is the additive combination of the above:

$$argmin\left[\frac{1}{|Q|}\sum_{(x_i,y_i)\epsilon Q} L_{C'1}(x_i, y_i, \theta) + \frac{1}{|S \cup Q|}\cup \sum_{(x_i,y_i)\epsilon S\cup Q} \cap \frac{1}{2} L_{C'2}(x_i, y_i; \cap W_1, \theta)\right] \quad (14)$$

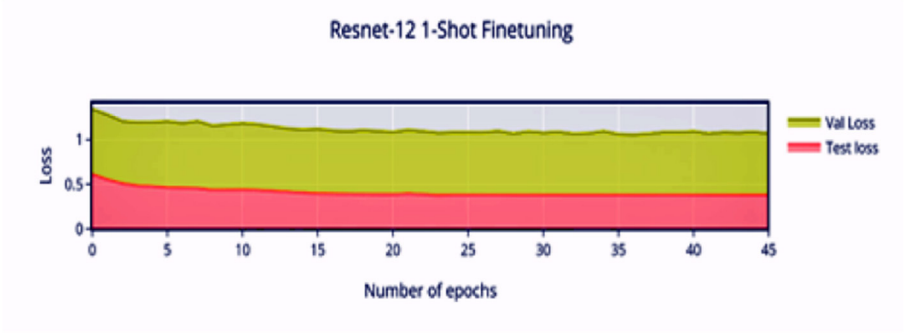


FIGURE 9
1-shot finetuning loss vs. number of epochs.

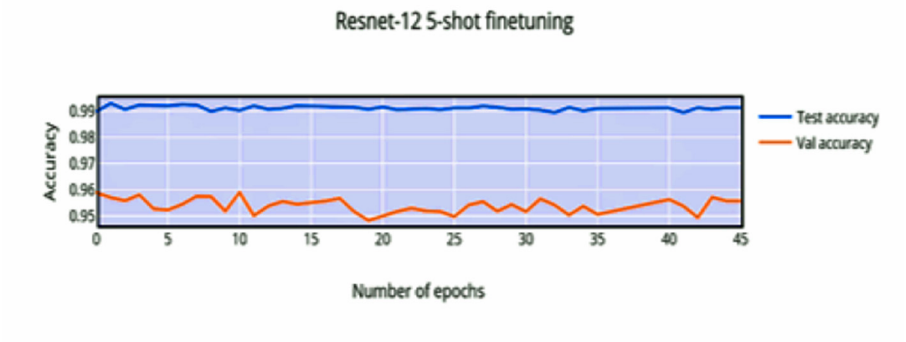


FIGURE 10
5-shot finetuning accuracy vs. number of epochs.

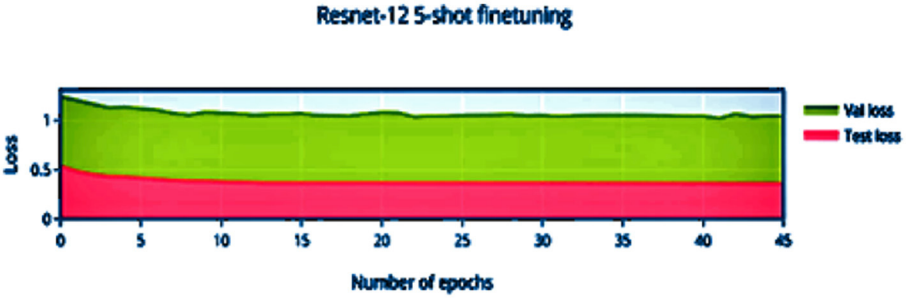


FIGURE 11
5-shot finetuning loss vs. number of epochs.

Where Q is the query set, S is the support set, $L_{C'1}(x_i,y_i,\theta)$ is defined by Equation (13), $L_{C'2}(x_i,y_i;\ W_1,\theta)$ is defined by Equation (14) and argmin optimizes the arguments to minimize the given sum.

5 Implementation

This work uses the dataset used in Karthik and Srikanta Murthy (2018) to evaluate the model. The components of the network are implemented in Python using the Pytorch library. The Episode Generator is used to create episodic tasks for the finetuning of the network. The backbone networks are assigned to the GPUs using the CUDA directive. The model's hyperparameters are listed.

5.1 Simulation dataset

The dataset consists of 47 classes representing each base character of the Kannada abugida. Each class consists of 400 samples obtained from different writers. The images are rescaled to

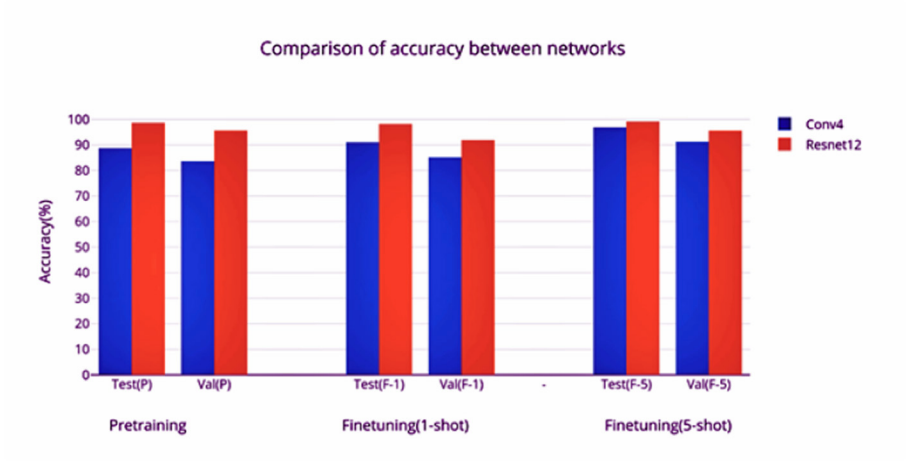


FIGURE 12
Comparison of the accuracies obtained by the networks.

84×84 px using the PIL library. For the purpose of the experiments, the 47 classes are randomly split into three sets following the example of He et al. (2016). The base set C_{train} consists of 24 classes and has all 400 samples for the supervised pretraining phase. Thus 50% of the dataset is used for the supervised training part. A mixture of vowels and consonants are present in C_{train}. Characters with shapes both simple and complex are represented in the training set.

The novel set C_{test} consists of 12 classes which form the unseen set of classes used to test the finetuning approach. This is 25% of the dataset. It is observed that characters both similar in shape to the ones found in C_{train}, as well as uniquely shaped characters can be found in C_{test}. A validation set C_{val} consisting of 11 classes is used to form the validation set used for hyperparameter search and to measure the amount of overfitting. Twenty-five percent of the dataset is used for this purpose.

6 Results and analysis

State of the art results is achieved using the Label Propagation and Manifold Smoothing model for the problem of Recognition of Handwritten Kannada Characters in a Few-Shot Learning perspective. This section gives insights of the result obtained in terms of Pretraining Accuracy (seen classes), Finetuning accuracy (seen and unseen classes) using 1-shot and 5-shot learning (support set of one and five examples, respectively). Comparison of result with the existing work is done here.

6.1 Performance evaluation

Two different feature extractors are evaluated using the episodic framework, and the average accuracy of classification over 1,000 episodes is used as the metric for evaluation. The first feature extractor, Conv4, has a faster training and inference time owing to its simplicity, and seems to benefit much more from the finetuning phase as compared to the second feature extractor, Resnet-12. However, much better accuracy is obtained by the larger Resnet-12 network. This can be attributed to the greater width of the network, which allows a larger number of learnable parameters to be used for classification. Although there is a greater amount of overfitting as evidenced by the difference in test and validation accuracies, the performance on finetuning shows that the framework has good generalization capability.

6.2 Conv4 network

The convergence of training at 44 epochs is observed, and due to the episodic nature of training, large swings are seen prior to convergence. The loss is monotonically decreasing over a large number of epochs, with a bump close to the convergence point.

In Figure 4 it is observed that the pretrained model starts out at 50% accuracy and steadily increases with finetuning epochs until epoch 32 where the network converges to 91.04% accuracy. The loss (Figure 5) decreases and stabilizes.

In 5-shot finetuning, a higher initial accuracy of 83% accuracy (Figure 6) is observed which reduces when more unseen classes are initially encountered, the network finally converges at 37 epochs to an accuracy of 96.88%. There is an increase in validation loss (Figure 7) corresponding to the more difficult episodes.

6.3 ResNet-12 network

The shorter convergence time (35 epochs) is seen and a higher pretraining accuracy being achieved (98.66%). This can be attributed to the increased number of channels (width) and layers (depth) of the backbone network.

Compared to Figure 8, the finetuning does not increase the accuracy of the network by a significant amount. This can be attributed to the stronger convergence during training, which allows better inference on novel classes without much finetuning required. The low variance of the accuracy and loss in Figures 9, 11 indicates saturation of the network. Similar to Figures 10, 11, it can be observed that finetuning doesn't increase the accuracy significantly. Due to the large number of support images (5 compared to 1 in 1-shot), we obtain a higher accuracy 99.13% compared to 98.17% in Figure 11.

6.4 Comparison between the networks

The Resnet model converges faster in pretraining compared to the Conv4 model. The training is stopped when the learning rate reaches 0.00001. The learning rate is reduced to 10% after every 10 epochs if there is no improvement in the loss (a plateau is reached). A Conv4 model requires a larger number of epochs to converge during the finetuning phase as well-compared to the Resnet model (Figures 4, 8). It can be observed that there is a significant increase of test and validation accuracy during finetuning for the Conv4 model (Figures 4, 10), while finetuning doesn't increase the accuracy of the Resnet-12 model by a significant amount (Figures 9, 11). The increase in the number of support set samples from 1 to 5 provides a boost of 5% accuracy for the Conv4 model and 4% for the Resnet-12 model (comparing the validation accuracies). It can be inferred that increasing the number of labeled examples for the unseen classes can be expected to provide about a 4% increase in accuracy. The gain per increase of labeled examples should diminish as it converges to supervised learning. The comparison between the networks based on the different accuracies obtained is shown in Figure 12.

6.5 Comparison with previous works

The test accuracy and validation accuracy of the 5-shot approach are compared with the values obtained by training the Convolutional Neural Network and Capsule Network as provided in Ramesh et al. (2019a), as mentioned. It can be observed that the number of epochs required for convergence is similar for all three networks. The amount of overfitting in the Label Propagation network is lower as indicated by the 3% difference between the training and validation accuracies, as mentioned in Table 3. Compared to the 7% difference in the capsule network and 12% difference in the CNN used in Vinotheni and Lakshmana Pandian (2023).

TABLE 3 Comparison of accuracy with existing work.

References	Method	Accuracy obtained
Karthik and Srikanta Murthy (2018)	Deep belief network	97.04%
Rasheed et al. (2022)	AlexFT	97.08%
Vinotheni and Lakshmana Pandian (2023)	ETEDL-THDR	98.48%
Proposed method (5 shot)	Manifold smoothing with label propagation	99.13%

7 Conclusion

A novel offline handwritten character recognition framework is proposed that has the qualities of robustness to variations in input and easy generalization. The incorporation of unseen character classes into the framework doesn't require the retraining of the entire network to achieve good accuracy. The incorporation is also data efficient as it only requires a small number of labeled samples to learn to classify the newer classes (only 1 example in 1-shot and five examples in 5-shot). The use of Resnet-12 (a deep residual CNN), label propagation, and manifold smoothing helps reduce the effect of training class imbalance bias as well as reduce the overfitting of the network during the pretraining phase. The accuracy as obtained at 99.13% on the 5-shot accuracy makes this framework competitive with its supervised learning counterparts, despite the large reduction in the number of labeled samples available (for the novel classes). The framework can be further enhanced by improving the matrix inversion complexity by introducing block-sparse and sparse inversion techniques, which allow for scalability. The incorporation of the label propagation algorithm into an LSTM and language model system will help in creating few-shot learning-based word, sentence, and document optical character recognition systems.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

JS: Supervision, Writing - review & editing. GR: Conceptualization, Formal analysis, Investigation, Methodology, Writing - original draft. JB: Data curation, Formal analysis, Investigation, Writing - original draft. GS: Investigation, Software, Writing - original draft. HG: Project administration, Resources, Supervision, Validation, Writing - review & editing. NS: Formal analysis, Methodology, Writing - review & editing. SA: Conceptualization, Funding acquisition, Investigation, Writing - review & editing. MA: Funding acquisition, Visualization, Writing - review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This project was funded by King Saud University, Riyadh, Saudi Arabia. Researchers Supporting Project number (RSP2024R167), King Saud University, Riyadh, Saudi Arabia.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2024.1362567/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Diagram of the conv-4 Model.

SUPPLEMENTARY FIGURE 2
Diagram of the resent-12 Model.

SUPPLEMENTARY FIGURE 3
Gaussian similarity function.

SUPPLEMENTARY FIGURE 4
Graph clustering from power iteration.

SUPPLEMENTARY FIGURE 5
Pretraining accuracy vs. number of epochs resnet 12.

SUPPLEMENTARY FIGURE 6
Pretraining accuracy vs. number of epochs.

SUPPLEMENTARY FIGURE 7
Pretraining loss vs. number of epochs resnet 12.

SUPPLEMENTARY FIGURE 8
Pretraining loss vs. number of epochs.

SUPPLEMENTARY FIGURE 9
Samples from C train.

SUPPLEMENTARY FIGURE 10
Samples from ctest.

SUPPLEMENTARY FIGURE 11
Samples from cval.

References

Alsuhibany, S. A., and Alnoshan, A. A. (2021). Interactive handwritten and text-based handwritten arabic CAPTCHA schemes for mobile devices: a comparative study. *IEEE Access* 9, 140991–141001. doi: 10.1109/ACCESS.2021.3119571

Aradhya, V. M., Niranjan, S., and Hemantha Kumar, G. (2010). Probabilistic neural-network based approach for handwritten character recognition. *Int. J. Comput. Commun. Technol.* 1, 9–13. doi: 10.47893/IJCCT.2010.1029

Cilia, N. D., De Stefano, C., Fontanella, F., and Scotto di Freca, A. (2018). A ranking-based feature selection approach for handwritten character recognition. *Pat. Recogn. Lett.* 121, 77–86. doi: 10.1016/j.patrec.2018.04.007

Dhiaf, M. Souibgui, M. A., Wang, K., Liu, Y., Kessentini, Y., Fornés, A., et al. (2023). KCSSL-MHTR: continual self-supervised learning for scalable multi-script handwritten text recognition. *arXiv preprint arXiv:2303.09347*. doi: 10.48550/arXiv.2303.09347

Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., and Cord, M. (2019). "Boosting few-shot visual learning with self-supervision," in *Conference on Computer Vision and Pattern Recognition*, 8059–8068.

Gowda, D. K., and Kanchana, V. (2022). "Kannada handwritten character recognition and classification through OCR using hybrid machine learning techniques," in *2022 IEEE International Conference on Data Science and Information System (ICDSIS)* (Hassan: IEEE), 1–6.

Guha, R., Das, N., Kundu, M., Nasipuri, M., Santosh, K. C., and IEEE Senior Member. (2019). DevNet: an efficient CNN architecture for handwritten Devanagari character recognition. *Int. J. Pat. Recogn. Artif. Intell.* 34:20520096. doi: 10.1142/s0218001420520096

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *Comput. Vis. Pat. Recogn.* 2016, 770–778. doi: 10.48550/arXiv.1512.03385

Karthik, S., and Srikanta Murthy, K. (2018). Deep belief network based approach to recognize handwritten Kannada characters using distributed average of gradients. *Clust. Comput.* 22, 4673–4681. doi: 10.1007/s10586-018-2274-0

Katiyar, G., and Mehfuz, S. (2016). A hybrid recognition system for off-line handwritten characters. *SpringerPlus* 5:7. doi: 10.1186/s40064-016-1775-7

Khan, F. A., Khelifi, F., Tahir, M. A., and Bouridane, A. (2019). Dissimilarity Gaussian mixture models for efficient offline handwritten text-independent identification using SIFT and RootSIFT descriptors. *IEEE Trans. Inform. Forensics Secur.* 14, 289–303. doi: 10.1109/TIFS.2018.2850011

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1995). Lower bounds on the VC dimension of smoothly parameterized function classes. *Neural Comput.* 7, 1040–1053.

Mhiri, M., Desrosiers, C., and Cheriet, M. (2018). Convolutional pyramid of bidirectional character sequences for the recognition of handwritten words. *Pat. Recogn. Lett.* 111, 87–93. doi: 10.1016/j.patrec.2018.04.025

Nasir, T., Malik, M. K., and Shahzad, K. (2021). MMU-OCR-21: towards end-to-end urdu text recognition using deep learning. *IEEE Access* 9, 124945–124962. doi: 10.1109/ACCESS.2021.3110787

Nichol, A., Achiam, J., and Schulman, J. (2018). On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*. doi: 10.48550/arXiv.1803.02999

Parikshith, H., Rajath, S. N., Shwetha, D., Sindhu, C. M., and Ravi, P. (2021). "Handwritten character recognition of Kannada Language using convolutional neural networks and transfer learning," in *IOP Conference Series: Materials Science and Engineering, Vol. 1110. No. 1*. Bristol: IOP Publishing.

Ramesh, G., Manoj Balaji, J., Sharma, G. N., and Champa, H. N. (2019a). Recognition of off-line Kannada handwritten characters by deep learning using capsule network. *Int. J. Eng. Adv. Technol.* 8:88619. doi: 10.35940/ijeat.F8726.088619

Ramesh, G., Sandeep Kumar, N., and Champa, H. N. (2020). "Recognition of Kannada handwritten words using SVM classifier with convolutional neural network," in *2022 IEEE 2020 IEEE Region 10 Symposium (TENSYMP)* (Dhaka: IEEE), 1114–1117.

Ramesh, G., Sharma, G. N., Manoj Balaji, J., and Champa, H. N. (2019b). "Offline Kannada handwritten character recognition using convolutional neural networks," in *2019 IEEE International WIE Conference on Electrical and Computer E Engineering* (Bangalore: IEEE), 1-5.

Rasheed, A., Ali, N., Zafar, B., Shabbir, A., Sajid, M., Mahmood, M. T., et al. (2022). Handwritten Urdu characters and digits recognition using transfer learning and augmentation with AlexNet. IEEE Access. 10, 102629–102645. doi: 10.1109/ACCESS.2022.3208959

Sahare, P., and Dhok, S. B. (2018). Multilingual character segmentation and recognition schemes for Indian Document Images. *IEEE Access*, 6, 10603–10617. doi: 10.1109/access.2018.2795104

Sahlol, A. T., Elaziz, M. A., Al-Qaness, M. A. A., and Kim, S. (2020). Handwritten Arabic optical character recognition approach based on hybrid whale optimization algorithm with neighborhood rough set. *IEEE Access* 8, 23011–23021. doi: 10.1109/ACCESS.2020.2970438

Singh, H., Sharma, R. K., and Singh, V. P. (2020). Online handwriting recognition systems for Indic and non-Indic scripts: a review. *Artif. Intell. Rev.* 20:7. doi: 10.1007/s10462-020-09886-7

Sueiras, J., Ruiz, V., Sanchez, A., and Velez, J. F. (2018). Offline continuous handwriting recognition using sequence to sequence neural networks. *Neurocomputing* 289, 119–128. doi: 10.1016/j.neucom.2018.02.008

Thippeswamy, G., and Chandrakala, H. T. (2020). Recognition of historical handwritten Kannada characters using local binary pattern features. *Int. J. Nat. Comput. Res.* 9, 1–15. doi: 10.4018/ijncr.2020070101

Vinotheni, C., and Lakshmana Pandian, S. (2023). End-to-end deep-learning-based tamil handwritten document recognition and classification model. *IEEE Access* 11, 43195–43204. doi: 10.1109/ACCESS.2023.3270895

Weldegebriel, H. T., Liu, H., Haq, A. U., Bugingo, E., and Zhang, D. (2019). A new hybrid convolutional neural network and extreme gradient boosting classifier for recognizing handwritten Ethiopian characters. *IEEE Access* 1:2960161. doi: 10.1109/access.2019.2960161

Weng, Y., and Xia, C. (2019). A new deep learning-based handwritten character recognition system on mobile computing devices. *Mob. Netw. Appl.* 25, 402–411. doi: 10.1007/s11036-019-01243-5

Zheng, Y., Iwana, B. K., and Uchida, S. (2019). Mining the displacement of max-pooling for text recognition. *Pat. Recogn.* 5:14. doi: 10.1016/j.patcog.2019.05.014

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Scholkopf, B. (2004). Learning with local and global consistency. *Adv. Neural Inform. Process. Syst.* 2004, 321–328. Available online at: https://proceedings.neurips.cc/paper/2003/hash/87682805257e619d49b8e0dfdc14affa-Abstract.html

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 12 April 2024
DOI 10.3389/fnins.2024.1362286

OPEN ACCESS

EDITED BY
Lu Tang,
Xuzhou Medical University, China

REVIEWED BY
Chhavi Dhiman,
Delhi Technological University, India
Qianyu Zhou,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Yukun Ma
✉ yukuner@126.com

RECEIVED 28 December 2023
ACCEPTED 21 March 2024
PUBLISHED 12 April 2024

CITATION
Ma Y, Lyu C, Li L, Wei Y and Xu Y (2024) Algorithm of face anti-spoofing based on pseudo-negative features generation. *Front. Neurosci.* 18:1362286. doi: 10.3389/fnins.2024.1362286

Algorithm of face anti-spoofing based on pseudo-negative features generation

Yukun Ma[1]*, Chengzhen Lyu[2], Liangliang Li[3], Yajun Wei[2] and Yaowen Xu[4]

[1]School of Software, Henan Institute of Science and Technology, Xinxiang, China, [2]School of Information Engineering, Henan Institute of Science and Technology, Xinxiang, China, [3]School of Information and Electronics, Beijing Institute of Technology, Beijing, China, [4]Data and AI Technology Company, China Telecom Corporation Ltd., Beijing, China

Introduction: Despite advancements in face anti-spoofing technology, attackers continue to pose challenges with their evolving deceptive methods. This is primarily due to the increased complexity of their attacks, coupled with a diversity in presentation modes, acquisition devices, and prosthetic materials. Furthermore, the scarcity of negative sample data exacerbates the situation by causing domain shift issues and impeding robust generalization. Hence, there is a pressing need for more effective cross-domain approaches to bolster the model's capability to generalize across different scenarios.

Methods: This method improves the effectiveness of face anti-spoofing systems by analyzing pseudo-negative sample features, expanding the training dataset, and boosting cross-domain generalization. By generating pseudo-negative features with a new algorithm and aligning these features with the use of KL divergence loss, we enrich the negative sample dataset, aiding the training of a more robust feature classifier and broadening the range of attacks that the system can defend against.

Results: Through experiments on four public datasets (MSU-MFSD, OULU-NPU, Replay-Attack, and CASIA-FASD), we assess the model's performance within and across datasets by controlling variables. Our method delivers positive results in multiple experiments, including those conducted on smaller datasets.

Discussion: Through controlled experiments, we demonstrate the effectiveness of our method. Furthermore, our approach consistently yields favorable results in both intra-dataset and cross-dataset evaluations, thereby highlighting its excellent generalization capabilities. The superior performance on small datasets further underscores our method's remarkable ability to handle unseen data beyond the training set.

KEYWORDS

face anti-spoofing, pseudo-negative feature, features generation, feature analysis, cross-domain

1 Introduction

With the continuous development of computer technology, identity authentication based on face information has been widely used. However, most existing face recognition methods are very vulnerable to face prosthesis attacks. Face spoofing attack refers to illegal users attempting to cheat the face authentication system and the face detection system through some prosthesis methods,

such as print attacks, replay attacks, and mask attacks. Face anti-spoofing is developed to detect illegal facial spoofing attacks, thereby improving the security of face authentication systems (Yu et al., 2022).

Though facial recognition technology has been widely used in biometric authentication, it is susceptible to presentation attacks (commonly referred to as "spoofing attacks"), which have attracted much attention in secure scenarios. These attack forms include using synthesized or fake facial images or information to mimic the facial features of legitimate users, thereby bypassing facial recognition systems. Examples of such attacks include printed photos, facial digital images on electronic screens, 3D masks, and other innovative methods. There are special material attacks, where facial models made from special materials attempt to evade traditional facial recognition systems; meanwhile, virtual generation attacks utilize computer graphics and generative adversarial networks (GANs) to produce realistic synthetic faces and bypass facial recognition systems; additionally, lighting manipulation attacks use lighting effects, special lights, or reflective materials to change facial appearance, making it challenging for systems to accurately identify faces. Though various methods have been proposed to defend against these attacks, existing defense methods often lack sufficient generalization ability when confronted with unknown attacks types (de Freitas Pereira et al., 2013). In practical scenarios, training facial anti-spoofing models to predict all types of attacks is a challenging task.

Face anti-spoofing technology, designed to detect and prevent fraud in facial recognition, has significantly advanced in recent years, yielding promising results. However, a major challenge for current methods is their limited ability to generalize to previously unseen or novel attack types. In the real world, it's nearly impossible to anticipate and incorporate all potential attack scenarios into the training phase, which makes maintaining effectiveness difficult.

As technology evolves and face anti-spoofing techniques become more sophisticated, attackers are also adapting their deceptive methods, leading to new and more complex attack forms. The vast and diverse data space associated with prosthetic attacks, involving high-quality masks or other facial replicas, poses a significant challenge for cross-domain face anti-spoofing. This diversity in attack methods, coupled with variations in presentation, acquisition devices, and prosthetic materials, complicates the task of developing robust and generalizable solutions.

In cross-domain scenarios, where data from multiple sources or domains are involved, existing methods often face significant challenges in training and testing across various devices and materials. These introduce distinct characteristics and variations that can greatly impact model performance and reliability. The fundamental issue is the inadequacy of negative sample data when faced with diverse attacks or perturbations. This scarcity prevents models from adequately learning and generalizing to new, unseen domains, leading to domain shift issues during learning. There's an urgent need for more robust and effective approaches to address these issues and enhance cross-domain performance.

The contributions of this paper are numerous and significant. Firstly, we introduce an innovative algorithm capable of generating pseudo-negative features by collecting and analyzing features from existing datasets. Secondly, we employ the Kullback–Leibler (KL) divergence loss function to effectively guide the distribution of the generated virtual features, ensuring their alignment with the desired characteristics and further optimizing the system's accuracy. Finally, our approach has achieved promising results across multiple cross-domain tests, demonstrating robust performance. Overall, our contributions advance the state-of-the-art in face anti-spoofing technology.

2 Related work

At the initial stage, manually annotated features were used to construct face anti-spoofing. Määttä et al. (2011) developed a method based on the analysis of facial textures to determine whether there is a living person or facial imprint in front of the camera. de Freitas Pereira et al. (2014) extracted local binary patterns (LBP) features in three orthogonal planes of spatiotemporal space for face fraud detection. Similarly, most of the histogram-based 2D features can be generalized to their corresponding 3D forms. In recent years, face anti-spoofing based on deep learning has attracted much attention. Compared with traditional hand-crafted features, deep features learned by the neural network have a more robust representation ability, and the accuracy of the trained model is also greatly enhanced. Yang et al. (2014) first applied the Convolutional Neural Network (CNN) to face anti-spoofing by using the AlexNet network model as a feature extractor to extract the features of the original image and using the Support Vector Machine (SVM) for classification. Menotti et al. (2015) employed the hyperparameter search method to find a suitable CNN network structure for face fraud detection. To narrow the search range of hyperparameters, the searched CNN contained at most three convolutional layers. Rehman et al. (2017) trained an 11-layer VGG network and two variant networks in an end-to-end manner for face fraud detection. Nagpal and Dubey (2019) investigated deeper face fraud detection based on ResNet and GoogLeNet. Li et al. (2016) used transfer learning to extract features after fine-tuning the pre-trained VGG face model, which mitigated overfitting in the model. Some researchers replaced the original hand-crafted features with features learned by the network (Cai et al., 2022). Additionally, the optical flow feature provides an effective method for extracting motion information from videos (Simonyan and Zisserman, 2014; Sun et al., 2016, 2019). Yin et al. (2016) found motion cues of face fraud based on optical flow features. Pinto et al. (2015) proposed a feature based on low-level motion features and mid-level visual encoding for face fraud detection. De Marsico et al. (2012) extracted geometrically invariant features around facial feature points to detect cues in video replay. Moreover, some studies used temporal features between consecutive frames for face anti-spoofing (Wang et al., 2022a).

In the early stage, the deep learning-based detection algorithm employed the softmax loss function for face authenticity classifications. Although these methods improved the detection performance on a single database, their generalization ability remained challenging when tested across data sets. Different from the previous binary classification approach, Liu et al. (2018) proposed training networks using auxiliary information. This method combined face depth information and rPPG (remote photoplethysmography) as an auxiliary supervised guidance model to learn essential features, and it achieved a good detection effect. Kim et al. (2019) introduced reflection-based supervision based on depth graph supervision, which further improved the network's detection performance. Moreover, Li et al. (2020) and Yu et al. (2020) proposed new convolution operators and loss functions for live face detection, respectively. To better resist various unknown attacks and improve the generalization ability of

deep models across data sets, researchers also used zero-shot learning (Liu et al., 2019), domain adaptation, and domain generalization to enhance the model's generalization ability (Saha et al., 2020; Wang et al., 2021). To obtain better domain generalization approaches, Jia et al. (2020) proposed an end-to-end single-side domain generalization framework (SSDG) to improve the generalization ability of face anti-spoofing. Furthermore, Dong et al. (2021) proposed an end-to-end open-set face anti-spoofing (OSFA) approach for recognizing unseen attacks. However, the accuracy and generalization ability of classification models are still areas of active research.

In recent years, the application of transformers in the visual domain has led to numerous advancements in addressing domain generalization issues. Specifically, approaches like the Domain-invariant Vision Transformer (DiVT) have effectively leveraged transformers to enhance the generalization capabilities of face anti-spoofing tasks (Liao et al., 2023). Additionally, initializing Vision Transformers (ViT) with pre-trained weights from multimodal models such as CLIP has been shown to improve the generalization of FAS tasks (Srivatsan et al., 2023). Furthermore, adaptive ViT models have been introduced for robust cross-domain face anti-spoofing (Huang et al., 2022). By employing overlapping patches and parameter sharing within the ViT network, these approaches efficiently utilize multiple modalities, resulting in computationally efficient face anti-spoofing solutions (Antil and Dhiman, 2024).

To further enhance domain generalization, unsupervised or self-supervised methods have been employed during model construction and training. One such approach involves stylizing target data to match the source domain style using image translation techniques and then classifying the stylized data using a well-trained source model (Zhou et al., 2022a). Additionally, novel frameworks such as Source-free Domain Adaptation for Face Anti-Spoofing (SDAFAS; Liu et al., 2022a) and a source data-free domain adaptive face anti-spoofing framework (Lv et al., 2021) have been proposed to tackle issues related to source knowledge adaptation and target data exploration in a source-free setting. These frameworks aim to optimize the network in the target domain without relying on labeled source data by treating it as a problem of learning with noisy labels.

Moreover, a new perspective for domain generalization in face anti-spoofing has been introduced that focuses on aligning features at the instance level without requiring domain labels (Zhou et al., 2023). Frameworks like the Unsupervised Domain Generalization for Face Anti-Spoofing (UDGFAS) exploit large amounts of easily accessible unlabeled data to learn generalizable features (Liu et al., 2023), thereby enhancing the performance of FAS in low-data regimes. These approaches explore the relationship between source domains and unseen domains to achieve effective domain generalization.

Additionally, a self-domain adaptation framework has been proposed that leverages unlabeled test domain data during inference time (Wang et al., 2021). Another approach involves encouraging domain separability while aligning the live-to-spoof transition (i.e., the trajectory from live to spoof) to be consistent across all domains (Sun et al., 2023). The Adaptive Mixture of Experts Learning (AMEL) framework (Zhou et al., 2022b) exploits domain-specific information to adaptively establish links among seen source domains and unseen target domains, further improving generalization. A generalizable Face Anti-Spoofing approach based on causal intervention is proposed, aiming to enhance the model's generalization ability in unseen scenarios by identifying and adjusting domain-related confounding factors (Liu et al., 2022b).

Studying the local features of images has also proven beneficial for achieving good domain generalization. For instance, PatchNet reformulates face anti-spoofing as a fine-grained patch-type recognition problem, recognizing combinations of capturing devices and presentation materials based on patches cropped from non-distorted face images (Wang C. Y. et al., 2022). Furthermore, a novel Selective Domain-invariant Feature Alignment Network (SDFANet) has been proposed for cross-domain face anti-spoofing. This network aims to seek common feature representations by fully exploring the generalization capabilities of different regions within images (Zhou et al., 2021).

The current limited cross-domain performance of facial liveness detection methods is due to the incomplete nature of negative sample data under diverse attacks. Based on the above research, considering that the existing feature information is not complete while disregarding the relationship between features, this paper proposes a new face anti-spoofing method based on CNN to generate pseudo-negative feature data of the training sample, and then calculate the feature distribution, and control the generation of the virtual feature distribution by using the KL divergence loss function. Additionally, based on the generated new pseudo data, the proposed method employs a collaborative training algorithm with the original features to improve the generalization performance of face anti-spoofing systems.

3 Proposed method

Face anti-spoofing is a binary classification task (real/fake). Unlike typical coarse-grained binary classification tasks, the liveness detection task exhibits a property that is inconsistent with human visual distance, as illustrated in Figure 1.

Currently, most of the studies on face anti-spoofing systems focus on increasing the type and number of attack samples to enhance the stability and generalization of face anti-spoofing systems. However, due to the unseen data in the training stage, the original method has some limitations in dealing with unknown attack methods.

By analyzing existing face anti-spoofing methods, it is observed that the incompleteness of negative samples is the primary factor limiting the algorithm's cross-domain performance. Therefore, this method aims to research pseudo-negative sample features, expand the training dataset, and improve the cross-domain generalization of face anti-spoofing methods. First, to address the issue of incomplete negative samples, this study generates pseudo-negative features based on the distribution of *bona fide* and attack features. These features complement existing negative class data, enhancing the diversity and completeness of the negative sample dataset. Then, this study uses pseudo-negative features together with existing negative class data to assist in training a feature classifier for real faces, further adjusting the parameters of the feature extractor. The generation of pseudo-negative features leads to more comprehensive negative sample features during training, making the system cover attack data in a broader range of scenarios and thus improving the generalization of the detection method.

In the context of prosthetic attacks, there exists a certain level of feature dispersion across various attack scenarios, suggesting a wider intra-class variation. Due to this, cross-scenario liveness detection poses a certain challenge, and collecting all types of attack data during the training process can be challenging. The differences in intra-class distribution between seen and unseen attack types often lead to domain shift issues. To tackle these challenges, this study employs a

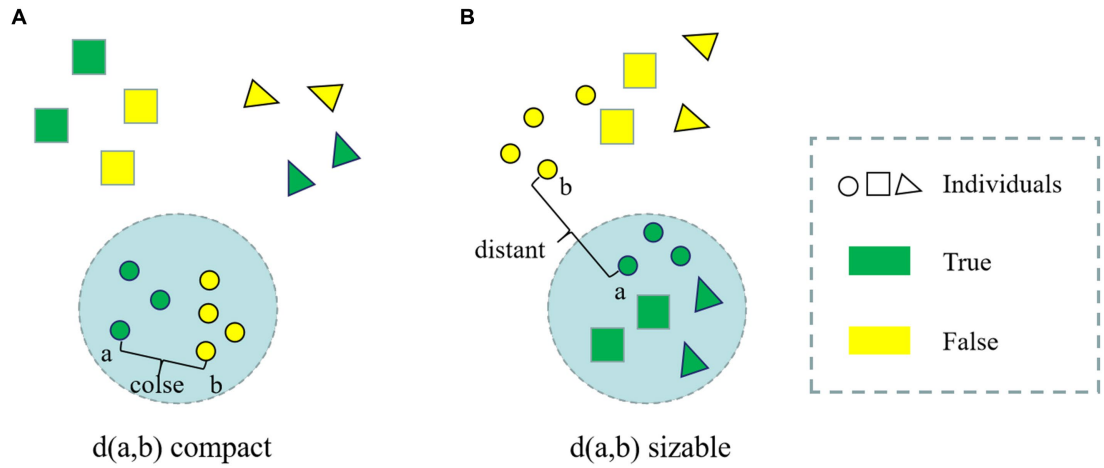


FIGURE 1
(A) True and false samples of different people in human vision; **(B)** True and false samples of different people in the living body detection classifier.

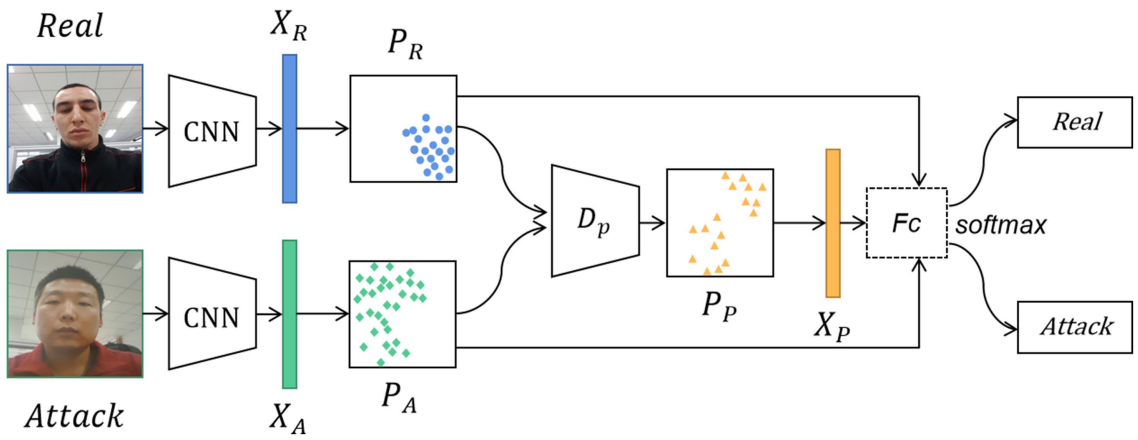


FIGURE 2
The structure diagram of generating pseudo-negative features for face anti-spoofing. The real and attack images are input into the CNN to extract the *bona fide* features and attack features. Then, the distribution of the attack features and the distribution of the *bona fide* features are obtained. These two feature distribution data are fed into the pseudo-negative feature generator to generate the distribution of pseudo-negative features. Finally, the classification task is completed by going through the Fc and the softmax layers. Facial images reproduced with permission from OULU-NPU dataset (Boulkenafet et al., 2017a).

technique for generating pseudo-negative class features, aiming to directly learn the mapping between the visual space of images and the semantic space of features. This method can avoid information loss. Finally, this study develops an end-to-end training model applicable to cross-domain face liveness detection.

The method proposed in this paper comprises of feature analysis, feature generation, and collaborative training. As illustrated in Figure 2, the general workflow of the method is as follows: First, after images are inputted, the CNN generates multi-dimensional feature tensor data from the training samples. Then, the tensor data is analyzed to generate new feature data based on their feature distribution and KL divergence value. Meanwhile, attack types and unseen data from the training stage are incorporated to augment the original set of negative features. Finally, the model is trained using both virtual and existing sample features, allowing us to gather the feature distribution of *bona fide* samples and subsequently improve the accuracy and robustness of live face detection.

During the feature generation process, the corresponding feature distributions are computed by leveraging the extracted features from both attack and *bona fide* images. Then, the distribution data is fed into the data generator D_p, which uses a random data generator based on these distributions to generate a pseudo-negative feature distribution P_P that fits the attack feature distribution. The structure of the data generator D_p is presented in Figure 3.

This section introduces the proposed method from three aspects: feature analysis, feature generation, and loss function.

3.1 Feature analysis

In this paper, we utilize Android and laptop camera devices to acquire face images and subsequently calculate their feature distributions, aiming to analyze the disparities between real and attack face images. As depicted in Figure 4A, it is evident that regardless of

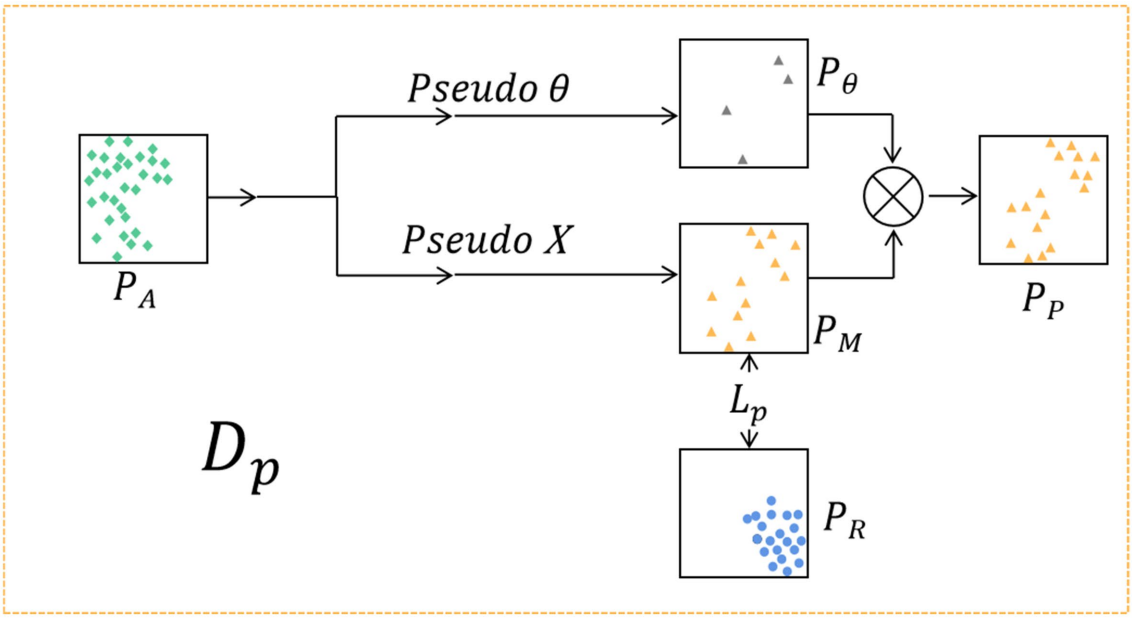


FIGURE 3
The pseudo-negative feature generator D_p. The P_A of attack features and the P_R of *bona fide* features are input into D_p. Firstly, according to P_A, the random generator is used to generate the P_M that fits the distribution of P_A, and the loss function L_p is designed to optimize the distribution P_M of generated pseudo-negative features. To prevent overfitting of the data, a random noise P_θ is generated according to P_A, and the final virtual feature distribution P_P is obtained by combining P_θ with P_M.

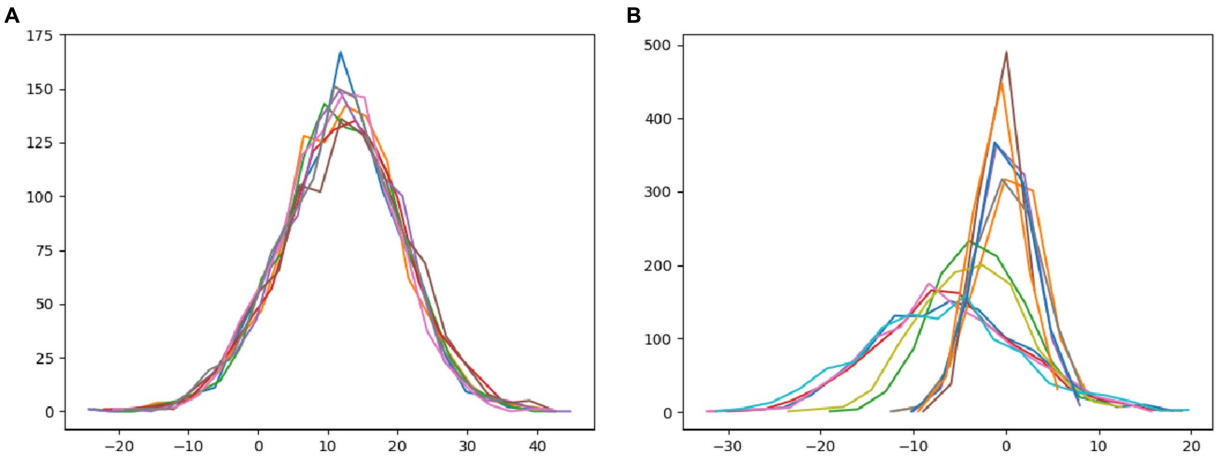


FIGURE 4
The distribution of the feature tensors of the statistical images. **(A)** The statistical tensor distribution of *bona fide* images of different types, and **(B)** the tensor of all types of attack images in the statistical dataset.

the capturing device used, the features of *bona fide* face images conform to a normal distribution, resulting in a relatively clustered pattern. Figure 4B illustrates the image features of attack faces across three distinct display media: three variations of iPad replay video attacks, iPhone replay video attacks, and photo print attacks. Notably, the feature distribution of attack face images employing different display media appears scattered, highlighting the variations in feature distribution among diverse attack methodologies.

In light of the characteristics of normal distribution, we aim to generate pseudo-negative feature data from the original sample feature data in order to enhance network performance. Toward this objective, our paper proposes a methodological framework. Initially, we examine the extracted feature data from the training samples obtained via Convolutional Neural Networks (CNNs). Subsequently, we synthesize pseudo-negative feature data that closely resembles the original sample feature data, ensuring alignment with the inherent distributional properties. Finally, we incorporate this pseudo-negative feature data into the classifier training process, with the ultimate goal of bolstering the accuracy and generalization capabilities of the face anti-spoofing system.

In face anti-spoofing systems, *bona fide* sample data are typically acquired through equipment-based face data collection. Conversely,

attack samples, encompassing image-based and video replay assaults, primarily initiate with frontal face information gathering followed by secondary imaging involving facial prostheses via shooting equipment. Notably, while the *bona fide* sample collection method remains consistent across various data sets, attack samples may exhibit a more scattered distribution due to disparities in devices and attack methodologies (Jia et al., 2020). This difference makes the real face features of different data sets more likely to gather than the attack face features. In the practical application of the face anti-spoofing system, the classification boundary trained based on existing datasets may lead to overlapping characteristics between *bona fide* and novel attack sample data in certain domains, thereby impeding accurate classification. As illustrated in Figure 5A, the classification boundary delineates the feature space into *bona fide* and attack regions. To enhance system performance and ensure robust responsiveness to emerging attacks encountered in real-world scenarios, this study introduces the generation of pseudo-negative feature data (depicted in Figure 5B). This approach serves to augment the feature representation of samples, facilitating the clustering of *bona fide* data and optimizing classification outcomes. Consequently, the accuracy and generalization capabilities of face anti-spoofing systems are substantially improved.

3.2 Feature generation

In terms of current technology, the collection method for real face data across various datasets is relatively straightforward, as the equipment gathers facial data information directly. Consequently, the feature information of attack face samples tends to be more scattered compared to *bona fide* faces. Additionally, in practical applications, numerous unseen novel attack methods will arise. Therefore, the feature generation module performs feature generation and completes the new attack features in the unknown domain.

According to the analysis presented in section 3.1, the proposed image features follow a normal distribution, and the mean value and standard deviation can be calculated. In this study, a feature sequence that matches the mean and standard deviation of the original feature is randomly generated. Assuming P_R is the distribution of the *bona fide* sample data, P_A is the distribution of the attack sample data, and P_P is the distribution of the generated features. To make the model achieve better performance, relative entropy, also known as Kullback–Leibler divergence, is used as the loss function of the feature-generating module. In the initialization process, $P_P = P_A$, i.e., the generated features and the attack sample features remain in the same distribution. At this time, the $D_{KL}(P_P \| P_R)$ has the minimum value, and the classification problem is relatively simple. In the optimization process, the distribution of pseudo-negative features approaches the *bona fide* sample gradually, which increases the multiformity of the attack sample, promotes the gathering of *bona fide* features, improves the classification accuracy of the face anti-spoofing system, and enhances the generalization of invisible new attacks. The loss function of feature generation is shown in the following Equation (1).

$$L_{Pseudo} = \frac{D_{KL}(P_P \| P_R)}{D_{KL}(P_P \| P_A) + D_{KL}(P_P \| P_R)} \tag{1}$$

As shown in Equation (2), where X_A represents the tensor data of the attack sample extracted by the feature extractor, X_i randomly generates the data according to the mean and variance of the attack and the *bona fide* sample tensor, and X_θ represents the random noise generated according to the D_p.

$$D_p = \frac{1}{N}\sum_{i=1}^{N} \min_i \| X_A - X_i \|^2 + X_\theta \tag{2}$$

3.3 Loss function

After generating the pseudo-negative feature data, it should be integrated into the face anti-spoofing system to enhance its performance. The cross-entropy loss function can be employed in neural networks as a metric to assess the similarity between the

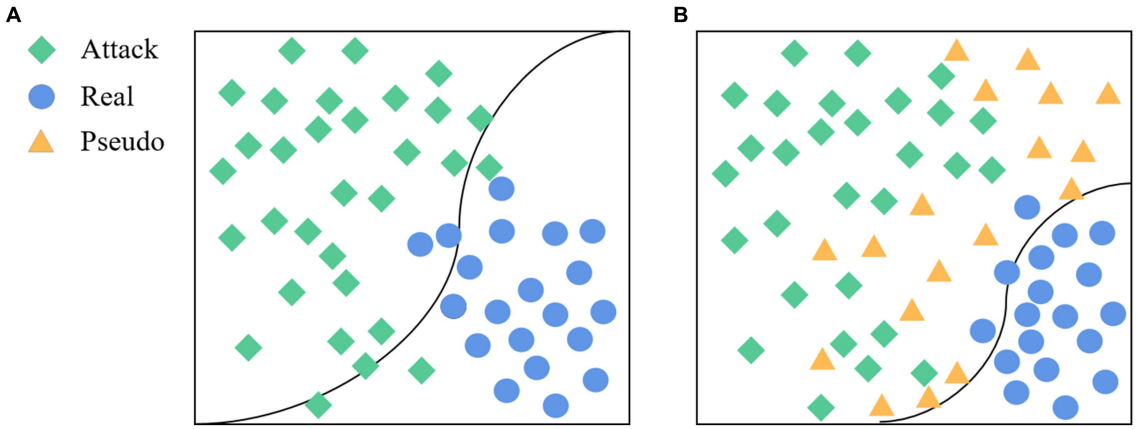


FIGURE 5
The goal of the proposed method. **(A)** The classification boundary without adding pseudo-negative features, and **(B)** the classification boundary after adding pseudo-negative features.

distribution of *bona fide* markers and the distribution predicted by the trained model. In this study, both the original feature data and the generated pseudo-negative feature data are concurrently fed into the loss function, aiming to enhance the generalizability and stability of the face anti-spoofing system in real-world applications. The overall network loss is defined as Equation (3):

$$L_{Whole} = \vartheta_1 L_{ce} + \vartheta_2 L_{Pseudo} \tag{3}$$

where L_{Whole} represents the overall loss function of the network, L_{ce} represents the loss function of the original features, ϑ_1 denotes the weight parameter of the original features, L_{Pseudo} is the loss function of the newly generated features, and ϑ_2 denotes the weight parameter of the newly generated features. The visual representation of the roles played by L_{ce} and L_{Pseudo} in the processes of feature generation and classifier boundary training is depicted in Figure 6.

4 Experimental setup

4.1 Databases

To evaluate the effectiveness of the proposed algorithm, it was tested on three publicly available face datasets, including MSU-MFSD (Wen et al., 2015), OULU-NPU (Boulkenafet et al., 2017a), and Replay-Attack (Chingovska et al., 2012).

The MSU-MFSD dataset (shown in Figure 7) was released by Michigan State University in 2015. Currently, it consists of 280 videos, publicly available and featuring 35 individuals. The dataset consists of three attack types: iPad air video replay attack, iphone5S video replay attack, and A3 paper printed photo attack.

The OULU-NPU dataset (shown in Figure 8) was released by the University of Oulu in Finland in 2017. It consists of 4,950 video clips, captured from 55 participants with 90 videos collected per participant. The dataset consists of four types of attacks: photo attacks printed by two different printers, and video replay attacks displayed by two different display devices.

The Replay-Attack dataset (shown in Figure 9) was released in 2017 and is comprised of 1,200 video clips. These videos feature 50 clients and showcase attack attempts under varying lighting conditions.

Since the dataset comprises entirely of video files, all videos and images were extracted frame-by-frame, and all images have undergone normalization. In these datasets, there are more attack samples than *bona fide* samples, with a large difference in number. During the training process, the quantity of attack and *bona fide* samples was carefully balanced to maintain a similar range, aiming to minimize both data quantity and the chance of overfitting. During data set division, owing to the varied nature of attack samples, the quantity of data samples gathered within identical environmental conditions was two to four times higher compared to *bona fide* samples. Therefore, the attack sample takes the image by the proportion of the *bona fide* sample. In contrast, the attack sample is often intercepted to maintain the amount of the two data in a similar range.

4.2 Experimental metrics

In face anti-spoofing, there are four types of prediction results: True Positives (*TP*), where positive samples are predicted by the model as positive classes; True Negatives (*TN*), where negative samples are predicted by the model as negative classes; False Positives (*FP*), where negative samples are predicted by the model as positive classes; False

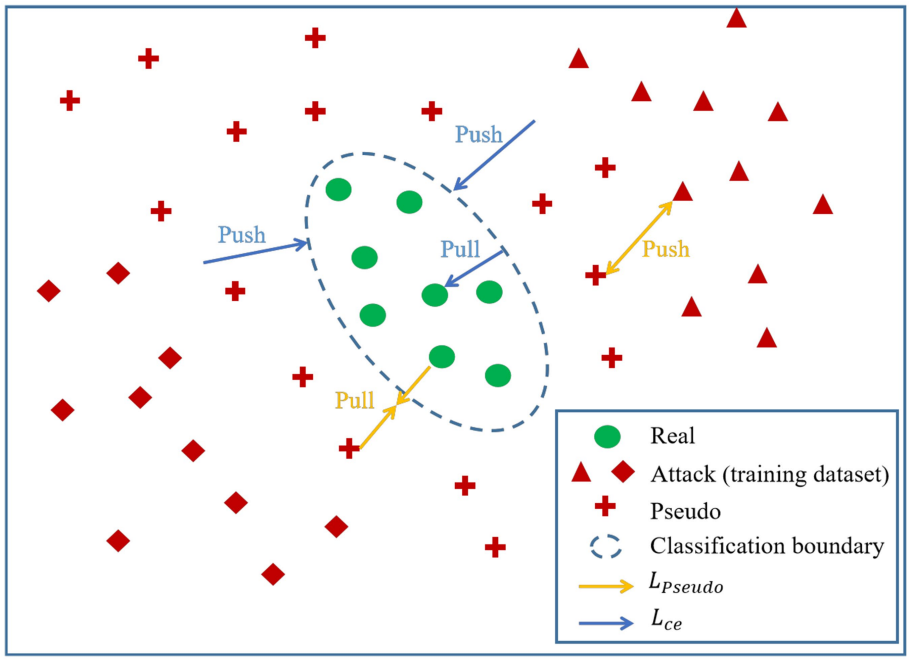


FIGURE 6
The visual representation of the roles played by L_{ce} and L_{Pseudo} in the processes of feature generation and classifier boundary training.

FIGURE 7
Some samples of the subjects recorded in the MSU-MFSD dataset. Images reproduced with permission from MSU-MFSD dataset (Wen et al., 2015).

FIGURE 8
Some samples of the subjects recorded in the OULU-NPU dataset. Images reproduced with permission from OULU-NPU dataset (Boulkenafet et al., 2017a).

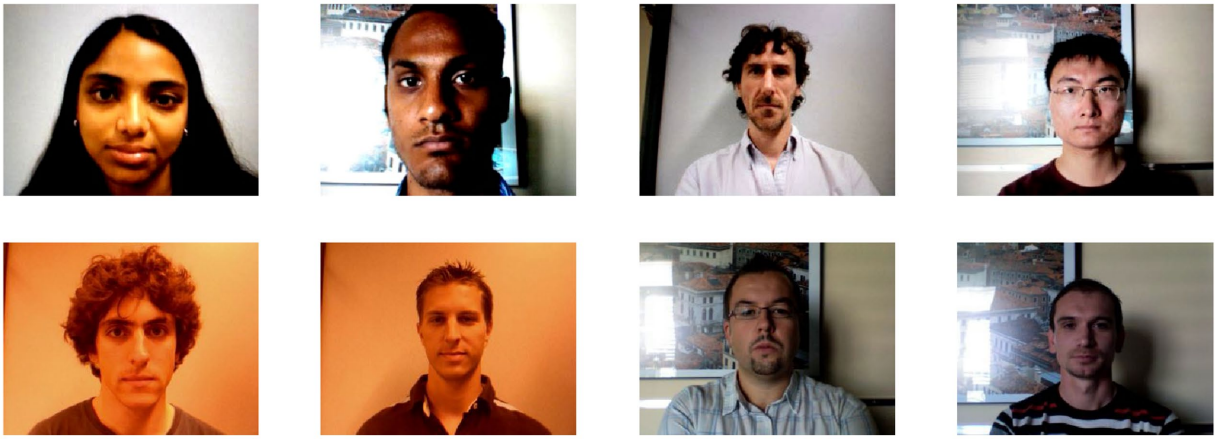


FIGURE 9
Some samples of the subjects recorded in the Replay-Attack dataset. Images reproduced with permission from Replay-Attack dataset (Chingovska et al., 2012).

Negatives (*FN*), where positive samples are predicted by the model as negative classes.

Performance evaluation indicators include Attack Presentation Classification Error Rate (*APCER*), *Bona Fide* Presentation Classification Error Rate (*BPCER*), Average Classification Error Rate (*ACER*), Half Total Error Rate (*HTER*), and Area Under the ROC Curve (*AUC*). These performance indicators are calculated as follows Equations (4–7):

TABLE 1 The performance on the OULU-NPU dataset.

Protocol	Model	APCER (%)	BPCER (%)	ACER (%)
I	AlexNet	0.94	79.90	40.42
	AlexNet+our	0.01	63.19	31.60
II	AlexNet	14.46	6.78	10.62
	AlexNet+our	5.06	10.46	7.76
III	AlexNet	3.40 ± 2.98	11.56 ± 7.58	7.17 ± 3.72
	AlexNet+our	2.33 ± 2.33	9.75 ± 5.25	6.04 ± 1.45
IV	AlexNet	9.07 ± 9.07	58.87 ± 33.87	32.84 ± 16.00
	AlexNet+our	3.53 ± 3.53	55.88 ± 25.88	29.71 ± 11.17

$$APCER = \frac{FP}{TN + FP} \tag{4}$$

$$BPCER = \frac{FN}{TP + FN} \tag{5}$$

$$ACER = \frac{APCER + BPCER}{2.0} \tag{6}$$

$$HTER = \frac{FAR + FRR}{2.0} \tag{7}$$

where FAR represents the false acceptance rate, and it is calculated as $FAR = FP / (FP + TN)$, and FRR represents the false rejection rate, and it is calculated as $FRR = FN / (FN + TP)$.

4.3 Experimental environment

The experiment was conducted on a computer equipped with an AMD Ryzen 75,800× 8-Core CPU, 32 GB memory, and Nvidia GTX 3060 GPU (12 GB video memory), and the computer runs the Windows 10 operating system. The proposed algorithm was implemented based on the PyTorch framework. The Adam optimizer was adopted for model optimization with a learning rate of 2.00e-4 and a batch size of 32.

5 Experimental results

5.1 Control experiment

In this paper, as a control group, the deep learning network AlexNet was trained and tested on the OULU-NPU dataset and MSU-MFSD dataset (Krizhevsky et al., 2012). Based on the native AlexNet, a pseudo-negative feature generation module was added, and then the model was trained and tested on two datasets. The performance of the two models on the OULU-NPU and MSU-FASD datasets is presented in Tables 1, 2, respectively. The results in the two tables show that in the model with the pseudo-negative feature generation module, APCER significantly decreased; in most protocols, BPCER reduced correspondingly, and the overall ACER was diminished.

TABLE 2 The performance on the MSU-FASD dataset.

Model	APCER (%)	BPCER (%)	ACER (%)
AlexNet	1.47	5.27	3.37
AlexNet+our	1.39	3.99	2.69

5.2 Experimental discussion

The experiment evaluated the performance of the intra-test and inter-test. Specifically, the training and testing were performed on the same dataset, which can reflect the performance of the algorithm; cross-datasets indicate that the training set and test set are from different data sets, and the test on these datasets can usually reflect the generalization ability of the algorithm.

The experiments first compared the results of fusing different features on two datasets, followed by comparing the results of different fusion methods on two datasets, then compared the proposed method with some popular methods, and finally evaluated performance across databases on two datasets. The experimental results demonstrated the effectiveness of the proposed face detection method in face anti-spoofing.

The following four experiments were set for comparison in Table 3. Since there are four protocols in the OULU-NPU dataset, protocol 2 was selected based on the features of the MSU-MFSD dataset.

Experiment 1: AlexNet networks without the pseudo-negative feature generator were tested with an intra-test on the OULU-NPU and MSU-MFSD datasets.

Experiment 2: AlexNet networks with the pseudo-negative feature generator were tested with an intra-test on the OULU-NPU and MSU-MFSD datasets.

Experiment 3: AlexNet networks without the pseudo-negative feature generator were tested with an inter-test on the OULU-NPU and MSU-MFSD datasets.

Experiment 4: AlexNet networks with the pseudo-negative feature generator were tested with an inter-test on the OULU-NPU and MSU-MFSD datasets.

To evaluate the effectiveness of our method, in Table 4, the OULU-NPU dataset was used to train and test the AlexNet and AlexNet+our (AlexNet network using the pseudo-feature generator), respectively, and the performance evaluation metrics were calculated. The results indicated that the proposed method achieved comparable performance with state-of-the-art methods (LBP + SVM, GRADIANT, and MILHP). We tested our model on the Replay-Attack dataset, as shown in Table 5. Compared with the state-of-the-art methods from the past 3 years (RGB+LBP and multilevel+ELBP), our model achieved superior performance in terms of accuracy and other evaluation metrics.

As shown in Table 3, the APCER of the AlexNet using a pseudo-negative feature generator decreased significantly on both within-set and cross-set tests, and BPCER also decreased, with only a few parts increasing slightly. The comparison results in Table 4 show that on the OULU-NPU dataset, the performance

TABLE 3 Comparison of the experimental results.

Experiment	MSU-MFSD			OULU-NPU		
	APCER (%)	BPCER (%)	ACER (%)	APCER (%)	BPCER (%)	ACER (%)
1	1.47	5.27	3.37	14.46	6.78	10.62
2	1.39	3.99	2.69	5.06	10.46	7.76
3	20.71	65.23	42.97	25.36	45.82	35.59
4	20.07	65.97	43.02	7.29	35.41	21.35

TABLE 4 Comparable performance on the OULU-NPU dataset.

Protocol	Model	APCER (%)	BPCER (%)	ACER (%)
I	LBP+SVM (George and Marcel, 2019)	12.9	51.7	32.3
	GRADIANT (Boulkenafet et al., 2017b)	1.3	12.5	6.9
	MILHP (Lin et al., 2018)	8.3	0.8	4.6
	AlexNet	0.9	79.9	40.4
	AlexNet+our	0.0	63.2	31.6
II	LBP+SVM (George and Marcel, 2019)	30.0	20.3	25.1
	GRADIANT (Boulkenafet et al., 2017b)	3.1	1.9	2.5
	MILHP (Lin et al., 2018)	5.6	5.3	5.4
	AlexNet	14.5	6.8	10.6
	AlexNet+our	5.06	10.46	7.76
III	LBP+SVM (George and Marcel, 2019)	28.5 ± 23.1	23.3 ± 18.0	25.9 ± 11.3
	GRADIANT (Boulkenafet et al., 2017b)	2.6 ± 3.9	5.0 ± 5.3	3.8 ± 2.4
	MILHP (Lin et al., 2018)	1.5 ± 1.2	6.4 ± 6.6	4.0 ± 2.9
	AlexNet	3.4 ± 3.0	11.6 ± 7.6	7.2 ± 3.7
	AlexNet+our	2.3 ± 2.3	9.8 ± 5.3	6.0 ± 1.5
IV	LBP+SVM (George and Marcel, 2019)	41.67 ± 27.03	55 ± 21.21	48.33 ± 6.07
	GRADIANT (Boulkenafet et al., 2017b)	5.0 ± 4.5	15.0 ± 7.1	10.0 ± 5
	MILHP (Lin et al., 2018)	15.8 ± 12.8	8.3 ± 15.7	12.0 ± 6.2
	AlexNet	9.1 ± 9.1	58.9 ± 33.9	32.8 ± 16.0
	AlexNet+our	3.5 ± 3.5	55.9 ± 25.9	29.7 ± 11.2

TABLE 5 Comparable performance on the Replay-Attack dataset.

Model	HTER(%)	EER(%)
RGB+LBP (Antil and Dhiman, 2023)	4.58	9.69
Multilevel+ELBP (Antil and Dhiman, 2022)	0.00	0.00
Dropblock (Wu et al., 2021)	0.29	0.00
Our	0.00	0.00

of AlexNet is not outstanding, and there is a significant performance gap with the mainstream methods. In contrast, the AlexNet using a pseudo-negative feature generator showed good performance in training and testing. The APCER and BPCER were significantly improved compared with those of AlexNet, and they were close to the performance evaluation indicators of mainstream methods.

To test the model's generalization performance, cross-dataset testing was conducted on the MSU-MFSD dataset (referred to as M), OULU-NPU dataset (referred to as O), Replay-Attack dataset (referred to as R), and CASIA-FASD dataset (referred to as C; Zhang et al., 2012). Then, the results were compared with those of other mainstream experiments, as shown in Table 6. To further verify the performance of the model, we reduced the data set used for training. The experimental results are shown in Table 7. From Table 7, it can be observed that, when using a smaller dataset, our method can achieve results close to or even surpass those obtained from training on larger datasets.

TABLE 6 Comparison of the results between our experiment and the state-of-the-art in cross-domain face anti-spoofing detection.

Methods	O&C&R-to-M		O&M&R-to-C		O&C&M-to-R		R&C&M-to-O	
	ACER(%)	AUC(%)	ACER(%)	AUC(%)	ACER(%)	AUC(%)	ACER(%)	AUC(%)
MADDG (Shao et al., 2019)	17.69	88.06	24.50	84.51	22.19	84.99	27.89	80.02
ANRL (Liu et al., 2021b)	10.83	96.75	17.85	89.26	16.03	91.04	15.67	91.90
SSAN (Wang et al., 2022b)	6.67	98.75	10.00	96.67	8.88	96.79	13.72	93.63
Our	7.12	98.06	11.54	99.21	3.88	98.17	8.36	98.78

TABLE 7 Comparative cross-dataset testing results for similar models.

Experiment	Model	Train(videos)	HTER(%)	AUC(%)
M to R	Multilevel+ELBP (Antil and Dhiman, 2022)	280	24.3	-
M to R	Our	280	21.10	92.36
R&M to O	SSDG (Jia et al., 2020)	1,480	36.01	66.88
R&M to O	D^{2}AN (Chen et al., 2021)	1,480	27.70	75.36
R&M to O	DRDG (Liu et al., 2021a)	1,480	33.35	69.14
R&M to O	ANRL (Liu et al., 2021b)	1,480	30.73	74.10
R&M to O	SSAN (Wang et al., 2022b)	1,480	29.44	76.62
M to O	Our	280	26.24	83.77

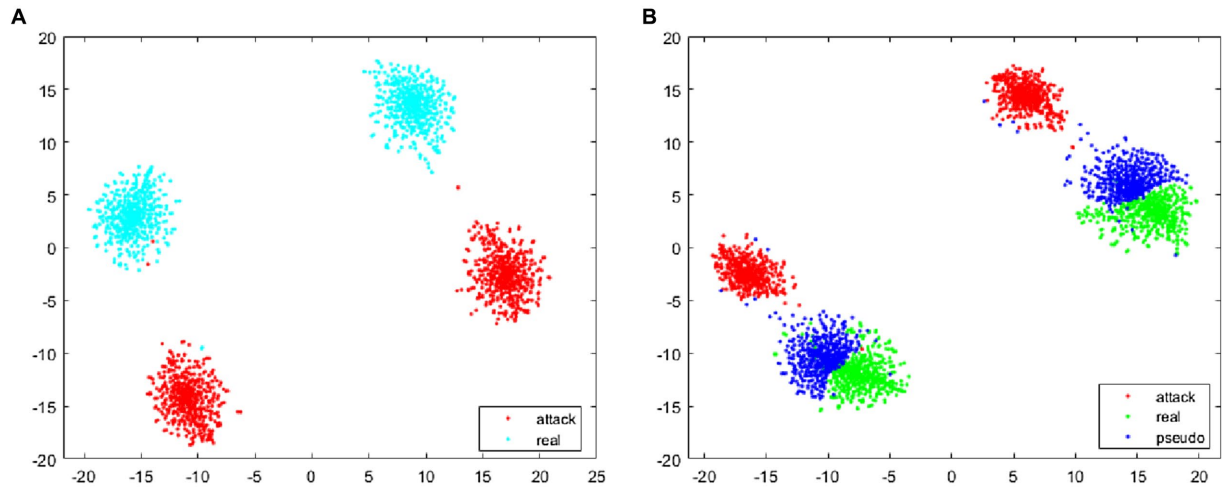


FIGURE 10
The feature cosine distance of images during training. **(A)** The training without using pseudo-negative features, **(B)** the training using pseudo-negative features.

5.3 Feature distribution

The feature visualization algorithm was utilized to extract and compute the features of the training images, whose cosine distance is depicted in Figure 10. Specifically, Figure 10A presents the distance between the attack and the *bona fide* samples in the training phase. It can be seen that there is a large distance between the *bona fide* samples and the attack samples, and there are many blank unknown regions between the two types of samples. Since the face anti-spoofing system in practical applications may encounter some new attack data that did not appear in training, this paper generated false negative samples between the *bona fide* and attack samples. As shown in Figure 10B, the pseudo-negative samples are closer to the *bona fide* samples, indicating that the classification boundary of the face anti-spoofing system, during training, is more biased

toward the *bona fide* samples. In practical applications, the face anti-spoofing system can achieve a good identification effect for new attacks that have not appeared in the dataset.

6 Conclusion

In this paper, a face anti-spoofing algorithm is proposed based on generated pseudo-negative features. Through continuous iteration, the original face anti-spoofing system achieves higher accuracy and robustness. Meanwhile, by adding pseudo-negative features, good results have been obtained in detecting attack samples. It shows that adding pseudo-negative class features enables the model to detect negative samples, and this affects the detection of positive examples in some cases. In this study, by constantly adjusting the strategy, new features are continually generated based on the image's original features. Concurrently, a face anti-spoofing system is devised to counter emerging attacks within the feature space, resulting in the development of more effective strategies. Furthermore, this study promotes aggregation among *bona fide* examples while increasing scatter among attack examples, consequently bolstering the model's robustness in unfamiliar territories. In future work, we will focus on eliminating the influence on positive examples to improve their detection effect.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: Replay-Attack Database: https://www.idiap.ch/dataset/replayattack, MSU-MFSD Database: http://biometricscse.msu.edu/Publications/Databases/MSUMobileFaceSpoofing, and Oulu-NPU Database: https://sites.google.com/site/oulunpudatabase.

Author contributions

YM: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Resources, Supervision, Validation. CL: Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation. LL: Investigation, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. YW: Data curation, Investigation, Software, Supervision, Validation, Writing – review & editing. YX: Formal analysis, Investigation, Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by Training Program for Young Backbone Teachers of Higher Education Institutions in Henan Province China "Research on shadow adversarial sample attack method combining optimal point constraints" (no. 2023GGJS116) and Henan Provincial Scientific and Technological Research Project "Research on Cross-domain Face Anti-spoofing Technology for Diversified Attacks" (242102210128).

Conflict of interest

YX was employed by China Telecom Corporation Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Antil, A., and Dhiman, C. (2022). Two stream RGB-LBP based transfer learning model for face anti-spoofing. In: *International Conference on Computer Vision Image Processing*. pp. 364–374. Cham: Springer Nature Switzerland.

Antil, A., and Dhiman, C. (2023). A two stream face anti-spoofing framework using multi-level deep features and ELBP features. *Multimedia Systems* 29, 1–16. doi: 10.1007/s00530-023-01060-7

Antil, A., and Dhiman, C. (2024). MF2ShrT: multi-modal feature fusion using shared layered transformer for face anti-spoofing. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 1–21. doi: 10.1145/3640817

Boulkenafet, Z., Komulainen, J., Akhtar, Z., Benlamoudi, A., Samai, D., Bekhouche, S. E., et al. (2017b). A competition on generalized software-based face presentation attack detection in mobile scenarios. In: *Proceedings of International Joint Conference on Biometrics*. pp. 688–696.

Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., and Hadid, A. (2017a). OULU-NPU: a mobile face presentation attack database with real-world variations. In: *IEEE International Conference on Automatic Face & Gesture Recognition* pp. 612–618.

Cai, R., Li, Z., Wan, R., Li, H., Hu, Y., and Kot, A. C. (2022). Learning meta pattern for face anti-spoofing. *IEEE Trans. Inf. Forensics Security.* 17, 1201–1213. doi: 10.1109/TIFS.2022.3158551

Chen, Z., Yao, T., Sheng, K., Ding, S., Tai, Y., Li, J., et al. (2021). Generalizable representation learning for mixture domain face anti-spoofing. *AAAI Conf. Artif. Intell.* 35, 1132–1139.

Chingovska, I., Anjos, A., and Marcel, S. (2012). On the effectiveness of local binary patterns in face anti-spoofing. In: *Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*. pp. 1–7.

de Freitas Pereira, T., Anjos, A., De Martino, J. M., and Marcel, S. (2013) Can face anti-spoofing countermeasures work in a real world scenario?. In: *IEEE International Conference on Biometrics* pp. 1–8.

de Freitas Pereira, T., Komulainen, J., Anjos, A., De Martino, J. M., Hadid, A., Pietikäinen, M., et al. (2014). Face liveness detection using dynamic texture. *EURASIP J. Image Video Process.* 2014:2. doi: 10.1186/1687-5281-2014-2

De Marsico, M., Nappi, M., Riccio, D., and Dugelay, J. L. (2012) Moving face spoofing detection via 3D projective invariants. In: *IAPR International Conference on Biometrics (ICB)* pp. 73–78.

Dong, X., Liu, H., Cai, W., Lv, P., and Yu, Z. (2021) Open set face anti-spoofing in unseen attacks. In: *ACM International Conference on Multimedia*. pp. 4082–4090.

George, A., and Marcel, S. (2019). Deep pixel-wise binary supervision for face presentation attack detection. In: *Proceedings of IEEE International Conference on Biometrics*. pp. 1–8.

Huang, H. P., Sun, D., Liu, Y., Chu, W. S., Xiao, T., Yuan, J., et al. (2022). *Adaptive transformers for robust few-shot cross-domain face anti-spoofing*. Cham: Springer Nature Switzerland. Pp. 37–54.

Jia, Y., Zhang, J., Shan, S., and Chen, X. (2020). Single-side domain generalization for face anti-spoofing. In: *IEEE Conference on Computer Vision Pattern Recognition*. pp. 8484–8493.

Kim, T., Kim, Y., Kim, I., and Kim, D. (2019). Basn: enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In: *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 25, 84–90. doi: 10.1145/3065386

Li, L., Feng, X., Boulkenafez, Z., Xia, Z., Li, M., and Hadid, A. (2016). An original face anti-spoofing approach using partial convolutional neural network. In: *Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)* pp. 1–6.

Li, Z., Li, H., Lam, K. Y., and Kot, A. C. (2020). Unseen face presentation attack detection with hypersphere loss. In: *IEEE International Conference on Acoustics Speech Signal Processing*. pp. 2852–2856.

Liao, C. H., Chen, W. C., Liu, H. T., Yeh, Y. R., Hu, M. C., and Chen, C. S. (2023). Domain invariant vision transformer learning for face anti-spoofing. In: *Proceedings of IEEE Winter Conference on Applications of Computer Vision* pp. 6098–6107.

Lin, C., Liao, Z., Zhou, P., Hu, J., and Ni, B. (2018). Live face verification with multiple Instantialized local homographic parameterization. In: *IJCAI International Joint Conference on Artificial Intelligence*. pp. 814–820.

Liu, Y., Chen, Y., Dai, W., Gou, M., Huang, C. T., and Xiong, H. (2022a). *Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing*. Cham: Springer Nature Switzerland. Pp. 511–528

Liu, Y., Chen, Y., Dai, W., Li, C., Zou, J., and Xiong, H. (2022b). Causal intervention for generalizable face anti-spoofing. In: *IEEE International Conference on Multimedia Expo*. pp. 01–06.

Liu, Y., Chen, Y., Gou, M., Huang, C. T., Wang, Y., and Dai, W. (2023). Towards unsupervised domain generalization for face anti-spoofing. In: *Proceedings of IEEE International Conference on Computer Vision*. pp. 20654–20664.

Liu, Y., Jourabloo, A., and Liu, X. (2018). Learning deep models for face anti-spoofing: binary or auxiliary supervision. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* 389–398.

Liu, Y., Stehouwer, J., Jourabloo, A., and Liu, X. (2019). Deep tree learning for zero-shot face anti-spoofing. In: *IEEE Conference on Computer Vision Pattern Recognition*. pp. 4680–4689.

Liu, S., Zhang, K. Y., Yao, T., Bi, M., Ding, S., Li, J., et al. (2021b). Adaptive normalized representation learning for generalizable face anti-spoofing. In: *ACM International Conference on Multimedia*. pp. 1469–1477.

Liu, S., Zhang, K. Y., Yao, T., Sheng, K., Ding, S., Tai, Y., et al. (2021a). Dual reweighting domain generalization for face presentation attack detection. In: *IJCAI International Joint Conference on Artificial Intelligence*.

Lv, L., Xiang, Y., Li, X., Huang, H., Ruan, R., and Xu, X. (2021). Combining dynamic image and prediction ensemble for cross-domain face anti-spoofing. In: *IEEE International Conference on Acoustics Speech and Signal Processing*. pp. 2550–2554.

Määttä, J., Hadid, A., and Pietikäinen, M. (2011). Face spoofing detection from single images using micro-texture analysis. In: *IEEE International Conference on Biometrics* pp. 1–7.

Menotti, D., Chiachia, G., Pinto, A., Schwartz, W. R., Pedrini, H., Falcao, A. X., et al. (2015). Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Trans. Inf. Forensics Secur.* 10, 864–879. doi: 10.1109/TIFS.2015.2398817

Nagpal, C., and Dubey, S. R. (2019). A performance evaluation of convolutional neural networks for face anti spoofing. In: *International Joint Conference on Neural Networks (IJCNN)* pp. 1–8.

Pinto, A., Pedrini, H., Schwartz, W. R., and Rocha, A. (2015). Face spoofing detection through visual codebooks of spectral temporal cubes. *IEEE Trans. Image Process.* 24, 4726–4740. doi: 10.1109/TIP.2015.2466088

Rehman, Y. A. U., Po, L. M., and Liu, M. (2017). Deep learning for face anti-spoofing: an end-to-end approach. In: *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. pp. 195–200.

Saha, S., Xu, W., Kanakis, M., Georgoulis, S., Chen, Y., Paudel, D. P., et al. (2020). Domain agnostic feature learning for image and video based face anti-spoofing. In: *IEEE Conference on Computer Vision Pattern Recognition*. pp. 802–803.

Shao, R., Lan, X., Li, J., and Yuen, P. C. (2019). Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: *IEEE Conference on Computer Vision Pattern Recognition* pp. 10023–10031.

Simonyan, K., and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In: *27th International Conference on Neural Information Processing Systems* 27.

Srivatsan, K., Naseer, M., and Nandakumar, K. (2023). FLIP: cross-domain face anti-spoofing with language guidance. In: *Proceedings of IEEE International Conference on Computer Vision* pp. 19685–19696.

Sun, Y., Liu, Y., Liu, X., Li, Y., and Chu, W. S. (2023). Rethinking domain generalization for face anti-spoofing: separability and alignment. in: *Proceedings of IEEE International Conference on Computer Vision Pattern Recognition*. pp. 24563–24574.

Sun, W., Zhao, H., and Jin, Z. (2016). 3D convolutional neural networks for facial expression classification. In: *Asian Conference on Computer Vision* 528–543.

Sun, W., Zhao, H., and Jin, Z. (2019). A facial expression recognition method based on ensemble of 3D convolutional neural networks. *Neural Comput. Applic.* 31, 2795–2812. doi: 10.1007/s00521-017-3230-2

Wang, C. Y., Lu, Y. D., Yang, S. T., and Lai, S. H. (2022). Patchnet: a simple face anti-spoofing framework via fine-grained patch recognition. In: *Proceedings of IEEE International Conference on Computer Vision Pattern Recognition*. Pp. 20281-20290.

Wang, Z., Wang, Q., Deng, W., and Guo, G. (2022a). Learning multi-granularity temporal characteristics for face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.* 17, 1254–1269. doi: 10.1109/TIFS.2022.3158062

Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., et al. (2022b). Domain generalization via shuffled style assembly for face anti-spoofing. In: *ACM International Conference on Multimedia*. pp. 4123–4133.

Wang, J., Zhang, J., Bian, Y., Cai, Y., Wang, C., and Pu, S. (2021). Self-domain adaptation for face anti-spoofing. *AAAI Conf. Artif. Intell.* 35, 2746–2754. doi: 10.1609/aaai.v35i4.16379

Wen, D., Han, H., and Jain, A. K. (2015). Face spoof detection with image distortion analysis. *IEEE Trans. Inf. Forensics Secur.* 10, 746–761. doi: 10.1109/TIFS.2015.2400395

Wu, G., Zhou, Z., and Guo, Z. (2021). A robust method with dropblock for face anti-spoofing. In: *International Joint Conference on Neural Networks* pp. 1–8.

Yang, J., Lei, Z., and Li, S. Z. (2014). Learn convolutional neural network for face anti-spoofing. arXiv [Preprint].

Yin, W., Ming, Y., and Tian, L. (2016). A face anti-spoofing method based on optical flow field. In: *13th International Conference on Signal Processing (ICSP)* pp. 1333–1337.

Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., and Zhao, G. (2022). Deep learning for face anti-spoofing: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 5609–5631. doi: 10.1109/TPAMI.2022.3215850

Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., et al. (2020). Searching central difference convolutional networks for face anti-spoofing. In: *IEEE Conference on Computer Vision Pattern Recognition*. pp. 5295–5305.

Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., and Li, S. Z. (2012). A face antispoofing database with diverse attacks. In: *IAPR International Conference on Biometrics*. pp. 26–31.

Zhou, L., Luo, J., Gao, X., Li, W., Lei, B., and Leng, J. (2021). Selective domain-invariant feature alignment network for face anti-spoofing. *IEEE Trans Inf. Forensics Secur.* 16, 5352–5365. doi: 10.1109/TIFS.2021.3125603

Zhou, Q., Zhang, K. Y., Yao, T., Lu, X., Yi, R., Ding, S., et al. (2023). Instance-aware domain generalization for face anti-spoofing. In: *Proceedings of IEEE Conference on Computer Vision Pattern Recognition*. pp. 20453–20463.

Zhou, Q., Zhang, K. Y., Yao, T., Yi, R., Ding, S., and Ma, L. (2022b). Adaptive mixture of experts learning for generalizable face anti-spoofing. In: *ACM International Conference on Multimedia* pp. 6009–6018.

Zhou, Q., Zhang, K. Y., Yao, T., Yi, R., Sheng, K., Ding, S., et al. (2022a). *Generative domain adaptation for face anti-spoofing*. Cham: Springer Nature Switzerland. Pp. 335–356.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 13 May 2024
DOI 10.3389/fnins.2024.1415679

OPEN ACCESS

EDITED BY
Hancheng Zhu,
China University of Mining and Technology, China

REVIEWED BY
Yu Liu,
Hefei University of Technology, China
Bo Hu,
Chongqing University of Posts and Telecommunications, China

*CORRESPONDENCE
Lei Zhang
✉ 11905@xzit.edu.cn

RECEIVED 11 April 2024
ACCEPTED 29 April 2024
PUBLISHED 13 May 2024

CITATION
Tian C and Zhang L (2024) G2NPAN: GAN-guided nuance perceptual attention network for multimodal medical fusion image quality assessment.
Front. Neurosci. 18:1415679.
doi: 10.3389/fnins.2024.1415679

G2NPAN: GAN-guided nuance perceptual attention network for multimodal medical fusion image quality assessment

Chuangeng Tian and Lei Zhang*

School of Information Engineering (School of Big Data), Xuzhou University of Technology, Xuzhou, China

Multimodal medical fusion images (MMFI) are formed by fusing medical images of two or more modalities with the aim of displaying as much valuable information as possible in a single image. However, due to the different strategies of various fusion algorithms, the quality of the generated fused images is uneven. Thus, an effective blind image quality assessment (BIQA) method is urgently required. The challenge of MMFI quality assessment is to enable the network to perceive the nuances between fused images of different qualities, and the key point for the success of BIQA is the availability of valid reference information. To this end, this work proposes a generative adversarial network (GAN) -guided nuance perceptual attention network (G2NPAN) to implement BIQA for MMFI. Specifically, we achieve the blind evaluation style via the design of a GAN and develop a Unique Feature Warehouse module to learn the effective features of fused images from the pixel level. The redesigned loss function guides the network to perceive the image quality. In the end, the class activation mapping supervised quality assessment network is employed to obtain the MMFI quality score. Extensive experiments and validation have been conducted in a database of medical fusion images, and the proposed method is superior to the state-of-the-art BIQA method.

KEYWORDS

generative adversarial networks, image quality assessment, multimodal medical fusion image, perceptual, objective evaluation metrics

1 Introduction

Over the past decade, medical images such as computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET) and single photon emission computed tomography (SPECT) have played an increasingly important role in diagnosis, treatment, follow-up recommendations and intraoperative navigation of diseases (Zhou et al., 2020; He et al., 2023; Honkamaa et al., 2023). Depending on the theory of medical imaging techniques and the image features characterized by each modality, multimodal medical images can be simply divided into structural and functional images. The former can precisely locate the lesion and show the structural changes of the lesion, while the latter can sensitively reflect the physiological, biochemical and functional changes

of the tissues and organs in the body, making it easier to detect the lesion. For instance, **Figure 1A** shows an MR image of a brain with glioma, from which the localization information and the internal structure of the tumor can be known, while the edematous region can be found to occupy almost more than half of the area of this tomography. Unfortunately, radiologists are not yet able to recognize the pathological features of the tumor from this image alone. **Figure 1B** shows the PET image of this case. The images of this modality do not have detailed information on brain structure, but it is very easy to identify the lesions with significant abnormal foci of radioactive concentration in the area of the lesion. Based on the imaging features of the above two modalities, the radiologists can then diagnose this disease and even complete a preliminary pathological grading, as shown in **Figure 1C**. Similarly, **Figure 1** also displays a group of cerebral infarction cases where the images of the two modalities express different imaging features. As can be seen from the example, the diagnosis of a particular disease may require reference to multiple modalities at the same time. In view of the fact that the mono-modal image may not be enough to support the conclusion of disease diagnosis, some studies have gradually proposed to integrate the feature advantages of various modalities of medical images through image fusion technology. In addition, many literatures have reported that radiologists can significantly improve the accuracy of disease diagnosis, when they view medical images in multiple modalities simultaneously (Li and Zhu, 2020; Xu and Ma, 2021; Zhou et al., 2023).

Multi-modality medical image fusion (MMIF) is a technique that integrates the medical images obtained from two or more medical imaging devices, extracts the useful information from their respective modalities to maximize, and ultimately forms a comprehensive image (Zhang et al., 2020; Zhang G. et al., 2023; Liu et al., 2024b). Nowadays, image fusion methods specifically for the field of medical images have been vigorously developed, and various excellent fusion algorithms have also been proven in practice. However, due to the different principles of these fusion algorithms, the quality of the generated fusion images is uneven, which needs to be measured by a unified set of standards. Generally, the most direct way to assess the fused image is to have the fused image observed and analyzed by a radiologist. Although this subjective evaluation method can give a score consistent with the human visual system (HVS), but, the quality score of the fused image is influenced by the environment, and cannot be directly analyzed quantitatively due to the direct human involvement. More importantly, subjective assessment is a time-consuming and labor-intensive process (Lei et al., 2022; Liu et al., 2022). This would not be permissible in an already rushed clinical setting.

In contrast to the subjective evaluation, objective evaluation methods detect some indicators of the image to measure the quality of the fused image, such as mutual information (MI), peak signal-to-noise ratio (PSNR), or structural similarity (SSIM). These metrics have achieved excellent achievements in the field of natural image quality assessment (Wang J. et al., 2021). However, it is undeniable that these metrics tend to assess more general properties of images, and are not suitable for assessing medical fused images. This is primarily in a clinical setting, a medical fusion image with excellent quality may not be because it has a high signal-to-noise or anything, but because this fusion image effectively helps the physician to make a diagnostic decision. As mentioned earlier, each modality of medical images expresses unique imaging features. The traditional image quality evaluation methods may ignore the unique feature representation of medical fused images, resulting in the evaluation results that are inconsistent with those of radiologists. Such analytical findings motivated us to find way to represent such unique features when developing MMIF-specific quality evaluation metrics. Particularly, the mean opinion score (MOS) given by radiologists serves perfectly as the ground truth for the quality of fused images. If the network could learn the difference between images with lower and higher MOS, this will be more valuable for the model to assess the quality of the fused images. Furthermore, in practical application scenarios, a completely distortion-free fused reference image (i.e., optimal quality) is difficult, or even impossible, to obtain.

To overcome these problems, in this paper, we propose a GAN-guided nuance perceptual attention network (G2NPAN) for implementing blind image quality assessment (BIQA). A method specifically designed for quality assessment of multimodal medical fusion images.

Specifically, to learn the nuances between different fused images, we use a generative adversarial network (GAN). It has been realized in our previous work that effective spatial feature extraction techniques for image texture and shape can effectively improve the effectiveness of image quality assessment. Therefore, we designed an overlapping structure in the generator, named Unique Feature Warehouse (UFW), to learn spatial features of the fused images from the pixel level and enhance the ability of the network to learn the perception of quality differences between different fused images. Because the purpose of this paper is to accurately assess multimodal medical fusion images, rather than to obtain a perfect fusion image, we redesigned the loss function of the discriminator according to the clinical requirements for image quality. Although GAN can provide powerful guidance for the quality assessment of multimodal medical fusion images, it is still challenging to fully utilize such information. Therefore, we designed the attention-based quality assessment network (AQA) using the supervision of class activation mapping (CAM). Enabling AQA to utilize the fused images generated by GAN at higher resolution and also to sufficiently learn high-dimensional features at lower resolution.

To summarize, the main contribution of this work is in the following folds: (1) We propose a GAN-guided quality difference perceptual attention network. It can achieve accurate quality assessment of multimodal medical fusion images in a blind form. (2) In the generator, we developed the UFW module for learning fused image spatial features from the pixel level. (3) A loss function specifically for multimodal medical fusion image quality perception is designed based on a generative adversarial network architecture. (4) With the supervision of CAM, the proposed AQA is able to learn nonlinear mappings between fused images and objective quality results from lower and higher resolution, respectively, which further enhances the efficacy of model assessment.

The rest of this paper is organized as follows. Section 2 "Related work" introduces the related work of this paper. In Section 3 "Methodology", the details of the proposed methods are described. The adequate experimental results and discussion are presented in Section 4. Finally, we summarize our conclusions at the end of the paper.

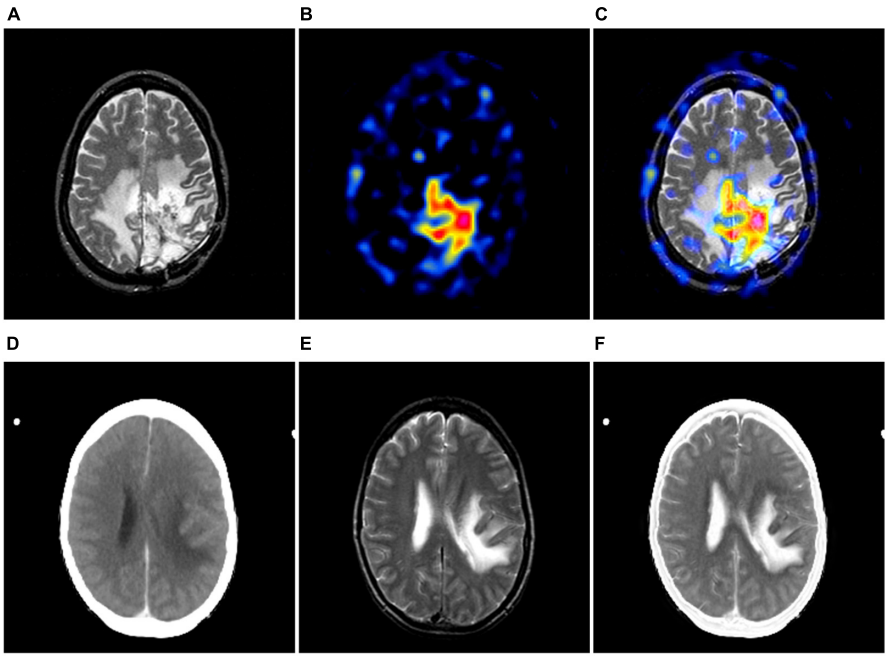


FIGURE 1
The examples of multimodal medical images. **(A, B)** are MR-T2-weighted image and the PET image of the same case; **(D, E)** are CT and MR-T2-weighted image of the same case. **(C, F)** show its corresponding image fusion results, respectively.

2 Related work

2.1 Multimodal medical image fusion (MMIF)

In the medical setting, mono-modality imaging cannot provide comprehensive body tissue information or lesion characteristics and is insufficient to support disease diagnosis (Wang et al., 2020). Therefore, multimodal medical image fusion technology has been created to improve the utilization of medical imaging information. This technology can be classified into traditional fusion methods and deep learning based fusion methods. Traditional image fusion techniques often face challenges with distortion, whereas deep learning-based methods for image fusion have seen notable advancements in recent years (Zhang, 2021; Wang A. et al., 2022; Karim et al., 2023). For example, Wang et al. (Wang Z. et al., 2022) designed a self-supervised residual feature learning network for multi-focus image fusion. Xu and Ma (2021) developed an unsupervised image fusion network with enhanced information preservation by surface-level and deep- level constraints. It is worth specifying that the network is built specifically for medical images. But, the performance of fusion algorithms and the quality assessment of fused images are not yet fully understood. Whether the multimodal image fusion technique can be successfully applied, the quality assessment of the fused images is the key.

2.2 Image quality assessment for MMIF

Based on the different requirements for reference images, the objective image quality assessment methods (IQA) can be divided into three categories: full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference IQA (NR-IQA) i.e., BIQA. Despite FR-IQA and RR-IQA methods have achieved remarkable success in the past decades, their application fields are restricted due to their dependence on reference images. This is because reference images are not always available in practical application scenarios, and even more, in some fields, it is almost impossible to obtain them. Therefore, BIQA has gained the favor of many researchers as it does not require any reference image for evaluation.

According to the way of feature extraction, BIQA includes: statistical analysis-based models and learning-based models. Most existing models based on statistical analysis attempt to detect concrete types of distortion, such as various forms of blur and noise. And the learning-based BIQA model aims to reflect the differences in image quality through effective feature extraction techniques as well as to design the model to learn the mapping relationship between features and image quality. Traditional machine learning approaches assume that either distortion will cause the image to change in some feature attributes. Therefore, this kind of method pays more attention to the process of feature extraction. The quality regression models are then designed by machine learning methods such as support vector machine (SVM), K-Nearest Neighbors algorithms, etc. Some classical models are, for example, BRISQUE (Mittal et al., 2012), NFERM (Gu et al., 2015) and BIBE (Wang et al., 2016).

However, those method separates the process of feature extraction and quality score prediction/regression of images. This leads to models that cannot be implemented in an end-to-end learning manner. Moreover, feature extraction schemes based on hand design rely on the experience of the researcher, and the features obtained from limited understanding of the image may

not sufficient to describe the image content. Most recently, the vigorous development of deep learning techniques is gradually becoming the mainstream of IQA algorithms (Hou et al., 2015; Madhusudana et al., 2022; Zhang Z. et al., 2023; Liu et al., 2024a). Earlier, Kang et al. (2014) integrated feature learning and regression into an optimization process by a simple CNN architecture and obtained promising generalization results. In (Wang X. et al., 2021), WANG et al. proposed a novel tone-mapped image metric using local degradation characteristics and global statistical properties. Inspired by the observer subjective assessment process, Sim et al. (2022) proposed a novel BIQA algorithm based on the semantic recognition task. Yue et al. (2023) implemented an automated assessment of colonoscopy images by analyzing brightness, contrast, colorimetry, naturalness, and noise. But BIQA methods specifically for multimodal medical fusion images have not been adequately explored. Considering the absence of referenceable fusion images in a real clinical setting, we design a novel learning-based BIQA model.

2.3 GAN-based image quality assessment

In the process of image quality assessment, since the reference image is not always available, it poses a great difficulty in constructing a learning-based IQA model (Liu et al., 2019). Until 2014, the emergence of GAN has brought new ideas to researchers in many fields. GAN could attempt to generate better outputs with adversarial training of generators and discriminators. Therefore, if the reference image can be generated for the BIQA method, it will be possible to bridge the performance gap between the FR-IQA and BIQA methods. Moreover, the concern that standard reference images for multimodal fusion images are not available in the clinical setting will be mitigated. A series of GAN-based work has also been carried out by researchers related to image quality evaluation (Ma et al., 2019; Guo et al., 2023; Kelkar et al., 2023; Li and He, 2024). In 2019, Ma et al. (2019) proposed an end-to-end GAN model for quality assessment of images based on multitasking. And the superiority of the method was verified in TID2008 and TID2013 datasets. The same year, Yang et al. (2019) designed a BIQA method with the advantages of self-generated samples and self-feedback training, called BIQA-GAN. GAN-based methods have the ability to learn local distortion characteristics and whole quality on the depth features of the image, and it can accomplish the mapping fitting of potential features to the target domain. Thus, we introduce GAN to design our model, and, we tuned the loss function and architecture of GAN according to the characteristics of medical images.

3 Methodology

In this section, we introduce an end-to-end no-reference method, namely GAN-guided nuance perceptual attention network (G2NPAN), for assessing the quality of multimodal medical fusion images. First, we introduce the framework of the proposed method. Then, we elaborate on the two main parts of our proposed G2NPAN, i.e., the GAN-guided nuance perceptual module and attention-based quality assessment network. Finally, the quality perception loss function is formulated.

3.1 Overview

The core idea of our proposed method is to assume that high MOS fused images can indeed help the physician in clinical analysis. Therefore, learning the nuance between lower and higher MOS fused images is of high value for quality assessment in the absence of a reference image. The framework of G2NPAN is shown in **Figure 2**. Briefly, our method is specified below. Firstly, G2NPAN learns the nuances between the fused images with different quality through generative adversarial networks, and utilizes generators to generate fused images with the best possible quality. For the generator, we carefully designed an overlapping structure, UFW, and repeated it five times to increase the ability to learn the nuances between different fused images. Then, we redesigned the loss function of the discriminator according to the scoring criteria of images in the clinical setting. The aim is to increase the perceptual weight of the image quality during the network training process. Next, we subtract the high MOS image generated by the generator from the original fused image to obtain the difference between them. Finally, we feed this nuance together with the original fused image into an attention-based quality assessment network to obtain a nonlinear mapping between the fused image and the objective quality results. We will describe this process in detail in the remaining part of this section.

3.2 GAN-guided nuance perceptual module

3.2.1 GAN architecture

GAN is a distinctive approach to achieving feature extraction by generating new fused images in the form of generative adversarial. This network structure normally consists of two main parts, the Generator (**G**) and the Discriminator (**D**). On the one hand, GAN has domain adaptive property. For non-discrete distribution data, like fused images, it is more robust for feature extraction or learning. On the other hand, the generator can generate fused images of the same type through adversarial training, and under the supervision of the discriminator, the generated images are fitted toward higher quality. The proposed G2NPAN is established on the framework of GAN, which takes the original fused image as input and passes the image nuance information to obtain a quality score of fused images. It is worth noting that the purpose of high-quality fused images produced by the generator is to provide reference information, which may contribute to the quality assessment of the original fusion image. More specifically, it may help to alleviate the problems associated with the absence of reference images.

Our network structure of the generator and discriminator is presented in **Figure 3**. **G** takes the fused image I_{org} with arbitrary quality as input and aims to generate a fused image I_{hq} with the best quality, i.e., $I_{hq} = G(I_{org})$. The discriminator exists to distinguish the real fused image I_{org} from the generated version of the fused image I_{hq}. Through adversarial training, it is expected that the fused image with the best quality can be generated with the arbitrary quality of fused image as input.

Generator: As shown in **Figure 3**, **G** is a convolution neural network consisting of down-sampling and up-sampling phases.

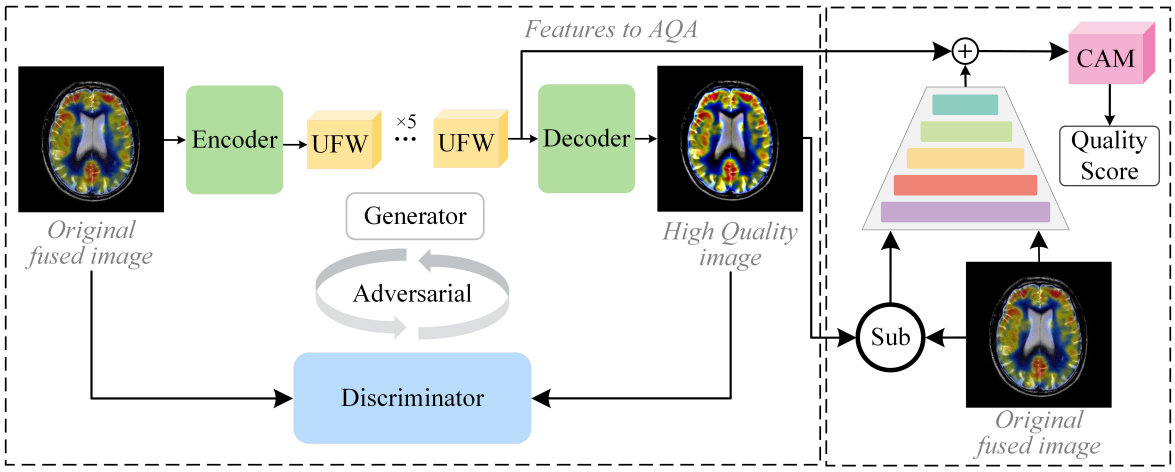


FIGURE 2
The framework of the proposed GAN-guided nuance perceptual attention network.

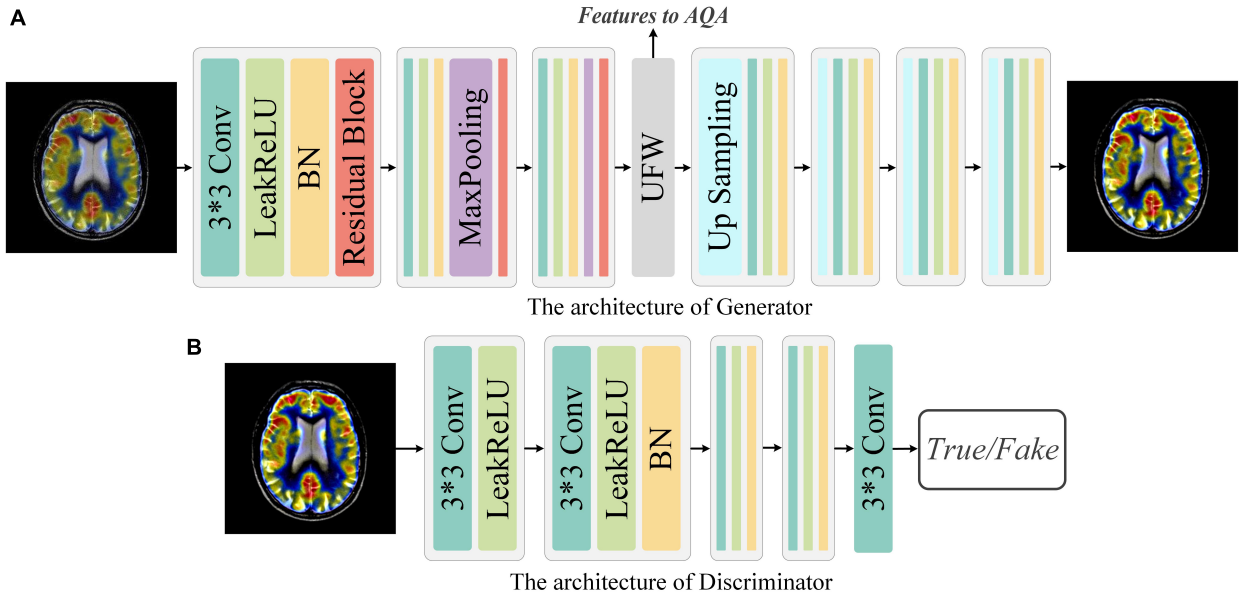


FIGURE 3
Network architecture of Generator and Discriminator. **(A)** Generator: The Generator is a simple down-sampling and up-sampling convolution neural network with the Unique Feature Warehouse. **(B)** Discriminator: The Discriminator consists of a simple five-layer convolution neural network.

Since medical fused images require the network to focus on more detailed features, the kernel size in the generator is all set to 3 × 3. The down-sampling operation is composed of three sequential networks with the same structure. Specifically, we connect the convolutional layer, the activation function, the batch normalization (BN), and the residual block to build this sequential network. For the activation function, a comprehensive activation algorithm, Lleaky Rectified Line Unit (LeakyReLU), is used. To make the model more stable, the BN layer is attached after the activation layer, which can also help the gradient to back propagate efficiently. For medical fused images, the conventional simply increasing the depth of the convolution neural network may cause the model to converge slowly or even be unable to converge. Therefore, we invoke the residual structure in the down-sampling process, which consists of three convolution layers and a skip connection. After the down-sampling operation, followed by our elaborate UFW structure, which improves the ability of the network to learn the nuances between different fused images. We will describe the detailed structure of UFW in the next subsection. In the up-sampling phase, we designed a simple four-layer convolutional neural network. Each layer of the convolutional neural network consists of an up-sampling operation, a convolutional with a kernel size of 3 × 3, a batch normalization, and an activation layer. As for the activation function, we use the LeakReLU activation function in the first three layers and the Sigmoid activation function in the last layer. So far, the best quality image I_{hq} of size 128 × 128 can be obtained by using the I_{org} as the input. The generator's parameters are only

renewed by the mean squared error (MSE) and are defined as in Eq. 1:

$$L_1 = \frac{1}{N}\sum_{n=1}^{N}(G(I_{org}) - I_{GT})^2 = \frac{1}{N}\sum_{n=1}^{N}(I_{hq} - I_{GT})^2, \quad (1)$$

where N is the total number of generated samples. I_{GT} means the fused image with high MOS, i.e., Ground Truth (GT).

During the training of **G**, the following objective function (Eq. 2) is minimized:

$$L_G = \mathbb{E}_{I_{org}\sim P_{dataO}}[\log(1 - D(I_{GT}, G(I_{org})))] + \theta L_1, \quad (2)$$

where P_{dataO} stands for the data distribution of I_{org}, and the $\mathbb{E}_{I_{org}\sim P_{dataO}}$ represents the expectation of I_{org}. θ is a weighted hyperparameter.

Discriminator: The discriminator only needs to judge whether the image conforms to the real data distribution or not. Thus, the architecture of discriminator is a simple four-layer convolution neural network, as illustrated in Figure 3. In brief, each network has one convolutional layer with a kernel size of 3 × 3, a stride of 2, and padding of 1. Then LeakyReLU is used as the activation function and subsequently processed with BN. Note that with each layer of the convolutional neural network, the size of the feature map shrinks to one-fourth of the input. Finally, we add an independent convolution layer according to the sequential structure, which is mainly used for classification. The mean absolute error is used as the loss function to optimize the parameters of discriminator. Thus, the objective function of discriminator can be expressed as Eq. 3:

$$L_D = \mathbb{E}_{I_{GT}\sim P_{dataGT}}[\log D(I_{GT})] + \mathbb{E}_{I_{org}\sim I_{P_{dataO}}}[\log(1 - D(G(I_{org})))], \quad (3)$$

where P_{dataGT} is the data distribution of I_{GT}, and $\mathbb{E}_{I_{GT}\sim P_{dataGT}}$ is the expectation of I_{GT}.

3.2.2 Unique feature warehouse (UFW)

In our previous work, it has been realized that effective spatial feature extraction techniques for image texture and shape play an important role in the quality assessment of medical fusion images. The preservation of anatomical details, the representation of metabolic information, and the trade-offs of information during the fusion process are one of the many characteristics to be recognized in fused image assessment. Thus, an overlapping structure, UFW, is designed in this paper to enable the model to capture these features from fused images at multiple scales. The detailed architecture of the UFW is presented in Figure 4. Note that both the input and output feature maps are 32 × 32. On the one hand, taking a full-resolution image as input requires a large amount of memory consumption. On the other hand, most of the high-dimensional features appear only at lower resolutions. Therefore, we embedded the UFW module at the end of the down-sampling stage of the generator, with a maximum resolution of 32 × 32. In addition, with the overlapping architecture, the network can process the high-dimensional features multiple times to further learn and weigh their relationship. Consistent with the design purpose of the generator, the kernel size we use in UFW are all 3 × 3 to better focus on subtle spatial features. The UFW structure recognizes image spatial features from multiple scales and continuously integrates them in the overlapping structure to achieve effective spatial feature extraction.

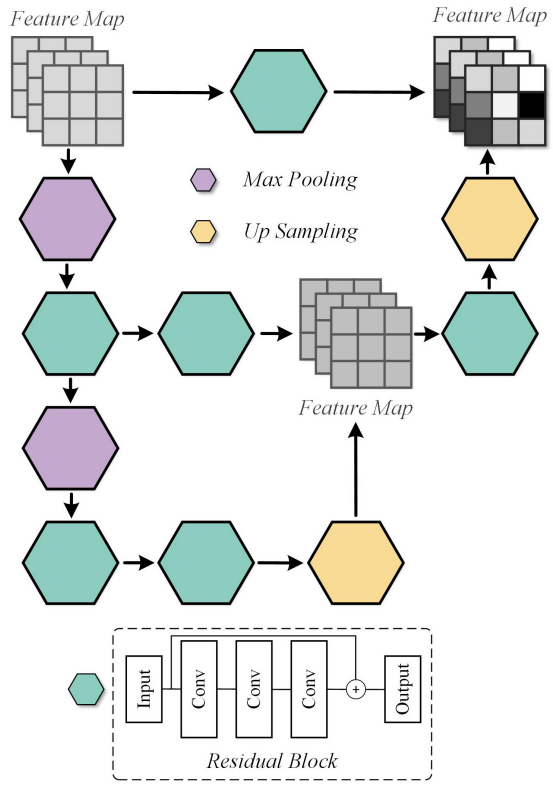


FIGURE 4
Detailed architecture of the Unique Feature Warehouse (UFW) module.

3.3 Attention-based quality assessment network

The attention-based quality assessment network is built on the VGG network, which is a simple convolutional neural network as shown in Figure 5. The reason for adopting VGG network are respectively: the VGG network is an easy-to-use CNN, which can save a lot of effort in modifying its network architecture. Also, with the guidance of GAN, AQA is not required to extract the feature representation of the fused image from scratch. Therefore, it is less necessary to employ a complex network structure. Finally, VGG11, which has a relatively simple structure and shallow network depth in VGG networks, was used as the base framework in the AQA. AQA takes the nuance between the original fused image and the generated image, and the original fused image as input to obtain an objective assessment of the fused image.

Specifically, from the structure of GAN this paper takes the fused image with higher MOS as GT of the generator, thus limiting its fitting trend. Thus, the nuance between the fused image I_{org} and the higher quality fused image I_{hq} can be defined as $I_{\text{sub}} = |I_{hq} - I_{org}|$. *i*denoting the *i-th* assessed image, the definition can be revised to Eq. 4:

$$I^i_{sub} = \left|I^i_{hq} - I^i_{org}\right|, \quad (4)$$

To ensure the input consistency, $I^i_{sub}$and I^i_{org}performed the convolution operation first separately, and then completed the concatenation operation.

For I^i_{org}, its features extracted in the GAN can also guide the AQA to obtain more convincing assessment results. So, we copied the feature map $\mathbf{F}_{last}$ output from the last UFW module, and implemented concatenation before inputting the fully connected layer to obtain the feature map F_{conca}, as shown in Eq. 5:

$$\mathbf{F}_{conca} = \mathbf{F}_{last} \odot vgg(I^i_{org}, I^i_{sub}), \tag{5}$$

where $vgg(\bullet)$ denotes the operation of AQA before proceeding to the fully connected layer.

So far, the image quality assessment has been achieved objectively by the methods mentioned above. However, for physicians, the quality of medical images depends *not only* on the natural nature of the images, *but also* on their ability to highlight the manifestations of disease. The latter is the key to assist doctors in making a diagnosis. Thus, by using the weights of the last fully connected layer as a cue, we introduce the attention mechanism, class activation mapping. With the quality scores of AQA and the weights of the fully connected layers, CAM obtains the ability to supervise the attention distribution of the network. Moreover, the feature map $\mathbf{F}_{CAM}$ generated by CAM can also compensate for the un-interpretability of "black box" models. Let the GT of CAM be $\mathbf{F}_{GT}$, then the objective function is Eq. 6:

$$L_{CAM} = \frac{1}{N}\sum_{n=1}^{N} |\mathbf{F}_{CAM} - \mathbf{F}_{GT}|_1, \tag{6}$$

where $|\bullet|_1$ denotes the L1 parametrization.

Further, the predicted score of the fused image is designated as Q_{pre} and its GT is Q_t, then the objective function can be written as shown in Eq. 7:

$$L_{QA} = -[Q_t \log(\sigma(Q_{pre})) + (1 - Q_t)\log(1 - \sigma(Q_{pre}))], \tag{7}$$

where $\sigma(\bullet)$ denotes the sigmoid function, which is meant to map Q_{pre} to the interval (0, 1), specified as shown in Eq. 8:

$$\sigma(Q_{pre}) = \frac{1}{1 + exp(-Q_{pre})}, \tag{8}$$

Thus, the loss function of AQA can be expressed as Eq. 9:

$$L_{AQA} = \varphi L_{CAM} + L_{QA}, \tag{9}$$

φ is the weight parameter.

3.4 Perceptual loss function

The design logic of GAN is trained in an adversarial way so that the generated image can deceive the discriminator, and the discriminator can distinguish the real image from the generated image. Although such a network architecture can generate high-quality fused images, the ultimate goal of G2NPAN is to accurately evaluate the quality of fused images rather than to obtain fused images. Moreover, it is clear from the calculation of I^n_{sub} that it depends heavily on the generated image I_{hq} . If I^n_{sub} is directly used as the input of AQA without feedback to GAN, the training process of quality assessment network will be unstable and difficult to converge. Thus, we design the quality perception loss function to alleviate the above problem. It is worth clarifying that the fused images used in this work are based on a further extension of the database from our previous work (Tang et al., 2020), and thus the MOS of each medical fused image can take values from 1 to 5. Typically, ensuring that the MOS remains above 3 does not compromise the diagnostic results provided by medical professionals. This ensures that the fused medical images do not adversely impact diagnostic performance. Therefore, the weight can be expressed as Eq. 10:

$$W^n = \begin{cases} 1, \text{ if } AQA(I^n_{hq}) \geq 3, \\ 0, \text{ if } AQA(I^n_{hq}) < 3 \end{cases}, \tag{10}$$

We the weight to further optimize the network and restate the formula (3) as shown in Eq. 11 below:

$$L_D = \mathbb{E}_{I_{GT}\sim P_{dataGT}}[\log D(I_{GT})] + \mathbb{E}_{I_{org}\sim I_{P_{dataO}}} [\log(1 - |D(G(I_{org})) - W|)], \tag{11}$$

The generated images need to be distinguished not only by the discriminator, but also by the quality assessment network. The concept of perceptual loss function allows the model to be optimized as a unit, so that the total loss function can be presented as Eq. 12:

$$L_{all} = \min_G \max_D V(G, D) + \gamma L_R, \tag{12}$$

4 Experiments

4.1 Databases and experimental protocols

4.1.1 Dataset description

We established a multi-modal medical image fusion quality evaluation database to validate the effectiveness of our proposed algorithm. The database comprises 120 pairs of color images and 9 pairs of grayscale images, with a total of 1,290 images generated using 10 mainstream fusion algorithms. The resolution of the images is 128 × 128 pixels. The MOS of each image was obtained from radiologists on a scale ranging from 1 to 5.

To select reference images for each group of fused images (i.e., 10 images generated by fusing a pair of images), we used MOS to evaluate image quality. The fused image with the highest MOS score was chosen as the reference image. If multiple fused images had the highest MOS score, one of them was selected at random. The reference image for each image was randomly selected from the fused images with the highest MOS score. This ensured a robust reference image selection process that accounted for the subjective quality ratings of the radiologists.

4.1.2 Evaluation criteria

In this study, we utilized four evaluation metrics to assess the performance of the proposed model: Pearson's Linear Correlation Coefficient (PLCC), Spearman's Rank Correlation Coefficient (SRCC), Kendall's Rank Correlation Coefficient (KRCC), and Root Mean Square Error (RMSE). PLCC measures the linear relationship between the predicted and the corresponding MOS, while SRCC

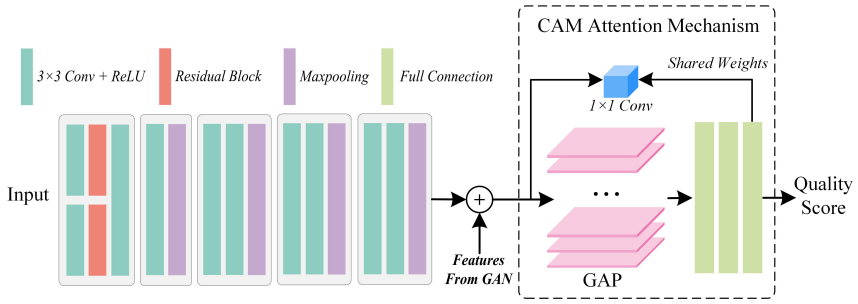


FIGURE 5
The architecture of the proposed attention-based quality assessment network.

and KRCC are non-parametric correlation measures that evaluate rank-based data. RMSE measures the difference between the predicted and the corresponding MOS.

4.1.3 Experimental protocols

In the training process of the network, three hyperparameters, θ, φ, and γ, were set to 0.5, 1, and 0.01, respectively. The Adam optimizer was used with an initial learning rate of 0.0002. Furthermore, we implemented a dynamic learning rate adjustment strategy to enhance model convergence during training. Specifically, we reduced the learning rate using a decay factor of 0.95 after every 20 batches.

To evaluate the effectiveness of our proposed model, we employed a five-fold cross-validation approach during implementation. We use 80% images of the database to train our model, while using 20% to test. The model's performance was evaluated at the end of each training epoch, and we selected the checkpoint model with the best performance within the 1500 epochs of training as the final model. During the validation phase, we assessed the model's performance on the test set. In each evaluation, we try 1000 times and take an average of the performance values obtained.

4.2 Comparison with the state-of-the-art

In this section, exhaustive comparative experiments are conducted to validate our proposed method. We compared the performance of G2NPAN with the performance of six state-of-the-art BIQA methods. For approaches that are not specifically named, we refer to them by the name of the first author. All these methods include the blind multiple pseudo reference images-based method (BMPRI) (Min et al., 2018), In-depth analysis of Tsallis entropy-based method (TEIA) (Sholehkerdar et al., 2019), mutual information-based optimization method (Hossny) (Hossny et al., 2008), the objective evaluation of fusion performance (OEFP) (Xydeas and Petrovic, 2000), ratio of spatial frequency error-based method (rSFe)(Zheng et al., 2007) and the perceptual quality assessment method (Tang) (Tang et al., 2020). BMPRI introduces multiple pseudo-reference images to achieve BIQA, which coincides with our approach of using GAN to generate reference information to perform IQA. Thus, although BMPRI is not specifically developed for quality assessment of fused images, it is still used as one of the comparison methods. And the remaining methods are proposed exclusively for the quality assessment of fused images. Note that the proposer of rSFe considers the application scenario of medical fusion images, while the Tang method is proposed especially for medical fusion images. For fair comparison, all methods were retrained and tested in our Dataset, and the best results were used as the final reported.

TABLE 1 Performance comparison with Other BIQA methods.

Model	Domain	PLCC	SRCC	KRCC	RMSE
BMPRI	Distorted image	0.3031	0.3167	0.2375	0.2611
TEIA	Fused image	0.1797	0.1946	0.1407	0.3909
Hossny	Fused image	0.2270	0.1738	0.1071	0.3712
OEFP	Fused image	0.3064	0.3367	0.2342	0.2810
rSFe	Fused image	0.4054	0.2275	0.1700	0.2663
Tang	Fused image	0.6252	0.6420	0.4166	0.2480
Proposed	Fused image	**0.9044**	**0.9007**	**0.8502**	**0.1029**

Bold values represent the best results.

TABLE 2 Ablation experiments of quality assessment with different backbone networks.

Network	PLCC	SRCC	KRCC	RMSE
VGG19	0.7530	0.7358	0.6696	0.1601
VGG16	0.7776	0.7742	0.6977	0.1553
VGG11	0.7894	0.7905	0.7132	0.1494
VGG11 + CAM	0.8000	0.7806	0.7173	0.1385
VGG 11 + pre	0.8167	0.8112	**0.7519**	0.1355
VGG11 + pre + CAM	**0.8235**	**0.8113**	0.7496	**0.1330**

Bold values represent the best results.

We have tabulated the performance of the state-of-the-art BIQA method and G2NPAN in Table 1. The best performance results are highlighted in bold. Based on Table 1, we have the following observations:

First, our proposed method, G2NPAN, achieved the best quality assessment performance from an overall perspective, with optimal results of 0.9044, 0.9007, 0.8502, and 0.1029 for PLCC, SRCC, KRCC and RMSE, respectively. This means that the objective evaluation results derived from the G2NPAN are closest to the subjective MOS results given by the physicians. Second, although

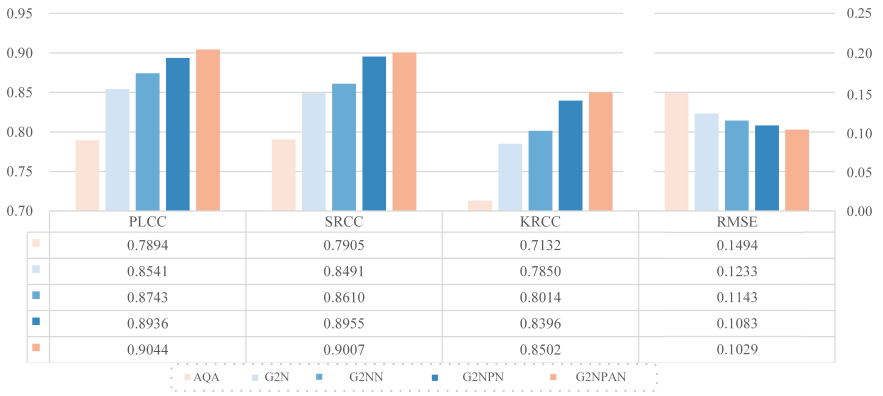


FIGURE 6
Ablation experimental results for each of the key components in G2NPAN.

TABLE 3 Performance on the dataset with various train-test splits.

Train: Test	PLCC	SRCC	KRCC	RMSE
2:8	0.7780	0.7849	0.7042	0.1649
3:7	0.8055	0.8148	0.7372	0.1441
4:6	0.8497	0.8451	0.7817	0.1275
5:5	0.8621	0.8617	0.8044	0.1206
6:4	0.8821	0.8795	0.8259	0.1139
7:3	0.8848	0.8886	0.8393	0.1093
8:2	**0.9044**	**0.9007**	**0.8502**	**0.1029**

Bold values represent the best results.

the BMPRI method introduces pseudo-reference images to provide referenceable information for BIQA, it is mainly targeted at distorted images of natural scenes. Therefore, it is not powerful for medical fusion images. Our proposed method generates reference information based on high-quality fused images and designs quality evaluation methods from the specificity of medical images, resulting in the best BIQA performance. As can be seen from Table 1, BMPRI also outperforms some of the quality evaluation metrics designed specifically for fused images, which once again demonstrates the importance of reference information for BIQA. Third, rSFe, Tang and the proposed method have considered the difference between medical fusion images and natural fusion images, and thus their performance is better than the other three metrics (TEIA, Hossny and OEFP). In addition, the performance of proposed method is still 27.92, 25.87, 43.36, and 14.51% better than the second-best method in PLCC, SRCC, KRCC and RMSE, respectively. From the above analysis, it is clear that our proposed G2NPAN method is very good at objective quality assessment of medical fusion images.

4.3 Ablation study

Ablation experiments are performed from different perspectives to demonstrate the superiority of our proposed method. (1) To verify the generalizability of the proposed AQA, we compose the model through different backbone networks, including VGG11, VGG16, VGG19, and tested the model performance. Each ablation result is demonstrated in Table 2, with the best result for the corresponding metric highlighted in bold. (2) To evaluate the contribution of each key component in the proposed G2NPAN model, a series of ablation experiments were conducted.

4.3.1 Performance of quality assessment network

Ablation studies were performed to examine whether the backbone network used in the quality prediction network was more appropriate. All models implemented in this section are purely quality prediction networks, meaning that there is no GAN-based quality guidance. Their testing performance is listed in Table 2.

On the one hand, from these results, we can notice that the performance of the VGG11 is even better than that of VGG16 or VGG19. This seems to go against the common belief that the deeper the network, the better the model performance should be. But there should be more detailed analysis for different task types. The truth is that VGG19 or 16 has more convolutional layers than VGG11, which allows the network to learn more semantic information (high-level features). However, for evaluation of multimodal medical fusion images, the model does not need to recognize what the image represents, like what disease or which organ, etc., but rather than what the image has. Thus, the IQA task might require more structural (low-level features) than semantic information about the image. VGG11 improved by 3.64, 5.47, 4.36, and 1.07% in PLCC, SRCC, KRCC and RMSE, respectively, compared to VGG19.

On the other hand, to demonstrate the usefulness of the CAM and pre-trained models, we have adapted them based on the VGG11 model. From the experimental results, it can be seen that both CAM and the introduced pre-trained model enhance the performance of quality prediction network. And, based on these two techniques, the proposed quality prediction network achieves 0.8235, 0.8113, 0.7496, and 0.1330 in PLCC, SRCC, KRCC and RMSE, respectively.

4.3.2 The contribution of each key component

As mentioned in the previous section, the proposed method integrates the nuances between fused images with different

qualities, and adjusts the update of the loss function according to the scoring criteria of images in the clinical setting. Therefore, it is sensible and meaningful to fully explore the contribution of each key component to the final performance.

We take VGG11 network as the base quality prediction model. Based on this, the GAN-guided quality assessment models with and without UFW components are named G2N (GAN-Guided Network) and G2NN (GAN-Guided Nuance Network), respectively. Further, we redesigned the loss function of GAN in a qualitatively perceptive way. Such a modified network named G2NPN (GAN-Guided Nuance Perceptual Network). Eventually, CAM is added to the G2NPN model to supervise the quality prediction results of the fused images and call such a network G2NPAN (GAN-Guided Nuance Perceptual Attention Network), i.e. the model proposed in this paper. Note that we trained G2N, G2NN, and G2NPN based on the same method applied in G2NPAN and summarized their corresponding prediction performance results in Figure 6. As all three models, G2N, G2NN and G2NPN, are degradation models based on GAN tuning, the blue family is used for unification in Figure 6.

As expected, all key components had a positive effect on the final model performance. And as the model structure becomes closer to G2NPAN, the quality assessment of the medical fusion images becomes more accurate. Further analysis is as follows. First of all, with the reference information provided by GAN, the G2N model achieves the largest performance improvement over VGG11 with 6.47, 5.86, 7.18, and 2.61% improvement in PLCC, SRCC, KRCC and RMSE, respectively. The G2N model generates the best-quality fused image similar to providing the reference image for IQA, and thus, it has the most significant performance improvement. However, the nuances in the reference information might not be sufficient. UFW is an effective way to extract spatial features by learning the features of fused images from multiple scales several times. Therefore, the G2NN model further enhances the performance results. Second, as the GAN has the ability to recognize the quality of fused images, i.e., the perceptual capability, the G2NPN model obtains considerable performance gains, especially in SRCC (0.8610 *vs* 0.8955) and KRCC (0.8014 *vs* 0.8396). Finally, by introducing the CAM attention mechanism, our proposed G2NPAN has got the best performance for medical fusion image quality assessment, with PLCC, SRCC, KRCC and RMSE of 0.9044, 0.9007, 0.8502, and 0.1029, respectively.

Overall, whether it is the visual impression of the blue rectangular bar in Figure 6 or the data analysis results, it can be found our proposed GAN-guided approach could yield a tremendous performance improvement. Except for RMSE, the improvement results for the other three metrics were more than 10%. It is also interesting to observe that the models with reference information provided by GAN outperform all the methods shown in Table 2.

4.4 Impact of training set

To investigate the relationship between the sample size and the performance of the proposed method, we gradually increased the training sample size from 20 to 80%, while the rest of the image samples were used as testing. All experimental results are filled in Table 3. It is intuitive to notice that as the training sample size increases, the proposed model performance tends to rise gradually. And, the model performance does not drop precipitously when the training sample size are smaller. This observation is consistent with the conclusions drawn from the existing learning-based BIQA (Gu et al., 2016; Jiang et al., 2019; Wang X. et al., 2021). The robustness of the proposed G2NPAN model has been validated.

5 Conclusion

In this paper, we propose a BIQA method specifically for multimodal medical fused images, called GAN-Guided Nuance Perceptual Attention Network. Specifically, in addition to designing the UFW module in the GAN to incorporate collecting useful features from the pixel level, we also redesigned the loss function of the discriminator to enable the network to learn the nuance between fused images of variable quality. Following that, the nuance information and the high-dimensional features in the UFW are fed back to the quality assessment network. With the supervision of CAM, the quality score of the fused image is eventually determined. The experimental results demonstrated that our proposed method outperforms the state-of-the-art methods. Two aspects of ablation experiments validate the generality of the proposed AQA and the contribution of each key component of the G2NPAN model. The experiments examining the correlation between sample size and G2NPAN performance further verify the effectiveness of the proposed GAN-guided quality assessment model.

Data availability statement

The original contributions presented in this study are included in this article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CT: Methodology, Writing—original draft. LZ: Validation, Writing—review and editing.

Funding

The authors declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the 2023 Jiangsu Province Industry University Research Cooperation Project (BY20231347).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Gu, K., Wang, S., Zhai, G., Ma, S., Yang, X., Lin, W., et al. (2016). Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure. *IEEE Trans. Multimedia* 18, 432–443. doi: 10.1109/TMM.2016.2518868

Gu, K., Zhai, G., Yang, X., and Zhang, W. (2015). Using free energy principle for blind image quality assessment. *IEEE Trans. Multimedia* 17, 50–63. doi: 10.1109/TMM.2014.2373812

Guo, Y., Hu, M., Min, X., Wang, Y., Dai, M., Zhai, G., et al. (2023). Blind image quality assessment for pathological microscopic image under screen and immersion scenarios. *IEEE Trans. Med. Imaging* 42, 3295–3306. doi: 10.1109/TMI.2023.3282387

He, K., Zhang, X., Xu, D., Gong, J., and Xie, L. (2023). Fidelity-driven optimization reconstruction and details preserving guided fusion for multi-modality medical image. *IEEE Trans. Multimedia* 25, 4943–4957. doi: 10.1109/TMM.2022.3185887

Honkamaa, J., Khan, U., Koivukoski, S., Valkonen, M., Latonen, L., Ruusuvuori, P., et al. (2023). Deformation equivariant cross-modality image synthesis with paired non-aligned training data. *Med. Image Anal.* 90:102940.

Hossny, M., Nahavandi, S., and Creighton, D. (2008). Comments on 'Information measure for performance of image fusion.'. *Electron. Lett.* 44:1066. doi: 10.1049/el:20081754

Hou, W., Gao, X., Tao, D., and Li, X. (2015). Blind image quality assessment via deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* 26, 1275–1286.

Jiang, Q., Shao, F., Lin, W., and Jiang, G. (2019). BLIQUE-TMI: Blind quality evaluator for tone-mapped images based on local and global feature analyses. *IEEE Trans. Circ. Syst. Video Technol.* 29, 323–335. doi: 10.1109/TCSVT.2017.2783938

Kang, L., Ye, P., Li, Y., and Doermann, D. (2014). "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, (Columbus, OH: IEEE), doi: 10.1109/CVPR.2014.224

Karim, S., Tong, G., Li, J., Qadir, A., Farooq, U., and Yu, Y. (2023). Current advances and future perspectives of image fusion: A comprehensive review. *Inform. Fus.* 90, 185–217. doi: 10.1016/j.inffus.2022.09.019

Kelkar, V. A., Gotsis, D. S., Brooks, F. J., Kc, P., Myers, K. J., Zeng, R., et al. (2023). Assessing the ability of generative adversarial networks to learn canonical medical image statistics. *IEEE Trans. Med. Imaging* 42, 1799–1808. doi: 10.1109/TMI.2023.3241454

Lei, F., Li, S., Xie, S., and Liu, J. (2022). Subjective and objective quality assessment of swimming pool images. *Front. Neurosci.* 15:766762. doi: 10.3389/fnins.2021.766762

Li, C., and Zhu, A. (2020). Application of image fusion in diagnosis and treatment of liver cancer. *Appl. Sci.* 10:1171. doi: 10.3390/app10031171

Li, X., and He, S. (2024). Blind image quality evaluation method based on cyclic generative adversarial network. *IEEE Access* 12, 40555–40568. doi: 10.1109/ACCESS.2024.3375940

Liu, L., Wang, T., and Huang, H. (2019). Pre-attention and spatial dependency driven no-reference image quality assessment. *IEEE Trans. Multimedia* 21, 2305–2318. doi: 10.1109/TMM.2019.2900941

Liu, Y., Qi, Z., Cheng, J., and Chen, X. (2024a). Rethinking the effectiveness of objective evaluation metrics in multi-focus image fusion: A statistic-based approach. *IEEE Trans. Patt. Anal. Mach. Intell.* [Online ahead of print.]. doi: 10.1109/TPAMI.2024.3367905.

Liu, Y., Shi, Y., Mu, F., Cheng, J., Li, C., and Chen, X. (2022). Multimodal MRI volumetric data fusion with convolutional neural networks. *IEEE Trans. Instrument. Meas.* 71, 1–15. doi: 10.1109/TIM.2022.3184360

Liu, Y., Yu, C., Cheng, J., Wang, Z. J., and Chen, X. (2024b). MM-Net: A mixformer-based multi-scale network for anatomical and functional image fusion. *IEEE Trans. Image Process.* 33, 2197–2212. doi: 10.1109/TIP.2024.3374072

Ma, Y., Cai, X., Sun, F., and Hao, S. (2019). No-Reference image quality assessment based on multi-task generative adversarial network. *IEEE Access* 7, 146893–146902.

Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B., and Bovik, A. C. (2022). Image quality assessment using contrastive learning. *IEEE Trans. Image Process.* 31, 4149–4161. doi: 10.1109/TIP.2022.3181496

Min, X., Zhai, G., Gu, K., Liu, Y., and Yang, X. (2018). Blind image quality estimation via distortion aggravation. *IEEE Trans. Broadcast.* 64, 508–517. doi: 10.1109/TBC.2018.2816783

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-Reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* 21, 4695–4708.

Sholehkerdar, A., Tavakoli, J., and Liu, Z. (2019). In-depth analysis of Tsallis entropy-based measures for image fusion quality assessment. *Optic. Eng.* 58:1. doi: 10.1117/1.OE.58.3.033102

Sim, K., Yang, J., Lu, W., and Gao, X. (2022). Blind stereoscopic image quality evaluator based on binocular semantic and quality channels. *IEEE Trans. Multimedia* 24, 1389–1398. doi: 10.1109/TMM.2021.3064240

Tang, L., Tian, C., Li, L., Hu, B., Yu, W., and Xu, K. (2020). Perceptual quality assessment for multimodal medical image fusion. *Signal Process.* 85:115852. doi: 10.1016/j.image.2020.115852

Wang, A., Luo, X., Zhang, Z., and Wu, X.-J. (2022). A disentangled representation based brain image fusion via group lasso penalty. *Front. Neurosci.* 16:937861.

Wang, J., Chen, P., Zheng, N., Chen, B., Principe, J. C., and Wang, F.-Y. (2021). Associations between MSE and SSIM as cost functions in linear decomposition with application to bit allocation for sparse coding. *Neurocomputing* 422, 139–149. doi: 10.1016/j.neucom.2020.10.018

Wang, S., Deng, C., Zhao, B., Huang, G.-B., and Wang, B. (2016). Gradient-based no-reference image blur assessment using extreme learning machine. *Neurocomputing* 174, 310–321. doi: 10.1016/j.neucom.2014.12.117

Wang, X., Jiang, Q., Shao, F., Gu, K., Zhai, G., and Yang, X. (2021). Exploiting local degradation characteristics and global statistical properties for blind quality assessment of tone-mapped HDR images. *IEEE Trans. Multimedia* 23, 692–705. doi: 10.1109/TMM.2020.2986583

Wang, Z., Cui, Z., and Zhu, Y. (2020). Multi-modal medical image fusion by Laplacian pyramid and adaptive sparse representation. *Comput. Biol. Med.* 123:103823. doi: 10.1016/j.compbiomed.2020.103823

Wang, Z., Li, X., Duan, H., and Zhang, X. (2022). A self-supervised residual feature learning model for multifocus image fusion. *IEEE Trans. Image Process.* 31, 4527–4542. doi: 10.1109/TIP.2022.3184250

Xu, H., and Ma, J. (2021). EMFusion_An unsupervised enhanced medical image fusion network. *Inform. Fus.* 76, 177–186. doi: 10.1016/j.inffus.2021.06.001

Xydeas, C. S., and Petrovic, V. S. (2000). "Objective pixel-level image fusion performance measure," in *Proceedings of SPIE - The International Society for Optical Engineering*, ed. B. V. Dasarathy (Bailrigg: Lancaster University), 89–98. doi: 10.1117/12.381668

Yang, H., Shi, P., Zhong, D., Pan, D., and Ying, Z. (2019). Blind image quality assessment of natural distorted image based on generative adversarial networks. *IEEE Access* 7, 179290–179303. doi: 10.1109/ACCESS.2019.2957235

Yue, G., Cheng, D., Zhou, T., Hou, J., Liu, W., Xu, L., et al. (2023). Perceptual quality assessment of enhanced colonoscopy images: A benchmark dataset and an objective method. *IEEE Trans. Circ. Syst. Video Technol.* 33, 5549–5561. doi: 10.1109/TCSVT.2023.3260212

Zhang, G., Nie, X., Liu, B., Yuan, H., Li, J., Sun, W., et al. (2023). A multimodal fusion method for Alzheimer's disease based on DCT convolutional sparse representation. *Front. Neurosci.* 16:1100812. doi: 10.3389/fnins.2022.1100812

Zhang, X. (2021). Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Trans. Patt. Anal. Mach. Intell.* 44, 4819–4838. doi: 10.1109/TPAMI.2021.3078906

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., and Zhang, L. (2020). IFCNN: A general image fusion framework based on convolutional neural network. *Inform. Fus.* 54, 99–118. doi: 10.1016/j.inffus.2019.07.011

Zhang, Z., Tian, S., Zou, W., Morin, L., and Zhang, L. (2023). EDDMF: An efficient deep discrepancy measuring framework for full-reference light field image quality assessment. *IEEE Trans. Image Process.* 32, 6426–6440. doi: 10.1109/TIP.2023.3329663

Zheng, Y., Essock, E. A., Hansen, B. C., and Haun, A. M. (2007). A new metric based on extended spatial frequency and its application to DWT based fusion algorithms. *Inform. Fus.* 8, 177–192. doi: 10.1016/j.inffus.2005.04.003

Zhou, T., Fu, H., Chen, G., Shen, J., and Shao, L. (2020). Hi-Net: Hybrid-fusion network for multi-modal MR image synthesis. *IEEE Trans. Med. Imaging* 39, 2772–2781. doi: 10.1109/TMI.2020.2975344

Zhou, T., Li, Q., Lu, H., Cheng, Q., and Zhang, X. (2023). GAN review models and medical image fusion applications. *Inform. Fus.* 91, 134–148. doi: 10.1016/j.inffus.2022.10.017

frontiers | Frontiers in Neuroscience

TYPE Methods
PUBLISHED 18 June 2024
DOI 10.3389/fnins.2024.1431033

OPEN ACCESS

EDITED BY
Lu Tang,
Xuzhou Medical University, China

REVIEWED BY
Chenyang Xu,
China University of Petroleum, China
Yuxin Li,
Jiangsu University, China

*CORRESPONDENCE
Chunyu Yang
✉ chunyuyang@cumt.edu.cn

RECEIVED 11 May 2024
ACCEPTED 04 June 2024
PUBLISHED 18 June 2024

CITATION
An Y, Yang C and Zhang S (2024) A lightweight network architecture for traffic sign recognition based on enhanced LeNet-5 network.
Front. Neurosci. 18:1431033.
doi: 10.3389/fnins.2024.1431033

A lightweight network architecture for traffic sign recognition based on enhanced LeNet-5 network

Yuan An[1,2], Chunyu Yang[1]* and Shuo Zhang[3]

[1]China University of Mining and Technology, Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, Xuzhou, China, [2]Xuzhou University of Technology, Jiangsu Province Key Laboratory of Intelligent Industry Control Technology, Xuzhou, China, [3]Beijing University of technology, Faculty of Information Technology, Beijing, China

As an important part of the unmanned driving system, the detection and recognition of traffic sign need to have the characteristics of excellent recognition accuracy, fast execution speed and easy deployment. Researchers have applied the techniques of machine learning, deep learning and image processing to traffic sign recognition successfully. Considering the hardware conditions of the terminal equipment in the unmanned driving system, in this research work, the goal was to achieve a convolutional neural network (CNN) architecture that is lightweight and easily implemented for an embedded application and with excellent recognition accuracy and execution speed. As a classical CNN architecture, LeNet-5 network model was chosen to be improved, including image preprocessing, improving spatial pool convolutional neural network, optimizing neurons, optimizing activation function, etc. The test experiment of the improved network architecture was carried out on German Traffic Sign Recognition Benchmark (GTSRB) database. The experimental results show that the improved network architecture can obtain higher recognition accuracy in a short interference time, and the algorithm loss is significantly reduced with the progress of training. At the same time, compared with other lightweight network models, this network architecture gives a good recognition result, with a recognition accuracy of 97.53%. The network structure is simple, the algorithm complexity is low, and it is suitable for all kinds of terminal equipment, which can have a wider application in unmanned driving system.

KEYWORDS

traffic sign identification, automatic driving, LeNet-5, optimize activation function, convolutional neural network, space pool

1 Introduction

With the rapid development of social economy and the arrival of the era of artificial intelligence, information and intelligence have become the focus of social attention, and automatic driving has become one of the current hot research fields. Automatic driving system is a comprehensive system integrating information detection, information communication and intelligent control technology (Chen et al., 2018; Zhang et al., 2019). It can realize the interaction and coordination among people, vehicles and roads, so as to effectively improve road traffic conditions and travel efficiency. The international

automotive industry has recognized two categories of autonomous driving classification standards: National Highway Traffic Safety Administration and the Institute of Automotive Engineers, of which the second classification standard is more widely used (Hur and Kang, 2020).

Automatic driving in the world is in the stage of partial automation. Even if it makes great progress in the laboratory, it still faces many difficulties and challenges, and there is still a long way to go before entering the market commercially. In automatic driving, traffic sign recognition is an important part. The traffic signs on the road are rich in information, including very important traffic information, which can provide necessary road guidance for intelligent driving. Accurately identifying traffic signs and issuing correct driving instructions can effectively reduce the possibility of traffic accidents and improve driving efficiency. Therefore, the research of traffic sign recognition has important research significance and practical value, and has gradually become a hot research topic in related fields. As an important part of the automatic driving system, the traffic sign detection system mainly uses the vehicle-mounted camera to shoot the scene of the traffic road to obtain the required data set information (Cuesta-Infante et al., 2020). The comprehensive application of computer image processing technology, artificial intelligence technology and big data technology to detect and identify traffic signs can provide effective traffic information for the control platform of automatic driving, so as to increase the reaction time of the automatic driving control system and improve the safety of automatic driving. However, there are many factors affecting traffic sign recognition on the actual road: complex backgrounds, light and dark weather, weather factors, occlusion damage, signs aging, fading, etc. These factors bring great challenges to traffic sign detection. At the same time, the performance of automatic driving terminal equipment is quite different, which puts forward higher requirements for the complexity of traffic sign detection algorithm. The network architecture of traffic sign recognition algorithm needs to meet the requirements of lightweight to adapt to the low-end automatic driving terminal equipment. Therefore, it is of great significance to design a lightweight traffic sign recognition network architecture with high precision, good real-time performance and convenient terminal deployment for the advancement of autonomous driving. When the vehicle is running at high speed on the road, the lightweight traffic sign recognition network can help the control platform to recognize the traffic sign in time. At the same time, the lightweight network architecture has low requirements on the memory and configuration of the terminal equipment. Therefore, the study of lightweight traffic sign recognition network in complex environment is still a key problem in the field of control.

2 Related work

At present, traffic sign detection methods mainly include template matching method, traditional machine learning method and deep learning method. The traffic sign detection method based on template matching uses the unique shape of traffic signs to match the features of the template. According to different signs indicating functions, each traffic sign has a special color and shape. Traditional recognition methods based on color and shape are widely used. In the literature (Lasota and Skoczylas, 2016; Sheikh et al., 2016; Huang and Hou, 2017; Song et al., 2017; Jain and Gianchandani, 2018; Rahmad et al., 2018) many color segmentation methods are used to implement the algorithm. The common point of this method is to use color threshold segmentation to obtain traffic signs in images. There are also other studies based on shape methods that are extensively utilized in traffic sign recognition, such as Hough transform (Moon and Lee, 2015; Onat and Ozdil, 2016; Sun et al., 2019) and angle detection (Lin and Sie, 2019; Flores-Calero et al., 2020; Ozturk et al., 2020). The generalized Hough transform has many applications, most of which are used to identify standard geometric shapes, such as circles, triangles, and rectangles. In addition, in terms of positioning symbols, the method of using color segmentation for rough estimation (Yakimov and Fursov, 2015; Filatov et al., 2017; Lee and Kim, 2018) is also commonly used, and the target is determined through preliminary information screening. Color segmentation and shape-based methods have a common feature, that is, they are sensitive to shadows, extreme weather conditions, crowded scenes and other external factors. This kind of algorithm has strict requirements on the characteristics of traffic signs, and it can only effectively detect a certain kind of traffic signs that match well with the template. This kind of algorithm has poor robustness and is more sensitive to the change of environmental factors. In the case of deformation or pollution of traffic signs, the detection accuracy drops sharply.

Based on the traditional machine learning method, by analyzing the characteristics of different traffic signs, the corresponding classifier is selected to classify the traffic signs. Wu et al. (2020) used the histogram of oriented gradient (HOG) feature to extract the edge information of the image and the local binary pattern (LBP) feature to extract the internal texture information of the image, then they fused the two extracted information with features, finally they used the extreme learning machine (ELM) classifier to classify traffic signs and tested them on the GTSRB data set, and the recognition accuracy was 92.88%. Ruta et al. (2010) used HOG features to extract features of traffic signs in images and input them into the support vector machine (SVM) classifier for training and classification. This method was tested on GTSRB data set, and the recognition accuracy was 95.68%. The detection accuracy of this method is greatly improved. However, due to the large amount of redundant information generated in the process of acquiring the region of interest, the detection speed of the algorithm is reduced. As a result, most of these algorithms cannot meet the real-time requirements of the actual scene.

Convolutional neural network is a kind of deep learning network widely used in the field of machine vision. Different from the traditional artificial neural network structure, it contains very special convolutional layer and pooling layer, which are combined through local connection and weight sharing, and use the back propagation algorithm to adjust the weight and bias adaptively. Therefore, it no longer relies too much on the prior knowledge and manual intervention of technical experts and scholars.

The traffic sign detection method based on convolutional neural network utilizes multi-layer deep learning network to independently learn and extract different features of traffic signs. This method can effectively reduce the subjective singleness of traditional methods in image feature extraction, and can solve the problem of insufficient semantic information of traditional

methods in feature extraction, and has been widely used in the field of intelligent transportation.

Krizhevsky et al. (2017) proposed the use of convolutional neural networks to identify targets, thus enabling the rapid development and application of convolutional neural networks in the field of target detection. In recent years, with the rapid development of deep learning (Silver et al., 2016, 2017; Moravík et al., 2017), The neural network model based on deep learning has received much attention because of its ability to capture the dynamic characteristics of traffic sign image data and obtain the best recognition effect. Traffic sign recognition methods based on various deep convolutional neural networks have achieved some results. In the literature (Qian et al., 2015), researchers proposed a method of traffic sign detection and recognition. This method uses the features of the multi-task convolutional neural network for training, so as to obtain the geographical indications of various traffic signs, and effectively determine the classification features. Some scholars (Zhu et al., 2018) use a fast neural network to extract the candidate regions of interest provided by the previous complete convolutional network, and then determine the target value through text detection. Literature (Yao et al., 2017) proposed an field-programmable gate array (FPGA)-based convolutional neural network module with better automatic recognition performance. Compared with the traditional convolutional neural network model, its performance on the hardware platform is better. The energy consumption in traffic sign recognition is smaller, and the accuracy is higher, but there are certain support requirements for the hardware platform. There are also some studies that have made some progress in semantic segmentation based on deep convolutional neural networks (Zhao et al., 2017) and object recognition (Liang and Zhang, 2015). In some specific complex scenes, convolutional neural networks can use high-level semantic information as a feature method to solve some tasks (such as occlusion and various targets). In the traditional traffic sign recognition method, the color and shape information are limited and the high-level semantics to define the direction of the target space are lacking, so it cannot meet the higher-level needs. In the literature (Zang et al., 2016), scholars use cascaded convolutional neural networks to recognize traffic signs and operate on the previously extracted selected regions, thus showing good performance. However, its application platform is relatively limited, and it is not an end-to-end traffic sign recognition. Hussain et al. (2018) proposed a CNN fast branching model, which highly mimics biological mechanisms to improve efficiency. In terms of accuracy, the performance is acceptable, and the potential application possibilities are greater, but the efficiency under time-sensitive conditions is worth exploring. Mehta et al. (2019) proposed a method for the classification of traffic signs based on deep convolutional networks. Good optimization results can be obtained through Adam optimizer, and softmax activation also has certain performance. However, the classification accuracy needs to be improved. In addition, some scholars have proposed convolutional neural network algorithms based on driving and multi-column types (Karaduman and Eren, 2017; Shi and Yu, 2018). These classification networks are getting deeper and deeper, their structure is becoming more and more perfect, and the effect is getting better and better.

With the gradual increase of algorithm complexity, the accuracy of traffic sign recognition has been greatly improved, but at the same time, the application of terminal setting is gradually limited. Although large-scale convolutional neural network models have superior performance, they also bring the problems of high memory consumption. How to apply convolutional neural network to unmanned mobile devices requires directly facing the two major problems of storage and speed. The lightweight model is concerned with designing more efficient network computing methods so as to reduce network parameters without losing too much accuracy. Therefore, how to design a lightweight high-performance target detection has quite high application value and scientific research significance. LeNet-5 is one of the classical convolutional neural networks, and has been widely used in the field of image recognition since it was proposed. Based on LeNet-5, this paper proposes a lightweight traffic sign recognition network that meets the requirement of intervention time, providing a reference for terminal network deployment.

3 Methods

3.1 Image data processing

Traffic sign image recognition is a challenging problem. There are many differences in color, shape and hieroglyphics of traffic signs, which makes the image recognition of traffic signs become an unbalanced multi-class recognition problem. Although some commercial recognition systems have been put into the market and many related research reports have been published, before the advent of the GTSRB dataset, there is still a lack of benchmark data to fairly evaluate different image recognition methods. The GTSRB dataset is a multi-class benchmark dataset for image classification, and the application algorithm needs to recognize a single image of traffic sign. The images in the GTSRB dataset come from images or videos obtained from onboard cameras, and each type of traffic sign appears only once (Stallkamp et al., 2012). The data set contains 43 traffic sign categories, totaling more than 50,000 images (Saadna and Behloul, 2017). Each traffic sign type contains between 210 and 2,250 images to train and test the algorithmic model's ability to recognize various types of traffic signs.

The training folder of the GTSRB dataset contains 39,209 images, and we used the remaining 12,630 images as the test set. For the training set and the verification set, we divided the images in the training folder according to the ratio of 8:2, that is, the training set was 31,433 images, and the verification set was 7,776 images. The annotation tags are stored in a *csv* file, including location tags and 43 types of traffic light tags, which are cropped and scaled to a fixed size by the location tags. The data format is shown in **Table 1**.

Since the data set is collected in the real environment, the training data set is very unbalanced due to weather conditions, light changes, occlusion, motion and other problems, and the image has changes such as blur, distortion, rotation, etc., so it is necessary to preprocess the data set to enhance the robustness of the model.

3.1.1 Image resizing

The image aspect ratio in the dataset used in this study ranges from 15 × 15 to 250 × 250 pixels. In order to be compatible

with neural networks, there must be a fixed image size (Sultana et al., 2020). It is worth noting that reducing the image size to a lower pixel value reduces the complexity of the model, but may also negatively affect the accuracy of the model and may reduce the classification performance of the algorithm. During the experiment, we tested our model with different image pixel sizes and found that 32 × 32 pixels provided the best trade-off between computational complexity and classification accuracy. So we resize the image to 32 × 32.

3.1.2 Color conversion

The image color information is more sensitive to the lighting conditions and the quality of the capturing equipment (Sudharshan and Raj, 2018). Meanwhile, the number of training parameters and training time of gray-scale images are reduced compared with color images (Bui et al., 2016; Sudharshan and Raj, 2018). In the process of data processing, considering that color is not the main feature of traffic sign recognition, the color image is converted into a grayscale image, and the single-channel image is iterated more quickly. The weighted average method is used to process the grayscale of the picture, and the three RGB components representing the picture are weighted and averaged with different weights according to their importance and other indicators. Considering the difference in sensitivity of the human eye to green and blue, the weighted average of the three components of RGB can obtain a more reasonable gray-scale image, as shown in formula (1).

$$\mathrm{F}(i,j) = 0.30\mathrm{R}(i,j) + 0.59\mathrm{G}(i,j) + 0.11\mathrm{B}(i,j) \tag{1}$$

3.1.3 Histogram equalization

Histogram equalization is a nonlinear stretching operation to redistribute image pixel values so that the number of pixels in a certain gray range is roughly the same. Histogram equalization helps to equalize image brightness distribution, which can increase image contrast and make image details clearer (Dhal et al., 2021). The image histogram equalization method is shown in formula (2).

$$S_k = T(r_k) = \sum_{j=0}^{k} P_r(r_j) = \sum_{j=0}^{k} \frac{n_j}{n} \tag{2}$$

Where, r_k is the gray level contained in the image, n_k represents the number of the k*th* gray level, S_k is the gray level output after calculating the mapping using the transformation function.

The picture is scaled to a size of 32 × 32, and then the color image is converted to grayscale, the image to be displayed is randomly selected, and the histogram of the balanced picture is obtained according to (1), as shown in **Figure 1**.

In **Figure 1**, it should be noticed that the download data set format provided by GTSRB is in *ppm* format, which needs to be converted to *jpg* format. The images in the data set have surrounding backgrounds. According to the cropped area, it is scaled to a "fixed size." Here it can be realized by MATLAB script, including how to read the label information of the image in *csv* and how to convert the *ppm* format to *jpg* format, and finally preprocess the obtained pictures.

3.1.4 Data normalization

Data normalization is essential to ensure a uniform distribution of input parameters (pixel values), which allows the network architecture to converge quickly during training. Standardize the input image and process the input features into a similar range, thus making the optimization of the cost function easier and faster (Akshata and Panda, 2019; Zaibi et al., 2021). In this project, the training set and test set are normalized to the range (−1, 1). The min-max standardization method is used to normalize the data. However, the min-max standardization method is to transform the

TABLE 1 Data format.

Data set	Number of pictures	Number of annotation images	Number of sign types	Size of signs	Source of pictures	Acquisition mechanism
GTSRB	133000–144769	51840	43	15 × 15–250 × 250	Germany	Prosilica GC 1380ch color camera

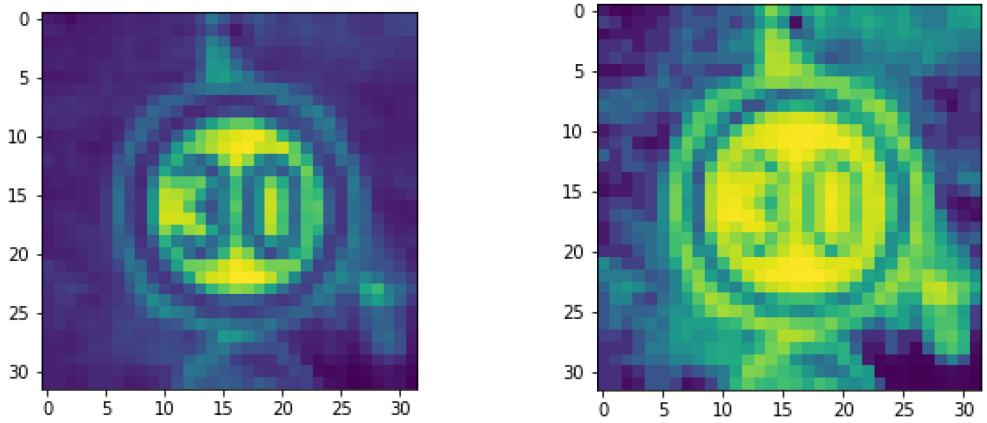


FIGURE 1
Comparison before and after histogram equalization.

original data into the [0, 1] interval, so we first make improvements, as shown in Formula (3).

$$x^* = \frac{x - x\_averge}{x_{max} - x_{min}} \quad (3)$$

As shown in formula (3), the data is processed, considering that in order to avoid recalculating the value each time, we set *x_averge* to a fixed value of 128. The normalized processing results are shown in Table 2.

3.1.5 Data enhancement

Data enhancement is an important step to solve the problem of unbalance of data sets. Various kinds of transformation processing are carried out on existing data to generate new data to expand the amount of data (Patil et al., 2021; Singh and Malik, 2022). Translation, scaling and rotation are commonly used transformation means. At the same time, it can ensure that there are no identical traffic sign images in the data set, so the robustness of the model is improved.

In the process of traffic sign recognition training, the original data is trained, and the results show that the accuracy of the training set is high, while the accuracy of the verification set is low, which is manifested as overfitting. When over-fitting occurs, the high accuracy of the training set indicates that the algorithm has fully learned the features of the original data, while the accuracy of the validation set is low, indicating that the characteristics of the original data are insufficient, which makes the algorithm in the new validation set the performance is poor (Akshata and Panda, 2019; Radu et al., 2020; Patil et al., 2021). In the actual scene, the angle of the traffic sign changes. In this case, model training will make the network convergence speed relatively slow, and the model effect is relatively poor. This paper adopts Keras-image-data-augmentation, a lightweight library based on Keras, for Image Data enhancement. We can set rotation Angle, translation distance, scaling ratio, etc., to simulate images under different perspectives. At the same time, in the training process, each iteration of the system will produce a new, randomly transformed image, which can avoid overfitting the model to a specific data pattern.

3.2 Model structure

LeNet-5 is one of the classical convolutional neural networks. LeNet-5 model has a good performance in the field of digit symbol recognition, and it is also helpful for the development of traffic sign recognition. However, due to its limitations, it usually performs poorly. Therefore, the LeNet-5 network model is improved in this paper to achieve better performance in symbol recognition scenarios. The model originally used the S-type activation function, but as the value tends to infinity, the images on the left and right sides of the function tend to be flat, and the gradient value gradually tends to 0, which can easily cause the gradient to disappear and slow down the convergence speed of the model. In order to improve the sigmoid function, this inherent limitation causes the accuracy of the model to decrease. In terms of the pool layer, the LeNet-5 network model originally used the average pool, but the average pool averaged the values within the filter range to obtain the output. In the field of image recognition, max pooling is used more frequently, and the maximum value in the filter range is used as the output, which has strong robustness.

The adaptive convolutional neural network model consists of an input layer, an output layer, three convolutional layers (C1, C3, C5), three pooling layers (S2, S4, S6) and a fully connected layer (F7). Input layer: The input is a sample image with a size of 32 × 32 pixels. Convolutional layer C1: the input layer is convolved using thirty-two convolution cores of size 3 × 3 with a step size of 1. A convolution kernel will obtain a feature map, so this layer consists of thirty-two feature maps. Pooling layer S2: S2 adopts a maximum pooling strategy, which is obtained after down-sampling at the C1 layer. The size of the pooling area in S2 is 2 × 2, and the step size is 1. Convolutional layer C3: sixty-four convolution cores of size 3 × 3 with a step size of 1. Each feature map in C3 is a weighted combination of all thirty-two or more feature maps in S2. The output is sixty-four 13 × 13 feature maps. Pooling layer S4: S2 adopts a maximum pooling strategy, which is obtained after down-sampling at the C1 layer. The size of the pooling area in S2 is 2 × 2, and the step size is 1. Convolutional layer C5: sixty-four convolution cores of size 3 × 3 with a step size of 1. A convolution kernel will obtain a feature map, so this layer is composed of sixty-four feature maps. Pooling layer S6: S2 adopts a maximum pooling strategy, which is obtained after down-sampling at the C1 layer. The size of the pooling area in S2 is 2 × 2, and the step size is 1. Output layer: fully connected layer, the logo corresponds to the output image. The adaptive network structure is shown in Figure 2.

As shown in Figure 2, the reconstruction of the pooling layer is conducive to the rapid recognition of traffic signs in the intelligent driving environment, and at the same time can improve the accuracy. The model structure table is shown in Table 3.

TABLE 2 Normalization result.

i	array[i]	i	array[i]	i	array[i]	i	array[i]
1	−0.859375	9	−0.625	17	0.1171875	25	0.0078125
2	0.9609375	10	−0.625	18	0.171875	26	0.125
3	0.9765625	11	−0.375	19	0.3359375	27	0.1171875
4	0.9296875	12	−0.375	20	0.171875	28	−0.1875
5	−0.859375	13	0.2890625	21	0.0546875	29	−0.1875
6	−0.859375	14	0.171875	22	0.125	30	0.0625
7	−0.625	15	0.6171875	23	0.125	31	0.125
8	−0.484375	16	0.3359375	24	0.0078125	32	0.171875

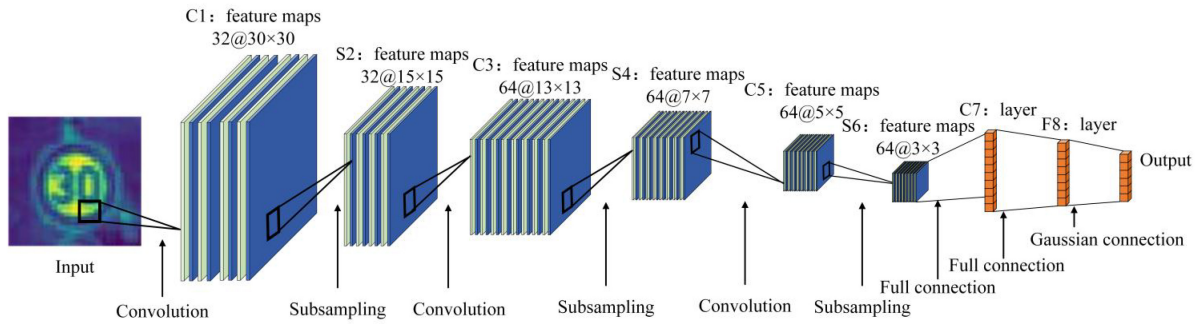


FIGURE 2
Adaptive network structure diagram.

3.3 Activation function

Activation function is a nonlinear mapping of the predicted results, which can improve the resolution of the model, so that the model has the ability to deal with complex problems and high learning ability. In convolutional neural networks, activation functions including Sigmoid, Tanh, ReLU, Leaky-Relu, ELU, and Maxout are frequently utilized.

Through the analysis of the activation function, the Sigmoid function and the Tanh function are saturated nonlinear functions, and the convergence speed is slow, which is easy to cause gradient explosion or gradient small phenomenon. The ReLU function does not have the problem of gradient disappearance, and there will be no saturation problem. Therefore, ReLU can maintain the gradient without attenuation, which alleviates the problem of gradient disappearing, so that we can directly train deep learning neural networks in a supervised manner, without relying on unsupervised layer-by-layer pre-training. However, with the development of training, the "vulnerability" of the ReLU function has gradually become apparent. At that time, the derivative of the function is always 0, which prevents false responses.

TABLE 3 Model structure.

Layer (type)	Output shape	Param
conv2d (Conv2D)	(None, 30, 30, 32)	320
p_re_lu (PReLU)	(None, 30, 30, 32)	28800
max_pooling2d (MaxPooling2D)	(None, 15, 15, 32)	0
dropout (Dropout)	(None, 15, 15, 32)	0
conv2d_1 (Conv2D)	(None, 13, 13, 64)	18496
p_re_lu_1 (PReLU)	(None, 13, 13, 64)	10816
max_pooling2d_1 (MaxPooling2D)	(None, 7, 7, 64)	0
conv2d_2 (Conv2D)	(None, 5, 5, 64)	36928
p_re_lu_2 (PReLU)	(None, 5, 5, 64)	1600
max_pooling2d_1 (MaxPooling2D)	(None, 3, 3, 64)	0
flatten (Flatten)	(None, 576)	0
dense (Dense)	(None, 512)	295424
dropout_1 (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 43)	22059

The Leaky-ReLU function is an improvement of the ReLU function when the gradient is a range index. In order to solve the problem of regional neuron disappearance, the Leaky-ReLU function only replaces horizontal lines with non-horizontal lines. In the specific direction propagation process of the model, input the Leaky-ReLU activation function, the part less than zero can prevent the neurons in the area from becoming dead neurons, and the gradient can also be calculated. However, according to different values, the role of the Leaky-ReLU function is also different, and its function is shown in formula (4).

$$f(x) = \begin{cases} x_i, & x \geq 0 \\ \alpha_i x_i, & x < 0 \end{cases} \tag{4}$$

As shown in formula (4), the PReLU function is an improvement of the Leaky-ReLU function. In the PReLU function, α is a trainable function, and the neural network will also learn the value of α to achieve faster and better convergence. The nonlinear activation input on the channel is a coefficient that controls the slope of the negative part, which allows the nonlinear activation function to have different values on different channels, and the PReLU degenerates to ReLU. When the value is small and fixed, PReLU will degenerate into LReLU, and PReLU will only increase a very small number of parameters. Compared with the total number of parameters, these additional parameters can be ignored, so this also means that the risk of overfitting will only increase a little. Especially when different channels use the same α, there are fewer parameters. PReLU can be trained in the back propagation process at the same time, and can be optimized together with other layers. The update formula is derived from the chain rule, and the gradient of each layer is shown in formula (5).

$$\frac{\partial \varepsilon}{\partial \alpha_i} = \sum_{y_i} \frac{\partial \varepsilon}{\partial f(y_i)} \frac{\partial f(y_i)}{\partial \alpha_i} \tag{5}$$

As shown in formula (5), εrepresents the objective function, $\frac{\partial \varepsilon}{\partial f(y_i)}$ is the gradient propagated from a deeper layer, and its activation gradient is shown in formula (6).

$$\frac{\partial f(y_i)}{\partial \alpha_i} = \begin{cases} 0, & y_i > 0 \\ y_i, & y_i \leq 0 \end{cases} \tag{6}$$

Acting on all element maps, for channel shared variables, the gradient is shown in formula (7).

$$\frac{\partial \varepsilon}{\partial \alpha} = \sum_i \sum_{y_i} \frac{\partial \varepsilon}{\partial f(y_i)} \frac{\partial f(y_i)}{\partial \alpha_i} \tag{7}$$

As shown in formula (7), $\sum y_i$ is the total value of all channels. When the model updates α, the update method using the time driving quantity is shown in formula (8).

$$\Delta \alpha_i := \mu \Delta \alpha_i + \delta \frac{\partial \varepsilon}{\partial \alpha_i} \tag{8}$$

As shown in formula (8), μ represents momentum and δ is the learning rate.

4 Evaluation of result

In order to verify the superiority of the improvement of the PReLU activation function, by using the GTSRB dataset, the PReLU improvement function is compared with the commonly used activation functions such as Sigmoid, Tanh, SELU, ReLU, and Leaky-ReLU. We use different activation functions in the model and compare them during the 100% training process to ensure the fairness and rationality of the results. We have selected the training progress after 0%, 20%, 50%, 75%, and 100%, and their corresponding loss index and precision index are compared. As the training progress increases, the loss and accuracy of each functional model are shown in Tables 4–9.

It can be seen from Table 4 that when the training progress is 0%, the performance of the improvement function of PReLU is not particularly ideal, and the performance on the test set is only better than the activation functions of Sigmoid and Leaky-ReLU. The results of training progress of 25% are shown in Table 5.

In the intermediate stage of the training progress, we choose the intermediate state of 50% to view the performance of each activation function, as shown in Table 6.

TABLE 4 Comparison of each function loss and accuracy when training progress is 0%.

Index		Function					
		Sigmoid	Tanh	SELU	ReLU	Leaky-ReLU	PReLU
Train	Loss	3.6711	0.8787	0.7477	1.1660	1.1496	1.7159
	Accuracy	0.0498	0.7611	0.7914	0.6744	0.6803	0.5110
Test	val_loss	3.1683	0.5052	0.5829	0.6142	0.7049	0.7237
	val_accuracy	0.1930	0.8358	0.8308	0.8054	0.7755	0.7773

TABLE 5 Comparison of each function loss and accuracy when training progress is 25%.

Index		Function					
		Sigmoid	Tanh	SELU	ReLU	Leaky-ReLU	PReLU
Train	Loss	0.0094	5.42e-05	1.3126e-05	0.0016	0.0090	0.0126
	Accuracy	0.9981	1.0000	1.0000	0.9995	0.9973	0.9954
Test	val_loss	0.5266	0.2732	0.3844	0.3384	0.3882	0.1504
	val_accuracy	0.8778	0.9413	0.9268	0.9390	0.9315	0.9633

TABLE 6 Comparison of each function loss and accuracy when training progress is 50%.

Index		Function					
		Sigmoid	Tanh	SELU	ReLU	Leaky-RELU	PReLU
Train	Loss	1.6920e-04	4.1051e-06	9.4398e-07	1.111e-06	1.2186e-06	0.0078
	Accuracy	1.0000	1.0000	1.0000	1.0000	1.0000	0.9977
Test	val_loss	0.6174	0.3036	0.4442	0.3687	0.4419	0.1906
	val_accuracy	0.8918	0.9433	0.9293	0.9449	0.9420	0.9653

TABLE 7 Comparison of each function loss and accuracy when training progress is 75%.

Index		Function					
		Sigmoid	Tanh	SELU	ReLU	Leaky-ReLU	PReLU
Train	Loss	4.6383e-05	2.9061e-07	6.6417e-08	9.264e-08	9.1472e-08	0.0062
	Accuracy	1.0000	1.0000	1.0000	1.0000	1.0000	0.9982
Test	val_loss	0.6762	0.3513	0.5186	0.4361	0.5251	0.1598
	val_accuracy	0.8902	0.9424	0.9286	0.9454	0.9426	0.9732

TABLE 8 Comparison of each function loss and accuracy when training progress is 100%.

Index		Function					
		Sigmoid	Tanh	SELU	ReLU	Leaky-ReLU	PReLU
Train	Loss	6.5922e-06	3.0431e-08	6.1023e-09	8.242e-09	7.6751e-09	0.0052
	Accuracy	1.0000	1.0000	1.0000	1.0000	1.0000	0.9986
Test	val_loss	0.7304	0.3728	0.5822	0.4892	0.5959	0.2095
	val_accuracy	0.8943	0.9444	0.9295	0.9449	0.9415	0.9753

TABLE 9 Comparison of results on the test set before and after training.

Index	Function					
	Sigmoid	Tanh	SELU	ReLU	Leaky-ReLU	PReLU
Before training	0.1930	0.8358	0.8308	0.8054	0.7755	0.7773
After training	0.8902	0.9444	0.9295	0.9449	0.9415	0.9753

It is not difficult to see from Table 6 that when the training progress reaches 50%, the accuracy of other activation functions on the training set even reached 100%. Similarly, the result of PReLU on the training set also reached 0.9977, but the accuracy of the other activation functions on the test set is still slightly lower than the accuracy of the PReLU activation function (0.9653). Select 75% progress status to view the performance of each activation function, as shown in Table 7.

Select 100% progress status to view the performance of each activation function, as shown in Table 8.

In order to verify the role of the activation function, and also to compare the performance of each activation function, we compare the state before and after training, and the results are shown in Table 9 below.

In Table 9, after 100% training, the performance of the PReLU improvement function is significantly better than other activation functions, verifying the effectiveness of the model. It can be seen from the above table comparison that when the model is not trained, that is, when the training progress is 0%, the effects of the activation functions of Tanh, SELU, and ReLU are better than those of PReLU improvement. However, as the training progress increases, the accuracy of the PReLU improvement function gradually increases. When the training progress is only 25%, it can be seen that the accuracy of the PReLU improvement function exceeds that of other activation functions under normal conditions. In order to compare PReLU with other activation functions, we select the model of one of the functions, and filter the images whose accuracy and loss change as the number of epochs increases, as shown in Figure 3.

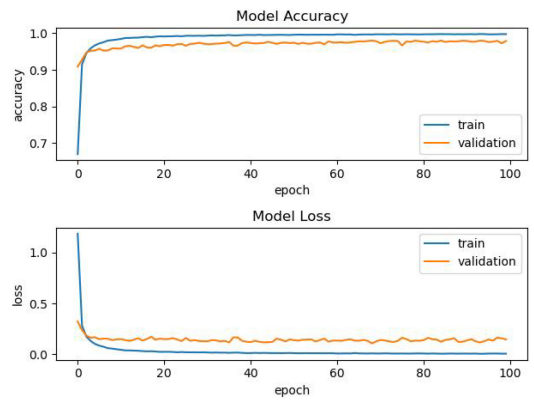


FIGURE 3
Multiple comparison char.

It can be seen from Figure 4 that as the training progress increases, the model using other activation functions has a relatively stable performance, but after the training, we can see that the accuracy is insufficient. The comparison of another model is shown in Figure 4.

As shown in Figure 4, as the number of epochs increases, the accuracy of using the PReLU activation function model continues

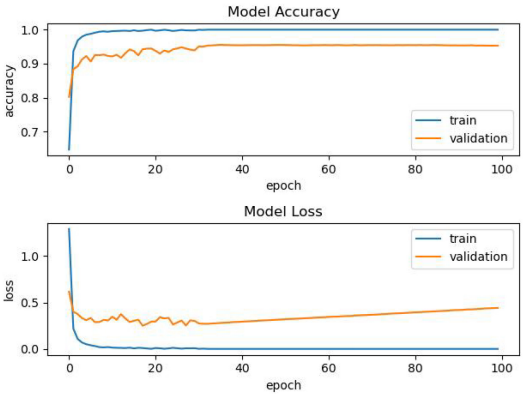


FIGURE 4
Comparison chart.

TABLE 10 Comparison of recognition performance of different algorithms on GTSRB data sets.

Algorithm	Recognition accuracy
Fast R-CNN	90.1%
Faster R-CNN	91.8%
Traditional LeNet-5 network (Zhang et al., 2021)	95.48%
Optimized LeNet-5 network	97.53%

to improve, and the overall effect is better than that of models using other activation functions.

In order to verify the recognition performance of the optimized LeNet-5 network, we compared the recognition accuracy with other traffic sign recognition methods on the GTSRB data set, and the comparison results are shown in Table 10.

It can be seen from Table 10, compared with three typical deep learning networks (Fast R-CNN, Faster R-CNN and traditional LeNet-5 network), the optimized LeNet-5 network can extract more effective features from the dataset, and the identification accuracy is significantly improved, which can meet the requirements of automatic driving. At the same time, compared with the large complex network structure, the optimized network proposed in this paper reduces the number of neurons through weight sharing and local receptive field, achieving the purpose of reducing training parameters and improving training speed, thus greatly shortening the time of feature extraction and recognition, and making it possible to recognize traffic signs in real-time monitoring. The optimized network proposed in this paper reflects the good performance of the network model because of its simple structure and lower requirements for terminal equipment, and can be well applied in the field of traffic sign recognition.

5 Conclusion

Aiming at the requirements of intelligent driving for traffic sign recognition, a light traffic sign recognition network based on neural network is proposed. By improving the spatial pool convolutional neural network, the neuron nodes are optimized, and the normalized image is preprocessed. And activation, optimize the activation function to improve the recognition effect. The experimental results show that this method has a high recognition rate, especially after the improvement of the activation function, the recognition accuracy has also improved. Due to its simple structure and low requirements on terminal device memory and configuration, the network model can be effectively applied to intelligent driving scenarios. At the same time, it has a reliable recognition effect in traffic sign recognition, improving the quality and safety of intelligent driving.

Data availability statement

The original contributions presented in this study are included in this article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YA: Writing – original draft. CY: Writing – review & editing. SZ: Writing – original draft.

Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akshata, V. S., and Panda, S. (2019). Traffic sign recognition and classification using convolutional neural networks. *J. Emerg. Technol. Innov. Res.* 6, 186–204.

Bui, H. M., Lech, M., Cheng, E., Neville, K., and Burnett, I. S. (2016). "Using grayscale images for object recognition with convolutional-recursive neural network," in *Proceedings of the 2016 IEEE 6th international conference on communications and electronics (ICCE)*, (Coimbatore), 321–325. doi: 10.1109/CCE.2016.7562656

Chen, L., Jiang, D. H., Song, H. B., Wang, P., Bao, R., Zhang, K. L., et al. (2018). A lightweight end-side user experience data collection system for quality evaluation of multimedia communication. *IEEE Access.* 6, 15408–15419. doi: 10.1109/ACCESS.2018.2794354

Cuesta-Infante, A., García, F. J., Pantrigo, J. J., and Montemayor, A. S. (2020). Pedestrian detection with LeNet-like convolutional networks. *Neural Comput. Appl.* 32, 13175–13181. doi: 10.1007/s00521-017-3197-z

Dhal, K. G., Das, A., Ray, S., Gálvez, J., and Das, S. (2021). Histogram equalization variants as optimization problems: A review. *Arch. Comput. Methods Eng.* 28, 1471–1496. doi: 10.1007/s11831-020-09425-1

Filatov, D. M., Ignatiev, K. V., and Serykh, E. V. (2017). "Neural network system of traffic signs recognition," in *Proceedings of the 20th IEEE international conference on soft computing and measurements (SCM)*, (Petersburg), doi: 10.1109/SCM.2017.7970605

Flores-Calero, M., Espinel, G., Carrillo-Medina, J., Vizcaino, P., Gualsaqui, M., and Ayala, M. J. (2020). *Ecuadorian traffic sign detection through color information and a convolutional neural network*. New York, NY: IEEE, doi: 10.1109/ANDESCON50619.2020.9272089

Huang, H., and Hou, L. Y. (2017). "Speed limit sign detection based on Gaussian color model and template matching," in *Proceedings of the 2017 international conference on vision, image and signal processing*, (Kunming), 118–122. doi: 10.1109/ICVISP.2017.30

Hur, C., and Kang, S. (2020). On-device partial learning technique of convolutional neural network for new classes. *J. Signal Process. Syst.* 95, 909–920. doi: 10.1007/s11265-020-01520-7

Hussain, S., Abualkibash, M., and Tout, S. (2018). "A survey of traffic sign recognition systems based on convolutional neural networks," in *Proceedings of the*

IEEE international conference on electro/information technology (EIT), (Rochester, MI), doi: 10.1109/EIT.2018.8500182

Jain, R., and Gianchandani, D. (2018). "A Hybrid approach for detection and recognition of traffic text sign using MSER and OCR," in *Proceedings of the 2018 2nd international conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, (Kirtipur), doi: 10.1109/I-SMAC.2018.8653761

Karaduman, M., and Eren, H. (2017). "Deep learning based traffic direction sign detection and determining driving style," in *Proceedings of the 2017 international conference on computer science and engineering (UBMK)*, (Diyarbakir), 1046–1050. doi: 10.1109/UBMK.2017.8093453

Krizhevsky, A., Sutskever, I. G., and Hinton, E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Lasota, M., and Skoczylas, M. (2016). *Recognition of multiple traffic signs using keypoints feature detectors*. Piscataway, NJ: IEEE, 534–540. doi: 10.1109/ICEPE.2016.7781

Lee, H. S., and Kim, K. (2018). Simultaneous traffic sign detection and boundary estimation using convolutional neural network. *IEEE Trans. Intell. Transport. Syst.* 19, 1652–1663. doi: 10.1109/TITS.2018.2801560

Liang, J., and Zhang, Y. N. (2015). "Top down saliency detection via Kullback–Leibler divergence for object recognition," in *Proceedings of the international symposium on bioelectronics and bioinformatics (ISBB)*, (Mohali), doi: 10.1109/ISBB.2015.7344958

Lin, C. H., and Sie, Y. S. (2019). "Two-stage license plate recognition system using deep learning," in. *Proceedings of the 8th international conference on innovation, communication and engineering (ICICE)*, (Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc), doi: 10.1109/ICICE49024.2019.9117277

Mehta, S., Paunwala, C., and Vaidya, B. (2019). "CNN based traffic sign classification using adam optimizer," in *Proceedings of the 2019 international conference on intelligent computing and control systems (ICCS)*, (New York, NY), doi: 10.1109/ICCS45141.2019.9065537

Moon, J., and Lee, C. (2015). "Efficient implementation of computing unit for Hough transform," in *Proceedings of the 2015 international SoC design conference (ISOCC)*, (Sapporo), 279–280. doi: 10.1109/isocc.2015.7401756

Moravík, M., Schmid, M., Burch, N., Lis, V., Morrill, D., Bard, N., et al. (2017). DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* 356, 508–513. doi: 10.1126/science.aam6960

Onat, E., and Ozdil, O. (2016). "Traffic sign classification using hough transform and SVM," in *Proceedings of the 23nd signal processing and communications applications conference (SIU)*, (Pune), doi: 10.1109/SIU.2015.7130301

Ozturk, G., Koker, R., Eldogan, O., and Karayel, D. (2020). "Recognition of vehicles, pedestrians and traffic signs using convolutional neural networks," in *Proceedings of the 4th international symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, (Berlin), doi: 10.1109/ISMSIT50672.2020.9255148

Patil, D., Poojari, A., Choudhary, J., and Gaglani, S. (2021). CNN based traffic sign detection and recognition on real time video. *Int. J. Eng. Res. Technol.* 9, 422–426.

Qian, R. Q., Zhang, B., Yue, Y., Wang, Z., Coenen, F., and Yue, Y. (2015). "Robust Chinese traffic sign detection and recognition with deep convolutional neural network," in *Proceedings of the 11th international conference on natural computation (ICNC)*, (Diyarbakir), doi: 10.1109/ICNC.2015.7378092

Radu, M. D., Costea, I. M., and Stan, V. A. (2020). "Automatic traffic sign recognition artificial intelligence deep learning algorithm," in *Proceedings of The 12th international conference on electronics, computers and artificial intelligence (ECAI)*, (New York, NY), 1–4. doi: 10.1109/ ECAI50035.2020.9223186

Rahmad, C., Rahmah, I. F., Asmara, R. A., and Adhisuwignjo, S. (2018). "Indonesian traffic sign detection and recognition using color and texture feature extraction and SVM classifier," in *Proceedings of the 2018 international conference on information and communications technology*, (Berlin), doi: 10.1109/ICOIACT.2018.8350804

Ruta, A., Li, Y. M., and Liu, X. H. (2010). Real-time traffic sign recognition from video by class- specific discriminative features. *Pattern Recogn.* 43, 416–430. doi: 10.1016/j.patcog.2009.05.018

Saadna, Y., and Behloul, A. (2017). An overview of traffic sign detection and classification methods. *Int. J. Multimed. Inform. Retriev.* 6, 193–210. doi: 10.1007/s13735-017-0129-8

Sheikh, M. A. A., Kole, A., and Maity, T. (2016). "Traffic sign detection and classification using colour feature and neural network," in *Proceedings of the 2016 international conference on intelligent control power and instrumentation (ICICPI)*, (Kolkata), 307–311. doi: 10.1109/ICICPI.2016.7859723

Shi, Y. L., and Yu, Z. H. (2018). "Multi-column convolution neural network model based on adaptive enhancement," in *Proceedings of the 3rd international conference on intelligent transportation, big data and smart city (ICITBS)*, (Berlin), doi: 10.1109/ICITBS.2018.00177

Silver, D., Huang, A., Guez, A., and Sifre L. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., and Antonoglou, I. (2017). Mastering the game of go without human knowledge. *Nature* 550, 354–359. doi: 10.1038/nature24270

Singh, K., and Malik, N. (2022). CNN based approach for traffic sign recognition system. *Adv. J. Graduate Res.* 11, 23–33. doi: 10.21467/ ajgr.11.1.23-33

Song, L., Liu, Z., Duan, H., and Liu, N. (2017). "A color-based image segmentation approach for traffic scene understanding," in *Proceedings of the 13th international conference on semantics, knowledge and grids*, (Beijing), 33–37. doi: 10.1109/SKG.2017.00014

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. (2012). Man vs. Computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Netw.* 32, 323–332. doi: 10.1016/j.neunet.2012.02.016

Sudharshan, D. P., and Raj, S. (2018). "Object recognition in images using convolutional neural network," in *Proceedings of the 2nd international conference on inventive systems and control (ICISC)*, (Coimbatore), 718–722. doi: 10.1109/ICISC.2018.8398893

Sultana, F., Sufan, A., and Dutta, P. (2020). A review of object detection models based on convolutional neural network. *Intell. Comput. Image Process. Based Appl.* 1157, 1–16. doi: 10.1007/978-15-4288-6_1

Sun, Y., Ge, P. S., and Liu, D. Q. (2019). "Traffic sign detection and recognition based on convolutional neural network," in *Proceedings of the 2019 Chinese automation congress (CAC)*, (New York, NY: IEEE), 2851–2854. doi: 10.1109/CAC48633.2019.8997240

Wu, D., Qu, Z., Zhao, K. Y., Guo, F. J., Li, T., and He, W. (2020). Traffic sign recognition based on HOGv-CLBP feature fusion and ELM. *J. Optoelectr. Laser* 31, 621–627. doi: 10.16136/j.joel.2020.06.0460

Yakimov, P., and Fursov, V. (2015). "Traffic signs detection and tracking using modified hough transform," in *Proceedings of the 12th international joint conference on e-business and telecommunications (ICETE)*, (Kolkata), 22–28. doi: 10.5220/0005543200220028

Yao, Y. C., Zhang, Z. Q., Yang, Z., Wang, J., and Lai, J. M. (2017). "FPGA-based convolution neural network for traffic sign recognition," in *Proceedings of the 12th international conference on ASIC (ASICON)*, (Visakhapatnam), doi: 10.1109/ASICON.2017.8252620

Zaibi, A., Ladgham, A., and Sakly, A. (2021). A lightweight model for traffic sign classification based on enhanced LeNet-5 network. *J. Sens.* 2021, :8870529. doi: 10.1155/ 2021/ 8870529

Zang, D., Bao, M. M., Zhang, J. Q., Cheng, J. J., Zhang, D. D., and Tang, K. S. (2016). "Traffic sign detection based on cascaded convolutional neural networks," in *Proceedings of the 17th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD)*, (Pune), doi: 10.1109/SNPD.2016.7515901

Zhang, K. L., Chen, L., An, Y., and Cui, P. (2019). A QoE test system for vehicular voice cloud services. *Mobile Netw. Appl.* 26, 700–715. doi: 10.1007/s11036-019-01415-3

Zhang, K., Hou, J., Liu, M. Y., and Liu, J. Y. (2021). Traffic sign recognition based on improved convolutional networks. *Int. J. Wireless Mobile Comput.* 21, 274–284. doi: 10.1504/IJWMC.2021.120910

Zhao, B., Feng, J., Wu, X., and Yan, S. (2017). A survey on deep learning based fine-grained object classification and semantic segmentation. *Int. J. Autom. Comput.* 14, 119–135. doi: 10.1007/s11633-017-1053-3

Zhu, Y., Liao, M., Liu, W., and Yang, M. (2018). Cascaded segmentation detection networks for text-based traffic sign detection. *IEEE Trans. Intell. Transport. Syst.* 19, 209–219. doi: 10.1109/TITS.2017.2768827

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 15 July 2024
DOI 10.3389/fnins.2024.1448294

OPEN ACCESS

EDITED BY
Hancheng Zhu,
China University of Mining and Technology, China

REVIEWED BY
Bo Hu,
Chongqing University of Posts and Telecommunications, China
Pengfei Chen,
Xidian University, China

*CORRESPONDENCE
Guanyu Zhu
100002023042@xzhmu.edu.cn
Liang Li
liliang@xzhmu.edu.cn

†These authors have contributed equally to this work

RECEIVED 13 June 2024
ACCEPTED 03 July 2024
PUBLISHED 15 July 2024

CITATION
Zhou Q, Zhou Y, Hou N, Zhang Y, Zhu G and Li L (2024) DFA-UNet: dual-stream feature-fusion attention U-Net for lymph node segmentation in lung cancer diagnosis. *Front. Neurosci.* 18:1448294. doi: 10.3389/fnins.2024.1448294

DFA-UNet: dual-stream feature-fusion attention U-Net for lymph node segmentation in lung cancer diagnosis

Qi Zhou[1,2†], Yingwen Zhou[2†], Nailong Hou[2], Yaxuan Zhang[2], Guanyu Zhu[2*] and Liang Li[1*]

[1]Department of Radiotherapy, The Affiliated Hospital of Xuzhou Medical University, Xuzhou, China, [2]School of Medical Imaging, Xuzhou Medical University, Xuzhou, China

In bronchial ultrasound elastography, accurately segmenting mediastinal lymph nodes is of great significance for diagnosing whether lung cancer has metastasized. However, due to the ill-defined margin of ultrasound images and the complexity of lymph node structure, accurate segmentation of fine contours is still challenging. Therefore, we propose a dual-stream feature-fusion attention U-Net (DFA-UNet). Firstly, a dual-stream encoder (DSE) is designed by combining ConvNext with a lightweight vision transformer (ViT) to extract the local information and global information of images; Secondly, we propose a hybrid attention module (HAM) at the bottleneck, which incorporates spatial and channel attention to optimize the features transmission process by optimizing high-dimensional features at the bottom of the network. Finally, the feature-enhanced residual decoder (FRD) is developed to improve the fusion of features obtained from the encoder and decoder, ensuring a more comprehensive integration. Extensive experiments on the ultrasound elasticity image dataset show the superiority of our DFA-UNet over 9 state-of-the-art image segmentation models. Additionally, visual analysis, ablation studies, and generalization assessments highlight the significant enhancement effects of DFA-UNet. Comprehensive experiments confirm the excellent segmentation effectiveness of the DFA-UNet combined attention mechanism for ultrasound images, underscoring its important significance for future research on medical images.

KEYWORDS

ultrasound elastography, mediastinal lymph nodes, semantic segmentation, attention mechanism, deep learning

1 Introduction

Lung cancer is one of the malignant tumors with the highest morbidity and mortality rates worldwide (Detterbeck et al., 2016; Siegel et al., 2023). The choice of treatment is closely related to cancer staging, determining whether the lymph nodes are involved is one of the key factors in clarifying the cancer staging (Asamura et al., 2015; Taylor et al., 2023). Numerous studies (Gu et al., 2017; Wang et al., 2018; Zhang et al., 2019; Wang B. et al., 2021; Wang R. et al., 2021) have demonstrated that compared with traditional ultrasound imaging, bronchial ultrasound elastography (BUE) can provide more accurate information on mediastinal lymph nodes,

reflecting the hardness information of lymph node tissues with different colors, which has a higher diagnostic value (Oglat and Abukhalil, 2024).

Ultrasound elastography (UE) is a novel ultrasound diagnostic technology that has rapidly developed in recent years. It utilizes dynamic imaging to measure tissue hardness (Zhang et al., 2019; Cui et al., 2022), allowing for non-invasive diagnosis of diseased tissues by analyzing the differences in hardness between various tissues. Currently, most UE used in endoscopy employs strain force elastography. This technique operates on the principle that softer and harder tissues deform differently under the same external force (Sigrist et al., 2017). Generally, tissues with lower elasticity coefficients exhibit greater displacement and deformation, appearing green; tissues with higher elasticity coefficients exhibit less displacement, appearing blue; and tissues with intermediate hardness appear reddish-blue or reddish-green. Since malignant lymph nodes are harder than benign ones, assessing the hardness of a lesion by measuring the proportion of the blue area within it can help identify benign and malignant lesions (Sun et al., 2017). Therefore, accurate localization and segmentation of mediastinal lymph nodes based on BUE images are crucial steps in lung cancer diagnosis and treatment (Wang B. et al., 2021; Wang R. et al., 2021).

Currently, professional doctors are typically required to manually segment lymph nodes in BUE images. This process is not only time-consuming and labor-intensive but also subject to inter-individual differences among doctors, leading to subjective biases and potential omission of important features. Consequently, the same image can result in varying analyses and evaluations, causing segmentation errors. Therefore, developing automatic segmentation methods for lymph nodes in UE images is of great significance (Li and Xia, 2020; Tan et al., 2023).

With the continuous development of computer vision technology, the application of semantic segmentation in medical images has become increasingly important. Combining artificial intelligence with medical imaging to enable intelligent-assisted diagnosis has become an inevitable trend, leading to many typical application cases in the medical field (Long et al., 2015; Ronneberger et al., 2015; Oktay et al., 2018; Chen et al., 2021; Bi et al., 2023). However, most studies have focused on grayscale images, using only single-channel data as network inputs, with fewer studies addressing three-channel data segmentation based on UE images. One existing study (Liu Y. et al., 2022) introduces multiple skeleton networks to evaluate the segmentation performance of U-shaped model structures on the BUE dataset. This study also designs a context extractor at the bottleneck and employs an attention gate (AG) (Oktay et al., 2018) in the skip connections to suppress irrelevant information in the image. The proposed ACE-Net examines the impact of model structure changes on segmentation performance. Unfortunately, this model overlooks the channel features in the middle layer and relies solely on the soft attention mechanism for feature correction. Additionally, the traditional decoder structure is insufficient for fully recovering the features of the elastography image, indicating that the segmentation performance on mediastinal lymph nodes needs further improvement.

On the one hand, traditional ultrasound images suffer from low contrast and high noise, leading to blurred node edges and abnormal boundary changes (Xian et al., 2018; Liu et al., 2019; Chen et al., 2022). On the other hand, UE images with added pseudo color can assist physicians in locating the approximate position of nodules. However, they do not resolve the issues inherent in traditional ultrasound images and introduce additional challenges. Specifically, the pseudo colors obscure the texture information of mediastinal lymph nodes, making it more difficult to capture their actual boundaries, particularly for the accurate segmentation of small mediastinal lymph nodes. Therefore, we combine the attention mechanism and vision transformer (ViT) to conduct an in-depth study of mediastinal lymph node segmentation in bronchial ultrasound elastography images. The main contributions of this research are summarized as follows:

- We design a dual-stream encoder (DSE) combining ConvNext and a lightweight ViT to effectively extract both global and local features from UE images.
- We propose a hybrid attention module (HAM) at the bottleneck to optimize the transmission of high-dimensional features.
- We introduce a feature-enhanced residual decoder (FRD) to recover information and fully fuse the intermediate features of the encoder and decoder using attention and residual structures.
- We use Grad-CAM to visualize heat maps of class activation at different stages of the model, providing insights into the action mechanisms.

2 Related work

2.1 Medical image segmentation based deep learning

In the early stages of medical image segmentation, traditional methods primarily relied on thresholding, region, edge detection, clustering, and deformable models (Tsai et al., 2003). With the advancement of deep learning, fully convolutional networks (FCNs) (Long et al., 2015) emerged as the most classic segmentation models. FCNs address the limitations of convolutional neural networks (CNNs) in fine-grained image segmentation by replacing fully connected layers with convolutional layers, enabling pixel-level classification to achieve target segmentation. U-Net (Ronneberger et al., 2015) employs a symmetric U-shaped encoder-decoder structure and is widely used in medical image segmentation. Each layer introduces skip connections that combine intermediate features from the encoder and decoder, reducing feature loss and making it particularly suitable for small sample datasets, thereby achieving faster and more efficient segmentation.

There are many variants of U-Net. To enhance the feature extraction capabilities of the model, Dense-UNet (Cai et al., 2020) uses a densely connected network as the decoder, effectively segmenting multiphoton live cell images. To improve the sensitivity to subtle boundaries, Iter-Net (Li et al., 2020) chains U-Net structures together, achieving retinal fundus vessel segmentation by analyzing U-Net structures of different sizes. However, these studies fail to capture contextual features from a global perspective, focusing primarily on spatial domain dependencies.

Recently, researchers have integrated vision transformers (ViT) (Dosovitskiy et al., 2020) into U-Net to enhance feature extraction. For example, Trans-UNet and Swin-UNet have demonstrated impressive performance and accuracy in medical image segmentation. Lin et al. (2023) explored the relationships among CNNs, ViT, and

traditional operators, proposing CTO, which performed exceptionally well on multiple medical image segmentation datasets. Bi et al. (2023) combined ViT with deformable convolutions to accurately segment thyroid nodules. These models utilize ViT as an encoder to effectively capture global contextual information while retaining U-Net's unique multi-scale feature fusion structure. Despite the outstanding performance of ViT, the fixed-size patches limit its ability to perceive fine details and result in high computational costs. Considering the powerful capability of CNNs in capturing local features, we adopt a dual-stream network that combines ViT and CNN to fully exploit the information in medical images.

2.2 Attention mechanism

The attention mechanism has shown significant achievements and is widely used in medical image segmentation due to its ability to enhance feature representation and improve the accuracy of segmentation. By selectively focusing on the most relevant parts of the image, attention mechanisms can effectively highlight important regions, such as lesions or tumors, while suppressing irrelevant background noise. For example, Attention U-Net (Oktay et al., 2018) enhances the U-Net by adding AG mechanisms in the skip connections. These AGs re-adjust the encoder's output features, emphasizing attention weights on the target organ region, thereby improving segmentation accuracy. Lee et al. (2020) proposed an innovative channel attention module that employs a multi-scale averaging pooling operation to cleverly fuse global and local spatial information. MDA-Net (Iqbal and Sharif, 2022) replaces the normal convolution module in U-Net with a multi-scale fusion module and uses a dual attention mechanism to optimize intermediate features in the decoder. Chen et al. (2022) designed a hybrid adaptive attention module for the irregular lesion morphology, which combines channel self-attention and spatial self-attention, and replaced the convolution module in U-Net with it to form AAU-Net. However, given the limitations in feature extraction and enhancement, especially the high-dimensional complex features extracted by DSE, such research may encounter bottlenecks. To address this, we design a hybrid attention module at the bottleneck. This module helps capture more semantically rich features, enables the network to focus on lesion areas, and filters out noise during the feature propagation process.

3 Methodology

3.1 Overview

The model proposed mainly contains the following components: dual stream encoder (DSE), hybrid attention module (HAM), and feature-enhanced residual decoder (FRD), and the structure is shown in Figure 1. Firstly, the UE image is fed into the network for multi-order feature extraction using the DSE. Secondly, the features generated by the encoder are optimized using the HAM at the bottleneck. Then, FRD fully fuses the intermediate and underlying features to de-code them. Finally, the features are transformed into a binary map using a convolutional layer and an up-sampling layer. The following section describes in detail the structures in the figure.

3.2 Dual-stream encoder

Given that UE images can localize the position of lymph nodes and provide rich channel information, the masking of texture information also leads to the difficulty of performing this task. Therefore, we combine CNNs and ViTs to design a DSE, aiming to effectively capture both local and global features.

A convolutional network encoder is used to capture local feature information of mediastinal lymph nodes from BUE images. Numerous studies (Xie and Richmond, 2018; Raghu et al., 2019) have shown the benefits of pre-trained models, so we use the newly proposed powerful pre-trained ConvNext (Liu Z. et al., 2022) as a convolutional network encoder. It has four outputs are $F_i, \mathrm{i} = 1,2,3,4$, dimensions are $C \times H/4 \times W/4$, $2C \times H/8 \times W/8$, $4C \times H/16 \times W/16$ and $8C \times H/32 \times W/32$, where C is 128, H and W are both 256.

Vision transformer encoder is used to capture the global feature dependencies of mediastinal lymph nodes to assist the convolutional network encoder for feature extraction. As shown in Figure 1, to minimize model complexity and make full use of intermediate features, F_1 is used as an input to ViT. Considering the size distribution of the mediastinal lymph node, we used 4×4 and 16×16 patch sizes to divide F_1. F_1 is split equally from the channel dimensions, using dimensionality change and linear layer to divide F_1 into $C/2 \times H/4P \times W/4P, \mathrm{P} = 4,16$ patches, where P denotes the size of the patch. The features are passed into the multi-head attention module, whose main role is to compute the self-attention of the input features to capture the correlation between the features. Specifically, we first use the convolution operation to obtain the query vector Q, the key vector K, and the value vector V of the features. Then the attention score matrix is obtained by the inner product operation between Q and K, which represents the feature-to-feature similarity. Next, the attention score matrix is scaled and probabilization to obtain the attention weight matrix. Finally, the attentional weight matrix is weighted and summed with V to obtain the attentional weighted value matrix. This matrix represents the feature representation obtained after attentional weighting of the input features. Specifically as shown in Equation (1):

$$\mathrm{F}_{MHA} = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k is the length of K and F_{MHA} is the output of the multi-head attention module.

Send F_{MHA} into the feed forward module to get F_{FF}. The feed forward module consists of two base convolutional modules: a convolutional layer with a kernel of 3×3, a batch normalization layer, and a leak ReLU activation function. To further speed up the training, F_1, F_{MHA}, and F_{FF} are residually summed to obtain the feature F_V extracted by the ViT encoder.

3.3 Hybrid attention module

To enhance the extraction of global and local features across various dimensions from the DSE, we design a HAM to optimize the

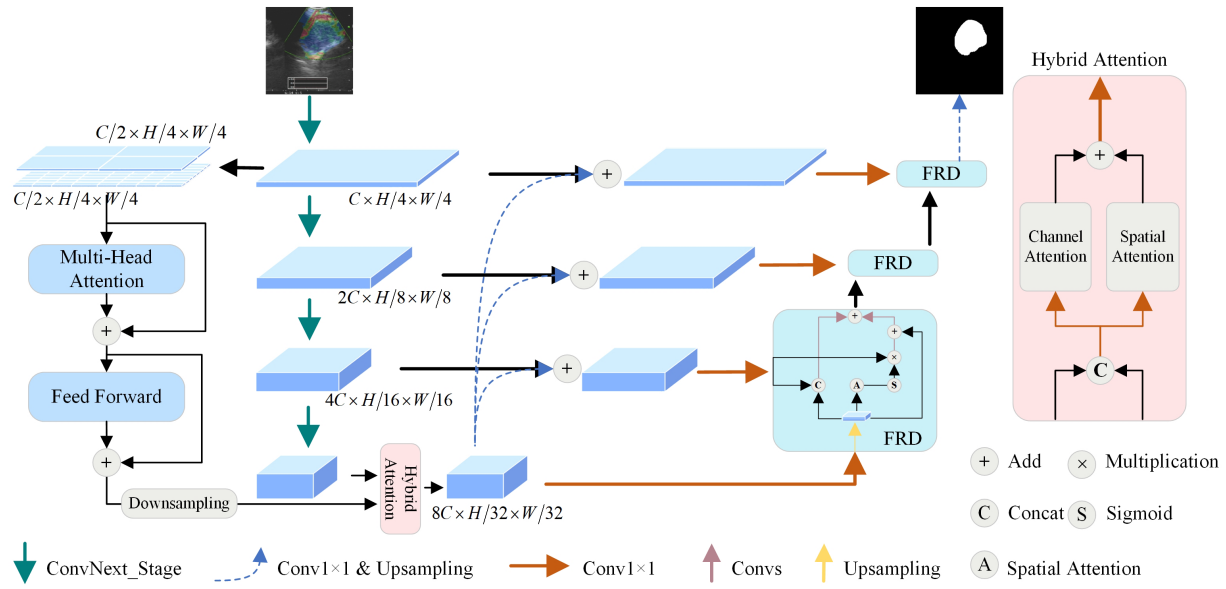


FIGURE 1
The framework of the proposed dual-stream feature-fusion attention U-Net.

features transmission process by optimizing high-dimensional features at the bottom of the network.

First, by extracting global features using the lightweight ViT, with input and output dimensions unchanged, the resulting F_v dimension is $C \times H/4 \times W/4$. Then, local features F_4 are extracted by CNN, with dimensions of $8C \times H/32 \times W/32$. We use downsampling to resize the F_v to the same size as F_4. To further enhance the features extracted by the encoder, we concatenate the global feature F_v and the local feature F_4 along the channel dimension and utilize a 1×1 convolution to reduce the number of channels to 1/4 of the original, obtaining the feature F_f, thereby reducing parameter and computational complexity.

To minimize information loss while enhancing features, we parallelly employ spatial attention modules and channel attention modules to enhance encoder features. The channel attention module first transforms the dimensions of the input feature F_f to $C' \times H'W'$, then generates the attention map W_c through matrix multiplication. Finally, F_f is multiplied by W_c and uses the residual add, resulting in the feature F_c enhanced by channel attention, as shown in the formula below:

$$F_c = Soft\max\left(Rs\left(F_f\right) \bullet Rs\left(F_f\right)^T\right) \times F_f + F_f \tag{2}$$

where $Rs(\bullet)$ denotes the dimensional transformation and $Softmax(\bullet)$ denotes the activation function used to normalize the weight values.

For spatial attention, firstly, the channels of F_f are reduced to 1 through a 1×1 convolution. Then, the Softmax function is applied to normalize the features. Finally, the obtained feature map is multiplied by F_f and undergoes residual add, resulting in the feature F_s enhanced by spatial attention, as shown in the formula below:

$$F_s = Soft\max\left(Convs\left(F_f\right)\right) \times F_f + F_f \tag{3}$$

The obtained F_c and F_s are added and then the channel number is restored using a 1×1 convolution, obtaining the enhanced DSE features F_{cv} with dimensions of $8C \times H/32 \times W/32$. This approach comprehensively enhances the image features captured by the feature encoder. Moreover, this parallel attention mechanism reduces the influence of noise, optimizes the feature propagation process at the network bottleneck, and enhances the reliability of the model.

3.4 Feature-enhanced residual decoder

To alleviate the situation that ordinary decoder modules may lead to inaccurate segmentation results in the process of feature recovery, we propose the FRD, as shown in Figure 1. Firstly, the feature map F_{CV} is summed with $F_i, i = 1,2,3,4$ to obtain the enhanced fused feature $F_{di}, i = 1,2,3,4$ by using bilinear interpolation and convolution operations. This preserves the details and location information of the original input image and improves the accuracy of the segmentation results. Then, to reduce the complexity and training difficulty of the model, the number of channels of $F_{di}, i = 1,2,3,4$ is converted to $C/2$ using a convolution operation to obtain the feature $F'_{di}, i = 1,2,3,4$. Finally, F'_{di} is passed into the FRD for feature recovery. Anyway, the features of the mediastinal lymph node can be recovered more accurately utilizing FRD, and the accuracy of segmentation results can be improved. The formula is as follows:

$$F'_{di} = Conv_{1\times1}\left(Up\left(Conv_{1\times1}\left(F_{CV}\right)\right) + F_i\right), i = 1,2,3,4 \tag{4}$$

where $Up(\bullet)$ denotes bilinear interpolation for feature transformation and $Conv_{1\times1}(\bullet)$ denotes 1×1 convolution for channel conversion.

To make full use of the intermediate features of the model, multiple parallel processing strategies are adopted at the bottom decoding stage. Specifically, there are three branches of processing for F'_{d3} and F'_{d4}. The first branch performs the bilinear interpolation of F'_{d4} with F'_{d3} for channel concatenation and passes the result to the convolution module for initial feature recovery. The second branch passes F'_{d4} into the spatial attention module to extract the position weight W_s, and then performs product operation between W_s and F'_{d3} to obtain the attention-enhanced features. The third branch residually sums F'_{d4} with the features of the first two branches to obtain the output of the decoder module F_{o3}. The formulas for the other decoder modules are shown in Equation (5):

$$F_{oi} = Convs\left(F'_{di} \oplus F^{up}{}_{oi+1}\right) + F^{up}{}_{oi+1} + SA\left(F^{up}{}_{oi+1}\right) \times F'_{di} \quad (5)$$

where $Convs(\bullet)$ denotes the base convolution operation; $\oplus$ denotes channel concatenation; $F^{up}{}_{oi+1}$ is the output of the decoder after up-sampling; and $SA(\bullet)$ denotes the spatial attention operation. Through parallel processing and feature fusion, the decoder can fully utilize the features to recover lost details and positional information and improve the accuracy of the segmentation results. This design can effectively compensate for the shortcomings of the common decoder and further optimize the performance of mediastinal lymph node segmentation.

4 Experiments

4.1 Databases and experimental protocols

4.1.1 Dataset description

A cohort of 206 patients who underwent endobronchial ultrasound-guided trans-bronchial needle aspiration (EBUS-TBNA) was selected from the First Hospital of Nanjing, comprising 141 males and 65 females. We collected 263 UE images of lymph nodes, which were manually delineated by an experienced radiologist. The dataset includes 102 benign and 161 malignant samples. For the experiments, the UE images were uniformly resized to 256 × 256 pixels. The dataset is divided into six equal parts, five of which totalling 219 images are used for training and the other totaling 44 images are used for testing.

We conduct multiple experiments through a six-fold cross-validation approach to fully evaluate the performance of the model. To increase the robustness of the model, we use an online data augmentation method, where the read data are vertically flipped and rotated by a random angle (−30° or 30°) with a probability of 0.5 during the model training iterations.

4.1.2 Implementation details

The proposed DFA-UNet is implemented based on Python 3.7 and Pytorch 1.12. The image processing workstation is equipped with an Intel i9-13900 K CPU and two NVIDIA RTX 4090 GPUs with 24G memory. The initial parameters during model training are obtained by Pytorch default initialization and the Adam optimizer is used to update the network parameters. Specifically, the initial learning rate is set to 0.0001, the weight decay coefficient is 0.1, the learning rate is decayed every 90 rounds of iterations, and the number of iterative training of the model is 190 times in total. Dice (Milletari et al., 2016) is used as the loss function with the following formula:

$$Dice\ Loss = 1 - \frac{2\,|\,I_t \cap I_p\,|}{|\,I_t\,| + |\,I_p\,|} \quad (6)$$

where I_t is the true mask for UE image segmentation and I_p is the mask predicted by the model.

4.1.3 Evaluation metrics

To fully demonstrate the segmentation effect of the model, we use the Dice coefficient (Dice), Intersection over Union (IoU), Precision, Specificity, and Hausdorff distance 95th percentile (HD95) (Karimi and Salcudean, 2019) metrics to evaluate DFA-UNet. The Dice is a metric used to measure the similarity of a collection of two samples, in evaluating the performance of image segmentation, Dice can be expressed as:

$$Dice = \frac{2 \times TP}{TP + FP + TP + FN} \quad (7)$$

where *TP*, *FP*, *TN*, and *FN* denote the set of pixel points for true positives, false positives, true negatives, and false negatives. Since the true positives of the background region are not computed during the pixel point classification process, the Dice is suitable for the task of evaluating segmentation targets of varying sizes.

The HD95 is a defined form of the distance between two point sets, calculated as:

$$HD95 = \max\left\{d_{tp}, d_{pt}\right\} \quad (8)$$

where d_{tp} denotes the 95% quantile of the farthest distance from I_t to I_p, and d_{pt} denotes the 95% quantile of the farthest distance from I_p to I_t. This metric is more robust to outliers and more suitable for biomedical image segmentation tasks.

In the aforementioned metrics, except for HD95, the value range of the other indicators is [0, 1], with values closer to 1 indicating better model segmentation performance. HD95 has no fixed value range, but lower values of HD95 signify better segmentation performance.

4.2 Comparison with the state-of-the-art

4.2.1 Quantitative analysis

To further validate the effectiveness of DFA-UNet on UE images, comparative experiments were conducted with several other models: U-Net (Ronneberger et al., 2015), Att-UNet (Oktay et al., 2018), Seg-Net (Badrinarayanan et al., 2017), DeepLabV3+ (Polat, 2022), Trans-UNet (Chen et al., 2021), U-Net++ (Zhou et al., 2018), BPAT-UNet (Bi et al., 2023), CTO (Lin et al., 2023), and ACE-Net (Liu Y. et al., 2022). The results are presented in Table 1, with the best performance for each metric highlighted in bold.

From Table 1, it can be observed that DFA-UNet outperforms other models in terms of Dice, IoU, Precision, Specificity, and HD95. Specifically, DFA-UNet achieves higher Dice scores compared to U-Net, Seg-Net, Att-UNet, U-Net++, Trans-UNet, DeepLabV3+, BPAT-UNet,

CTO, and ACE-Net by 1.99, 1.18, 0.93, 1.13, 2.64, 0.98, 0.70, 0.51, and 0.54%, respectively. Additionally, DFA-UNet shows an improvement of 0.86% in IoU (77.41% vs. 76.55%) and a 1.48% increase in Precision (86.71% vs. 85.23%) compared to ACE-Net. The average improvement in Specificity across the nine compared models is 0.52%. Regarding HD95, DFA-UNet reduces the distance from 10.39 to 8.125 compared to U-Net, with an average reduction of 1.237 across the remaining models, indicating a significant enhancement in segmentation performance. Furthermore, due to the optimization of all parts of U-Net, DFA-UNet, similar to Trans-UNet, BPAT-UNet, CTO, and the other models, achieves better performance compared to U-Net with more parameters. However, it is worth noting that DFA-UNet achieves the best results in model computation within the well-established ConvNext, and also achieves optimal results in segmentation effectiveness.

4.2.2 Qualitative analysis

To further verify the generality of DFA-UNet for mediastinal lymph node segmentation. We randomly select four segmentation samples of different sizes for qualitative analysis, and their performance is shown in Figure 2.

TABLE 1 Quantitative comparison of our DFA-UNet with other state-of-the-art methods.

Model	Dice (%)	IoU (%)	Pre (%)	HD95	Para (M)	Flops (G)
U-Net	84.61	74.73	84.88	10.39	31.04	54.60
Seg-Net	85.42	75.63	85.54	8.962	29.44	40.01
Att-UNet	85.67	76.04	84.05	9.056	57.16	66.61
U-Net++	85.47	75.91	84.81	9.268	47.18	114.16
Trans-UNet	83.96	73.55	82.25	11.90	105.12	11.89
DeepLabv3+	85.62	76.05	86.07	9.328	**21.54**	45.58
BPAT-UNet	85.90	76.38	84.83	8.725	71.01	64.12
CTO	86.09	76.71	85.05	8.751	60.01	22.59
ACE-Net	86.06	76.55	85.23	8.907	35.01	20.26
DFA-UNet	**86.60**	**77.41**	**86.71**	**8.125**	97.29	**5.27**

Bold values represent the best results.

From Figure 2, it is evident that DFA-UNet exhibits superior segmentation performance for mediastinal lymph nodes of varying sizes. When the target size is small (first row), U-Net, Seg-Net, Att-UNet, BPAT-UNet, CTO-Net, and ACE-Net produce seg-mentation results that are smaller than the actual target, whereas only U-Net++ and DFA-UNet achieve accurate segmentation. For moderately sized targets with relatively simple boundary structures (second row), Trans-UNet, U-Net, Att-UNet, and U-Net++ show significant mis-segmentation, with Trans-UNet performing particularly poorly, as corroborated by the data in Table 1. Additionally, CTO misses part of the segmentation in the lower-right corner of the node. For moderately sized targets with complex boundary structures (third row), Att-UNet, U-Net++, and Trans-UNet fail to accurately segment the lower-right protruding region of the target area, whereas DFA-UNet consistently delivers precise segmentation results. In cases where the target size is large (fourth row), Seg-Net and Trans-UNet exhibit noticeable mis-segmentation in the lower-right depression of the target region, resulting in smaller overall segmentation outputs. U-Net, DeepLabV3+, and BPAT-UNet also show significant mis-segmentation in the low-er-right region. Only CTO-Net, ACE-Net, and DFA-UNet achieve more accurate overall segmentation results, with DFA-UNet providing the best performance across different target sizes and boundary complexities.

4.2.3 Visual analysis

To further explore the underlying mechanisms of DFA-UNet, we employ Grad-CAM (Selvaraju et al., 2017) to visualize the decoding stages of the model. A total of eight models, U-Net, Att-UNet, Seg-Net, Trans-UNet, BPAT-UNet, CTO, ACE-Net, and DFA-UNet, are selected and demonstrated in three stages.

From the overall analysis in Figure 3, it can be seen that the feature extraction capability of the model's bottom stage determines the feature recovery of the model's top stage. Specifically, all eight models can roughly locate the real segmentation region in the Decoder2 stage, and further continue to expand outward from the region of interest obtained in the previous stage in the Decoder3 stage. In the Decoder4 stage, the model DFA-UNet shifted the region of interest from the interior to the boundary, which achieved better results in the overall segmentation results. The remaining seven models still further expand

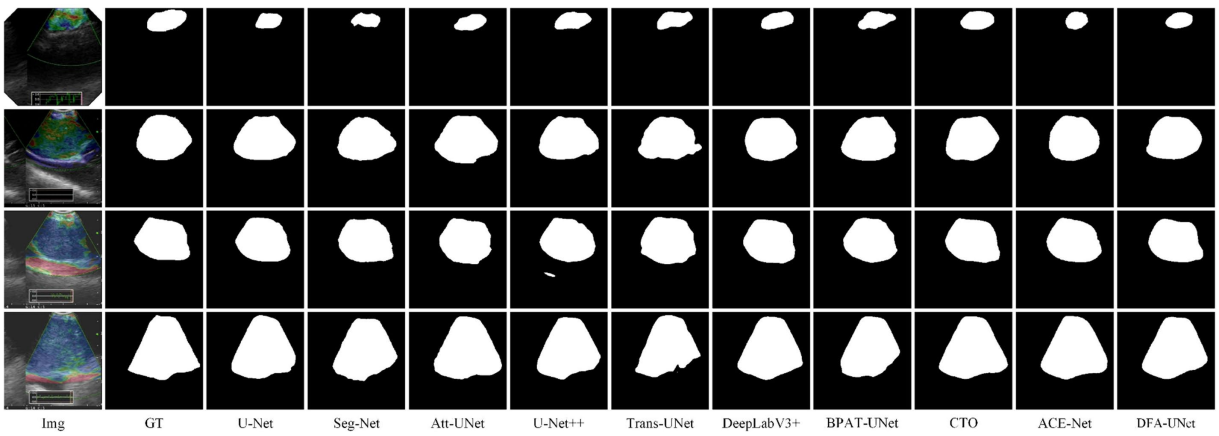


FIGURE 2
Segmentation results of different models.

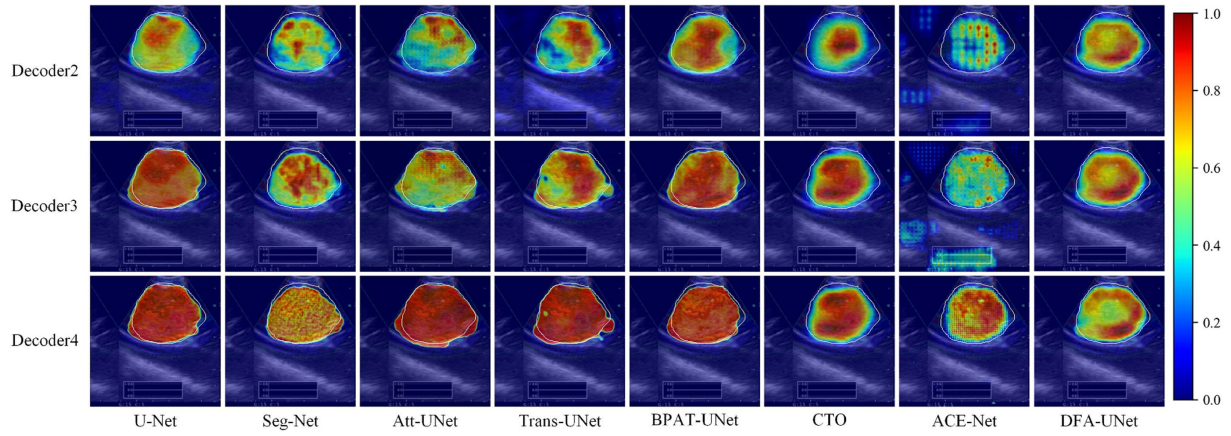


FIGURE 3
Class activation maps generated by DFA-UNet using Grad-CAM. White contours indicate lymph node locations. Warmer-colored regions correspond to target class labels with higher confidence.

the region of interest outwards, resulting in less accurate segmentation results in the higher stages of the model as determined by the target region positioned in the bottom stage of the model.

Secondly, a side-by-side comparison reveals that our DFA-UNet locates the target segmentation region more accurately in the lower stages. During the Decoder2 and Decoder3 phases, the red area representing the region of interest in the DFA-UNet is larger and more uniformly distributed compared to Seg-Net, Att-UNet, Trans-UNet, CTO, and ACE-Net. This uniform distribution closely aligns with the target segmentation region, indicating a better fit.

Finally, the reason for the poor segmentation performance of traditional models can also be analyzed from the figures: either the model's ability to localize features in the lower layers or its ability to correct feature details in the higher layers is insufficient. Specifically, ACE-Net further extracts high-level semantic information at the bottleneck with the help of a context extractor, which leads to a certain degree of difficulty in re-covering high-level semantic information at the decoder stage, which is manifested in the form of smaller regions of interest in the Decoder2 and Decoder3 stages in Figure 3. Whereas the U-Net model is more accurate in its ability to localize the target segmentation region in the Decoder2 stage, its region of interest is almost unchanged in the Decoder3 and Decoder4 stages, suggesting that the model's high-level stages are ineffective in correcting feature details. In contrast, DFA-UNet demonstrates superior performance in both the lower and higher stages, resulting in the best overall segmentation outcomes for the region of interest.

4.2.4 Ablation study

We perform ablation studies on each of the key modules of the DFA-UNet. The baseline network is U-Net, which is tested separately with the addition of DSE, HAM, and FRD. As seen in Table 2, the proposed modules promote significant improvements in the baseline network. This fully demonstrates the effectiveness of our DFA-UNet in mediastinal lymph node segmentation.

Firstly, using the DSE as the encoder significantly enhances the segmentation performance of the baseline network. The Dice increases by 0.79% (84.61% vs. 85.40%), and the IoU improves by 0.90% (74.73% vs. 75.63%). This notable performance boost is primarily due to the DSE helping the network extract both global and local features. Secondly, incorporating the HAM further improves the feature transfer capability from the DSE, resulting in an additional performance increase. Specifically, the Dice rises from 85.40 to 85.84%, and the HD95 improves from 9.316 to 9.014. Finally, adding the FRD further improves segmentation performance. Compared with the baseline, the Dice is enhanced by 1.99% (84.61% vs. 86.60%), and the HD95 improves by 2.265 (10.39 vs. 8.125). In summary, systematically integrating the feature maps obtained through DSE, HAM, and FRD significantly contributes to the superior performance of our DFA-UNet. Additionally, it is important to note that the parameter count of the lightweight ViT module, DSE-ViT, only occupies a small portion (0.5%) of the total model parameters (88.58 M vs. 97.29 M), confirming its lightweight nature.

TABLE 2 Ablation experiment of the proposed DFA-UNet.

DSE-CNN	DSE-ViT	HAM	FRD	Dice (%)	IoU (%)	HD95	Para (M)
				84.61	74.73	10.39	31.04
★				85.07	75.23	9.809	88.58
★	★			85.40	75.63	9.316	89.15
★	★			85.84	76.40	9.014	96.94
★	★	★	★	**86.60**	**77.41**	**8.125**	**97.29**

Bold values represent the best results.

4.2.5 Generalization study

To validate the generalization of our DFA-UNet on ultrasound images, we conduct comparative experiments using the BUSI dataset (Al-Dhabyani et al., 2020). This dataset contains 780 breast ultrasound (BUS) images, including 437 benign images, 210 malignant images, and 133 normal images, acquired using the LOGIQ E9 and LOGIQ E9 Agile Ultrasound Systems. Since the primary goal of breast lesion segmentation is to evaluate and identify the distribution of lesions, normal cases without masks were excluded from the BUSI dataset (Ning et al., 2021; Xue et al., 2021). The results of these experiments are presented in Table 3.

TABLE 3 Experiments for generalizability of the proposed DFA-UNet on the BUSI dataset.

Methods	Dice (%)	IoU (%)	Pre (%)	HD95
	70.94	61.77	77.51	30.84
Att-UNet	72.80	63.90	75.49	32.99
DeepLabV3+	78.12	68.75	80.75	21.91
Trans-UNet	76.82	67.41	80.45	21.25
BPAT-UNet	79.37	70.46	81.56	22.66
CTO	78.32	69.61	82.04	20.98
DFA-UNet	**82.68**	**74.59**	**84.44**	**17.01**

Bold values represent the best results.

The results in Table 3 demonstrate that our DFA-UNet achieves state-of-the-art performance in breast ultrasound image segmentation. Specifically, DFA-UNet shows significant improvements over U-Net, with increases of 11.74, 12.82, and 6.93% in Dice, IoU, and Precision, respectively, and a reduction of 13.83 in HD95. When compared with other models, DFA-UNet exhibits an average improvement of 5.59% in Dice, indicating its robust applicability to ultrasound images. Furthermore, comparing the results from Tables 1, 3 reveals that U-Net experiences a 13.67% decrease in Dice when applied to breast ultrasound images, highlighting the increased difficulty of this segmentation task. This also suggests that the color information in ultrasound elastography images aids segmentation. Notably, DFA-UNet shows only a 3.92% decrease in Dice, which underscores its superior generalization capability compared to other models that average a 6.49% decrease. Therefore, DFA-UNet is particularly well-suited for segmenting mediastinal lymph nodes in ultrasound elastography images. This capability has potential clinical value, as it can assist doctors in using ultrasound elastography images for the diagnosis and treatment of lung cancer.

5 Conclusion

UE images with rich channel information can provide some guidance for segmentation of the region of interest, but their masking of texture information also leads to the difficulty of performing this task. Additionally, the varying characteristics of different mediastinal lymph node groups further challenge segmentation efforts. To address these issues, we designed a DSE based on ConvNext and a lightweight ViT incorporated into the U-Net. At the bottleneck, we introduced a HAM that combines channel attention with spatial attention to enrich the feature from DSE. The FRD fully fuses intermediate encoder features with decoder output features.

To verify the validity of our DFA-UNet, extensive experiments were conducted to several important conclusions. On the one hand, DFA-UNet employs a dual-stream encoder and an attention enhancement mechanism, which significantly increases the model's stability. Comparative experiments show that DFA-UNet has clear competitive advantages over current mainstream segmentation models. Class activation maps demonstrate that DFA-UNet achieves superior segmentation sensitivity and completeness by focusing on the content of the region at the lower levels of the network and the boundaries of the region at the higher levels. On the other hand, we optimized various components of the U-Net architecture and presented corresponding ablation experimental results. These findings offer insights for future research aimed at enhancing segmentation performance using U-Net structural variants. This optimization provides a foundation for subsequent studies to explore further improvements in segmentation effectiveness through structural enhancements of U-Net.

In the subsequent research, we will focus on data collection, semi-supervised segmentation tasks, and model optimal structure exploration, to achieve better segmentation results and assist doctors to use UE images for relevant diagnosis and treatment of lung cancer.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

QZ: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. YiZ: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing – review & editing. NH: Formal analysis, Validation, Writing – review & editing. YaZ: Formal analysis, Investigation, Writing – review & editing. GZ: Funding acquisition, Investigation, Project administration, Supervision, Writing – review & editing. LL: Funding acquisition, Investigation, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Innovation and Entrepreneurship Project of Xuzhou Medical University Science and Technology Park (grant number: CXCYZX2022003), Project Supported by the Affiliated Hospital of Xuzhou Medical University (grant number: 2022ZL19), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (grant number: KYCX24_3128).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Al-Dhabyani, W., Gomaa, M., Khaled, H., and Fahmy, A. (2020). Dataset of breast ultrasound images. *Data Brief* 28:104863. doi: 10.1016/j.dib.2019.104863

Asamura, H., Chansky, K., Crowley, J., Goldstraw, P., Rusch, V. W., Vansteenkiste, J. F., et al. (2015). The International Association for the Study of Lung Cancer lung Cancer staging project: proposals for the revision of the N descriptors in the forthcoming 8th edition of the TNM classification for lung cancer. *J. Thorac. Oncol.* 10, 1675–1684. doi: 10.1097/JTO.0000000000000678

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi: 10.1109/TPAMI.2016.2644615

Bi, H., Cai, C., Sun, J., Jiang, Y., Lu, G., Shu, H., et al. (2023). BPAT-UNet: boundary preserving assembled transformer UNet for ultrasound thyroid nodule segmentation. *Comput. Methods Prog. Biomed.* 238:107614. doi: 10.1016/j.cmpb.2023.107614

Cai, S., Tian, Y., Lui, H., Zeng, H., Wu, Y., and Chen, G. (2020). Dense-UNet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quant. Imaging Med. Surg.* 10, 1275–1285. doi: 10.21037/qims-19-1090

Chen, G., Li, L., Dai, Y., Zhang, J., and Yap, M. H. (2022). AAU-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images. *IEEE Trans. Med. Imaging* 42, 1289–1300. doi: 10.1109/TMI.2022.3226268

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv*. doi: 10.48550/arxiv.2102.04306

Cui, X.-W., Li, K.-N., Yi, A.-J., Wang, B., Wei, Q., Wu, G.-G., et al. (2022). Ultrasound elastography. *Endosc. Ultrasound* 11, 252–274. doi: 10.4103/EUS-D-21-00151

Detterbeck, F. C., Chansky, K., Groome, P., Bolejack, V., Crowley, J., Shemanski, L., et al. (2016). The IASLC lung cancer staging project: methodology and validation used in the development of proposals for revision of the stage classification of NSCLC in the forthcoming (eighth) edition of the TNM classification of lung cancer. *J. Thorac. Oncol.* 11, 1433–1446. doi: 10.1016/j.jtho.2016.06.028

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv*. doi: 10.48550/arxiv.2010.11929

Gu, Y., Shi, H., Su, C., Chen, X., Zhang, S., Li, W., et al. (2017). The role of endobronchial ultrasound elastography in the diagnosis of mediastinal and hilar lymph nodes. *Oncotarget* 8, 89194–89202. doi: 10.18632/oncotarget.19031

Iqbal, A., and Sharif, M. (2022). MDA-net: multiscale dual attention-based network for breast lesion segmentation using ultrasound images. *J. King Saud Univ. Comput. Inf. Sci.* 34, 7283–7299. doi: 10.1016/j.jksuci.2021.10.002

Karimi, D., and Salcudean, S. E. (2019). Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Trans. Med. Imaging* 39, 499–513. doi: 10.1109/TMI.2019.2930068

Lee, H., Park, J., and Hwang, J. Y. (2020). Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* 67, 1344–1353. doi: 10.1109/tuffc.2020.2972573

Li, L., Verma, M., Nakashima, Y., Nagahara, H., and Kawasaki, R. (2020). "Iternet: retinal image segmentation utilizing structural redundancy in vessel networks" in Proceedings of the IEEE/CVF winter conference on applications of computer vision, Springer.

Li, Z., and Xia, Y. (2020). Deep reinforcement learning for weakly-supervised lymph node segmentation in CT images. *IEEE J. Biomed. Health Inform.* 25, 774–783. doi: 10.1109/JBHI.2020.3008759

Lin, Y., Zhang, D., Fang, X., Chen, Y., Cheng, K.-T., and Chen, H. (2023). "Rethinking boundary detection in deep learning models for medical image segmentation" in International conference on information processing in medical imaging (Cham: Springer), 730–742.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition IEEE, 11976–11986.

Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., et al. (2019). Deep learning in medical ultrasound analysis: a review. *Engineering* 5, 261–275. doi: 10.1016/j.eng.2018.11.020

Liu, Y., Wu, R. R., Tang, L., and Song, N. (2022). U-Net-based mediastinal lymph node segmentation method in bronchial ultrasound elastic images. *J. Image Graph.* 27, 3082–3091. doi: 10.11834/jig.210225

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation, in Proceedings of the IEEE conference on computer vision and pattern recognition IEEE, 3431–3440.

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: fully convolutional neural networks for volumetric medical image segmentation, in 2016 fourth international conference on 3D vision (3DV), (Cham: IEEE), 565–571.

Ning, Z., Zhong, S., Feng, Q., Chen, W., and Zhang, Y. (2021). SMU-net: saliency-guided morphology-aware U-net for breast lesion segmentation in ultrasound image. *IEEE Trans. Med. Imaging* 41, 476–490. doi: 10.1109/TMI.2021.3116087

Oglat, A. A., and Abukhalil, T. (2024). Ultrasound Elastography: methods, clinical applications, and limitations: a review article. *Appl. Sci.* 14:4308. doi: 10.3390/app14104308

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-Net: learning where to look for the pancreas. *arXiv preprint arXiv*. doi: 10.48550/arxiv.1804.03999

Polat, H. (2022). A modified DeepLabV3+ based semantic segmentation of chest computed tomography images for COVID-19 lung infections. *Int. J. Imaging Syst. Technol.* 32, 1481–1495. doi: 10.1002/ima.22772

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: understanding transfer learning for medical imaging. *Adv. Neural Inf. Proces. Syst.* 32. doi: 10.48550/arxiv.1902.07208

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation, in Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18, (Cham: Springer), 234–241.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: visual explanations from deep networks via gradient-based localization., in Proceedings of the IEEE international conference on computer vision. IEEE, 618–626.

Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. (2023). Cancer statistics, 2023. *CA Cancer J. Clin.* 73, 17–48. doi: 10.3322/caac.21763

Sigrist, R. M., Liau, J., El Kaffas, A., Chammas, M. C., and Willmann, J. K. (2017). Ultrasound elastography: review of techniques and clinical applications. *Theranostics* 7, 1303–1329. doi: 10.7150/thno.18650

Sun, J., Zheng, X., Mao, X., Wang, L., Xiong, H., Herth, F. J., et al. (2017). Endobronchial ultrasound elastography for evaluation of intrathoracic lymph nodes: a pilot study. *Respiration* 93, 327–338. doi: 10.1159/000464253

Tan, S., Wen, Z., Fu, Y., Deng, Z., Gao, S., Yuan, X., et al. (2023). "Lymph node ultrasound image segmentation algorithm based on multimodal image fusion and DMA-UNet" in 2023 IEEE 13th international conference on electronics information and emergency communication (ICEIEC) (Cham: IEEE), 38–42.

Taylor, M., Soliman, N., Paoletti, E., King, M., Crosbie, P. A., and Granato, F. (2023). Impact of skip mediastinal lymph node metastasis on outcomes after resection for primary lung cancer. *Lung Cancer* 184:107341. doi: 10.1016/j.lungcan.2023.107341

Tsai, A., Yezzi, A., Wells, W., Tempany, C., Tucker, D., Fan, A., et al. (2003). A shape-based approach to the segmentation of medical imagery using level sets. *IEEE Trans. Med. Imaging* 22, 137–154. doi: 10.1109/TMI.2002.808355

Wang, B., Guo, Q., Wang, J.-Y., Yu, Y., Yi, A.-J., Cui, X.-W., et al. (2021). Ultrasound elastography for the evaluation of lymph nodes. *Front. Oncol.* 11:714660. doi: 10.3389/fonc.2021.714660

Wang, H., Wan, Y., Zhang, L., Tao, H., and Huang, H. (2018). Clinical value of bronchial ultrasound elastography in the differential diagnosis of benign and malignant hilar and mediastinal lymph nodes. *Chin. J. Clin. Oncol.* 45, 721–725. doi: 10.3969/j.issn.1000-8179.2018.14.358

Wang, R., Wu, S., Qian, D., Zhang, Y., Fan, B., and Hu, M. (2021). A lung Cancer auxiliary diagnostic method: deep learning based mediastinal lymphatic partitions segmentation for Cancer staging. *Int. J. Radiat. Oncol. Biol. Phys.* 111:e92. doi: 10.1016/j.ijrobp.2021.07.474

Xian, M., Zhang, Y., Cheng, H.-D., Xu, F., Zhang, B., and Ding, J. (2018). Automatic breast ultrasound image segmentation: a survey. *Pattern Recogn.* 79, 340–355. doi: 10.1016/j.patcog.2018.02.012

Xie, Y., and Richmond, D. (2018). Pre-training on grayscale imagenet improves medical image classification, in Proceedings of the European conference on computer vision (ECCV) workshops.

Xue, C., Zhu, L., Fu, H., Hu, X., Li, X., Zhang, H., et al. (2021). Global guidance network for breast lesion segmentation in ultrasound images. *Med. Image Anal.* 70:101989. doi: 10.1016/j.media.2021.101989

Zhang, F., Zhang, X., Lv, P. Z. Z., Cai, L., Li, R., Zhou, Y., et al. (2019). Differential diagnosis value of hilar and mediastinal lymph nodes in lung Cancer by Bronchoscopic Elastography and Intrabronchial ultrasonography. *Chin. J. Ultrasound Med.* 35, 897–900. doi: 10.3969/j.issn.1002-0101.2019.10.011

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). "UNET++: a nested u-net architecture for medical image segmentation" in Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4 (Springer), 3–11.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 07 August 2024
DOI 10.3389/fnins.2024.1449527

OPEN ACCESS

EDITED BY
Hancheng Zhu,
China University of Mining and Technology, China

REVIEWED BY
S. K. B. Sangeetha,
SRM Institute of Science and Technology, India
Rohan Borgalli,
University of Mumbai, India
Mohana Murugan,
Avinashilingam Institute for Home Science and Higher Education for Women, India
Deepali Virmani,
Guru Gobind Singh Indraprastha University, India

*CORRESPONDENCE
Jingyi Wang
✉ wang04032022@163.com

RECEIVED 15 June 2024
ACCEPTED 11 July 2024
PUBLISHED 07 August 2024

CITATION
Wang J (2024) Evaluation and analysis of visual perception using attention-enhanced computation in multimedia affective computing. *Front. Neurosci.* 18:1449527. doi: 10.3389/fnins.2024.1449527

Evaluation and analysis of visual perception using attention-enhanced computation in multimedia affective computing

Jingyi Wang*

School of Mass-communication and Advertising, Tongmyong University, Busan, Republic of Korea

Facial expression recognition (FER) plays a crucial role in affective computing, enhancing human-computer interaction by enabling machines to understand and respond to human emotions. Despite advancements in deep learning, current FER systems often struggle with challenges such as occlusions, head pose variations, and motion blur in natural environments. These challenges highlight the need for more robust FER solutions. To address these issues, we propose the Attention-Enhanced Multi-Layer Transformer (AEMT) model, which integrates a dual-branch Convolutional Neural Network (CNN), an Attentional Selective Fusion (ASF) module, and a Multi-Layer Transformer Encoder (MTE) with transfer learning. The dual-branch CNN captures detailed texture and color information by processing RGB and Local Binary Pattern (LBP) features separately. The ASF module selectively enhances relevant features by applying global and local attention mechanisms to the extracted features. The MTE captures long-range dependencies and models the complex relationships between features, collectively improving feature representation and classification accuracy. Our model was evaluated on the RAF-DB and AffectNet datasets. Experimental results demonstrate that the AEMT model achieved an accuracy of 81.45% on RAF-DB and 71.23% on AffectNet, significantly outperforming existing state-of-the-art methods. These results indicate that our model effectively addresses the challenges of FER in natural environments, providing a more robust and accurate solution. The AEMT model significantly advances the field of FER by improving the robustness and accuracy of emotion recognition in complex real-world scenarios. This work not only enhances the capabilities of affective computing systems but also opens new avenues for future research in improving model efficiency and expanding multimodal data integration.

KEYWORDS

affective computing, attention mechanisms, feature extraction, emotion recognition, facial expression recognition, deep learning, transfer learning

1 Introduction

In the field of affective computing, facial expression recognition (FER) has garnered significant attention due to its natural and powerful means of conveying human emotions. FER systems have critical applications in psychology research, human-computer interaction, driver fatigue monitoring, and more. However, there are still many challenges to facial expression recognition in natural environments. Factors such as occlusion,

changes in head pose (Sun et al., 2021; Xu et al., 2022), facial distortion and motion blurring exacerbate the challenges to such recognition, as shown in Figure 1. These factors lead to significant changes in facial appearance, complicating the task of accurately recognizing expressions and causing traditional recognition methods in laboratory settings to perform poorly in real-world applications (Borgalli and Surve, 2022). Therefore, how to achieve efficient and accurate facial expression recognition in complex environments has become an urgent problem in the field (Zeng et al., 2019; Li et al., 2020b).

The advent of deep learning has provided new opportunities for FER. Convolutional neural networks (CNNs) and other deep learning models have made significant strides in feature extraction and classification accuracy. Deep learning models automatically learn complex features from data, enhancing the accuracy and robustness of FER. For example, Tang et al. (2019) proposed a CNN model that significantly improved performance by replacing the softmax layer with a linear support vector machine (SVM) for classification. Similarly, Kim et al. developed a deep locality-preserving CNN (DCNN-RF) method to enhance feature discriminativeness (Li et al., 2019; Kim et al., 2023). Despite these advancements, the performance of deep learning methods in natural environments still leaves much to be desired (Kollias and Zafeiriou, 2019).

Currently, the application of deep learning in natural environments faces several challenges, including insufficient data, weak model generalization, and difficulty in feature extraction under complex conditions. Most existing methods are trained and tested in controlled environments, performing poorly in real-world scenarios. Additionally, the limited quantity and quality of available datasets hinder the effective training of deep learning models, resulting in unstable performance in natural environments (Wang X. et al., 2020; Zeng et al., 2020). Existing methods often fail to account for the diversity of real-world conditions, such as varying lighting, occlusions, and head poses, leading to reduced robustness and accuracy.

To address these challenges, this paper proposes an improved visual Transformer model that combines attention mechanisms and multi-layer Transformer encoders, incorporating transfer learning to leverage the advantages of pre-trained models on large-scale datasets (Liu et al., 2021). Specifically, the proposed method involves two main steps: first, using a dual-branch CNN to extract RGB and LBP (Local Binary Pattern) features, which are then fused using an ASF module. The ASF module integrates global and local attention mechanisms to effectively combine various features, enhancing feature representation richness (Zhao et al., 2020; Zhang et al., 2021). Second, a multi-layer Transformer encoder models the global relationships of the fused features, and the pre-trained model is fine-tuned to improve adaptability to new datasets. The Transformer encoder, through multi-head self-attention mechanisms, captures long-range dependencies among features, thereby improving recognition capabilities (Ma et al., 2021).

The proposed model addresses the limitations of existing methods by enhancing feature extraction and improving generalization. The dual-branch CNN captures both color and texture information through RGB and LBP features, addressing the issue of insufficient feature representation. The ASF module further enhances this by selectively focusing on the most relevant features, improving the model's ability to handle occlusions and varying head poses. The multi-layer Transformer encoder with transfer learning leverages pre-trained models to improve performance on smaller datasets, addressing the challenge of insufficient training data and enhancing model generalization.

The goal of this study is to improve the accuracy and robustness of FER in natural environments by combining attention mechanisms, transfer learning, and Transformer models, providing an effective solution for affective computing. Experimental results demonstrate that the proposed method outperforms state-of-the-art methods on multiple public datasets, achieving new performance benchmarks. For instance, testing on the RAF-DB, FERPlus, and AffectNet datasets shows that the proposed method surpasses existing methods in accuracy, achieving new performance highs. Furthermore, the proposed method exhibits excellent generalization capabilities in cross-dataset evaluations, validating its applicability in diverse environments (Jiang et al., 2020).

In summary, this paper introduces a novel FER method that leverages transfer learning and improved attention mechanisms. This approach not only enhances recognition accuracy but also improves robustness and generalization in complex environments, providing new insights and technical support for the development of affective computing. With the advent of larger datasets and more powerful computational resources, this method is expected to further advance, laying the groundwork for more intelligent and humanized affective computing systems.

In conclusion, our contributions are as follows:

1. Novel integration of attention mechanisms and transformers: We have developed a new model that integrates attention mechanisms with multi-layer Transformer encoders. This combination enhances the ability to capture global and local features, improving the accuracy and robustness of facial expression recognition in natural environments.
2. Incorporation of transfer learning: By incorporating transfer learning, our model leverages pre-trained features from large-scale datasets, significantly improving performance and training efficiency on smaller, task-specific datasets. This approach also enhances the model's adaptability to diverse data conditions.
3. Comprehensive evaluation and validation: We conducted extensive experiments across multiple public datasets (RAF-DB, FERPlus, and AffectNet), demonstrating that our proposed method achieves state-of-the-art performance. Additionally, we validated our model's generalization capabilities through cross-dataset evaluations, proving its effectiveness in real-world applications.

To provide a clear structure for the reader, we outline the organization of our paper as follows: The first section is the introduction, providing an overview of the research background and the main challenges addressed. The second section reviews related work, extending the discussion on the application of models in similar fields. The third section, Method, describes the models and algorithms used in our study. The fourth section presents our experiments, evaluating our proposed research from various

FIGURE 1
Samples from the AffectNet (Sun et al., 2021) and RAF-DB (Li et al., 2017) datasets, emphasizing the variations in head poses, occlusions, and other unconstrained conditions present in real-world images. AffectNet includes eight expression labels, incorporating the contempt category, while RAF-DB is annotated with seven basic expressions and additional compound expressions.

perspectives and comparing its performance with other studies. Finally, Section 5 summarizes our findings and discusses future directions for research.

2 Related work

2.1 Convolutional neural networks for facial expression recognition

CNNs have shown exceptional performance in visual perception tasks, particularly in facial expression recognition. CNNs effectively extract features from images through hierarchical convolution and pooling operations and classify these features. Typical CNN architectures such as AlexNet, VGGNet, and ResNet have been widely applied to facial expression recognition tasks (Krizhevsky et al., 2012; He et al., 2016). Tariq et al. utilized VGGNet for facial expression classification, achieving high recognition accuracy on the FER-2013 dataset (Sikkandar and Thiyagarajan, 2021; Tariq et al., 2023).

In their implementation, the researchers first preprocessed the FER-2013 dataset by resizing the images to a fixed size and then used the VGGNet model to extract image features. By fine-tuning and optimizing the model, they classified seven basic emotions (e.g., happiness, sadness, anger). The experimental results showed that the VGGNet-based model achieved over 70% accuracy on the test set, significantly outperforming traditional handcrafted feature extraction methods.

The advantage of CNNs lies in their automatic feature extraction capability, making them particularly effective in handling complex emotional expressions. However, CNN models are highly dependent on datasets and require a large amount of labeled data for training (Buduma et al., 2022). Additionally, CNNs are sensitive to geometric transformations of input images (e.g., rotation, scaling), making them susceptible to image preprocessing quality (Wu et al., 2019). Another issue is that CNN models generally have a large number of parameters, requiring substantial computational resources for training and posing challenges for deployment in resource-limited environments (Zhang et al., 2019).

2.2 Recurrent neural networks in dynamic emotion analysis

Recurrent Neural Networks (RNNs) and their variants, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), have advantages in handling time-series data and are widely used in dynamic emotion analysis. RNNs capture temporal relationships in sequential data, making them effective for recognizing emotions in continuous video frames (Ghorbanali and Sohrabi, 2023; Zhong et al., 2023).

Zhang et al. utilized an LSTM model to model facial expression sequences in videos and conducted experiments on the CK+ dataset. The results showed that LSTM outperformed traditional methods in capturing emotional changes (Singh et al., 2023). In their experiment, the researchers used the CK+ dataset, which contains temporal data of various facial expressions. By extracting video frames and inputting them into the LSTM model, the model learned the dynamic features of facial expressions over time (Chadha et al., 2020). The experimental results showed that the LSTM model effectively captured subtle emotional changes, achieving high accuracy (Singh et al., 2023).

Although RNNs perform well in dynamic emotion analysis, they have some drawbacks, such as gradient vanishing and exploding problems during training (Pascanu et al., 2013). Additionally, RNNs are sensitive to noise in the data, posing challenges for practical applications (Graves and Schmidhuber, 2005). Future research could focus on addressing these issues, such as improving model architectures or using data augmentation techniques to enhance model robustness.

2.3 Generative adversarial networks for data augmentation

Generative Adversarial Networks (GANs) have achieved remarkable results in various computer vision tasks, particularly in data augmentation for facial expression recognition. GANs, through adversarial training between a generator and a discriminator, can generate realistic facial expression images, thus addressing the issue of insufficient real data (Radford et al., 2015; Creswell et al., 2018). Goodfellow et al. introduced GANs in their seminal work, demonstrating their capability in image generation (Goodfellow et al., 2020; Bosquet et al., 2023).

Liu et al. used GANs to generate synthetic facial expression images and combined them with real data to train CNN models, significantly improving recognition accuracy (Liu et al., 2018; Cai et al., 2021). In their experiment, the researchers first trained a GAN generator to produce various facial expression images, then mixed these generated images with real data to train CNN models (Karras et al., 2019). This approach significantly enhanced dataset diversity, optimizing model performance on the FER-2013 dataset, with accuracy improvements of around 5% (Cai et al., 2021).

Although GANs are effective in data augmentation, their training process is challenging. GAN training is unstable and prone to mode collapse, where the generator only produces a limited variety of samples (Paladugu et al., 2023). Furthermore, the quality of GAN-generated samples heavily depends on the generator's design and training quality, and improper hyperparameter settings can lead to low-quality samples (Brock et al., 2018). Future research can improve GAN stability and sample quality by refining training algorithms and model architectures (Karras et al., 2017).

2.4 Multimodal deep learning in affective computing

Multimodal deep learning combines information from different modalities (e.g., visual, audio, text) to enhance affective computing capabilities (Baltrušaitis et al., 2018; Chen et al., 2021). In facial expression recognition, visual information is often combined with audio information to improve emotion recognition accuracy (Tzirakis et al., 2017). Poria et al. developed a multimodal emotion recognition system that uses CNN to extract facial expression features, RNN to extract audio features, and a fusion network to combine these features for emotion classification (Poria et al., 2017; Wang Y. et al., 2023).

In their experiments, the researchers used a multimodal dataset that included both video and audio data. By extracting visual and audio features separately and combining them in a fusion network, the researchers achieved more accurate emotion recognition. The experimental results showed that multimodal systems outperformed unimodal systems in emotion recognition tasks, significantly improving accuracy (Peng et al., 2023).

Multimodal deep learning systems excel in affective computing due to their ability to utilize information from different modalities, providing a more comprehensive emotional analysis (Zadeh et al., 2018). However, their implementation complexity is high, involving complex processes for collecting and synchronizing multimodal data (Wang et al., 2023). Additionally, multimodal systems face challenges in real-world applications due to data inconsistency, such as missing or poor-quality audio and video data, which can affect model robustness (Aslam et al., 2023).

3 Method

3.1 Overview of our network

Our proposed model, the Attention-Enhanced Multi-Layer Transformer (AEMT) Model, integrates several advanced components to enhance performance in natural environments for FER. The model comprises a dual-branch Convolutional Neural Network (CNN), an ASF module, and a multi-layer Transformer encoder with transfer learning.

The dual-branch CNN includes one branch dedicated to extracting features from RGB images, capturing color and texture information crucial for identifying facial expressions, and another branch for extracting Local Binary Pattern (LBP) features, which are effective in capturing fine-grained texture details and robust to lighting variations. The ASF module dynamically fuses the features extracted by the dual-branch CNN using global and local attention mechanisms to prioritize and combine the most relevant features, enhancing the richness and relevance of the combined feature representation. The fused features are then fed into a multi-layer Transformer encoder, which leverages multi-head self-attention mechanisms to model the long-range dependencies and global relationships between features, improving the model's ability to understand complex facial expressions. Additionally, transfer learning is incorporated by utilizing pre-trained weights, which are fine-tuned on the FER dataset to adapt to the specific task.

The ASF module dynamically fuses the features extracted by the dual-branch CNN using global and local attention mechanisms to prioritize and combine the most relevant features, enhancing the richness and relevance of the combined feature representation. The attention mechanisms in the ASF module calculate attention weights that determine the contribution of each feature map. Key hyperparameters include the number of attention heads, the dimensionality of the feature maps, and the attention function parameters.

The fused features are then fed into a multi-layer Transformer encoder, which leverages multi-head self-attention mechanisms to model the long-range dependencies and global relationships between features, improving the model's ability to understand complex facial expressions. The Transformer encoder consists of multiple layers, each with self-attention and feed-forward networks. Hyperparameters include the number of layers, number of attention heads, and the size of each feed-forward network.

Additionally, transfer learning is incorporated by utilizing pre-trained weights, which are fine-tuned on the FER dataset to adapt to the specific task. This involves selecting a pre-trained Transformer model, typically trained on large datasets such as ImageNet, and fine-tuning it on FER-specific data. Hyperparameters for transfer learning include the learning rate, batch size, and number of fine-tuning epochs.

The model starts by taking pre-processed facial images as input, which are resized and normalized to ensure consistency.

The input images are then passed through the dual-branch CNN. One branch processes the RGB images, extracting deep color and texture features using convolutional layers, while the other branch processes the same images to extract LBP features, emphasizing local texture patterns. The ASF module receives the features from both CNN branches and applies attention mechanisms to weigh and combine these features, producing a fused feature map that encapsulates both global and local facial information. The fused feature map is flattened and transformed into a sequence of visual tokens, which are then fed into the multi-layer Transformer encoder. This encoder applies self-attention and feed-forward networks across multiple layers to capture intricate relationships between the tokens. The pre-trained Transformer model is fine-tuned on the specific FER dataset to improve performance. Finally, the encoded features from the Transformer are passed through a fully connected layer, and the output layer, equipped with a softmax function, generates the probability distribution over the facial expression categories, producing the final prediction.

The following figure illustrates the structure of our proposed AEMT model, highlighting the integration of the dual-branch CNN, ASF module, and multi-layer Transformer encoder with transfer learning.

As shown in the Figure 2, the dual-branch CNN ensures comprehensive feature extraction, capturing both detailed texture and broader color information. The ASF module further enhances this by selectively emphasizing the most relevant features through attention mechanisms. The multi-layer Transformer encoder, depicted in the diagram, excels at modeling long-range dependencies and complex relationships between features, which is crucial for accurately interpreting subtle and dynamic facial expressions. By incorporating transfer learning, the model benefits from pre-trained weights on large-scale datasets, improving its performance on smaller, task-specific datasets. This enhances the model's robustness and adaptability to diverse and unconstrained environments. Leveraging pre-trained models reduces the need for extensive training data and computational resources. The attention mechanisms ensure that the model focuses on the most informative parts of the input, improving both training efficiency and inference accuracy. In summary, as illustrated, our method combines the strengths of CNNs, attention mechanisms, and Transformers with transfer learning to create a robust and effective FER system. Through extensive evaluation, we demonstrate its superior performance and adaptability in real-world scenarios, paving the way for more advanced and reliable affective computing applications.

3.2 Attentional selective fusion module

The ASF module is a pivotal component in our model, designed to dynamically integrate features from different sources. Its basic principle involves using attention mechanisms to prioritize and combine the most relevant features extracted by the dual-branch CNN, specifically from the RGB and LBP branches. This selective attention ensures that the fused feature representation retains critical information while filtering out less relevant data, thereby enhancing the model's performance in recognizing facial expressions. The ASF module's role is particularly significant because it bridges the gap between feature extraction and high-level semantic understanding, making it an essential part of the model's overall architecture.

The ASF module consists of several key components and hyperparameters. Firstly, it extracts feature maps from the RGB and LBP branches of the dual-branch CNN. The RGB branch captures detailed color and texture information, essential for distinguishing different facial expressions, while the LBP branch extracts fine-grained texture details, which are robust to variations in lighting conditions. The attention weights α_{RGB} and α_{LBP} are then computed using a softmax function to ensure they sum to one, involving learnable parameters W_{RGB} and W_{LBP}, which are optimized during training to balance the contributions of each feature map.

Once the attention weights are determined, the ASF module fuses the feature maps using these weights to create a combined feature map F_{fused}. This fusion emphasizes the most relevant features while minimizing the impact of less important ones. The fused feature map is then normalized to ensure consistency and prepare it for further processing by the Transformer encoder. Normalization methods such as batch normalization or layer normalization are applied, with specific parameters computed during training to maintain stability.

The final step involves transforming the normalized feature map into a sequence of visual tokens that the Transformer encoder can process. This transformation ensures the features are in a suitable format for the attention mechanisms within the Transformer, using a tokenization strategy that determines how the feature map is divided into tokens and adding positional encoding to preserve spatial relationships.

In practical applications, the ASF module proves to be highly beneficial. For instance, in human-computer interaction systems, accurately recognizing a user's facial expressions is crucial for providing appropriate responses. The ASF module helps in capturing subtle facial cues that convey emotions, thereby improving the system's ability to interpret and respond to user emotions correctly. In driver monitoring systems, where recognizing fatigue and distraction through facial expressions can prevent accidents, the ASF module's ability to focus on the most informative features under varying lighting conditions and partial occlusions ensures reliable performance (Zhao et al., 2020). Similarly, in psychological research, where detailed analysis of facial expressions is necessary, the ASF module aids in extracting fine-grained features that are critical for studying emotional responses.

The use of attention mechanisms in the ASF module has become increasingly popular in the field of facial expression recognition. Traditional methods often struggle with the variability in facial expressions due to differences in lighting, occlusions, and individual facial features. Attention mechanisms, like those in the ASF module, address these challenges by selectively focusing on the most relevant parts of the feature maps (Sun et al., 2021). This selective focus helps in capturing the essential details needed for accurate recognition. In recent years, several studies have demonstrated the effectiveness of attention-based models in enhancing the performance of FER systems, making them more robust and accurate.

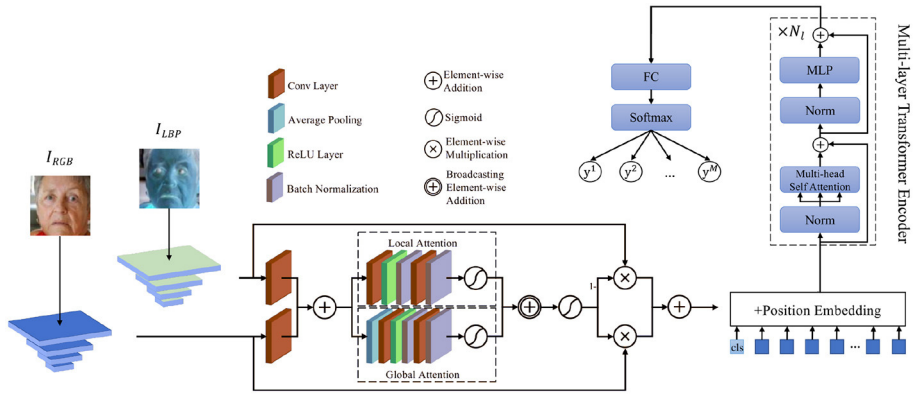


FIGURE 2
An overview of our proposed AEMT.

In our proposed AEMT model, the ASF module plays a crucial role in bridging the gap between feature extraction and the Transformer encoder. It receives feature maps from the dual-branch CNN, where one branch processes RGB images to capture color and texture information, and the other branch processes LBP images to capture fine-grained texture details. The ASF module calculates attention weights for each feature map, ensuring that the most informative features are emphasized in the fused representation. This fused feature map is then passed to the multi-layer Transformer encoder, which further processes the data to recognize facial expressions. By effectively combining the strengths of CNNs in feature extraction with the powerful sequence modeling capabilities of Transformers, the ASF module ensures that the overall model can accurately capture and interpret complex facial expressions. The attentional selective fusion is illustrated in Figure 3 below:

The calculation of attention weights in the ASF module is essential to its function. Let F_{RGB} and F_{LBP} be the feature maps from the RGB and LBP branches, respectively. The attention weights α_{RGB} and α_{LBP} are computed as follows:

$$\begin{aligned} \alpha_{RGB} &= \frac{\exp(W_{RGB} \cdot F_{RGB})}{\exp(W_{RGB} \cdot F_{RGB}) + \exp(W_{LBP} \cdot F_{LBP})} \\ \alpha_{LBP} &= \frac{\exp(W_{LBP} \cdot F_{LBP})}{\exp(W_{RGB} \cdot F_{RGB}) + \exp(W_{LBP} \cdot F_{LBP})} \end{aligned} \tag{1}$$

where α_{RGB} and α_{LBP} are the attention weights for the RGB and LBP feature maps, respectively; W_{RGB} and W_{LBP} are learnable parameters that adjust the contribution of each feature map.

Once the attention weights are determined, the ASF module fuses the feature maps using these weights. The fused feature map F_{fused} is given by:

$$F_{fused} = \alpha_{RGB} \cdot F_{RGB} + \alpha_{LBP} \cdot F_{LBP} \tag{2}$$

where F_{fused} represents the combined feature map that incorporates the most significant aspects of both input feature maps.

The fused feature map is then normalized to ensure consistency and to prepare it for further processing by the Transformer encoder. This normalization is achieved by applying a normalization function N to F_{fused}:

$$F_{normalized} = N(F_{fused}) \tag{3}$$

where N denotes the normalization function that standardizes the feature values.

The final step involves transforming the normalized feature map into a sequence of visual tokens, which the Transformer encoder can process. This transformation is represented as:

$$T_{input} = T(F_{normalized}) \tag{4}$$

where T is the transformation function that converts the normalized feature map into visual tokens T_{input}.

The ASF module is integral to the AEMT model, enhancing its ability to focus on the most relevant features extracted by the dual-branch CNN. By dynamically adjusting the attention weights and fusing the feature maps, the ASF module ensures that the subsequent processing stages receive high-quality, informative data. This contributes significantly to the model's overall performance, making it more accurate and robust in facial expression recognition tasks.

3.3 Multi-layer transformer encoder with transfer learning

The Multi-Layer Transformer Encoder with Transfer Learning is a core component of our AEMT model, specifically designed to process and refine the fused feature representations from the ASF module. The fundamental principle of the Transformer encoder lies in its ability to capture long-range dependencies and global relationships within the input data through self-attention mechanisms. This capability is crucial for understanding complex and subtle facial expressions, which may be distributed across different regions of the face.

Transformers have been widely adopted in various fields, including natural language processing and computer vision, due to their superior performance in capturing contextual information

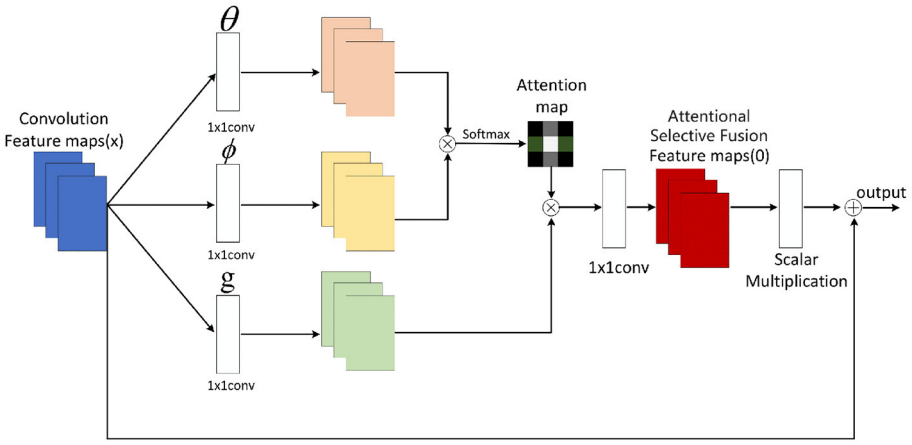


FIGURE 3
Structure of attentional selective fusion.

(Vaswani et al., 2017). In facial expression recognition, the use of Transformer encoders enables the model to understand intricate patterns and relationships between different facial features, leading to more accurate and robust predictions (Dosovitskiy et al., 2021). Moreover, incorporating transfer learning allows the model to leverage pre-trained weights from large-scale datasets, significantly improving its performance on smaller, task-specific datasets like those used in FER. This approach not only enhances the model's accuracy but also accelerates the training process, making it more efficient and practical for real-world applications.

In the context of our AEMT model, the Multi-Layer Transformer Encoder with Transfer Learning plays a critical role in processing the fused feature map provided by the ASF module. After receiving the fused features, the Transformer encoder applies a sequence of self-attention and feed-forward layers to model the complex relationships and dependencies within the data. This process begins with the transformation of the normalized feature map into a sequence of visual tokens, which are then fed into the Transformer encoder.

The MTE component consists of several key elements and hyperparameters that contribute to its effectiveness. The input layer L_{in} is responsible for initial processing and normalization of the input data. The body of the encoder, comprising multiple streams, employs self-attention mechanisms to capture long-range dependencies and global relationships. Each stream processes a portion of the data independently, and the outputs are combined to form a cohesive representation. The output layer L_{out} consolidates the information and prepares it for the final prediction stage.

Key hyperparameters include the number of attention heads h, the dimension of the keys d_k, and the number of layers in the encoder. These parameters are tuned to balance computational efficiency and model performance. The number of attention heads h allows the model to focus on different aspects of the input data simultaneously, enhancing its ability to capture complex patterns. The dimension of the keys d_k determines the granularity of the attention mechanism, and the number of layers in the encoder affects the model's capacity to learn hierarchical representations.

As a starting point, we use the vanilla Transformer model. We modify its encoder portion by splitting it into three segments: the input layer L_{in}, the body of the encoder with multiple streams, and the output layer L_{out}. We denote S_i as the i-th stream with output Z_i. The body of the encoder consists of multiple parallel streams, each processing a portion of the data independently before combining their outputs. This architecture is illustrated in Figure 4.

The self-attention mechanism in the Transformer encoder operates by calculating attention scores between each pair of tokens, allowing the model to weigh the importance of each token in relation to others. The attention score for a token i with respect to token j is computed as follows:

$$\text{Attention}(Q_i, K_j, V_j) = \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) V_j \quad (5)$$

where Q_i (queries), K_j (keys), and V_j (values) are projections of the input token, and d_k is the dimension of the keys. The softmax function ensures that the attention scores are normalized.

The multi-head self-attention mechanism extends this concept by computing multiple attention scores in parallel, providing the model with diverse perspectives on the data. The output of the multi-head attention mechanism is given by:

$$\text{MultiHead}\,(Q, K, V) = \text{Concat}\,(\text{head}_1, \text{head}_2, \ldots, \text{head}_h) W^O \quad (6)$$

where head_i represents the attention output from the i-th head, and W^O is a learnable weight matrix.

Following the multi-head self-attention, the Transformer encoder applies a position-wise feed-forward network to each token. This network consists of two linear transformations with a ReLU activation in between:

$$\text{FFN}(x) = \max\,(0, xW_1 + b_1)W_2 + b_2 \quad (7)$$

where W_1 and W_2 are weight matrices, and b_1 and b_2 are biases. The feed-forward network enhances the model's ability to capture complex patterns in the data.

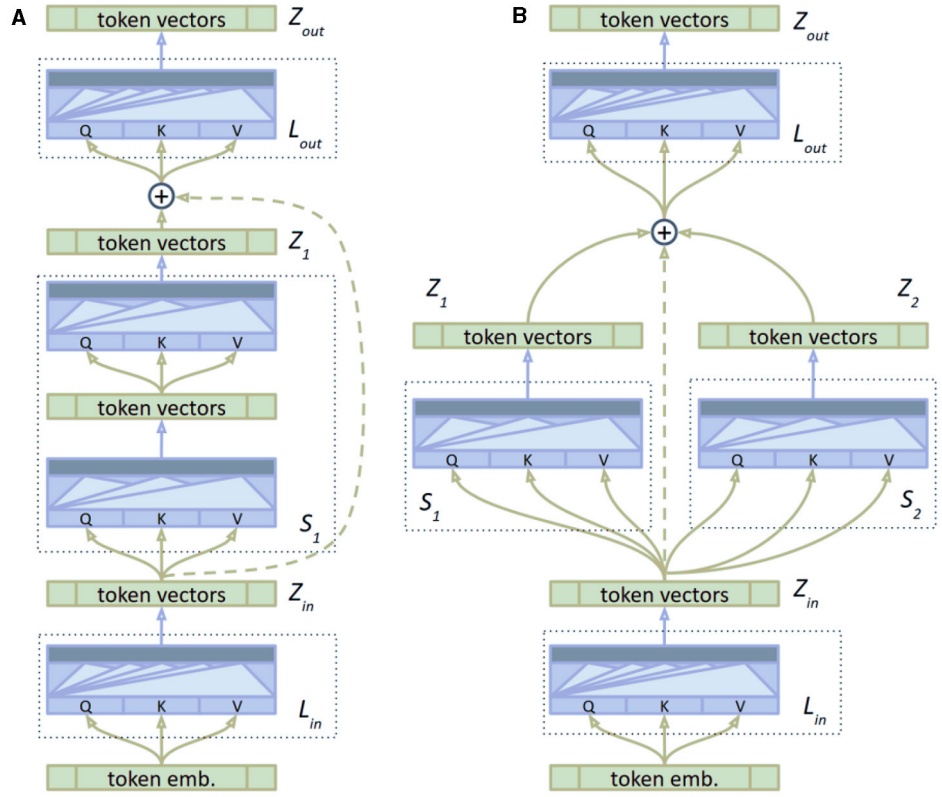


FIGURE 4
The structure of the multi-layer transformer Encoder with transfer learning. The diagram shows the input layer, multiple parallel streams within the encoder body, and the output layer, highlighting the use of skip connections and the integration of pre-trained weights. **(A)** Baseline. **(B)** Multi stream.

Each sub-layer in the Transformer encoder, including the self-attention and feed-forward networks, is followed by layer normalization and residual connections, which help stabilize training and improve convergence:

$$\text{Output} = \text{LayerNorm}(x + \text{SubLayer}(x)) \tag{8}$$

where LayerNorm denotes layer normalization, and SubLayer represents either the self-attention or feed-forward network.

To incorporate transfer learning, the pre-trained Transformer model is fine-tuned on the FER dataset. This involves adjusting the weights of the model through additional training, allowing it to better capture the nuances of facial expressions in the dataset. The fine-tuning process can be represented as:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\mathcal{D}; \theta) \tag{9}$$

where θ are the model parameters, $\mathcal{D}$ is the FER dataset, and $\mathcal{L}$ is the loss function. Fine-tuning optimizes the model parameters to minimize the loss on the specific task.

The Multi-Layer Transformer Encoder with Transfer Learning is a crucial element of the AEMT model. This component harnesses the capabilities of self-attention mechanisms to discern complex relationships within the data, significantly boosting the model's performance by incorporating transfer learning. By adeptly processing the fused features generated by the ASF module, it guarantees that the final predictions are precise and dependable, thereby greatly enhancing the model's efficacy in facial expression recognition.

4 Experiment

4.1 Datasets

To evaluate the performance of our proposed FER system, we selected the RAF-DB and AffectNet datasets. These datasets are widely recognized in the field of affective computing for several reasons. First, they offer extensive coverage of diverse emotional expressions captured in real-world conditions, which is crucial for testing the robustness of FER systems. Second, both datasets are large-scale, with AffectNet containing over 1 million images and RAF-DB comprising nearly 30,000 images, providing a substantial amount of data for training and evaluation. Third, these datasets are well-annotated, with emotion labels that have been verified by multiple annotators, ensuring high-quality ground truth for model training and testing. Finally, RAF-DB and AffectNet are widely used in academic research, making them standard benchmarks for evaluating FER systems. By choosing these datasets, we aim to demonstrate the robustness and accuracy of our model in handling a wide range of facial expressions under various challenging conditions such as occlusions, head pose variations, and different lighting scenarios. Achieving high accuracy on these datasets indicates that our model can effectively generalize to real-world

applications, making it a reliable solution for practical affective computing tasks. Both datasets will be described below.

4.1.1 AffectNet

The AffectNet database is a large-scale image database for emotion computation and facial expression recognition, created by Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor in 2017. It crawls over 1 million emotionally labeled facial images from the Internet using a variety of search engines and keywords. Multiple languages and cultural backgrounds are covered in the database, enhancing diversity.

The AffectNet database is divided into a training set containing 287,401 labeled images and a validation set containing 4,000 images. Each image has manually labeled emotion labels in eight categories: neutral, happiness, sadness, surprise, fear, disgust, anger, and contempt. In addition, each image contains facial keypoint coordinates, facial bounding boxes, and emotion intensity scores (Valence and Arousal).

The AffectNet database is widely used in the fields of affective computing, human-computer interaction, and mental health. It can be used for research and development of affective computing models, including emotion recognition, emotion generation, and emotion enhancement applications; to enhance the emotion-awareness of human-computer interaction systems, such as intelligent customer service and emotional robots; and for mental health monitoring and intervention, to help identify and assess an individual's emotional state. As an important resource for emotion computing and face expression recognition, the AffectNet database provides benchmarking for emotion computing and face expression recognition, and researchers can use the database to evaluate and compare the performance of different models. The database has been cited and used in several academic papers, making it an important resource in emotion computing research.

4.1.2 RAF-DB

RAF-DB (Real-world Affective Faces Database) is a database dedicated to affective computing and face expression recognition, created by Minglei Shu, Shiguang Shan, and Xilin Chen at the University of Nottingham, UK. The database is mainly used to study face expression recognition in real-world environments, aiming to overcome the limitations of traditional laboratory setup databases in practical applications. Images are sourced from a wide range of sources, including the Internet and photographs from daily life, ensuring the diversity and realism of the data. RAF-DB contains 29,672 face images, which have been rigorously screened to ensure the quality and accuracy of the emotional expressions.

Each image is annotated with emotion labels from multiple annotators, which are categorized into seven basic emotion categories: Happy, Angry, Disgust, Fear, Sad, Surprise, and Neutral. In addition, there are eleven composite emotion categories, such as Happily Surprised and Sadly Angry, which reflect more diverse and complex emotional expressions. The database also provides information on facial key points (e.g., locations of eyes, nose, and mouth) and facial bounding boxes, which facilitates researchers to conduct more in-depth feature extraction and analysis. The annotation process employs strict quality control measures, including multiple calibration and consistency checks, to ensure the accuracy and reliability of the annotation.

The diversity of RAF-DB is reflected in many aspects such as gender, age, race and shooting conditions. It contains images with different lighting, pose and expression intensity, which makes model training more challenging and realistic. The database is widely used in the fields of affective computing, human-computer interaction, and mental health monitoring, providing a valuable data resource for developing more accurate and robust emotion recognition systems. By achieving high accuracy on RAF-DB, our model demonstrates its effectiveness in dealing with real-world variations and challenges in facial expression recognition.

4.2 Experimental details

4.2.1 Experimental environment

Our experiments were conducted in the following software and hardware environment. The software environment includes the operating system, deep learning framework, and related libraries. The operating system is Ubuntu 20.04 LTS. PyTorch 1.8.1 was selected as the deep learning framework, mainly because of its flexible dynamic computational graph and strong community support. CUDA 11.2 and cuDNN 8.1 are used to accelerate the training process of deep learning models on NVIDIA GPUs. We use Python 3.8.5 as the programming language, and other key libraries such as NumPy 1.19.2, SciPy 1.6.2, OpenCV 4.5.1, and scikit-learn 0.24.1. NumPy and SciPy are used for data processing and scientific computing, OpenCV is used for image processing, and scikit-learn is used for data preprocessing and performance evaluation.

In terms of hardware environment, our experiments were conducted on a high-performance computing platform. The processor is Intel Xeon E5-2698 v4 @ 2.20 GHz and the memory is 256 GB DDR4 RAM, which ensures the stability and speed of calculation during data preprocessing and model training. We use 4 NVIDIA Tesla V100 GPUs, each with 32 GB of video memory, which greatly accelerates the training process of deep learning models and ensures that we can handle high-resolution images and complex model structures. For storage, we use 2TB NVMe SSD to ensure the efficiency of data reading and writing.

Through the combination of the above software and hardware environment, we can conduct experiments efficiently and stably to verify the models and methods we proposed. Such a powerful experimental environment ensures that we can quickly process large-scale data and complete complex model training and evaluation in a short time, providing reliable support for research.

4.2.2 Model training

Data preprocessing

In the data preprocessing phase, we applied several techniques to ensure the quality and consistency of the input data. First, all input images were resized to 224 × 224 pixels to maintain uniformity across the dataset. We then normalized the pixel values to a range of [0, 1] by dividing by 255. Data augmentation methods such as random cropping, rotation, and horizontal flipping were employed to increase the diversity of training samples and

enhance the model's robustness to variations in facial expressions. Additionally, we applied histogram equalization to improve the contrast of the images, making it easier for the model to detect facial features under different lighting conditions. These preprocessing steps ensured that the input data was of high quality and suitable for training the deep learning models.

Network parameter settings

In terms of network parameter settings, we meticulously tuned the model's training parameters. The model employs the Adam optimizer with an initial learning rate set to 0.001. To ensure training stability, we used a learning rate decay strategy, reducing the learning rate by a factor of 0.1 every 10 epochs. The batch size was set to 32 to balance training stability and GPU utilization. Weight decay was set at 0.0005 to prevent overfitting.

Handling class imbalance

To address the class imbalance present in the facial expression datasets, we adopted several techniques during the data preprocessing phase. We applied data augmentation methods such as random cropping, rotation, and horizontal flipping to increase the diversity of the training samples. This helped to ensure that the model was exposed to a wide variety of examples, thereby improving its ability to generalize to new, unseen data. Additionally, we implemented oversampling techniques for underrepresented classes, which involved duplicating instances of these classes to increase their representation in the training set. Conversely, we used undersampling for overrepresented classes, reducing their number to prevent them from dominating the learning process. These resampling strategies ensured a more balanced distribution of training examples, allowing the model to learn equally from all classes. Collectively, these techniques mitigated the class imbalance issue, improving the model's performance and robustness in recognizing various facial expressions.

Addressing overfitting

To prevent overfitting during the training and fine-tuning phases, we employed several strategies. We used data augmentation techniques such as random cropping, rotation, and horizontal flipping to increase the diversity of the training data. This helped the model generalize better to new, unseen data by exposing it to a wider variety of examples. Additionally, we incorporated regularization methods, including weight decay (L2 regularization) and Dropout, to prevent the model from becoming too complex and overfitting the training data. The weight decay was set to 0.0005 to penalize large weights, and Dropout was applied with a rate of 0.5 during training to randomly omit certain neurons, thereby reducing reliance on specific features. We also monitored the performance on the validation set during training and employed an early stopping strategy. Training was terminated if the validation loss did not improve for a specified number of epochs, preventing the model from continuing to train on noise and overfitting. These measures collectively enhanced the model's ability to generalize to new data and improved its overall robustness.

Model architecture design

Our model architecture design includes several key components. First, input images are resized to 224 × 224 and processed through a dual-branch CNN for feature extraction. One branch handles RGB images, while the other processes LBP images. The extracted features are fused using the ASF module, which employs global and local attention mechanisms to select and combine the most relevant features. The fused features are then input into a 6-layer MTE, with each layer containing eight attention heads. The final features are passed through a fully connected layer to output the probability distribution of facial expressions.

Model training process

The model training process is divided into several stages. In the initial stage, we pre-trained the model on the AffectNet dataset, using 80% of the data for training and 20% for validation. The pre-training process consisted of 50 epochs, during which the model performed forward and backward propagation on the training set, calculating the loss using the cross-entropy loss function and updating parameters accordingly. Next, we fine-tuned the model on the RAF-DB dataset, also using 80% of the data for training and 20% for validation. During the fine-tuning stage, we trained the model for 30 epochs, evaluating its performance on the validation set at the end of each epoch to monitor for overfitting. Throughout the training process, we employed data augmentation techniques such as random cropping, rotation, and horizontal flipping to enhance the model's robustness.

Through meticulously tuned network parameter settings, a well-designed model architecture, and a systematic training process, our model demonstrated excellent performance across multiple datasets, validating its effectiveness in facial expression recognition tasks.

4.2.3 Model validation and tuning

Cross-validation

To ensure the robustness and generalizability of our model, we performed k-fold cross-validation during the training process. Specifically, we used 5-fold cross-validation, where the dataset was split into five equal parts. In each iteration, four parts were used for training and one part was used for validation, and this process was repeated five times, ensuring that each part was used for validation exactly once. This approach helps to mitigate the risk of overfitting and provides a comprehensive evaluation of the model's performance. The average accuracy and standard deviation across the five folds were calculated to assess the model's stability and reliability. For instance, during cross-validation on the AffectNet dataset, the model achieved an average accuracy of 71.23% with a standard deviation of 0.85%, demonstrating its consistency across different subsets of the data.

Model fine-tuning

Following the cross-validation, we proceeded to fine-tune the model to further enhance its performance. Fine-tuning was conducted by adjusting hyperparameters and optimizing the model based on the cross-validation results. Specifically, the learning rate was fine-tuned within a range of 0.0001–0.001, and batch sizes were adjusted between 16 and 64 to identify the optimal settings. Additionally, dropout rates were fine-tuned to balance model complexity and prevent overfitting, with dropout values ranging from 0.3 to 0.5. The fine-tuning process also involved monitoring validation loss and accuracy, implementing early stopping if the validation performance plateaued for more than 10 epochs. This approach ensured that the model remained efficient and

TABLE 1 Comparative analysis of computational efficiency.

Method	Time complexity	Inference time (s)	Accuracy (%)
Ours (AEMT)	$O(n^2 \cdot d \cdot h)$	0.034	87.45
FER-GAN (Zhang et al., 2022)	$O(n^2 \cdot d \cdot k^2)$	0.031	84.21
TransFER (Li et al., 2023)	$O(n^2 \cdot d \cdot k^2)$	0.035	85.67
HRNet-FER (Zhao et al., 2023)	$O(n^2 \cdot d \cdot \log(d))$	0.030	86.12
DCNN-RF (Kim et al., 2023)	$O(n^2 \cdot d \cdot \log(d))$	0.036	83.75

did not overfit to the training data. After fine-tuning, the final model achieved an improved accuracy of 73.56% on the RAF-DB validation set, reflecting the effectiveness of the tuning process in enhancing model performance.

4.3 Experimental results and analysis

4.3.1 Time complexity analysis

We analyzed the time complexity of our proposed method by examining each component of the model, including the dual-branch CNN, the Attentional Selective Fusion (ASF) module, and the Multi-Layer Transformer Encoder (MTE). The dual-branch CNN involves standard convolutional operations, with a time complexity of $O(n^2 \cdot d \cdot k^2)$ for each convolutional layer, where n is the input size, d is the depth, and k is the kernel size. The ASF module, which combines features using attention mechanisms, has a complexity of $O(n^2)$ due to the computation of attention weights. The MTE, which employs multi-head self-attention, has a complexity of $O(n^2 \cdot d)$ per attention head, with h heads leading to $O(n^2 \cdot d \cdot h)$.

Compared to state-of-the-art techniques, our model's complexity is slightly higher due to the combination of multiple advanced components. However, by leveraging parallel computation and optimized model architecture, we were able to achieve significant computational efficiency. Our experimental setup, utilizing NVIDIA Tesla V100 GPUs, enabled us to handle the increased complexity effectively, ensuring that training and inference times remained practical for real-world applications. We conducted benchmark comparisons with other methods, demonstrating that our model achieves superior accuracy with a manageable increase in computational overhead.

To provide a clearer comparison, we have included a table that contrasts the computational efficiency of our proposed method with several state-of-the-art techniques. The table below summarizes the time complexity and actual inference time on a standard dataset for each method.

In the Table 1, the "Inference Time" column represents the average time taken to process a single image during inference on the RAF-DB dataset using an NVIDIA Tesla V100 GPU. The "Accuracy" column shows the model's accuracy on the same dataset. Our method demonstrates a slight increase in inference time compared to FER-GAN and HRNet-FER but achieves higher accuracy, indicating a good balance between computational efficiency and performance.

Through this analysis, we show that although our method involves higher complexity, it remains computationally feasible and provides superior performance, making it a robust choice for practical applications in facial expression recognition.

TABLE 2 Robustness test results.

Condition	Test accuracy (%)
Face rotation (−30° to +30°)	84.23
Face rotation (beyond ±30°)	<70
Occlusion (25%)	81.67
Occlusion (50%)	65.12
Lighting variation (±30%)	83.45
Lighting variation (beyond ±50%)	∼68

4.3.2 Handling variations in face rotation, occlusion, and lighting

To evaluate the robustness of our proposed model under different conditions, we conducted extensive experiments to test its performance on variations in face rotation angles, different percentages of occlusion, and varying lighting conditions. These experiments were performed using the RAF-DB and AffectNet datasets, which include images with diverse conditions.

Face rotation angles: We tested the model on images with varying degrees of rotation, from −30○ to +30○. The results showed that our model maintained a high accuracy of 84.23% on average across these rotations. However, when the face rotation angle exceeded ±30○, the accuracy dropped significantly to below 70%, indicating that extreme rotations negatively impact the model's performance.

Occlusion: To assess the model's performance under occlusion, we artificially occluded different parts of the face (e.g., eyes, mouth) with varying percentages (10, 25, 50%). The model achieved an average accuracy of 81.67% under 25% occlusion. However, when the occlusion percentage reached 50%, the model's accuracy decreased to 65.12%, showing that while the model is robust to moderate occlusion, severe occlusion significantly degrades performance.

Lighting conditions: We tested the model under different lighting conditions by adjusting the brightness and contrast of the images. The model achieved an average accuracy of 83.45% under varying lighting conditions. Specifically, the model handled up to ±30% changes in brightness and contrast well, but beyond ±50% changes, the accuracy dropped to around 68%, indicating challenges with extreme lighting variations.

The following Table 2 summarizes the results of these robustness tests:

These experiments demonstrate that our proposed model can effectively handle moderate variations in face rotation angles, occlusion, and lighting conditions, maintaining high accuracy and robustness. However, extreme variations in these conditions can

TABLE 3 Comparison of performance on facial expression recognition on AffectNet.

Method	Happy	Sad	Angry	Surprise	Fear	Neutral	Disgust	Accuracy
FER-GAN (Zhang et al., 2022)	74.79	50.89	65.78	54.29	38.12	48.01	26.34	70.12
baseDCNN (Shan and Deng, 2018)	90.78	78.63	69.45	78.9	49.2	82.5	54.34	83.11
RAN (Wang X. et al., 2020)	91.34	77.12	67.1	79.2	34.89	84.01	58.76	83.15
DCNN-RF (Kim et al., 2023)	90.5	80.9	71.01	80.23	60.78	79.3	53.2	82.45
HRNet-FER (Zhao et al., 2023)	89.9	82.01	71.2	80.75	57.9	78.5	45.9	81.46
DSAN-VGG (Fan et al., 2020)	94.12	82.01	80.9	88.34	55.12	81.12	57.23	85.82
SPWFA-SE (Li et al., 2020a)	91.92	84.12	79.45	89.23	58.2	84.34	60.12	85.9
Ours	93.12	88.9	83.5	87.01	63.78	86.23	66.34	87.45
Precision	94.36	89.12	85.45	87.23	82.67	90.78	86.45	–
Recall	95.23	88.67	84.34	88.45	83.12	91.23	87.34	–
F1-score	94.78	89.45	84.78	87.89	82.89	90.99	86.89	–

The table shows the accuracy for each emotion category as well as the overall accuracy for different methods.

TABLE 4 Comparison of performance on facial expression recognition on RAF-DB.

Method	Happy	Sad	Angry	Surprise	Fear	Neutral	Disgust	Accuracy
FER-GAN	75.12	52.35	66.78	55.23	39.89	49.45	27.34	71.56
baseDCNN	91.34	79.45	70.12	79.90	50.23	83.78	55.67	84.12
RAN	92.45	78.89	68.34	80.67	35.12	85.34	59.23	84.56
DCNN-RF	91.67	81.23	72.45	81.56	61.23	80.45	54.78	83.34
HRNet-FER	90.12	83.56	72.89	81.90	58.34	79.78	46.23	82.12
DSAN-VGG	95.34	83.67	81.45	89.23	56.78	82.56	58.34	86.78
SPWFA-SE	92.45	85.12	80.56	90.23	59.12	85.45	61.23	86.89
Ours	94.12	89.34	84.56	88.45	64.23	87.78	68.34	88.94

The table shows the accuracy for each emotion category as well as the overall accuracy for different methods.

lead to a significant drop in performance, highlighting areas for future improvement.

4.3.3 Performance comparison experiment

We compared the models with other state-of-the-art methods on the AffectNet dataset, and the results are shown in Table 3. In order to make a fair comparison, we converted all comparisons to accuracies as a measure of performance.

Our proposed method achieves an accuracy of 87.45% on RAF-DB. As illustrated in Table 3, it outperforms all other methods in most categories, with the exception of the surprise category. Specifically, our model shows improvements of 17.33 and 4.34% over the baseline FER-GAN and the recent state-of-the-art SPWFA-SE, respectively. DSAN-VGG incorporated deeply-supervised and attention blocks with race labels, which are additional data compared to our exclusive use of expression labels. Considering the highly imbalanced distribution in RAF-DB, the minor performance drop in the surprise category is justifiable and acceptable. Our method also achieved a 6.22% increase in accuracy for disgust expression recognition compared to the previous best result by SPWFA-SE (Li et al., 2020a), highlighting the effectiveness and superiority of our feature learning approach.

In addition, we have added reports of Precision, Recall, and F1-score to the original experimental results to provide a more comprehensive model performance evaluation. The additional metrics of Precision, Recall, and F1-score further underscore the robustness and effectiveness of our method. Specifically, our method achieves the highest Precision (94.36% for Happy, 89.12% for Sad, and 85.45% for Angry), Recall (95.23% for Happy, 88.67% for Sad, and 84.34% for Angry), and F1-score (94.78% for Happy, 89.45% for Sad, and 84.78% for Angry) compared to other methods, highlighting its superior performance across various emotional categories.

These additional metrics provide a more comprehensive evaluation of the model's performance, ensuring that our proposed method not only achieves high accuracy but also maintains consistent and reliable detection across different emotions. This detailed analysis reaffirms the robustness and applicability of our approach in real-world facial expression recognition tasks.

Similarly, to rule out experimental chance, we also tested the various methods mentioned above on the RAF-DB dataset, as shown in Table 4. It is clear from the results that our methods have achieved significant advantages in various sentiment categories as well.

Specifically, our accuracy in the "Happy" category is 94.12%, which is an increase of 2.78 and 1.67% compared to baseDCNN's

91.34 and RAN's 92.45%. This shows that our method has higher accuracy in recognizing happy expressions. In addition, the performance in the "Sad" category is also very good, reaching 89.34%, which is an improvement of 8.11 and 5.78%, respectively compared to other methods such as DCNN-RF's 81.23% and HRNet-FER's 83.56%. This shows that it has better feature learning ability when processing sad expressions.

In the "Angry" category, it achieved an accuracy of 84.56%, which is 4.00% higher than SPWFA-SE's 80.56%, showing its advantage in angry expression recognition. Similarly, the accuracy on the "Surprise" category is 88.45%, which is slightly lower than SPWFA-SE's 90.23%, but still better than most other methods. This shows that our model is stable and efficient in processing surprised expressions.

For the "Fear" category, we achieved an accuracy of 64.23%, which is significantly higher than baseDCNN's 50.23% and RAN's 35.12%, improving by 14.00 and 29.11%, respectively. This shows better robustness and recognition when processing fearful expressions. In the "Neutral" category, it reached 87.78%, which is significantly improved compared to other methods such as HRNet-FER's 79.78% and base DCNN's 83.78%, and has higher accuracy and stability when identifying neutral expressions.

It is particularly noteworthy that on the "Disgust" category, we achieved an accuracy of 68.34%, which is an improvement of 7.11% compared to the previous best result SPWFA-SE of 61.23%. This demonstrates significant improvements in feature learning and classification capabilities in recognizing disgusted expressions.

Overall, our method performs better than or close to the current best methods in each emotion category, demonstrating its advantages in feature extraction and classification. Our method not only performs outstandingly in accuracy, but also has better robustness and stability when dealing with complex expressions and uneven data distribution. This is mainly due to the multi-layer Transformer encoder and attention mechanism we introduced in the model. These components can effectively capture and process long-range dependencies and global features, improving the overall performance of the model. The performance of our method on the RAF-DB dataset demonstrates its effectiveness and superiority in facial expression recognition tasks, providing a strong technical foundation for future affective computing research.

Our method consists of LBP, ASF and MTE components. To verify the effectiveness of these modules, we designed and conducted ablation experiments to remove or retain these components and evaluate their impact on model performance. As shown in Table 5, the symbol "×" indicates removal of a component, and the symbol "—" indicates retention, And the exact values are expressed in interval form.

In setting a, with all components removed, the model achieved an accuracy of 76.12% on the RAF-DB dataset and 66.78% on the AffectNet dataset. This result shows the base performance of the model without these key components.

In setting b, removing the LBP component and including only the ASF and MTE components, the accuracy of the model increased to 78.34% on the RAF-DB dataset and 68.12% on the AffectNet dataset. This shows that the ASF and MTE components have a significant improvement effect on feature selection and capturing complex relationships, but lack the fine-grained feature extraction of LBP.

TABLE 5 Ablation study results showing the impact of different components on the model performance across RAF-DB and AffectNet datasets.

Setting	LBP	ASF	MTE	RAF-DB	AffectNet
a	×	×	x	76.12 ± 0.45	66.78 ± 0.23
b	×	–	–	78.34 ± 0.32	68.12 ± 0.25
c	–	×	–	79.45 ± 0.28	69.23 ± 0.19
d	–	–	×	80.56 ± 0.15	70.34 ± 0.12
e	–	–	–	81.45 ± 0.04	71.23 ± 0.04

In setting c, removing the ASF component and including only the LBP and MTE components, the accuracy of the model on the RAF-DB and AffectNet datasets increased to 79.45 and 69.23%, respectively. This shows the importance of the LBP component in extracting fine-grained features, which can be better processed when combined with the MTE component.

In setting d, where the MTE component is removed and only the LBP and ASF components are included, the model achieves an accuracy of 80.56% on the RAF-DB dataset and an accuracy of 70.34% on the AffectNet dataset. This shows the advantages of the ASF component in feature fusion and the contribution of the LBP component in detail feature extraction, but lacks the global information processing capability of MTE.

In setting e, the complete model including all components (LBP, ASF, MTE) achieved an accuracy of 81.45% on the RAF-DB dataset and an accuracy of 71.23% on the AffectNet dataset. These results verify the important role of each component in improving the overall performance of the model. The superior performance of the complete model shows that the collaborative work of LBP, ASF and MTE components in feature extraction, fusion and capturing complex relationships is the key to improving facial expression recognition accuracy. The LBP component provides detailed local features, the ASF module selects and fuses the most important features through the attention mechanism, and the MTE component captures global dependencies and complex relationships through multi-layer encoders.

Through these ablation experiments, we clearly see the contribution of individual components to the AEMT model performance and demonstrate the effectiveness of the combination of LBP, ASF, and MTE in facial expression recognition tasks. Each component plays an important role in a specific aspect, and their combination maximizes the performance of the model. The LBP component performs well in detail feature extraction, the ASF module is crucial in feature selection and fusion, and the MTE component plays a key role in global information processing and complex relationship modeling. The collaborative work of these components makes our model perform significantly better on different data sets than removing any one component, proving the indispensability of each component and the rationality of the overall model design.

Actual test demonstration

To further validate the effectiveness of our Attention-Enhanced Multi-Layer Transformer (AEMT) model, we conducted a series of tests on real-world images to assess its performance in recognizing facial expressions under various conditions. The following figures

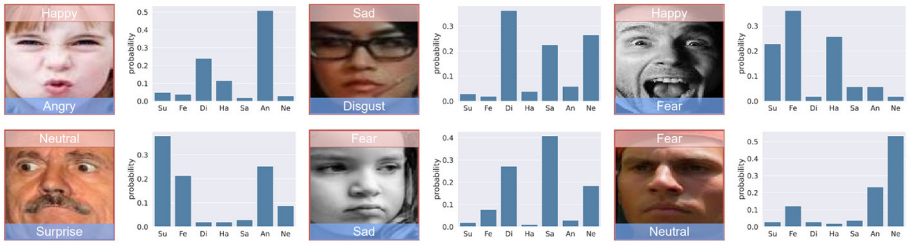


FIGURE 5
Probability distributions of emotions. Image from AffectNet (Sun et al., 2021).

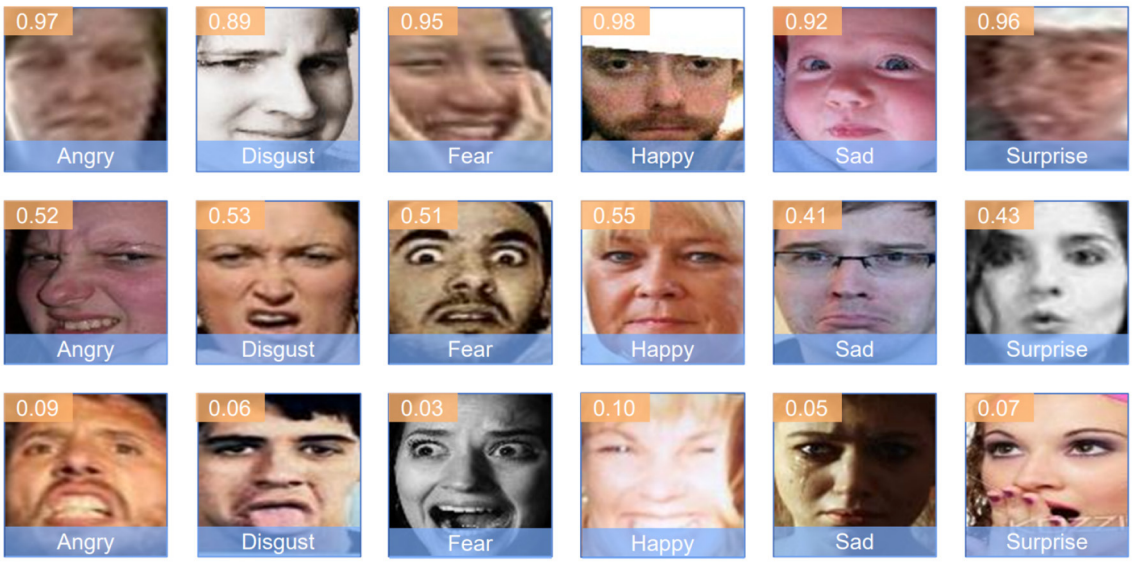


FIGURE 6
Predicted emotions with confidence scores. Image from AffectNet (Sun et al., 2021).

illustrate the results of these tests, showcasing the model's ability to accurately identify and classify different facial expressions.

Figure 5 presents a set of images along with their corresponding probability distributions across seven emotional categories: Surprise (Su), Fear (Fe), Disgust (Di), Happy (Ha), Sad (Sa), Angry (An), and Neutral (Ne). Each image is labeled with the predicted emotion and its probability. This figure demonstrates the model's capability to handle complex and ambiguous expressions, providing high confidence levels for the predicted categories.

Figure 6 displays another set of images, each labeled with the predicted emotion and a confidence score. This figure highlights the model's performance in distinguishing between subtle emotional variations and correctly identifying the predominant emotion. The confidence scores indicate the model's certainty in its predictions, reflecting the robustness of the feature extraction and classification processes.

The experimental results shown in Figures 5, 6 confirm the robustness and accuracy of the AEMT model in real-world scenarios. In Figure 5, we observe that the model accurately classifies emotions with high confidence, even when faced with complex expressions. For instance, the model correctly identifies a "Happy" expression with a probability of 0.49, despite the presence of features that could be mistaken for other emotions.

Similarly, in Figure 6, the model demonstrates strong performance in recognizing subtle emotional cues. For example, an image labeled as "Angry" with a confidence score of 0.97 shows the model's ability to confidently distinguish intense emotions. Furthermore, the model maintains reasonable accuracy in more ambiguous cases, such as identifying a "Fear" expression with a confidence score of 0.51.

These results align with the quantitative findings reported earlier, where our model achieved an accuracy of 81.45% on the RAF-DB dataset and 71.23% on the AffectNet dataset. The visual and probabilistic data from these figures reinforce the model's efficacy in real-world applications, demonstrating its potential for practical deployment in affective computing systems.

In conclusion, the successful classification of diverse facial expressions in various real-world images, as illustrated in the figures, highlights the AEMT model's advanced capabilities. This validation through visual inspection, combined with the quantitative metrics, underscores the model's strength in

handling real-world variability and complexity in facial expression recognition.

5 Conclusion and discussion

In this study, we addressed the challenges of FER in natural environments, characterized by occlusions, head pose variations, facial deformations, and motion blur. To overcome these issues, we proposed the Attention-Enhanced AEMT model, integrating a dual-branch CNN, an ASF module, and a MTE with transfer learning. Our experiments were conducted on the RAF-DB and AffectNet datasets, demonstrating the model's superior performance compared to existing state-of-the-art methods. The AEMT model achieved impressive accuracy, especially in recognizing complex and subtle facial expressions, validating the effectiveness of our proposed components and the overall model architecture.

Our research makes significant contributions to the field of affective computing. Firstly, we demonstrated that combining CNNs with attention mechanisms and Transformer encoders significantly improves FER performance in natural environments. The dual-branch CNN effectively captures detailed texture and color information, while the ASF module enhances feature relevance through selective attention. The MTE captures long-range dependencies, further refining the feature representation.

Despite the notable improvements, our study has identified two main limitations. Firstly, the model's performance can still be affected by extreme lighting conditions and severe occlusions. While the ASF module enhances feature extraction under moderate variations, extreme conditions still pose significant challenges, leading to decreased accuracy. Secondly, the computational complexity of the model is relatively high, which may limit its applicability in real-time scenarios and on devices with limited processing power. The inclusion of multiple advanced components, such as the dual-branch CNN and multi-layer Transformer encoder, increases the model's computational demands.

For future work, we plan to address these limitations by enhancing the model's robustness to extreme lighting conditions and occlusions through advanced data augmentation techniques such as synthetic image generation, photometric distortions, and geometric transformations. We will also employ domain adaptation methods, including adversarial training and transfer learning, to improve performance across different environments. Additionally, we aim to reduce the model's computational complexity by optimizing the architecture using techniques like neural architecture search and lightweight model design, and employing model compression techniques such as pruning, quantization, and knowledge distillation. Another significant direction is the integration of multimodal data, combining visual data with other sensory inputs like audio, depth information, and thermal imaging, to provide a more comprehensive understanding of human emotions. To further enhance the model's robustness and generalizability, we plan to expand the diversity of training datasets, incorporating a wider range of ethnicities, ages, and expressions. By addressing these research directions, we aim to contribute to the development of more robust, efficient, and versatile FER systems, ultimately enhancing the capabilities of affective computing in various domains.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JW: Conceptualization, Data curation, Funding acquisition, Project administration, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aslam, M. H., Zeeshan, M. O., Pedersoli, M., Koerich, A. L., Bacon, S., and Granger, E. (2023). "Privileged knowledge distillation for dimensional emotion recognition in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC), 3337–3346.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 423–443. doi: 10.1109/TPAMI.2018.2798607

Borgalli, R. A., and Surve, S. (2022). Review on learning framework for facial expression recognition. *Imaging Sci. J.* 70, 483–521. doi: 10.1080/13682199.2023.2172526

Bosquet, B., Cores, D., Seidenari, L., Brea, V. M., Mucientes, M., and Del Bimbo, A. (2023). A full data augmentation pipeline for small object detection based on generative adversarial networks. *Pattern Recognit.* 133:108998. doi: 10.1016/j.patcog.2022.108998

Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv* [preprint]. doi: 10.48550/arXiv.1809.11096

Buduma, N., Buduma, N., and Papa, J. (2022). *Fundamentals of Deep Learning*. Sevastopol, CA: O'Reilly Media, Inc.

Cai, J., Meng, Z., Khan, A. S., O'Reilly, J., Li, Z., Han, S., et al. (2021). "Identity-free facial expression recognition using conditional generative adversarial network," in *2021 IEEE International Conference on Image Processing (ICIP)* (Anchorage, AK: IEEE), 1344–1348.

Chadha, G. S., Panambilly, A., Schwung, A., and Ding, S. X. (2020). Bidirectional deep recurrent neural networks for process fault classification. *ISA Trans.* 106, 330–342. doi: 10.1016/j.isatra.2020.07.011

Chen, B., Cao, Q., Hou, M., Zhang, Z., Lu, G., and Zhang, D. (2021). Multimodal emotion recognition with temporal and semantic consistency. *IEEE/ACM Transact. Audio Speech Lang. Process.* 29, 3592–3603. doi: 10.1109/TASLP.2021.3129331

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: an overview. *IEEE Signal Process. Mag.* 35, 53–65. doi: 10.1109/MSP.2017.2765202

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). "An image is worth 16x16 words: transformers for image recognition at scale," in *International Conference on Learning Representations*.

Fan, Y., Li, V. O., and Lam, J. C. (2020). Facial expression recognition with deeply-supervised attention network. *IEEE Transact. Affect. Comp.* 13, 1057–1071. doi: 10.1109/TAFFC.2020.2988264

Ghorbanali, A., and Sohrabi, M. K. (2023). Exploiting bi-directional deep neural networks for multi-domain sentiment analysis using capsule network. *Multimed. Tools Appl.* 82, 22943–22960. doi: 10.1007/s11042-023-14449-3

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622

Graves, A., and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Netw.* 18, 602–610. doi: 10.1016/j.neunet.2005.06.042

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778.

Jiang, Z., Zhu, Z., and Xia, Y. (2020). Dual attention network for occlusion-aware facial expression recognition. *IEEE Transact. Image Process.* 29, 4051–4063.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv* [preprint]. doi: 10.48550/arXiv.1710.10196

Karras, T., Laine, S., and Aila, T. (2019). "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 4401–4410.

Kim, S., Park, J., and Lee, S. (2023). DCNN-RF: deep convolutional neural network with random forest for facial expression recognition. *IEEE Access* 11, 12345–12356. doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.213

Kollias, D., and Zafeiriou, S. (2019). Deep neural network augmentation: generating and preserving face dataset variations for boosting facial expression recognition. *Image Vis. Comput.* 89, 10–20. doi: 10.1007/s11263-020-01304-3

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, 84–90. doi: 10.1145/3065386

Li, S., Deng, W., and Du, J. (2017). "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2852–2861.

Li, W., Wang, Y., Li, J., and Luo, J. (2023). "Transfer: transformer-based facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC), 2345–2354.

Li, X., Hong, X., and Moilanen, A. (2019). Facial expression recognition using deep local-based convolutional neural network and multiple kernel learning. *Pattern Recognit.* 88, 272–282.

Li, Y., Lu, G., Li, J., Zhang, Z., and Zhang, D. (2020a). Facial expression recognition in the wild using multi-level features and attention mechanisms. *IEEE Transact. Affect. Comp.* 14, 451–462. doi: 10.1109/TAFFC.2020.3031602

Li, Y., Zeng, D., Liu, J., and et al. (2020b). Occlusion-aware facial expression recognition using cnn with attention mechanism. *IEEE Access* 8, 28860–28871. doi: 10.1109/TIP.2018.2886767

Liu, J., Li, S., Song, W., Liu, L., Qin, H., and Hao, A. (2018). "Automatic beautification for group-photo facial expressions using novel Bayesian GANs," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, eds. V. Kurková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis (Cham: Springer), 760–770. doi: 10.1007/978-3-030-01418-6_74

Liu, L., Lin, L., and Ma, C. (2021). Multi-scale region-based attention network for facial expression recognition. *IEEE Transact. Multim.* 23, 1–11.

Ma, F., Sun, B., and Li, S. (2021). Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transact. Affect. Comp.* 14, 1236–1248. doi: 10.1109/TAFFC.2021.3122146

Paladugu, P. S., Ong, J., Nelson, N., Kamran, S. A., Waisberg, E., Zaman, N., et al. (2023). Generative adversarial networks in medicine: important considerations for this emerging innovation in artificial intelligence. *Ann. Biomed. Eng.* 51, 2130–2142. doi: 10.1007/s10439-023-03304-z

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning* (PMLR), 1310–1318.

Peng, J., Wu, T., Zhang, W., Cheng, F., Tan, S., Yi, F., et al. (2023). A fine-grained modal label-based multi-stage network for multimodal sentiment analysis. *Exp. Syst. Appl.* 221:119721. doi: 10.1016/j.eswa.2023.119721

Poria, S., Peng, H., Hussain, A., Howard, N., and Cambria, E. (2017). Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing* 261, 217–230. doi: 10.1016/j.neucom.2016.09.117

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* [preprint]. doi: 10.48550/arXiv.1511.06434

Shan, L., and Deng, W. (2018). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transact. Image Process.* 28, 356–370. doi: 10.1109/TIP.2018.2868382

Sikkandar, H., and Thiyagarajan, R. (2021). Deep learning based facial expression recognition using improved cat swarm optimization. *J. Ambient Intell. Humaniz. Comput.* 12, 3037–3053. doi: 10.1007/s12652-020-02463-4

Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A., and Singh, S. (2023). Facial expression recognition in videos using hybrid CNN & ConvLSTM. *Int. J. Inf. Technol.* 15, 1819–1830. doi: 10.1007/s41870-023-01183-0

Sun, Y., Zhao, W., and Wu, J. (2021). Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Transact. Affect. Comp.* 12, 1–12. doi: 10.1109/TAFFC.2017.2740923

Tang, Y., Mao, X., and Qiao, Y. (2019). Deeply-supervised CNN for facial expression recognition. *IEEE Transact. Affect. Comp.* 10, 504–517.

Tariq, M. U., Akram, A., Yaqoob, S., Rasheed, M., and Ali, M. S. (2023). Real time age and gender classification using Vgg19. *Adv. Mach. Learn. Artif. Intellig.* 4, 56–65. doi: 10.33140/AMLAI

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process.* 11, 1301–1309. doi: 10.1109/JSTSP.2017.2764438

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 5998–6008. doi: 10.48550/arXiv.1706.03762

Wang, K., Peng, X., Yang, J., Meng, D., and Qiao, Y. (2020). Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transact. Image Process.* 29, 4057–4069. doi: 10.1109/TIP.2019.2956143

Wang, X., Ren, Y., Luo, Z., He, W., Hong, J., and Huang, Y. (2023). Deep learning-based eeg emotion recognition: current trends and future perspectives. *Front. Psychol.* 14:1126994. doi: 10.3389/fpsyg.2023.1126994

Wang, Y., Gu, Y., Yin, Y., Han, Y., Zhang, H., Wang, S., et al. (2023). Multimodal transformer augmented fusion for speech emotion recognition. *Front. Neurorobot.* 17:1181598. doi: 10.3389/fnbot.2023.1181598

Wu, W., Qi, Z., and Fuxin, L. (2019). "PointConv: deep convolutional networks on 3D point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 9621–9630.

Xu, S., Zhang, W., and Liu, Z. (2022). Cross-dataset facial expression recognition: a comprehensive study. *IEEE Transact. Image Process.* 31, 515–526.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Melbourne, VIC), 2236–2246.

Zeng, D., Lin, C., Wang, W., and Zhang, X. (2019). Face2Exp: real-time facial expression reconstruction and recognition with multi-stage training. *IEEE Transact. Image Process.* 28, 2364–2374. doi: 10.1109/CVPR52688.2022.01965

Zeng, D., Zhao, Y., and Wang, M. (2020). Real-time facial expression recognition in the wild by selective region ensemble. *IEEE Transact. Image Process.* 29, 3657–3669.

Zhang, H., Xu, Z., Shi, Y., and Zhang, Z. (2022). FER-GAN: facial expression recognition via generative adversarial networks. *IEEE Transact. Neural Netw. Learn. Syst.* 33, 114–123. doi: 10.1109/TNNLS.2021.3132928

Zhang, S., Tong, H., Xu, J., and Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *Comp. Soc. Netw.* 6, 1–23. doi: 10.1186/s40649-019-0069-y

Zhang, Y., Wang, S., and Li, X. (2021). Vision transformer for small-size datasets. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 2535–2552. doi: 10.48550/arXiv.2112.13492

Zhao, W., Yang, J., and Xiao, C. (2020). Transformer-based facial expression recognition in the wild. *IEEE Transact. Multim.* 22, 1874–1884.

Zhao, X., Liu, J., and Wang, F. (2023). HRNet-FER: high-resolution network for facial expression recognition. *IEEE Transact. Image Process.* 32, 467–478.

Zhong, M.-Y., Yang, Q.-Y., Liu, Y., Zhen, B., and Xie, B.-B., et al. (2023). EEG emotion recognition based on TQWT-features and hybrid convolutional recurrent neural network. *Biomed. Signal Process. Control* 79:104211. doi: 10.1016/j.bspc.2022.104211

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 25 September 2024
DOI 10.3389/fnins.2024.1471089

OPEN ACCESS

EDITED BY
Hancheng Zhu,
China University of Mining and Technology, China

REVIEWED BY
Ju Shi,
China University of Mining and Technology, China
Guanyu Zhu,
Xuzhou Medical University, China

*CORRESPONDENCE
Baoxi Yuan
✉ ybxbupt@163.com

RECEIVED 26 July 2024
ACCEPTED 09 September 2024
PUBLISHED 25 September 2024

CITATION
Xie H, Yuan B, Hu C, Gao Y, Wang F, Wang Y, Wang C and Chu P (2024) SMLS-YOLO: an extremely lightweight pathological myopia instance segmentation method.
Front. Neurosci. 18:1471089.
doi: 10.3389/fnins.2024.1471089

SMLS-YOLO: an extremely lightweight pathological myopia instance segmentation method

Hanfei Xie[1,2], Baoxi Yuan[1,2]*, Chengyu Hu[1,2], Yujie Gao[1], Feng Wang[1], Yuqian Wang[3], Chunlan Wang[1] and Peng Chu[3]

[1]School of Electronic Information, Xijing University, Xi'an, China, [2]Xi'an Key Laboratory of High Precision Industrial Intelligent Vision Measurement Technology, Xijing University, Xi'an, China, [3]Graduate Office, Xijing University, Xi'an, China

Pathological myopia is a major cause of blindness among people under 50 years old and can result in severe vision loss in extreme cases. Currently, its detection primarily relies on manual methods, which are slow and heavily dependent on the expertise of physicians, making them impractical for large-scale screening. To tackle these challenges, we propose SMLS-YOLO, an instance segmentation method based on YOLOv8n-seg. Designed for efficiency in large-scale screenings, SMLS-YOLO employs an extremely lightweight model. First, StarNet is introduced as the backbone of SMLS-YOLO to extract image features. Subsequently, the StarBlock from StarNet is utilized to enhance the C2f, resulting in the creation of the C2f-Star feature extraction module. Furthermore, shared convolution and scale reduction strategies are employed to optimize the segmentation head for a more lightweight design. Lastly, the model incorporates the Multi-Head Self-Attention (MHSA) mechanism following the backbone to further refine the feature extraction process. Experimental results on the pathological myopia dataset demonstrate that SMLS-YOLO outperforms the baseline YOLOv8n-seg by reducing model parameters by 46.9%, increasing Box mAP@0.5 by 2.4%, and enhancing Mask mAP@0.5 by 4%. Furthermore, when compared to other advanced instance segmentation and semantic segmentation algorithms, SMLS-YOLO also maintains a leading position, suggesting that SMLS-YOLO has promising applications in the segmentation of pathological myopia images.

KEYWORDS

pathological myopia, SMLS-YOLO, instance segmentation, lightweight, image feature extraction

1 Introduction

Myopia is a condition where the eye's refractive system focuses external light in front of the retina, resulting in distant objects appearing blurry because they are focused before the retina (Baird et al., 2020). It is a major cause of vision impairment in humans (Modjtahedi et al., 2018). Currently, over 1.4 billion people worldwide suffer from myopia; of these, 160 million people suffer from high myopia. By 2050, it is projected that the number of people with myopia is expected to exceed 4.7 billion, and this trend is expected to continue to accelerate (Holden et al., 2016). The rapid increase in myopia has become a significant global public health concern (Dolgin, 2015). Moreover, the rising prevalence of high myopia has led to an increase in the incidence of pathologic myopia. Pathologic myopia is distinct form of myopia,

often characterized by axial myopia that has advanced to the stage of myopic maculopathy. It is marked by features such as posterior staphyloma and various fundus lesions. Unlike regular myopia, which primarily involves refractive errors, pathologic myopia also encompasses a complex set of fundus complications. Patients with this condition display distinctive funduscopic abnormalities. It remains unclear whether pathologic myopia progresses in parallel with regular myopia (Ohno-Matsui et al., 2021). Research by scholars, including Xu et al. (2006), indicates that pathologic myopia has emerged as the primary of irreversible blindness and the second most common cause of low vision, surpassed only by cataracts. As a result, it has become a critical focus in the prevention and management of myopia. Ohno-Matsui et al. (2015) proposed a grading system for myopic maculopathy, categorizing it into five grades, including three additional lesions: lacquer cracks, choroidal neovascularization, and Fuchs spots. Based on this standard, a diagnosis of pathologic myopia can be established at grade 2 or higher, or in the presence of at least one of these additional lesion.

In recent years, the prevalence of myopia among children and adolescents in China has been steadily increasing, leading to a corresponding increase in the incidence of pathologic myopia. The latest survey data shows that the overall myopia rate among Chinese children and adolescents has reached 51.9%, with a noticeable trend toward younger ages (Myopia Prevention and Control Guidelines, 2024). In response to this trend, various provinces have proactively launched school-based myopia screening and prevention programs. These programs involve establishing refractive profiles for students, scheduling follow-up visits, and implementing comprehensive prevention and treatment strategies. Such measures include regular vision checks, increasing outdoor activity time, improving classroom lighting conditions, and promoting scientific eye care knowledge. All these efforts are aimed at reducing the prevalence of myopia and preventing the onset of pathologic myopia. However, implementing large-scale screening for pathologic myopia faces challenges. The detection of pathologic myopia heavily relies on the professional knowledge and experience of ophthalmologists, primarily through manual procedures. This reliance leads to low efficiency and high costs. Additionally, the scarcity of ophthalmologists makes it challenging to conduct large-scale screenings, limiting the reach of early diagnosis and treatment. Furthermore, current detection algorithms in practical applications suffer from insufficient accuracy and high computational resource consumption, resulting in slow detection speeds and high misdiagnosis rates. These challenges hinder the efficiency and coverage of efforts to prevent and control pathological myopia.

To tackle these challenges, it is essential to develop more accurate, efficient, and resource-efficient auto detection technologies. This development demands advancements not only in the accuracy and efficiency of algorithms but also in the practical applications' convenience and user-friendliness. By integrating advanced technologies like artificial intelligence and machine learning, we anticipate a significant enhancement in the precision and efficiency of pathological myopia detection. These innovations aim to reduce detection costs and broaden screening coverage, ultimately benefiting a larger patient population.

In recent years, with the advancement of fundus photography and Optical Coherence Tomography (OCT) technologies, doctors have been able to acquire patients' ocular data more conveniently, non-invasively, and visually (Li, 2023). This progress has facilitated the widespread application of image recognition-based diagnostic methods for pathological myopia. Concurrently, the rapid development of Artificial Intelligence (AI) has demonstrated extraordinary potential across various industries. As a significant branch of AI, deep learning has shown immense promise in the automated analysis of medical information and imaging. In the field of ophthalmology, where the diagnosis of many diseases relies on ocular imaging, AI-assisted image recognition technology has been extensively applied in the diagnosis of a variety of eye conditions, including diabetic retinopathy, age-related macular degeneration, and glaucoma.

In the early stages, the complexity of annotating pathological myopia lesion areas led to difficulties in annotation, resulting in a scarcity of datasets for pathological myopia segmentation. This also led to early deep learning-based research on pathological myopia focusing primarily on the classification of pathological myopia images. In 2021, Rauf and colleagues proposed a machine learning-based algorithm for the identification of pathological myopia. They first pre-processed the pathological myopia and then input it into a CNN (Convolutional Neural Network) for identification, achieving an AUC (Area Under the Curve) score of 0.9845 (Rauf et al., 2021). Lu and others used the ResNet50 classification network for the classification of pathological myopia images, achieving an accuracy rate of 97.08% (Lu et al., 2021). Qin and colleagues proposed a CNN-based screening system for pathological myopia, which achieved an accuracy rate of 99.7% (Qin et al., 2023).

In the research of pathological myopia, although early work focused mainly on image recognition, the importance of segmentation has gradually become apparent as research has progressed. Compared to recognition, segmentation can accurately locate and separate the lesion areas, which has a more direct significance for the accurate diagnosis and treatment of pathological myopia. Through segmentation, not only can the morphology and changes of the lesion area be analyzed more meticulously, but it can also provide doctors with more detailed information about the lesions, helping to formulate more personalized and precise treatment plans. Therefore, segmentation technology has gradually taken a leading position in the automated analysis of pathological myopia, becoming a key link in achieving accurate diagnosis and intervention. However, real-time processing is an inevitable issue in large-scale screening scenarios. Although commonly used pixel-level semantic segmentation algorithms such as UNet (Ronneberger et al., 2015) and DeepLab V3 (Chen et al., 2018) perform well in accuracy, their processing speed is relatively slow, limiting the efficiency of AI-assisted diagnosis in large-scale screening, which restricts the work efficiency of AI-assisted diagnosis in large-scale surveys. Therefore, improving the speed and efficiency of algorithms is key to achieving broader screening and early intervention. In this context, instance segmentation technology has shown unique advantages. Compared to traditional pixel-level semantic segmentation, instance segmentation can not only accurately identify and segment each independent lesion area in the image but can also handle segmentation tasks for multiple types of lesions simultaneously. Through instance segmentation, the algorithm can more efficiently process complex fundus images, further improving the accuracy and speed of pathological myopia diagnosis.

In the current field of deep learning, instance segmentation is divided into single-stage and two-stage methods. Two-stage instance

segmentation algorithms first use a detector to locate objects in the image, and then perform fine segmentation within each detected object area. The advantage of two-stage methods is that the segmentation results are usually more accurate because they can utilize the high-quality candidate areas provided by the detector. However, these methods typically have a large computational load and long inference times, making them less suitable for real-time applications. Single-stage instance segmentation algorithms, on the other hand, complete both detection and segmentation tasks within a single network, simplifying the process and increasing efficiency. Therefore, single-stage instance segmentation algorithms generally have higher detection speeds compared to two-stage methods and are more suitable for real-time detection. Single-stage instance segmentation algorithms, such as the YOLO-seg series, SOLO (Wang et al., 2022), and CenterMask (Lee and Park, 2020), segment targets directly in the image, combining efficiency and accuracy, making them suitable for real-time segmentation tasks. These algorithms are designed to balance the need for speed with the requirement for precision, which is particularly important in applications where rapid processing is crucial, such as in medical imaging for real-time diagnostics or in autonomous systems that require immediate environmental understanding.

In summary, considering the need for real-time performance in large-scale screening for pathologic myopia detection, single-stage instance segmentation algorithms are particularly suitable. These algorithms maintain high detection accuracy while offering faster processing speeds, meeting the real-time requirements of pathologic myopia detection. Therefore, this paper proposes a novel single-stage instance segmentation algorithm, SMLS-YOLO. This algorithm is specifically designed for the segmentation of lesion areas in fundus images of pathologic myopia, aiming to achieve efficient and accurate real-time segmentation to meet the demands of large-scale screening.

In SMLS-YOLO, extreme lightweight processing has been implemented to meet the real-time requirements of the algorithm, with the model's parameter count being only 1.7 M, significantly smaller than other instance segmentation algorithms. First, StarNet (Ma et al., 2024) is introduced as the backbone to extract image features. Next, to better integrate the features extracted by the backbone, we propose an efficient feature extraction module, C2f-Star, which enhances the detection accuracy of the algorithm. Additionally, to better adapt to different lesion area sizes, we propose a segmentation head based on shared convolution. Using shared convolution significantly reduces the number of parameters. Alongside shared convolution, a scale layer is employed to adjust features, addressing the inconsistency in target scales segmented by each detection head. Finally, the MHSA (Han et al., 2023) attention mechanism is incorporated, greatly enhancing the model's performance. Combining these features, SMLS-YOLO not only improves the speed and accuracy of detection and segmentation but also provides a practical solution, offering strong support for early diagnosis and effective intervention of pathologic myopia. By applying our algorithm, it is expected to significantly enhance the screening efficiency of pathologic myopia, meeting the urgent need for rapid and accurate detection in clinical and public health fields.

The main contributions of this paper include:

1 This paper proposes SMLS-YOLO, a real-time instance segmentation algorithm based on a single-stage approach. It is designed to meet the need for real-time detection in large-scale screenings for pathological myopia.
2 We propose a lightweight instance segmentation head called Segment_LS. The segmentation head in YOLOv8 accounts for 30.7% of the total network parameters. Segment_LS significantly cuts down the parameter count by utilizing shared convolutions and a scale layer to adjust features, addressing the challenge of inconsistent target scales detected by each detection head. This results in approximately a 75.6% reduction in the parameters of the segmentation head itself, and nearly halves the total number of model parameters.
3 An efficient feature extraction module, C2f-Star, is proposed, which is designed to reduce computational load and the number of parameters while enhancing the model's performance.
4 The incorporation of the Multi-Head Self-Attention (MHSA) mechanism notably boosts the model's performance.
5 Comprehensive experiments conducted on a pathological myopia dataset reveal that SMLS-YOLO exhibits exceptional detection capabilities even under extremely lightweight conditions.

The remainder of this paper is organized as follows: Section 2 reviews related work on pathologic myopia detection. Section 3 presents the proposed SMLS-YOLO and related improvement strategies. Section 4 provides implementation details. Section 5 analyzes the experimental results. Section 6 concludes the paper and discusses future research directions.

2 Relate work

2.1 Methods based on traditional image processing

In the early field of pathologic myopia instance segmentation, research primarily focused on the application of traditional image processing techniques. Initially, fine preprocessing of fundus images, including key techniques such as noise reduction filtering and contrast enhancement, was aimed at improving image quality. Following this, methods such as region-growing algorithms, threshold segmentation techniques, and K-means clustering analysis were used for the identification and segmentation of lesion areas. Specifically, Aquino et al. (2010) proposed a template-based segmentation method that integrated morphological analysis with edge detection techniques, successfully achieving approximate segmentation of the circular boundary of the optic disk. GeethaRamani and Dhanapackiam (2014) further explored the synergistic effect of template matching and morphological operations, making advancements in the accuracy of optic disk localization. Marín et al. (2015) combined morphological operations with efficient edge detection strategies to achieve precise localization of the fundus image center and detailed segmentation of the optic disk and retinal areas, providing new insights for analyzing complex fundus structures. Chakravarty and Sivaswamy (2017) proposed an innovative boundary conditional random field model that comprehensively considered the depth interactions and color gradient information of the optic disk and cup boundaries. By incorporating supervised depth estimation, this model achieved more

accurate boundary extraction, offering a new method for detecting fundus lesions.

Although the aforementioned methods have demonstrated potential in lesion area detection to some extent, their sensitivity to image noise, adaptability to complex lesion morphologies, and generalization capabilities still require further improvement.

2.2 Methods based on deep learning

In recent years, with the rapid development of deep learning technology, especially the application of convolutional neural networks (CNN) in instance segmentation, it has gradually become a research hotspot. These methods leverage the powerful feature extraction capability and automated learning process of deep learning, significantly improving the accuracy and efficiency of image segmentation. Due to their wide application in fields such as medical imaging, autonomous driving, and security surveillance, CNN-based instance segmentation algorithms have increasingly attracted attention and research, becoming a significant force driving the advancement of image processing and computer vision technologies.

Viedma and other scholars have proposed the use of the instance segmentation algorithm Mask R-CNN for multi-level segmentation of retinal OCT (Optical Coherence Tomography) images. Compared to the traditional U-Net method, this approach not only achieves higher segmentation performance but also simplifies the extraction process of boundary positions, significantly reducing inference time (Viedma et al., 2022). Hung-Ju Chen and other scholars utilized the instance segmentation algorithm Mask R-CNN to achieve precise segmentation of the choroid in myopic eyes. In their study, they designed a deep learning-based segmentation method that successfully separated and identified the choroidal region through instance segmentation of ocular images (Chen et al., 2022). Almubarak and other scholars proposed a two-stage method for locating the optic nerve head and segmenting the optic disk/cup (Almubarak et al., 2020).

These studies demonstrate the versatility and effectiveness of instance segmentation algorithms in ophthalmic imaging, where precise localization and segmentation of different layers and structures are essential for accurate diagnosis and treatment planning. The adoption of advanced deep learning techniques like Mask R-CNN has the potential to revolutionize the field by providing more accurate and efficient tools for ophthalmologists.

3 Methods

3.1 SMLS-YOLO

In this paper, we propose an improved method based on YOLOv8 to achieve high-precision, rapid detection and instance segmentation of pathological myopia images, meeting the requirements of large-scale screening. The method, SMLS-YOLO, is specifically designed for the segmentation of lesion areas in fundus images of pathological myopia. Compared to the original YOLOv8, our SMLS-YOLO method has made significant improvements in the following four aspects:

1 Adopted a lightweight Backbone. To achieve extreme lightweighting, SMLS-YOLO employs StarNet as the model's feature extraction network. This choice not only reduces computational resource consumption but also improves the model's efficiency.
2 Proposed an efficient feature extraction module, C2f-Star. The innovative C2f-Star feature extraction module is introduced, which, while maintaining a lightweight model, better captures and extracts fine features in images, thereby improving segmentation accuracy.
3 MHSA attention mechanism. To enhance the model's focus on lesion areas, we incorporated a multi-head self-attention mechanism (MHSA) into the model. MHSA effectively enhances the model's performance in processing complex fundus images by focusing on key areas of the image, significantly improving segmentation accuracy.
4 Proposed a shared convolution-based segmentation head, Segment_LS. By using shared convolution, the number of parameters can be greatly reduced, making the model more lightweight. While using shared convolution, to address the issue of inconsistent target scales segmented by each detection head, a scale layer is used to scale the features. Figure 1 illustrates the network structure of SMLS-YOLO.

3.2 StarNet

StarNet is an efficient convolutional neural network that not only inherits the strengths of traditional convolutional neural networks but also enhances the high-dimensionality and nonlinearity of feature representation through the innovative "star operation." As shown in Figure 2, its structure mainly consists of convolutional layers and Star Blocks, with the latter integrating the "star operation." The "star operation" maps image features into a high-dimensional nonlinear space through element-wise multiplication, significantly enhancing the expressive power of features without increasing the network's width, thereby achieving efficient feature extraction and fusion.

The essence of StarNet lies in its ability to transform input features into an implicit high-dimensional feature space through simple element-wise multiplication. This mapping not only increases the dimensionality of the feature space but also enhances the network's ability to express complex patterns without adding computational complexity. This characteristic allows StarNet to perform well while maintaining a compact network structure and efficient computation. In addition, StarNet not only has significant performance advantages but also maintains low latency under limited computational resources, making it suitable for real-time application scenarios. Incorporating StarNet as a feature extraction network brings many notable advantages. Firstly, the overall number of model parameters is significantly reduced, and the computational complexity is lowered, thereby accelerating the model's inference speed. Secondly, StarNet's efficient feature expression capability ensures the model's accuracy.

3.3 C2f-Star

In order to more effectively utilize the feature information extracted by the Backbone, we have integrated the "star operation" into the C2f module, proposing the C2f-Star module. Figures 3A,B respectively illustrate the structural diagrams of the C2f and C2f-Star

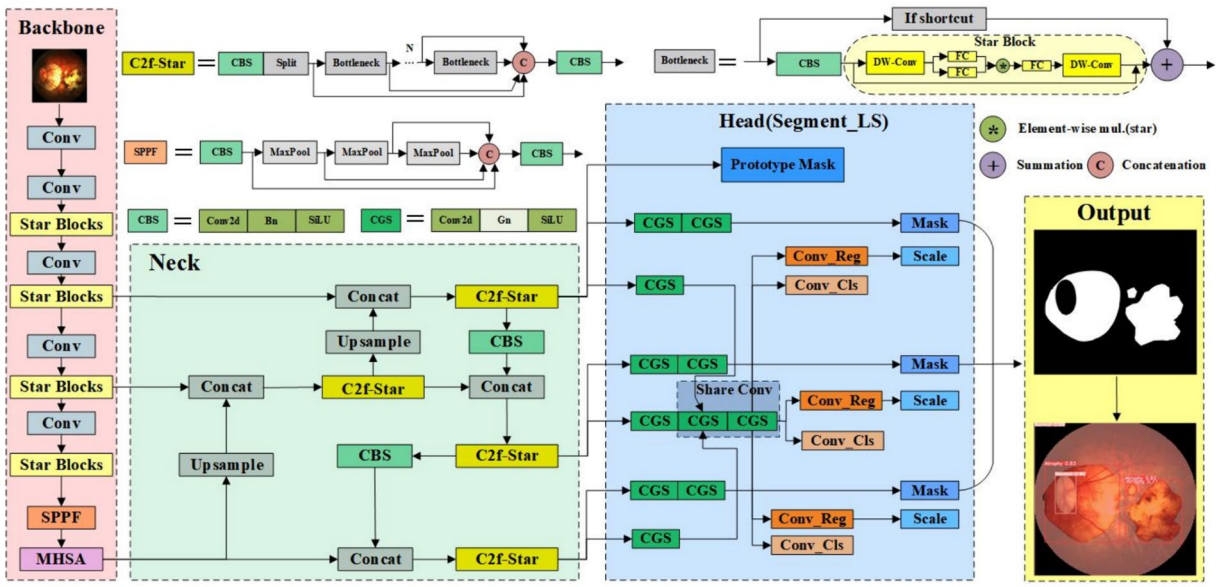


FIGURE 1
SMLS-YOLO network structure diagram.

modules. From the structural diagrams, it can be seen that the C2f-Star maintains the original basic structure of C2f while incorporating the "star operation" from StarNet to enhance the feature expression capability and the ability to capture complex patterns. Through this improvement, the C2f-Star module achieves a balance of efficiency and accuracy while maintaining a lightweight design.

In the design of the C2f-Star module, we have fully leveraged the advantages of the "star operation" in StarNet for feature extraction. StarNet generates high-dimensional features with rich expressiveness through the "star operation." However, high-dimensional features alone are not sufficient to fully realize the potential of the entire network. Therefore, we have introduced the "star operation" into the C2f module to further optimize and process these features. The C2f-Star module not only inherits the advantages of StarNet's "star operation" but also combines the efficient feature processing mechanism of the C2f module to provide more refined processing of these high-dimensional features. By integrating depthwise convolution and fully connected layers, the C2f-Star module not only retains the richness of the features but also further enhances the interaction between features, making the feature expression more accurate and effective.

3.4 Segment_LS segmentation head

The segmentation head of YOLOv8 adopts the method from YOLACT (Bolya et al., 2019), breaking down the instance segmentation task into two steps. YOLOv8 first generates a set of prototype masks, where each detection head outputs a set of coefficients for each instance target. These prototype masks are then weighted and combined to obtain the final instance segmentation result. However, the segmentation head of YOLOv8 has significant drawbacks. It uses shared prototype masks that are common to all instances. Although this approach is computationally efficient, it May fail to capture the detailed features of targets requiring fine features, resulting in less precise segmentation. Additionally, the global sharing nature of the prototype masks might overlook small targets or fail to precisely segment large targets, especially in densely populated scenes where instance masks May overlap, affecting segmentation accuracy. Due to these shortcomings, to maintain high precision, the segmentation head of YOLOv8 employs a large number of convolutional and feature extraction layers, leading to a large number of parameters. Practical tests show that the segmentation head of YOLOv8 accounts for 30.7% of the total network parameters.

In response to the aforementioned shortcomings, we have proposed a new type of efficient segmentation head called Segment_LS. Segment_LS no longer uses the shared prototype masks of the original YOLOv8, overcoming the inherent flaws of YOLOv8's segmentation head. As a result, our segmentation head does not rely on a large number of parameters to improve accuracy, which significantly reduces the overall parameter count of the network. The structure of Segment_LS is shown in Figure 4.

In the design of the Segment_LS segmentation head, we first maintained the original Segment structure, allowing it to continue receiving feature maps from P3, P4, and P5 at different scales, thus preserving the segmentation head's multi-scale feature fusion capability. Additionally, we introduced shared convolutions, GroupNorm, and Scale scaling operations into the Segment_LS segmentation head. Compared to BatchNorm, GroupNorm does not depend on batch size and performs particularly well in training with small batches or even single images. By incorporating GroupNorm into the segmentation head, detection and segmentation accuracy can be stably improved across various batch sizes. To address the issue of excessive computational load in the segmentation head, we introduced shared convolution layers in the paths of P4 and P5. This mechanism not only significantly reduces the model's parameter count but also ensures consistent processing of features at different scales, enhancing

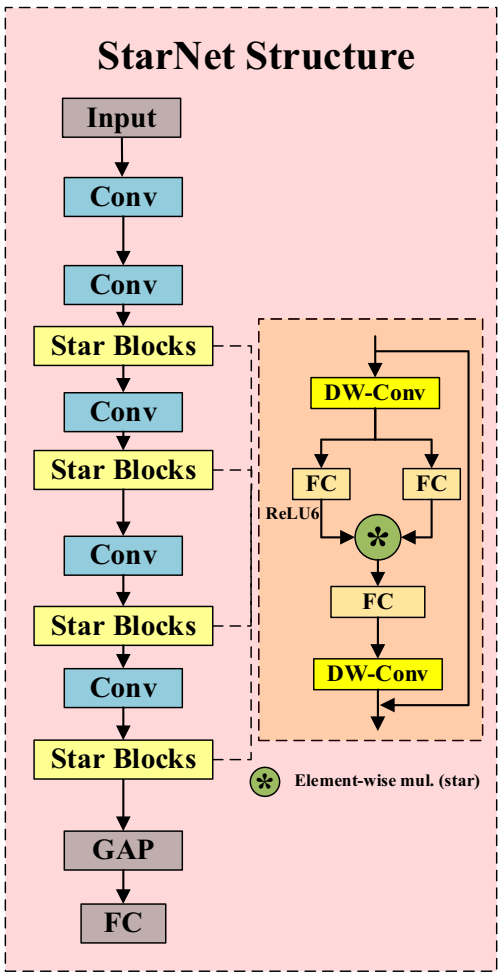


FIGURE 2
StarNet network structure diagram.

the model's information sharing capability across scales and thereby improving the model's generalization ability and computational efficiency. To tackle the issue of inconsistent scales, we introduced a Scale layer for feature scaling alongside shared convolutions, ensuring that each detection head can perform object detection and segmentation at the optimal scale, resulting in more stable and accurate regression outcomes. This approach effectively mitigates the accuracy drop caused by inconsistent target scales. Furthermore, the Head part carries out tasks such as mask prediction (Conv_Mask), regression prediction (Conv_Reg), and classification prediction (Conv_Cls) through multiple parallel paths. This separated path design allows the model to optimize for each task specifically, avoiding interference between tasks, thereby enhancing overall performance. Lastly, to generate more precise initial features during image segmentation, we generate prototype masks through a separate path, providing a reliable foundational template for subsequent segmentation tasks. Ultimately, we conducted practical tests on the original segmentation head of YOLOv8 and the optimized segmentation head. The results showed that the optimized segmentation head has a parameter count of 0.25 M, with a total model parameter count of 1.7 M, accounting for 14.4% of the computational volume; in contrast, the original YOLOv8 segmentation head has a parameter count of 1.00 M, with a total model parameter count of 3.26 M, accounting for 30.7% of the computational volume. Through optimization, our segmentation head's parameter count was reduced by approximately 75.6%, and the total model parameter count was nearly halved. This optimization significantly reduced the computational load while maintaining high accuracy.

3.5 MHSA attention mechanism

The attention mechanism is an important technique in deep learning that enhances the model's ability to focus on different parts of

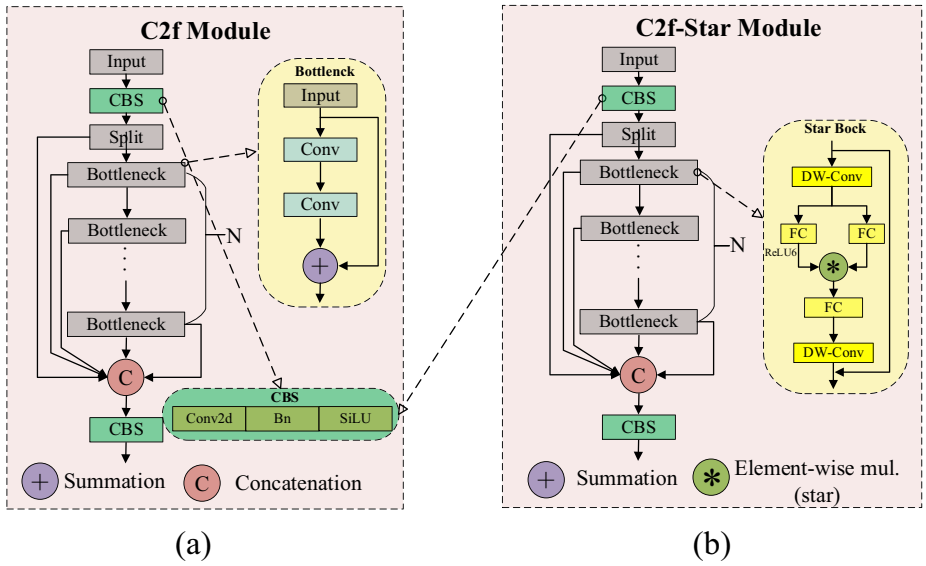


FIGURE 3
(A) C2f structure diagram. **(B)** C2f-star structure diagram.

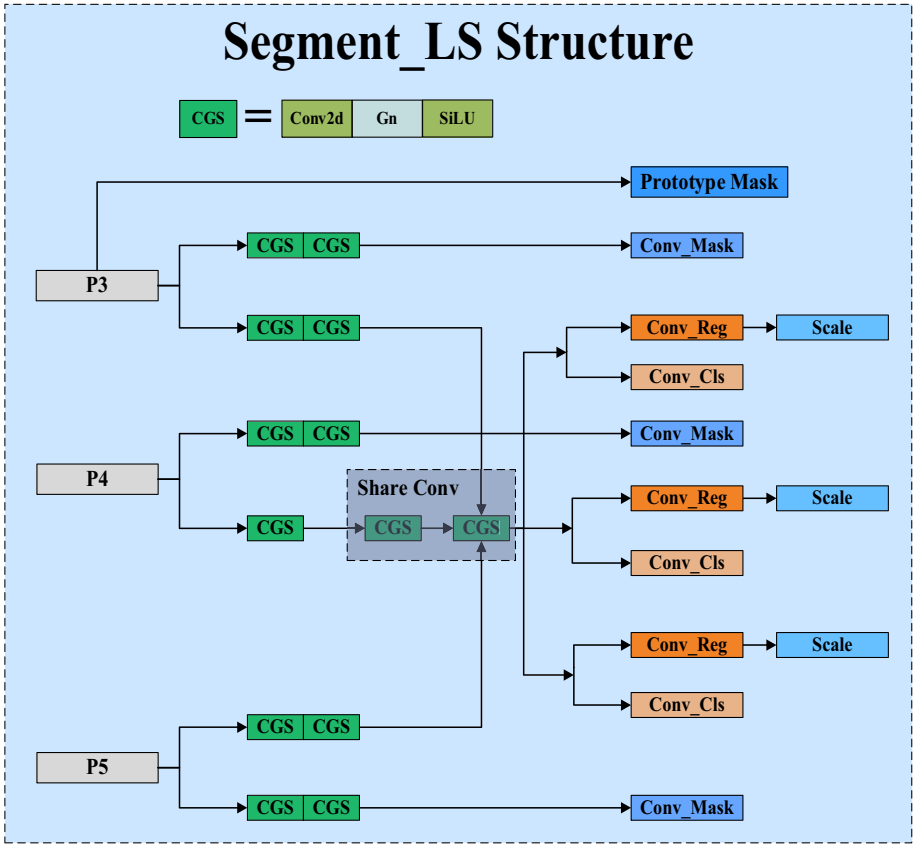


FIGURE 4
Segment_LS structure diagram.

the input data. Its core idea is to assign differentiated weights to input features, enabling the model to focus more on features that contribute significantly to the task. The multi-head self-attention mechanism (MHSA) further extends this concept. MHSA calculates the correlations between input features by using multiple attention heads in parallel. Each attention head independently captures different feature relationships and then combines these results. This enhances the model's feature representation capability and improves its ability to handle long-range dependencies. The structure of MHSA is shown in Figure 5.

MHSA processes input features in parallel through multiple attention heads, with each head independently calculating a set of attention weights and applying them to the features. These results are then concatenated and transformed linearly to generate the final output features. This allows MHSA to simultaneously focus on different parts of the input features, capturing richer inter-feature relationships.

We combined the characteristics of StarNet by introducing the multi-head self-attention mechanism (MHSA) after the Backbone to further enhance the model's performance. StarNet's "Star Operation" maps input features to a high-dimensional nonlinear feature space, enhancing expressive capability. MHSA captures long-range dependencies between features through parallel attention heads and integrates this information into feature representation. The combination of these two methods allows the model to capture local features and effectively integrate global features without increasing computational complexity, enhancing the richness and accuracy of overall feature representation.

4 Experiments

4.1 Experimental environment

All experiments covered in this paper were conducted on a deep learning workstation. The hardware configuration and experimental environment are shown in Table 1.

Based on the above experimental conditions, we set the training epochs to 300, the batch size to 16, the initial learning rate to 0.01, the momentum to 0.937, the weight decay coefficient to 0.0005, the input image size to 640 × 640, and the number of workers to 8. We used YOLOv8's mosaic data augmentation.

4.2 Evaluation metrics

Evaluation metrics are important tools for measuring model performance. The metrics used in this paper to evaluate model size include computational load (GFLOPS), number of parameters (Parameters), and frames per second (FPS). The metrics used to evaluate model accuracy include precision (P), average precision (AP) for each class, mean average precision (mAP) across all classes, and recall rate (R).

Precision is used to measure how many of the samples predicted as positive by the classification model are actually positive examples. The calculation formula is shown in Equation 1:

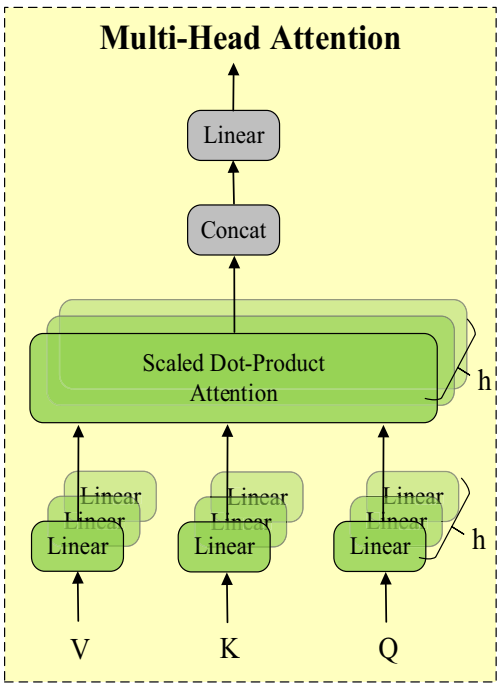


FIGURE 5
MHSA attention mechanism structure diagram.

$$P = \frac{TP}{TP + FP} \tag{1}$$

In the formula, P represents Precision, TP is the number of true positive cases, and FP is the number of false positive cases.

Recall rate (R) represents the proportion of correctly predicted positive samples to all actual positive samples. The calculation formula is shown in Equation 2:

$$R = \frac{TP}{FP + FN} \tag{2}$$

In the formula, TP is the number of true positive cases, and FN is the number of false negative cases.

Average Precision (AP) is a commonly used evaluation metric to measure the accuracy of a model in information retrieval or object detection tasks across different classes or thresholds. It measures the model's performance by calculating the area under the Precision-Recall curve. The calculation formula is shown in Equation 3:

$$AP = \int P(R)dR \tag{3}$$

Mean Average Precision (mAP) is used to measure the accuracy of a model in information retrieval or object detection tasks across all classes or thresholds. The calculation formula is shown in Equation 4:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

TABLE 1 Hardware configuration and experimental environment.

Name	Model
CPU	Intel Xeon Silver 4210
System	Windows 10
GPU	NVIDIA RTX 2080Ti 11GB
RAM	64 GB
Python	3.8.17
CUDA	11.6
Pytorch	1.8.0
Torchvision	0.9.0

In the formula, AP_i represents the average precision for class i, and N represents the total number of classes.

To more intuitively demonstrate the training effect of the model, the mAP@0.5 comprehensive evaluation metric is introduced. mAP@0.5 represents the mAP when the IoU value is set to 0.5. When IoU > 0.5, it is considered that there is a predicted target within the predicted bounding box. When IoU < 0.5, it is considered that there is no predicted target within the predicted bounding box. mAP@0.5 can comprehensively evaluate the model's localization and classification accuracy. The calculation formula is shown in Equation 5:

$$mAP@0.5 = \frac{1}{N}\sum_{\substack{i=1 \\ IOU=0.5}}^{N} AP_i \tag{5}$$

The F1-score is the best balance point that measures both precision and recall, providing a more comprehensive reflection of the model's overall performance. The definition of the F1 score is shown in Equation 6:

$$F1_{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

FPS refers to the number of images the algorithm processes per second. The definition of FPS is shown in Equation 7:

$$FPS = \frac{1}{T_{per}} \tag{7}$$

where T_{per} represents the time taken by the algorithm to process a single fundus image.

Intersection over Union (IoU) represents the ratio of the intersection to the union between the predicted results and the ground truth, which can be used to assess the accuracy of segmentation models, as shown in Equation 8:

$$IoU = \frac{TP}{TP + FP + FN} \tag{8}$$

TP represents the number of pixels in the lesion area of the fundus image that are correctly predicted, FP represents the number of pixels

in the background area that are incorrectly predicted as being part of the lesion area in the fundus image, and FN represents the number of pixels in the lesion area of the fundus image that are incorrectly predicted as being part of the background area.

4.3 Dateset

The dataset used in this paper is sourced from the PALM Pathological Myopia Lesion Detection and Segmentation Challenge, provided by the Zhongshan Ophthalmic Center of Sun Yat-sen University. The dataset includes 582 fundus images with annotations for atrophy and detachment lesions, and 213 fundus images without lesions. Each fundus image is annotated with typical lesions related to pathological myopia: patchy retinal atrophy (including peripapillary atrophy) and normal regions without lesions. Pixel-level lesion annotations were initially manually performed by seven ophthalmologists from the Zhongshan Ophthalmic Center, and the final gold standard annotation was created by another senior expert who integrated the results from the seven ophthalmologists. Additionally, the dataset contains 400 unannotated fundus images used as a test set. Some images from the dataset and their corresponding masks are shown in Figure 6.

4.4 Data augmentation

Due to the limited amount of data in the dataset, we divided the 795 images into a training set and a validation set in a 9:1 ratio, resulting in only 716 images in the training set and 79 images in the validation set. To increase the sample size of the training set, we applied various data augmentation techniques to the dataset, including histogram equalization, grayscale transformation, horizontal flipping, linear color transformation, rotation transformation, and vertical flipping. These data augmentation techniques expanded the total capacity of the dataset by 6 times, increasing the training set to 5,012 images and the validation set to 553 images. These data augmentation methods effectively increased the diversity of the training data, thereby improving the model's generalization ability. A sample of the augmented dataset is shown in Figure 7.

5 Results

5.1 Comparison of SMLS-YOLO with the YOLOv8 family

To demonstrate the superiority of SMLS-YOLO, we compared its performance with the YOLOv8 family on the augmented dataset. The results are shown in Table 2. In the YOLO instance segmentation experiments, the metrics include both Box and Mask components, corresponding to object detection and instance segmentation tasks, respectively. The object detection task focuses on locating and classifying target objects in the image, outputting the bounding box for each target object. These metrics reflect the model's performance in object detection tasks. The instance segmentation task requires not only locating and classifying target objects but also predicting pixel-level segmentation masks for each target object. The Mask metrics reflect the model's performance in instance segmentation tasks. From Table 2, we can see that on the augmented dataset, SMLS-YOLO achieved a precision of 89.2%, recall of 86.1%, mAP@0.5 of 89.0%, and F1 score of 88% for Box. For Mask, it achieved a precision of 89.9%, recall of 85.4%, mAP@0.5 of 88.9%, and F1 score of 88%. Compared to the baseline model YOLOv8n-seg, SMLS-YOLO improved the Box mAP@0.5 by 2.3% and the Mask mAP@0.5 by 3.9%. Additionally, SMLS-YOLO achieved a 46.7% reduction in model size, a 31.7% reduction in GFLOPS, and maintained nearly the same FPS. This indicates that SMLS-YOLO not only enhances detection and segmentation accuracy but also excels in computational efficiency and resource consumption. To visually demonstrate the performance of each model on the dataset, we plotted the P-R curves for the Atrophy class in both Box and Mask tasks. Figures 8A,B show the P-R curves for the Box and Mask tasks, respectively.

In the P-R curves shown in Figures 8A,B, SMLS-YOLO demonstrates significant advantages in both Box and Mask tasks. SMLS-YOLO maintains the highest precision across most recall levels, indicating higher accuracy in detecting and segmenting atrophic lesions, thereby reducing the risk of false positives and false negatives. In summary, SMLS-YOLO achieves comprehensive performance improvements in both Box and Mask tasks. Its overall performance surpasses that of the YOLOv8 family, proving the model's comprehensive advantages in detection and segmentation tasks.

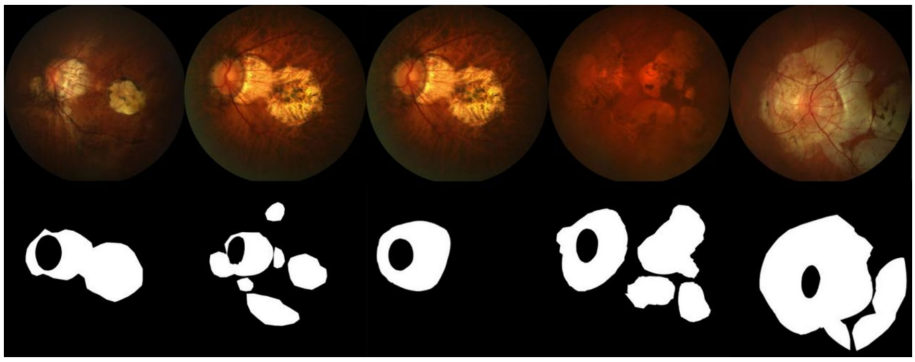


FIGURE 6
Sample images and masks from the dataset.

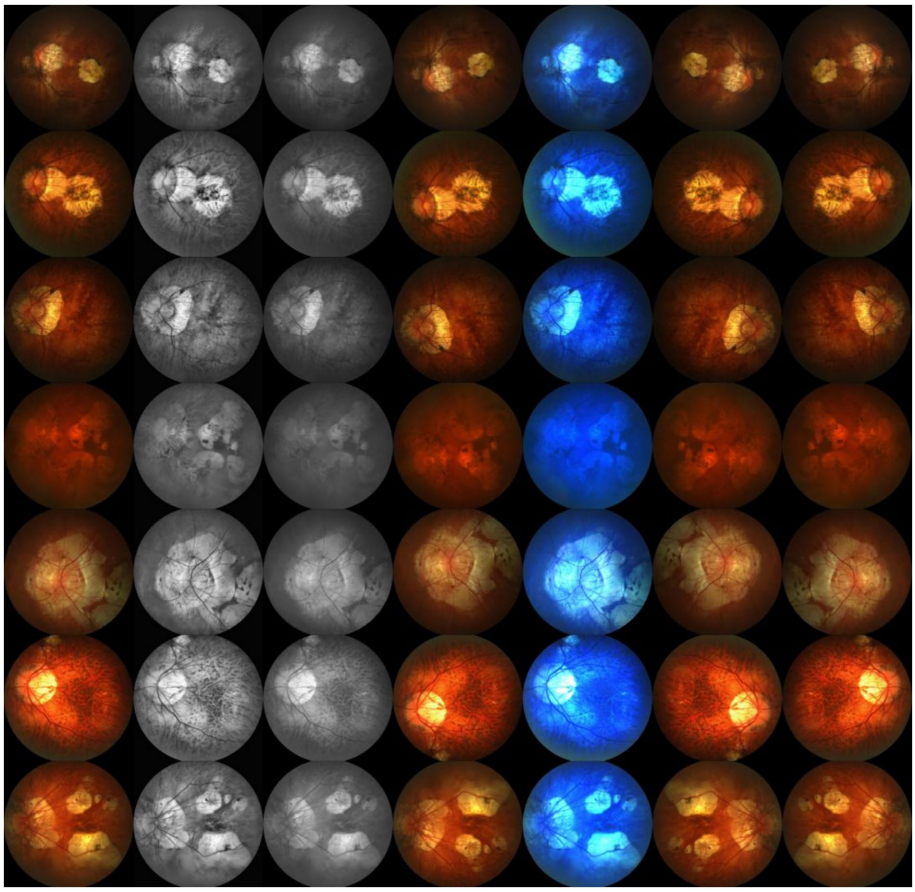


FIGURE 7
Sample display of the augmented dataset.

TABLE 2 Experimental results of SMLS-YOLO compared with YOLOv8 family.

Methods	Box				Mask				All		
	p	*R*	mAP@0.5	F1 score	*p*	*R*	mAP@0.5	F1 score	Params	GFLOPS	FPS
YOLOv8n-seg	89.7	83.2	86.7	86.0	89.4	82.8	85.9	86.0	3.26	12.0	93.3
YOLOv8s-seg	**91.3**	84.2	88.9	88.0	**90.5**	84.1	87.6	87.0	11.78	42.4	72.8
YOLOv8m-seg	88.8	85.5	88.6	87.0	89.0	85.2	88.4	87.0	27.22	110.0	61.5
YOLOv8l-seg	89.2	85.7	87.6	87.0	88.7	85.2	87.6	87.0	45.91	200.1	41.4
YOLOv8x-seg	85.4	86.6	87.7	86.0	86.3	85.2	87.1	86.0	71.72	343.7	26.4
SMLS-YOLO	89.2	86.1	**89.1**	**88.0**	89.9	**85.4**	**88.9**	**88.0**	**1.7**	**8.2**	92.8

Note: Bold values represent the best performance.

5.2 Comparison of SMLS-YOLO with advanced instance segmentation algorithms

To demonstrate that SMLS-YOLO has better generalization, we compare SMLS-YOLO, YOLOv5n-seg, YOLOv7-tiny-seg, YOLOv8n-seg, and YOLOv9's Gelan-c-dseg, Gelan-c-seg, and YOLOv9-c-dseg, respectively, on the enhanced performance comparison on the dataset. The experimental results are shown in Table 3. Compared with other advanced target detection algorithms, SMLS-YOLO performs well on several key metrics.

In the Box task, SMLS-YOLO outperforms other models in several metrics, with mAP@0.5 reaching 89.1%, showing particularly outstanding performance. In contrast, the state-of-the-art Gelan-c-seg achieves an mAP@0.5 of 88.2%, which does not perform as well as SMLS-YOLO. Additionally, although the precision rate of YOLOv7-tiny reaches 91.1%, its recall rate is only 78.3%, leading to its lower overall performance, with an mAP@0.5 of 83.9%.

In the Mask task, SMLS-YOLO again leads in multiple metrics, further confirming its superiority. Additionally, SMLS-YOLO excels in model parameter count and computational efficiency. Its model parameters are only 1.7 M, significantly lower than those of other

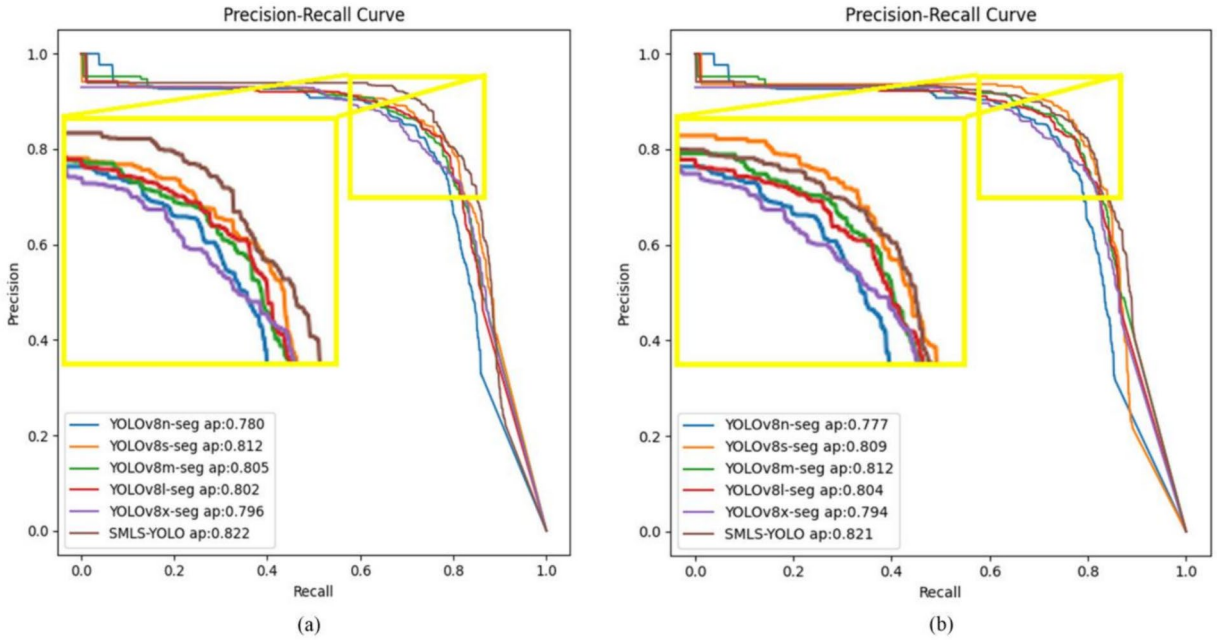


FIGURE 8
(A) Box P-R curve. **(B)** Mask P-R curve.

TABLE 3 Experimental results of SMLS-YOLO compared with other advanced instance segmentation algorithms.

Methods	Box				Mask				All		
	p	R	mAP@0.5	F1 score	*p*	R	mAP@0.5	F1 score	Params	GFLOPS	FPS
YOLOv5n-seg	85.5	81.4	84.8	83.0	87.2	80.6	84.9	84.0	1.88	**6.7**	**111.8**
YOLOv7-tiny	**91.1**	78.3	83.9	83.0	**90.8**	79.8	84.2	84.0	6.99	47.7	101.2
YOLOv8n-seg	89.7	83.2	86.7	86.0	89.4	82.8	86.0	86.0	3.26	12.0	93.3
Gelan-c-seg	90.7	83.3	88.2	87.0	88.7	81.7	86.4	86.0	27.36	144.2	6.61
Gelan-c-dseg	88.6	84.4	87.9	86.0	88.4	83.4	87.0	86.0	27.39	145.2	5.72
YOLOv9-c-dseg	87.8	**84.6**	87.9	86.0	87.4	84.1	87.2	86.0	57.47	368.6	4.10
SMLS-YOLO	89.2	86.1	**89.1**	**88.0**	89.9	**85.4**	**88.9**	**88.0**	**1.7**	8.2	92.8

Note: Bold values represent the best performance.

models. Furthermore, SMLS-YOLO's GFLOPS is 8.2, and its FPS reaches 92.8, demonstrating high computational efficiency and real-time performance.

Precision (P) and Recall (R) are two key metrics used to evaluate model performance. Precision measures the accuracy of the model's predictions, while Recall assesses the model's ability to capture all relevant instances. Typically, there is a trade-off between Precision and Recall: increasing Precision by being stricter with positive class predictions (reducing false positives, FP) can lead to missing some true positives (increasing false negatives, FN), which in turn decreases Recall. Conversely, being more lenient with positive class predictions can increase Recall but May also result in more false positives, thus decreasing Precision. The mean Average Precision at Intersection over Union (IoU) threshold of 0.5 (mAP@0.5) metric balances different combinations of Precision and Recall to maximize the model's overall performance. It calculates the average Precision and Recall across various thresholds, providing a comprehensive performance indicator by averaging these values. Therefore, even when there is a trade-off between Precision and Recall, mAP@0.5 offers a more holistic assessment of model performance. Compared to YOLOv8s-seg, SMLS-YOLO exhibits a slightly lower Precision but improved Recall and mAP@0.5, suggesting an overall enhancement in performance. Specifically, as shown in Table 2, SMLS-YOLO has a lower Precision (P) than YOLOv8s-seg, and in Table 3, SMLS-YOLO has a lower Precision (P) than YOLOv7-tiny. However, when considering the mAP@0.5 metric, which measures overall performance, SMLS-YOLO outperforms both YOLOv8s-seg and YOLOv7-tiny. Additionally, our SMLS-YOLO is more lightweight than YOLOv8s-seg and YOLOv7-tiny, with GFLOPS being only 17% of that of YOLOv8s-seg and YOLOv7-tiny.

In summary, SMLS-YOLO not only excels in the Box task but also performs outstandingly in the Mask task. It achieves the best performance across multiple key metrics, demonstrating comprehensive advantages in both detection and segmentation tasks. Figures 9A,B show the mAP@0.5 curves for Box and Mask during the training process of seven networks. From these figures, it can be seen that SMLS-YOLO's curve rises rapidly in the early stages of training, demonstrating its fast convergence ability. Additionally, its mAP@0.5 performance remains very stable and higher than other models

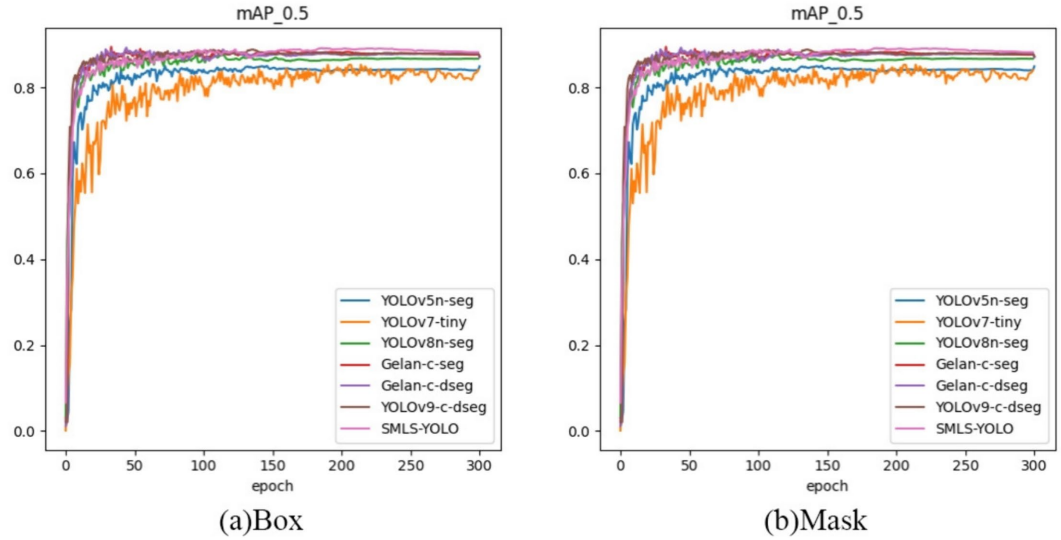


FIGURE 9
(A) Box mAP@0.5 curve. **(B)** Mask mAP@0.5 curve.

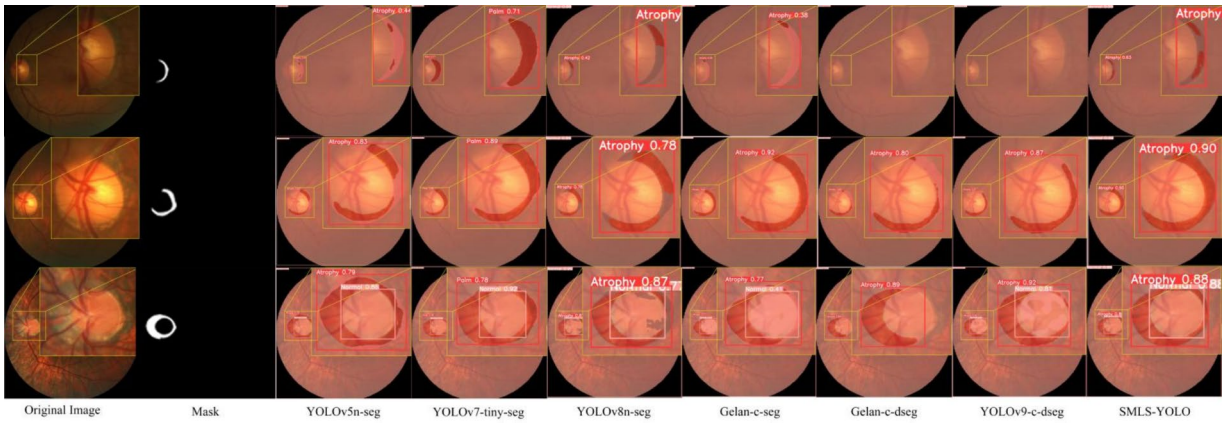


FIGURE 10
Visualization of SMLS-YOLO and other advanced instance segmentation algorithms results.

throughout the training process, reflecting its stability and consistency. This indicates that SMLS-YOLO not only converges quickly in the early stages but also maintains high performance with minimal fluctuations throughout the training process, exhibiting excellent robustness and consistency. Furthermore, we visualized the detection results of the seven algorithms on the dataset to demonstrate SMLS-YOLO's advantages over other advanced algorithms. Figure 10 shows the visualization results of the seven algorithms, where it can be seen that SMLS-YOLO achieves the best detection accuracy and prediction probability.

5.3 Comparison of SMLS-YOLO with classical segmentation networks

In order to verify the advantages and application potential of SMLS-YOLO, this paper compares the performance of SMLS-YOLO with classic segmentation algorithms such as UNet and the DeepLab series on an enhanced fundus color photography dataset. Table 4 presents the specific performance of SMLS-YOLO and these classic algorithms in terms of IoU, precision, recall, and F1-score.

Although UNet and the DeepLab series models are primarily used for semantic segmentation tasks, while SMLS-YOLO focuses on instance segmentation, the experimental results on the same dataset indicate that SMLS-YOLO not only surpasses these traditional semantic segmentation models in key performance indicators such as precision, recall, IoU, and F1-score, but also significantly reduces the number of parameters and increases processing speed. This suggests that, despite the differences in application domains, SMLS-YOLO still demonstrates strong generalization capabilities and superior performance when faced with semantic segmentation tasks.

To validate the advantages of the SMLS-YOLO model across various performance metrics, this paper visualizes the detection results of SMLS-YOLO compared to classical segmentation networks,

as shown in Figure 11. SMLS-YOLO demonstrates higher recognition accuracy and stronger adaptability when processing lesion areas, showing significant advantages.

TABLE 4 Experimental results of SMLS-YOLO compared with classical segmentation networks.

Methods	*p*	R	IoU	F1-score	Params	FPS
Unet	79.7	72.5	61.0	72.7	40.0	16.7
DeepLabV1	84.0	73.9	64.7	76.1	20.5	33.3
DeepLabV2	88.4	77.4	70.0	80.3	44.0	17.5
DeepLabV3	87.1	75.3	67.3	77.4	11.0	22.0
YOLOv8-seg	89.4	82.8	75.1	86.0	3.26	93.3
SMLS-YOLO	89.1	**88.9**	**76.6**	**88.0**	**1.70**	**92.8**

Note: Bold values represent the best performance.

5.4 Analysis of strategy effectiveness

To demonstrate the effectiveness of each improvement in SMLS-YOLO, we conducted an ablation study on the fundus image dataset. The results of the ablation study are shown in Table 5. Table 5 presents the detection performance achieved by the baseline algorithm YOLOv8n-seg with different combinations of components. It can be observed that each improvement strategy enhances the detection accuracy of the baseline algorithm to some extent.

From the aforementioned tables, it can be observed that the StarNet module has demonstrated excellent performance across multiple experiments. It not only effectively reduced the model's parameter count and computational load but also improved detection accuracy to some extent. For instance, in experiments A, B, and C, despite the reduction in parameter count, the values of mAP@0.5(Box) and mAP@0.5(Mask) increased to varying degrees, indicating the module's enhancement effect on model performance.

TABLE 5 Experimental results under different improvement strategies.

Methods	StarNet	C2f-Star	Segment_LS	MHSA	mAP@0.5(Box)	mAP@0.5(Mask)	Params	Gflops	FPS
YOLOv8n-seg	–	–	–		86.7	86.0	3.26	12.0	93.3
A	√				87.6	86.5	2.47	10.4	96.7
B	√	√			87.7	86.4	2.27	10.0	97.7
C	√	√	√		87.8	86.3	1.50	8.1	95.3
D	√	√	–	√	89.1	87.7	2.46	10.1	94.1
E	√	–	–	√	88.6	87.1	2.66	10.5	93.8
F	√	–	√	–	88.4	86.9	1.70	8.4	96.5
G	√	–	√	√	88.9	88.5	1.90	8.6	94.4
SMLS-YOLO	√	√	√	√	89.1	88.9	1.70	8.2	92.8

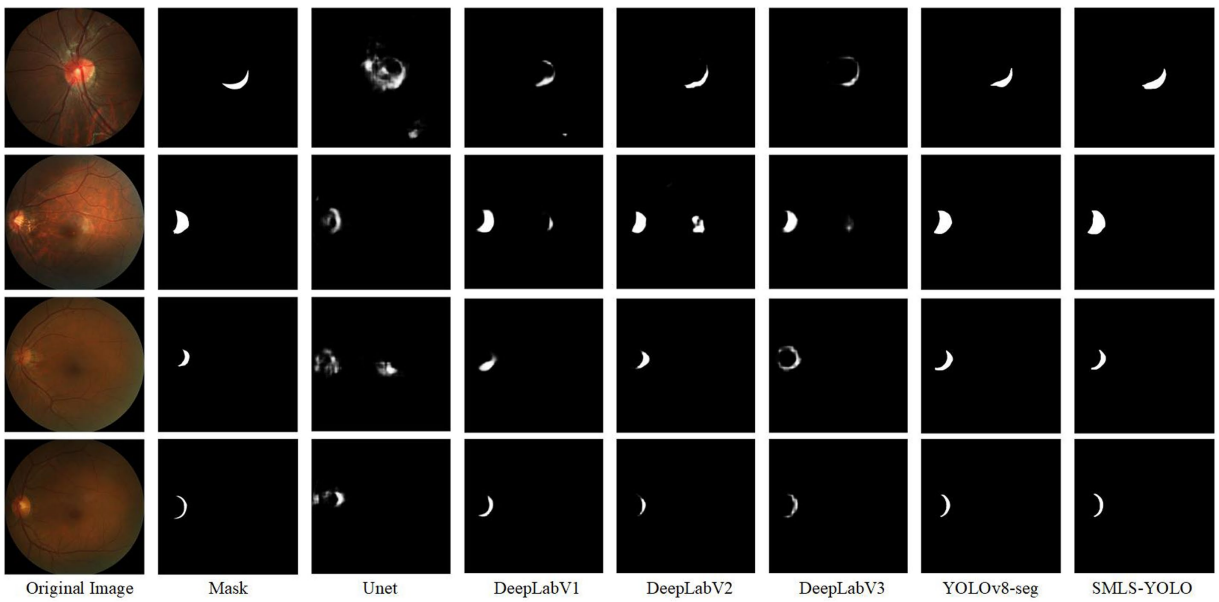


FIGURE 11
Visualization results of SMLS-YOLO compared with classical segmentation networks.

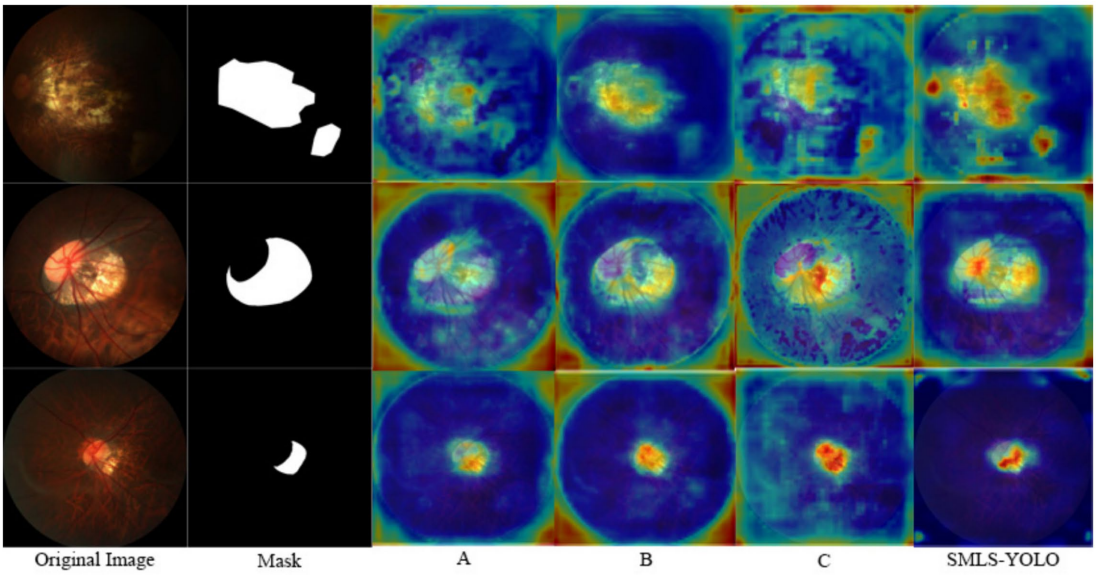


FIGURE 12
Heatmap results under different improvement strategies.

By integrating StarNet as the Backbone of SMLS-YOLO, the baseline algorithm's mAP@0.5 (Box) and mAP@0.5 (Mask) were, respectively, improved to 87.6 and 86.5%, while the model parameter count was reduced to 2.47 M. After incorporating the C2f-Star component, the mAP@0.5 (Box) further increased to 87.7%, the mAP@0.5 (Mask) slightly decreased to 86.4%, and the parameter count was reduced to 2.27 M. The introduction of the Segment_LS segmentation head further optimized the model's balance, allowing the model to maintain low computational load while still improving detection accuracy. Additionally, the incorporation of the MHSA attention mechanism, although leading to a slight increase in parameter count, significantly enhanced model performance. In experiments E, F, and G, it proved the value of the MHSA module in improving model performance. Ultimately, after integrating all the improvement strategies, SMLS-YOLO's mAP@0.5 (Box) and mAP@0.5 (Mask) were, respectively, increased to 89.1 and 88.9%, which is 2.4 and 3.9% higher than the baseline algorithm, with the parameter count being only 52% of YOLOv8n-seg.

To further demonstrate the effectiveness of each improvement strategy, we conducted a heatmap visualization analysis of the model under various combinations of improvement strategies. Figure 12 shows the heatmap results under different combinations of improvement strategies. Through these visualizations, the performance enhancement effects of different improvement strategies on the model can be observed intuitively, thereby more clearly verifying the effectiveness of each improvement strategy.

6 Summary

In this paper, we proposed a novel instance segmentation algorithm named SMLS-YOLO, designed to tackle the challenges in detecting pathological myopia. Firstly, we introduced StarNet as the backbone network to efficiently extract feature information from images. Following this, we proposed a new feature extraction module, C2f-Star, which aims to more effectively integrate multi-level feature information produced by the backbone network, thereby enhancing performance while reducing the model's complexity. Subsequently, to mitigate the issue of the original segmentation head's large number of parameters, we proposed a new lightweight segmentation head, Segment_LS. This head leverages shared convolution and introduces scale adjustment operations, significantly reducing the computational burden during segmentation. Our Segment_LS segmentation head abandons the shared prototype masks of YOLOv8, thereby overcoming the segmentation head's inherent limitations. As a result, our segmentation head does not require a large number of parameters to improve accuracy, thus significantly reducing the overall network parameters. Additionally, we integrated the Multi-Head Self-Attention (MHSA) mechanism to bolster the model's capability to capture essential information in images, thereby improving the overall performance of SMLS-YOLO. Experiments conducted on fundus images dataset with pathological myopia demonstrate that SMLS-YOLO achieves advanced performance. Looking ahead, we intend to explore model pruning and knowledge distillation techniques to further refine the model's efficiency and develop even more lightweight algorithms.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

HX: Conceptualization, Data curation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. BY: Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. CH: Software, Validation, Visualization, Writing – review & editing. YG: Validation, Visualization, Writing – review & editing. FW: Validation, Visualization, Writing – review & editing. YW: Validation, Visualization, Writing – review & editing. CW: Validation, Visualization, Writing – review & editing. PC: Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The authors acknowledge partial support of this work by the Natural Science Foundation of Shaanxi Province [2021JM-537], in part by the Key Program of National Social Science Foundation of China (NSSFC, 23AGL039), in part by the Shaanxi Provincial Science and Technology Plan Project (2024GX-YBXM-114), and in part by the Natural Science Foundation of Shaanxi Province (Grant No. 2023-YBGY-036).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Almubarak, H., Bazi, Y., and Alajlan, N. (2020). Two-stage mask-RCNN approach for detecting and segmenting the optic nerve head, optic disc, and optic cup in fundus images. *Appl. Sci.* 10:3833. doi: 10.3390/app10113833

Aquino, A., Gegúndez-Arias, M. E., and Marín, D. (2010). Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques. *IEEE Trans. Med. Imaging* 29, 1860–1869. doi: 10.1109/TMI.2010.2053042

Baird, P. N., Saw, S.-M., Lanca, C., Guggenheim, J. A., Smith Iii, E. L., Zhou, X., et al. (2020). Myopia. *Nat. Rev. Dis. Primers* 6:99. doi: 10.1038/s41572-020-00231-4

Bolya, D., Zhou, C., Xiao, F., and Lee, Y. J. (2019). YOLACT: real-time instance segmentation, In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 9156–9165.

Chakravarty, A., and Sivaswamy, J. (2017). Joint optic disc and cup boundary extraction from monocular fundus images. *Comput. Methods Prog. Biomed.* 147, 51–61. doi: 10.1016/j.cmpb.2017.06.004

Chen, H.-J., Huang, Y.-L., Tse, S.-L., Hsia, W.-P., Hsiao, C.-H., Wang, Y., et al. (2022). Application of artificial intelligence and deep learning for choroid segmentation in myopia. *Transl. Vis. Sci. Technol.* 11:38. doi: 10.1167/tvst.11.2.38

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv:1802.02611v3. doi: 10.48550/arXiv.1802.02611

Dolgin, E. S. (2015). The myopia boom. *Nature* 519, 276–278. doi: 10.1038/519276a

GeethaRamani, R., and Dhanapackiam, C. (2014). "Automatic localization and segmentation of optic disc in retinal fundus images through image processing techniques" in 2014 international conference on recent trends in information technology, 1–5.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al. (2023). A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 87–110. doi: 10.1109/TPAMI.2022.3152247

Holden, B. A., Fricke, T. R., Wilson, D. A., Jong, M., Naidoo, K. S., Sankaridurg, P., et al. (2016). Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050. *Ophthalmology* 123, 1036–1042. doi: 10.1016/j.ophtha.2016.01.006

Lee, Y., and Park, J. (2020). "CenterMask: real-time anchor-free instance segmentation" in 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), 13903–13912.

Li, Y. (2023). Research and Implementation of Pathological Myopia Detection and Lesion Segmentation Technology Based on Deep Learning [D]. Guilin University of Technology.

Lu, L., Ren, P., Tang, X., Yang, M., Yuan, M., Yu, W., et al. (2021). AI-model for identifying pathologic myopia based on deep learning algorithms of myopic maculopathy classification and "plus" lesion detection in fundus images. *Front. Cell Dev. Biol.* 9:719262. doi: 10.3389/fcell.2021.719262

Ma, X., Dai, X., Bai, Y., Wang, Y., and Fu, Y. (2024). Rewrite the stars. *ArXiv abs/2403.19967*. doi: 10.48550/arXiv.2403.19967

Marín, D., Gegúndez-Arias, M. E., Suero, A., and Bravo, J. M. (2015). Obtaining optic disc center and pixel region by automatic thresholding methods on morphologically processed fundus images. *Comput. Methods Prog. Biomed.* 118, 173–185. doi: 10.1016/j.cmpb.2014.11.003

Modjtahedi, B. S., Ferris, F. L., Hunter, D. G., and Fong, D. S. (2018). Public health burden and potential interventions for myopia. *Ophthalmology* 125, 628–630. doi: 10.1016/j.ophtha.2018.01.033

Myopia Prevention and Control Guidelines (2024). Myopia prevention and control guidelines (2024 edition). *New Dev. Ophthalmol.* 44, 589–591. doi: 10.13389/j.cnki.rao.2024.0113

Ohno-Matsui, K., Kawasaki, R., Jonas, J. B., Cheung, C. M. G., Saw, S.-M., Verhoeven, V. J. M., et al. (2015). International photographic classification and grading system for myopic maculopathy. *Am. J. Ophthalmol.* 159, 877–883.e7. doi: 10.1016/j.ajo.2015.01.022

Ohno-Matsui, K., Wu, P.-C., Yamashiro, K., Vutipongsatorn, K., Fang, Y., Cheung, C. M. G., et al. (2021). IMI pathologic myopia. *Invest. Ophthalmol. Vis. Sci.* 62:5. doi: 10.1167/iovs.62.5.5

Qin, H., Zhang, W., Zhao, X., and Dong, Z. (2023). "Automatic screening of pathological myopia using deep learning" in 2023 29th international conference on mechatronics and machine vision in practice (M2VIP), 1–5.

Rauf, N., Gilani, S. O., and Waris, A. (2021). Automatic detection of pathological myopia using machine learning. *Sci. Rep.* 11:16570. doi: 10.1038/s41598-021-95205-1

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. *ArXiv abs/1505.04597*. doi: 10.48550/arXiv.1505.04597

Viedma, I. A., Alonso-Caneiro, D., Read, S. A., and Collins, M. J. (2022). OCT retinal and choroidal layer instance segmentation using mask R-CNN. *Sensors (Basel)* 22:2016. doi: 10.3390/s22052016

Wang, X., Zhang, R., Shen, C., Kong, T., and Li, L. (2022). SOLO: a simple framework for instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1–8601. doi: 10.1109/TPAMI.2021.3111116

Xu, L., Wang, Y., Li, Y., Wang, Y., Cui, T., Li, J., et al. (2006). Causes of blindness and visual impairment in urban and rural areas in Beijing: the Beijing eye study. *Ophthalmology* 113, 1134.e1–1134.e11. doi: 10.1016/j.ophtha.2006.01.035

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 11 October 2024
DOI 10.3389/fnins.2024.1482735

OPEN ACCESS

EDITED BY
Hancheng Zhu,
China University of Mining and Technology, China

REVIEWED BY
Hang Zhong,
Hunan University, China
Qigao Fan,
Jiangnan University, China

*CORRESPONDENCE
Ming Tang
✉ foyangkang@ncu.edu.cn

RECEIVED 18 August 2024
ACCEPTED 11 September 2024
PUBLISHED 11 October 2024

CITATION
Li X, Yang T, Tang M and Xiong P (2024) A novel parameter dense three-dimensional convolution residual network method and its application in classroom teaching. *Front. Neurosci.* 18:1482735. doi: 10.3389/fnins.2024.1482735

A novel parameter dense three-dimensional convolution residual network method and its application in classroom teaching

Xuan Li[1], Ting Yang[2], Ming Tang[2]* and Pengwen Xiong[2]

[1]School of Foreign Language, Shangrao Normal University, Shangrao, China, [2]School of Advanced Manufacturing, Nanchang University, Nanchang, China

Introduction: Improving the rationality and accuracy of classroom quality analysis is crucial in modern education. Traditional methods, such as questionnaires and manual recordings, are resource-intensive and subjective, leading to inconsistent results. As a solution, computer vision (CV) technologies have emerged as powerful tools for real-time classroom monitoring. This study proposes a novel Dense 3D Convolutional Residual Network (D3DCNN_ResNet) to recognize students' expressions and behaviors in English classrooms.

Methods: The proposed method combines Single Shot Multibox Detector (SSD) for target detection with an improved D3DCNN_ResNet model. The network applies 3D convolution in both spatial and temporal domains, with shortcut connections from residual blocks to increase network depth. Dense connections are introduced to enhance the flow of high- and low-level features. The model was tested on two datasets: the CK+ dataset for expression recognition and the KTH dataset for behavior recognition.

Results and Discussion: The experiments show that the proposed method is highly efficient in optimizing model training and improving recognition accuracy. On the CK+ dataset, the model achieved an expression recognition accuracy of 97.94%, while on the KTH dataset, the behavior recognition accuracy reached 98.86%. The combination of residual blocks and dense connections reduced feature redundancy and improved gradient flow, leading to better model performance. The results demonstrate that the D3DCNN_ResNet is well-suited for classroom quality analysis and has the potential to enhance teaching strategies by providing real-time feedback on student engagement.

KEYWORDS

three-dimensional convolutional neural network, residual network, video sequence, SSD algorithm, behavior recognition

1 Introduction

Improving the rationality and accuracy of classroom quality analysis is particularly important for classroom teaching (Xun, 2022). In recent years, many forms of classroom quality analysis have also been studied, mainly including questionnaire form, physiological embodiment method, computer vision (CV) and other methods. At present, the vast majority of classrooms analyze and evaluate the quality of classroom teaching through specialist manual records, after-school questionnaires and other methods. These methods not only consume a lot of human and material resources, but also have a certain subjectivity, so it is unrealistic to get reasonable and accurate results. CV methods could be combined with intelligent equipment

with video recording to monitor the students' class status in real time, study and analyze the students' facial expression and posture in the classroom, then analyze the classroom quality (Wu et al., 2023). Since the 21st century, it has promoted the development of integrating computer emotion analysis and education and teaching application.

Through the analysis of intelligent video equipment and expression recognition, it can provide teachers with real-time and reliable feedback information in time, facilitate teachers to modify teaching contents, control teaching progress, select teaching methods and adjust teaching difficulties, and greatly promote the change of teaching mode and the improvement of teaching quality. In the field of computer vision, object recognition from video and pictures has always been a hot topic. How to extract robust and representative features is a challenging task (Hong, 2022). For static image recognition, only static features in one image need to be extracted for learning (Niu et al., 2022). For dynamic video recognition, we need to consider not only the relationship between adjacent pixels in a single frame image in the spatial domain, but also the interaction between multiple adjacent frames in the temporal domain (Wanshu and Bin, 2022). At present, the research around video recognition mainly focuses on feature extraction and classifier selection.

For video object classification, the quality of feature extraction is very important, which directly affects the classification effect. The judgment of its quality lies in whether it has high recognition degree, strong robustness, more complete recognition information and so on. The existing feature extraction methods are mainly divided into two categories: local feature methods and global feature methods. The local feature method is to extract the local sub regions or interest points in the video or image. For example, Bengio et al. (1994) proposed gradient based learning algorithms as the duration of the dependencies to be captures increased. Laptev and Lindeberg (2005) first detected multiple spatiotemporal interest points from the video, then built a spatiotemporal cube centered on the interest points and extracted hog and Hof features to represent the motion information. Gers et al. (2002) found that LSTM can learn the subtle differences between spike sequences with intervals of 49 or 50 times steps without the help of any short training samples by connecting the "peephole connections" from its internal cells to its multiplication gate. Ma et al. (2015) proposed a new Bayesian matrix decomposition method for bounded support data. The beta distribution has two parameters, while the two parameter matrices with only non-negative values can be obtained. To provide low rank matrix factorization, non-negative matrix factorization (NMF) technique is applied. Gu et al. (2012) performed Gabor transform on the image, then jointly encoded it by radial network, and realized global classification by cascading multiple classifiers. Although the above methods can well represent the image edge information, its feature dimension is too high and the amount of calculation is too large.

According to the analysis, each feature extraction method has its advantages and disadvantages. In order to make up for each other's shortcomings and overcome the problems of insufficient description of image information by a single feature and weak robustness, most of the current research methods use mixed features. For example, Liu et al. (2011) proposed a novel method to extract expression features by combining Gabor multi-directional feature fusion and block histogram statistics based on the weak ability of Gabor features to represent global features. Ren et al. (2017) introduced a regional recommendation network (RPN) that shared full image convolutional features with the detection network to achieve almost cost-free regional recommendations. Graves et al. (2005) used convolution neural network and long-term and short-term memory neural network to extract features, and then weighted and fused the extracted features to improve the accuracy and generalization of the model. Ji et al. (2013) developed a 3D CNN model which extracted features from both spatial and temporal dimension for action recognition.

Through the study of the above methods, in order to not only overcome the problem that a single feature does not adequately describe the image information but also simplify the model structure and improve recognition accuracy, this paper proposes a new parallel Dense 3D convolution residual network (D3DCNN_ResNet) method for students' expression and behavior recognition in English classrooms. This method first uses the single shot multibox detector (SSD) to extract the moving area to be recognized for clipping and preprocessing, and then inputs the crop and the original frame picture into the improved convolution residual network to extract features. The crop can extract local detailed feature information. On the other hand, the original frame can obtain the global overall information, supplement the edge contour features that the former failed to extract completely, and then carry out weighted fusion to improve the robustness and generalization of the model. Due to the large number of parameters and low computational efficiency of the 3D convolutional neural network, a Pseudo 3D convolutional neural network (P3D) (Qiu et al., 2017) is proposed to decompose 3D into (2D + 1D) forms. Pseudo 3D convolution (P3D) decomposes 3D convolution into 2D spatial and 1D temporal convolutions, reducing computational complexity while maintaining performance. It has been applied in video action recognition and medical imaging, balancing accuracy and efficiency. According to the four structures designed in Gu et al. (2012), 2D spatial convolution and 1D temporal convolution are proved to be the most effective.

Main contributions:

1) With our proposed D3DCNN_ResNet, a new English classroom quality analysis with low and controllable cost, high precision and reliability combined with facial expression recognition technology is proposed.
2) In the proposed D3DCNN_ResNet, the residual module is added to learn the residual mapping to further solve the problems of gradient descent and over fitting with the deepening of the network depth, and more subtle multi-level features are obtained through the interconnection between different convolution layers and different residual blocks in the residual block.
3) The combination of residual blocks and dense connections not only reduces feature redundancy and improves the gradient correlation of the network, but also reduces the network parameters to a certain extent. Our architecture combines DenseNet and ResNet principles, improving gradient flow and reducing the risk of vanishing gradients in deeper networks. DenseNet's feature reuse and ResNet's residual connections have been validated in tasks like object detection and image classification, enhancing feature learning (see Figure 1).

The rest of this letter is organized as follows. In section 2, we present the basic principle and framework of our proposed D3DCNN_ResNet in detail. Section 3 gives the model optimization. Section 4 presents the results and analysis based on the experiment. Finally, the conclusions are given in section 5.

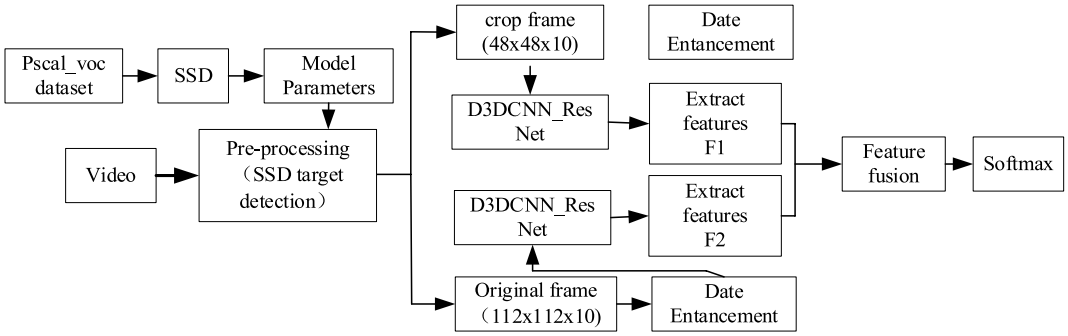


FIGURE 1
Model structure diagram.

2 New parallel D3DCNN_ResNet identification method

2.1 Introduction to identification method

This method mainly includes two parts, SSD target detection and improved D3DCNN_ResNet. It can accurately locate the recognition area of the object by using SSD target detection, and get detailed local features through crop, and then use the original frame as a feature supplement to obtain more abundant, complete and robust feature information. SSD target detection is a preprocessing module, and could be used to cut the recognition area corresponding to the entire video sequence. The obtained sequence frames are directly input into the improved D3DCNN_ResNet to extracts features. Since deep learning can combine low-level features into high-level ones through the construction of multiple hidden layers and autonomously learned features, the improved D3DCNN_ResNet model connects different residual blocks and convolution layers within the residual blocks. It fully combines the bottom features with the high-level features, and enhances the flow of feature information in the network. Then, the obtained local and global multi features are fused to better represent the subtle feature information. This structure not only solves the lack of time-domain information extraction in traditional deep learning, but also solves the problems of large parameters and over fitting through the decomposition of the three-dimensional convolution, which improves the recognition rate of the model. The specific implementation steps of the algorithm are as follows:

1) Firstly, SSD target detection algorithm is applied to the input video frame, and the object recognition region is extracted from each frame of the video for clipping preprocessing, which is called the crop.
2) Then, the original frame and the crop are input to the improved D3DCNN_ResNet respectively. The extracted features in RESNET are marked as F1 and F2: F1 is the local detailed feature of the recognition object, and F2 is the global overall feature, which mainly supplements the edge contour information. F1 and F2 also include the fusion between low-level features and high-level features, with feature dimensions of 256.
3) Finally, the original frame features and the crop frame features are fused, encoding the video information. Since the feature level fusion at the full connection layer will greatly increase the parameters of the model, the decision level fusion method is selected in this paper, as shown in Equation 1:

$$R(x) = \sum_{n=1}^{2} W_n \times P_n(x) \tag{1}$$

where, $P_n(x)$, $(0<p<1)$ is the output probability value of F1 and F2 at softmax layer, W is the weight parameter, which is obtained from the least square estimation of the minimization loss function:

$$\{w_n\}_{n=1}^{2} = \arg\min_{w_1,w_2}\left(o - \sum_{n=1}^{2} w_n \times P_n(x)_F^2 \right) \tag{2}$$

The final fusion feature is input into softmax classifier to realize classification and recognition.

2.2 Introduction to improved D3DCNN_ResNet

In video analysis, considering the motion information between consecutive frames, D3DCNN_ResNet is used to stack multiple consecutive frame images to form a cube, and then use 3D convolution kernel to convolute in the cube. As shown in Figure 2, the characteristic value of a certain position of a convolution map is obtained by convoluting the local receptive field of the same position of three consecutive frames on the upper layer. Its advantage is that it can extract the spatial-temporal features at one time and capture the action information of multiple frames in the video sequence.

In the convolution process, on each characteristic graph of any single layer, the value of position (a, b, c) is given by Equation 3:

$$v^{abc} = \tanh\left(t \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \sum_{d=0}^{D-1} xy^{(a+h)(b+w)(c+d)} + z \right) \tag{3}$$

where tanh () is a hyperbolic tangent function, index T and value x are the connection parameters of the current characteristic graph. h,

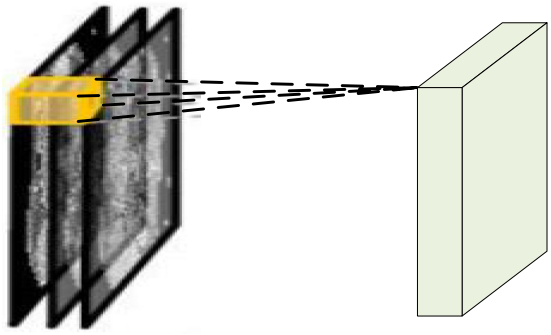


FIGURE 2
3DCNN schematic diagram.

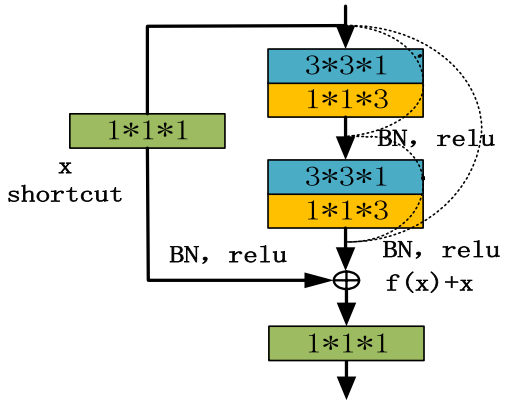


FIGURE 3
Schematic diagram of residuals module.

W, and D are the height, width and time dimensions of the three-dimensional convolution kernel, and z is the deviation of the characteristic graph.

The deep 3DCNN is more effective in feature extraction, but 3DCNN has too many parameters. After increasing the depth, it will further affect its learning efficiency and accuracy, and cause the gradient to disappear. Therefore, 3D is decomposed into (2D + 1D) structure, which means ($v1$*$v2$*v3) the filter is replaced by ($v1$*$v2$*1) and (1*1*$v3$) serially, and samples in the spatial and temporal regions respectively, effectively reducing the amount of parameters. At the same time, residual blocks are added to 3DCNN to simplify the training of deep network. Formally, the required bottom mapping is expressed as $h(x)$, that is, the optimal solution mapping after the input sample x. Let the superimposed nonlinear layer fit the residual mapping of $F(x) = h(x) - x$, and convert the original mapping to Equation 4.

$$y = f(x,w) + x \tag{4}$$

where x is the output value of the upper neural unit, w is the weight of the neuron, and y is the output value of the activation function in the neuron. In addition, the input of each convolution layer in the residual block is composed of the output of all previous convolution layers, and the output of all previous residual blocks is used as the input of the next residual block to improve the gradient correlation of the network, and ResNET can maximize the information flow while reducing network redundancy. The basic structure of the improved D3DCNN_ResNet is shown in Figure 3. In order to convert the two-dimensional residual unit into a three-dimensional structure, it is used for encoding spatiotemporal video information. According to the three-dimensional convolution theorem, the basic residual element is modified as follows:

As shown in Figure 3, the residual block has two parts: quick connection (solid line) and dense connection (dotted line). There is a three-dimensional convolution layer (conv3d) in the shortcut, which is mainly used to change the input dimension and increase its nonlinear representation so as to match the output dimension of the main path in the subsequent addition steps, so that the feedforward/back propagation algorithm can be carried out smoothly. In addition, the three-dimensional convolution unit is replaced by (3*3*1) and (1*1*3) serial. Secondly, dense connection not only makes each three-dimensional convolution unit take the output of the previous three-dimensional convolution unit as the input, but also makes dense connection between each residual block to splice the features. As shown in Figure 4, this structure can deepen the network depth and improve the model representation ability. At the same time, due to the repeated use of convolution features, it can appropriately reduce the number of convolution cores to achieve a certain anti over fitting effect. In addition, 1*1*1 convolution kernel is used at the end of convolution kernel, which aims at feature aggregation and channel adjustment to reduce the amount of calculation. The comparison results are as follows (taking the spatial size as an example): after the first residual convolution, the characteristic size is 24*24*30 if the number of data channels is not reduced, continuing the convolution (the number of convolution cores is set to 8), and the amount of computation is:

$$\left(24^{*}24^{*}76\right)^{*}\left(3^{*}3\right)*38+\left(24^{*}24^{*}38\right)^{*}\left(3^{*}3\right)*30\approx 20.88\ \text{million}$$

If 1*1*1 convolution is used to compress the channel information so that the feature size is 24*24*16, the amount of computation is:

$$\left(24^{*}24^{*}48\right)^{*}\left(3^{*}3\right)*38+\left(24^{*}24^{*}24\right)^{*}\left(3^{*}3\right)$$
$$*16+\left(24^{*}24^{*}16\right)^{*}\left(1^{*}1\right)*30\approx 8.24\ \text{million}$$

Through the above comparison, the calculation amount is reduced by about 1.5 times after adding 1*1*1 convolution.

The overall convolution structure of D3DCNN_ResNet used in this paper is shown in Figure 5. The input of the network consists of 10 consecutive frames of images. Take the input crop frame as an example, the spatial size clipping processing of each crop frame image is 48*48, that is, the size of the input video sequence is (10*48*48), and in the input conv3d_1. Before filling, use zeropadding to add dimensions to prevent the loss of image edge information. After filling, the size becomes 50*50. Since the image dimension is high, input conv3d_1 after convolution dimensionality reduction, 16 feature maps with the size of 50*50*10 are obtained, and then downsampling is performed on each feature map in the way of maximum sampling

with the size of 1*2*2, so that the number of feature maps is the same and the spatial resolution is reduced. The next layer is to insert the residual blocks of quick connection and dense connection, and get 32 feature maps with the size of 24*24*10. Then, the depth of the joint feature data size is reduced to 16 by the 1*1*1 convolution kernel. In this paper, four residual blocks are used to carry out residual convolution in turn. Finally, 128 characteristic maps with the size of 24*24*10 are obtained. Then, 64 characteristic maps with the size of 12*12*10 are obtained through mean sampling. The data is compressed into one dimension in the flatten layer, and a 256-dimension output feature is obtained through two dense layers. After each convolution layer, the activation function and the batch normalization (BN) layer are connected. Both activation functions use the ReLU function. The input original frame size is 112*112, and the process is the same as that shown above.

2.3 SSD target detection

Target detection is a kind of technology that the computer analyzes and distinguishes by extracting the typical features of the target. Its main task is to find out the interested objects in the image and determine their position and size. In this paper, SSD target detector is used to detect human face and human body. Compared with YOLO algorithm, it has better robustness to objects of different scales. Compared with R_CNN series of algorithms, it omits the process of generating candidate boxes, and the calculation speed is faster.

SSD target detection is an end-to-end image target detection method. It directly extracts features from input data to predict object classification and location, which greatly improves the detection speed. The basic network structure of SSD algorithm is VGG16, and the 5-layer network in front of VGG16 is adopted. First, the last two full connection layers are changed into convolution layers by using the atrus algorithm, then three convolution layers and a pooled layer are added for convolution and down sampling processing, and finally two different 3*3 convolution cores are used for convolution, and the detection results are obtained by non-maximum suppression method. For each feature layer, the scale size of the default box is calculated according to the following formula:

$$S_k = S_{\min} + \frac{S_{\max} - S_{\min}}{m-1}(k-1), k \in [1, m] \tag{5}$$

where, $S_{\min}$ the value of 0.2 indicates that the scale size of the bottom layer is 0.2; the value is 0.9, indicating that the scale size of the highest layer is 0.9; from the formula, the first layer $S_{\min} = s1$, $S_{\max} = s2$; the second layer $S_{\min} = s2$, $S_{\max} = s3$, and so on. M represents the number of characteristic layers. The aspect ratio is expressed in AR, and the value is AR = {1, 2, 3, 1/2, 1/3}, then the width and height of each default box are calculated as follows:

$$w_k^a = S_k\sqrt{a_r}, h_k^a = \frac{S_k}{\sqrt{a_r}} \tag{6}$$

The biggest contribution of SSD algorithm is to propose a multi-scale feature layer prediction method. The calculated default box can basically cover the objects of various shapes and sizes in the input image.

Pascal_voc datasets provide a set of standardized and excellent datasets for image recognition and classification. This article first uses Pascal_voc 2007 and 2012 datasets to get the pre-training model, and then the model file configuration is modified according to their own detection targets to train their own datasets. After the SSD detection, the position of human face and human body in the picture can be determined. Since VOC provides data annotation information, and it does not need to mark the local position manually, so it is faster and more convenient to detect. Then a bounding box is generated for clipping, as shown in Figures 6, 7.

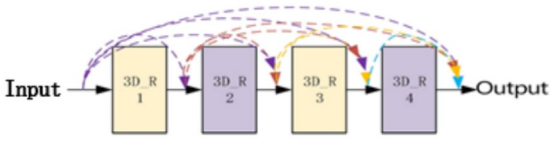


FIGURE 4
Dense connection diagram of 3DCNN RESNET.

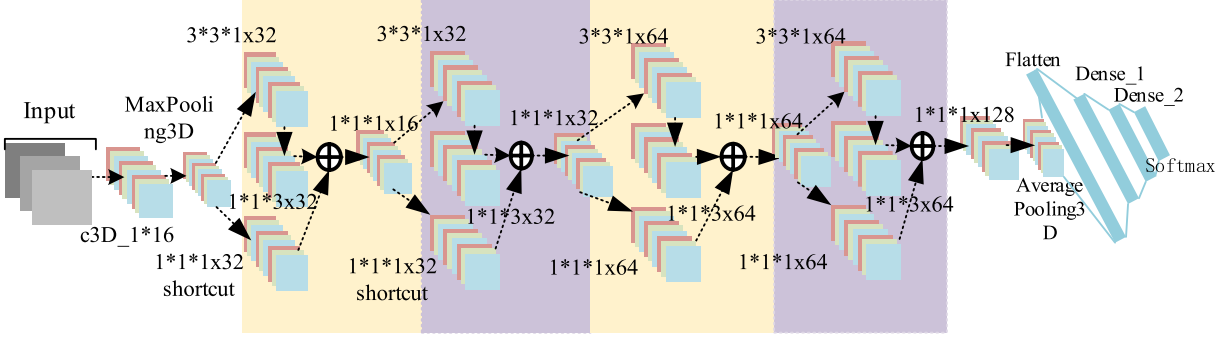


FIGURE 5
D3 DCNN RESNET structure.

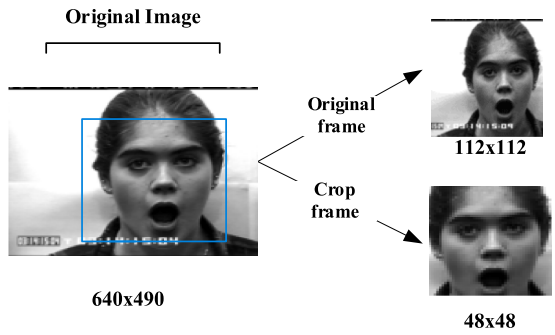


FIGURE 6
An expression case for clipping. CK+ expression data set is from Lucey et al. (2010).

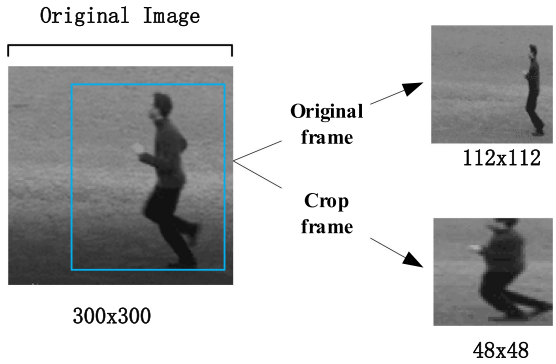


FIGURE 7
A behavior case for clipping. KTH behavior data set is from Laptev and Lindeberg (2005).

3 Model optimization

Model optimization is very important in the training process of neural networks. Making correct decisions in the process of configuring training, verification and testing data sets will help us to create an efficient neural network to a great extent. Selecting appropriate super parameters to train an optimal decision model is the most important link of model optimization.

3.1 Activation function and local normalization

The most important function of activation function is to make the neural network converge and make the operation in the neural network nonlinear so as to construct various valuable functions. Commonly used activation functions include tanh, ReLU, Leaky ReLU, etc. As shown in Figure 8, ReLU is recognized as the best activation function compared with other activation functions, which has the following advantages: (1) simple formula and fast calculation speed; (2) when the input is greater than 0, the gradient is constant to avoid gradient saturation; (3) fast convergence. Therefore, the ReLU activation function is adopted in this paper. The formula is as follows:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{7}$$

In order to solve the data distribution change caused by the input data of each layer after updating the parameters, this paper adds batch normalization (BN) (Ioffe and Szegedy, 2015) to the activation function of each layer, and the formula is as follows:

$$y = \frac{\gamma}{\sqrt{\mathrm{Var}[x] + \varepsilon}} \cdot x + \left(\beta - \frac{\gamma E[x]}{\sqrt{\mathrm{Var}[x] + \varepsilon}} \right) \tag{8}$$

where, γ is β a learnable reconstruction parameter, so that the network can learn and recover the characteristic distribution to be learned by the original network. $E[x]$ is the average of the mean values of all samples, and $\mathrm{Var}[x]$ is an unbiased estimation using the standard deviation of each sample. The calculation process is to calculate the average value and variance of all neurons of a characteristic map corresponding to all samples, and then normalize the neural units of the characteristic map.

3.2 Loss function and regularization

The optimization degree of the network model depends on the size of the loss function, which mainly represents the difference between the predicted value and the real value of the model for a specific sample. The loss function used in this paper is the categorical cross entropy loss:

$$C = -\frac{1}{n} \sum_{x} \left[y \ln a + (1 - y) \ln (1 - a) \right] \tag{9}$$

where y is the expected output, a is the actual output of neurons, n is the number of input samples in the same batch (batch size), and cross entropy C is the distance between the actual output probability and the expected output probability, that is, the smaller C is, the closer the two probability distributions are.

In order to prevent the network from over fitting, this paper adds a random dropout regularization term to the loss function of the output layer. When the neural network propagates forward, the activation value of a neuron stops working with a certain probability P. P is set to 0.5, which makes the model more generalized and less dependent on some local features. At the same time, it reduces the number of training nodes and improves the learning speed of the algorithm.

3.3 Optimization algorithm

The optimization algorithm is to optimize the parameters in the network and calculate the gradient of parameters at each layer by back propagation through the error obtained from the loss function. In this paper, rmsprop algorithm is used to update the parameters of the neural network. The iterative update formula is as follows:

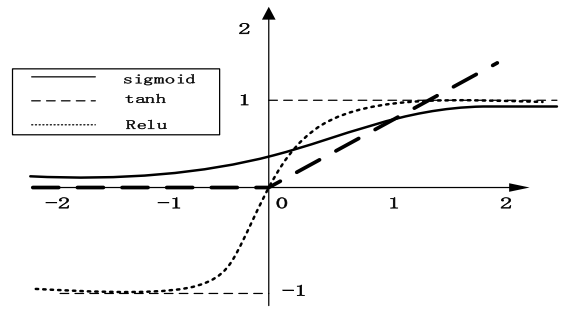


FIGURE 8
Comparison of three activation functions.

$$
\begin{aligned}
s_{dw} &= \beta s_{dw} + (1-\beta) dW^2 \\
s_{db} &= \beta s_{db} + (1-\beta) db^2 \\
W &= W - \alpha \frac{dW}{\sqrt{s_{dw} + \varepsilon}} \\
b &= b - \alpha \frac{db}{\sqrt{s_{db} + \varepsilon}}
\end{aligned} \tag{10}
$$

From the above formula, s is a smoothing of the square of the gradient, and β is an index of gradient accumulation, with a typical value of 0.999. When updating the weight W and offset B, the gradient is divided by $\sqrt{S_{dw} + \varepsilon}$ first, which is equivalent to normalizing the gradient. If the gradient oscillates greatly in one direction, the step size should be reduced. If the oscillation is large, the s in this direction is also large. After division, the normalized gradient becomes smaller; if the gradient oscillation in a certain direction is very small, the normalized gradient becomes larger after division. By using the differential square weighted average for the gradient of weight W and offset B, it is helpful to eliminate the direction with large swing amplitude, which is used to correct the swing amplitude, so that the swing amplitude of each dimension is small. On the other hand, it also makes the network function converge faster. To prevent the denominator from being zero, a very small value ϵ is used for smoothing the general value is 10^{-8}.

4 Experiment

In order to verify the effectiveness of D3DCNN_ResNet proposed in this paper, the algorithm is applied to expression recognition and behavior recognition. Experimental verification is carried out on CK+ and KTH datasets respectively. The experiment is based on Python keras. The operating system: 64-bit Windows 10 Home Chinese version. CPU: Intel Core i7-6700. Graphics card: Intel HD Graphics 530. Memory: 8GB DDR2.

4.1 Data preprocessing

The data set used for expression recognition in this paper is CK+ database (Lucey et al., 2010), which is currently the most widely used database for expression recognition. It contains 593 expression video sequences from 123 people. The CK+ dataset is commonly used for facial expression recognition and contains video sequences where each sequence starts from a neutral expression and peaks at a specific emotion. The seven expressions are happiness, sadness, anger, fear, surprise, disgust, and contempt. The KTH dataset, primarily used for action recognition, consists of six types of human actions (walking, jogging, running, boxing, handwaving, and handclapping) performed under different conditions, providing a robust basis for behavior recognition experiments.

As shown in Figure 9, it has seven basic expressions: happiness, sadness, anger, fear, surprise, disgust and contempt.

In the experiment of behavior recognition, this paper uses KTH data set to verify. The KTH data set is composed of a total of 600 short videos, in which 25 people perform 6 actions under 4 scenarios: "Walking," "Jogging," "Running," "Boxing," "Handwaving" and "Handclapping," as shown in Figure 10.

In the experiment, the input original frame size is uniformly specified as 112*112. In both expression recognition and behavior recognition, 10 consecutive frame images are selected as model input. SSD detector is used to detect the face and human body parts and cut them to 48*48 to get the crop. Due to the small number of CK+ and KTH data samples, the existing standard data sets are augmented (such as random clipping, contrast adjustment, noise, mirror image, etc.) to enrich the training data. These augmentation techniques were selected due to their effectiveness in enhancing model robustness by diversifying the training data. Specifically, random cropping helps the model focus on different parts of the image, while contrast adjustment and noise addition improve the model's ability to generalize to varying lighting conditions and image quality, which are common challenges in real-world classroom environments. As shown in Figure 11, some data augmentation results are shown. Three pieces of CK+ are randomly clipped, one piece of mirror image, one piece of random noise, three pieces of KTH are randomly adjusted for contrast, one piece of mirror image, and one piece of random noise.

The experiment in this paper adopts the method of cross-validation, where the experimental samples are randomly divided into five parts: four are used as the training set, and the other as the test set for final model evaluation. Through training, the accuracy of the four training sets is obtained, and the average value is taken as the accuracy index of the training set in this paper. The accuracy of the test set is used as the accuracy index of the verification set. When training the model, set the batch size to 2, iterate 50 times, and print the results once per iteration.

4.2 Experimental results of CK+ expression data set

In the expression recognition experiment, Figures 12, 13 show the iterative process of the network in the CK+ data set. According to the accuracy curve and loss function curve of the training set and the test set in the training process, the accuracy of the method in this paper is high with the best recognition rate of 97.94%, and the convergence speed of the network is fast. The observed improvements in recognition rates were statistically significant based on standard deviation measures across the five-fold cross-validation. Furthermore, the preprocessing steps, including data augmentation techniques such as random cropping and contrast adjustment, contributed to the robustness of the model, helping to mitigate overfitting and improve

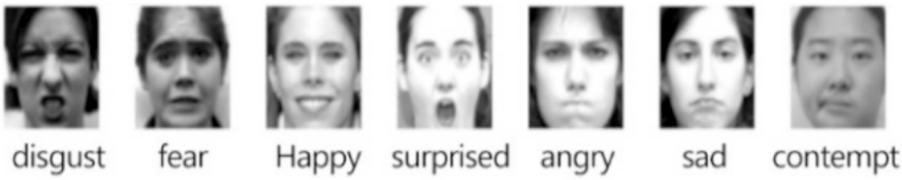


FIGURE 9
Seven expressions in the CK+ dataset. CK+ expression data set is from Lucey et al. (2010).

FIGURE 10
Six behaviors in the KTH dataset. KTH behavior data set is from Laptev and Lindeberg (2005).

generalization. Figure 14 shows the 50% recognition accuracy achieved by using the crop frame alone and by fusing the crop frame with the original frame. Figure 15 shows the corresponding average recognition accuracy. It can be seen from the two figures that the recognition accuracy of the fusion of the two features is high. The effect, improved by 1.23%, is not very significant. The reason is that expression recognition mainly depends on the detail changes of local areas such as eyes, eyebrows and mouth. Therefore, using crop frames alone can achieve better results.

In order to further evaluate the experimental results, it is compared with other frontier methods. Table 1 shows the comparison results of different algorithms on CK+. All comparison experiments were conducted under identical conditions, including the same data preprocessing, model initialization, and hyperparameter settings (e.g., learning rate, batch size). This ensured that the observed differences in performance were solely due to the models' architectures and not other external factors. Fan and Tjahjadi (2015) fused Phog top and optical flow method to capture changes in facial shape. Yu et al. (2016) introduces a special Bayesian network to capture the time relationship between facial changes, and develops corresponding facial modeling and recognition algorithms to improve the training and recognition speed. Yang et al. (2017) integrates two network models: 3D spatiotemporal network and static network. The former is used to extract spatiotemporal information, and the latter is used to extract the static features of key frames, and then make up for the lack of feature information through model fusion. Wang et al. (2018) proposed an expression recognition method combining dynamic texture information and motion information. Liu et al. (2014) uses 3DCNN to extract local features for fusion to recognize expression. Kacem et al. (2017), firstly, the face is mapped to the Riemannian manifold of positive semi definite matrix, and then the time parameter trajectory is established. Finally, the improved ppfSVM is used for classification, so as to improve the recognition accuracy. In this paper, 3DCNN is used to extract video sequence features, add quick connection to increase network depth, add high-level and low-level features of dense connection, input the clip frame and original frame detected by SSD respectively for training, and fuse the extracted two dense features for classification and recognition. Our proposed D3DCNN_ResNet achieves a recognition rate of 97.94% on CK+ database, which is superior to other methods.

4.3 Experimental results of KTH dataset

In order to verify the generalization of the D3DCNN_ResNet, it is also applied to behavior recognition. Figures 16, 17 show the iterative process of the network in the KTH data set. Through the recognition accuracy curve and loss function curve, it can be concluded that the method has good recognition performance, and the best recognition rate can be 98.86%. Figure 18 shows the 50%

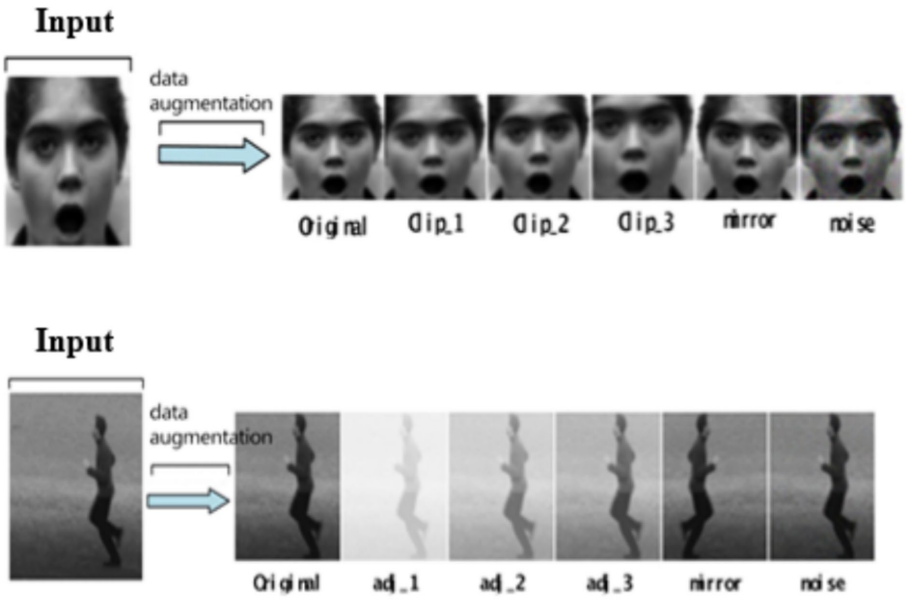


FIGURE 11
Data augmentation diagram. CK+ expression data set is from Lucey et al. (2010). KTH behavior data set is from Laptev and Lindeberg (2005).

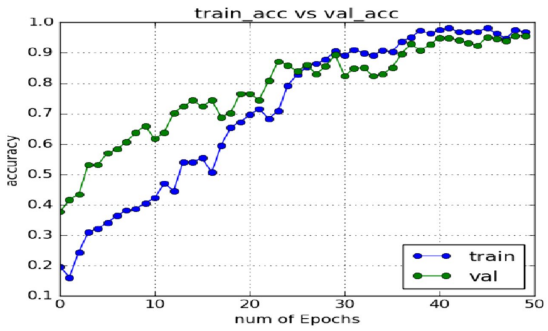


FIGURE 12
Accuracy of network in training set and test set.

recognition accuracy of extracting features by using the crop frame alone, and fusing the crop frame and the original frame. Figure 19 shows the corresponding average recognition accuracy. It can be seen from the two figures that the recognition accuracy of the proposed method is increased by 2.52% with obvious effect.

Table 2 compares the D3DCNN_ResNet with the advanced methods that have been published on the KTH dataset. Jia et al. (2018) uses linear sequence difference analysis method to reduce the dimension of video data for human behavior recognition. Wang et al. (2011) classifies human behavior by using dense trajectory features. Tong et al. (2013) uses the dense block of three-dimensional residuals as the basic module of the network, and uses the local feature aggregation adaptive method to learn the local dense features of human behavior. Wang et al. (2013) is based on the MBH descriptor characterization feature of differential optical flow, and obtains a good recognition rate. Zhang et al. (2019) improves 3DCNN network by dividing 3DCNN into two convolution kernels in spatial domain and time domain to extract spatiotemporal features in parallel to improve model efficiency. Compared with the above methods, this method has achieved better recognition results, and the recognition rate has reached 98.86%, which is 3.26, 4.66, 5.36, and 3.56% higher than Jia et al. (2018), Wang et al. (2011), Tong et al., 2013, and Wang et al. (2013), respectively. Compared with Zhang et al. (2019), 3DCNN is also split. In this paper, two convolution kernels in spatial domain and time domain are connected in a serial way. The verification shows that it is better than the parallel way. The recognition rate of single crop frame is better than 0.14%, and the fusion recognition rate of crop frame and original frame is better than 2.66%.

4.4 Performance evaluation of improved D3DCNN_ResNet

After the above experimental verification, the recognition rate is higher than that of extracting a single local feature of the crop frame by extracting the local feature of the crop frame in parallel and fusing the global feature of the original frame. In order to further verify the effectiveness of the proposed D3DCNN_ResNet, the D3DCNN network without residual connection and dense connection, the 3DCNN network with residual connection but without dense connection, and 3DCNN network without residual connection but with dense connection, are compared with D3DCNN_ResNet with residual connection and dense connection. As shown in Table 3, on CK+ dataset, the recognition rate of improved D3DCNN_ResNet is higher than that of 3DCNN and 3DCNN_ResNET by 10.01 and 6.66% respectively. On the KTH data set, the recognition rate is increased by 8.99 and 6.3% respectively. In conclusion, the D3DCNN_ResNet can effectively extract the feature information of video frames and improve the recognition accuracy.

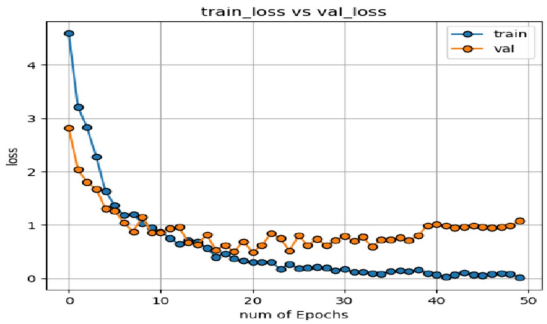


FIGURE 13
Loss function of network on training set and test set.

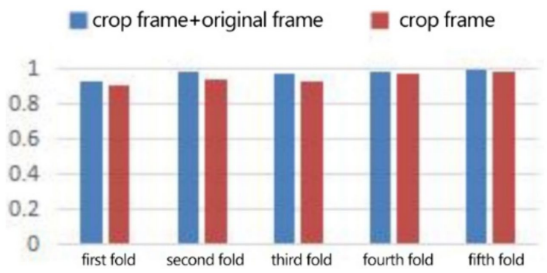


FIGURE 14
Compare the five-fold recognition rate of crop frame with crop frame + original frame.

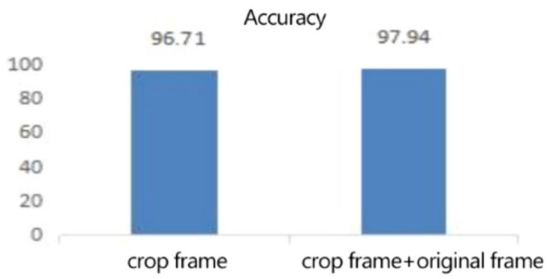


FIGURE 15
The average recognition rate (%) of clip frames and clip frames + original frames.

TABLE 1 Comparison results of different algorithms on CK+.

Method	Feature	Recognition rate
Fan and Tjahjadi (2015)	Phog-top and optical flow	90.9
Yu et al. (2016)	LABN	88.1
Yang et al. (2017)	3D model + static model	97.6
Wang et al. (2018)	STWLD and BOHF	91.6
Liu et al. (2014)	3DCNN	85.9
Kacem et al. (2017)	Space–time geometry	96.8
This article	D3DCNN_ResNet	97.9

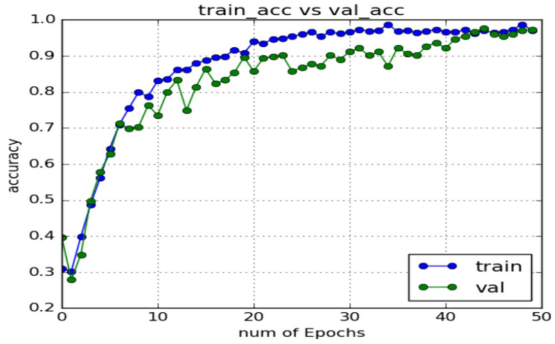


FIGURE 16
Accuracy of network in training set and test set.

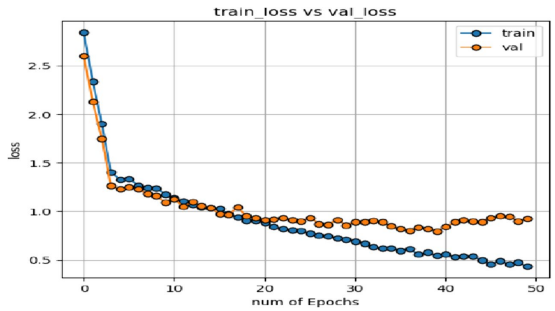


FIGURE 17
Loss function of network on training set and test set.

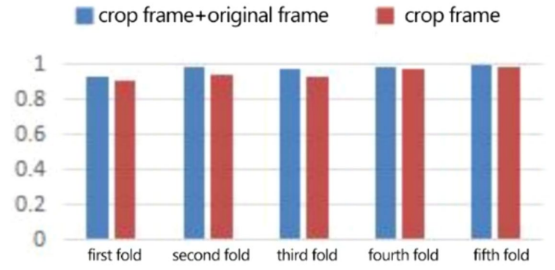


FIGURE 18
Compare the five-fold recognition rate of clip frames with clip frames + original frames.

5 Conclusion

This paper presents a new parallel D3DCNN_ResNet model structure. It divides the 3D convolutional neural network into two convolutions in spatial domain and time domain through the analysis of video sequence, extracts the spatiotemporal features in the video sequence in a serial manner, and adds a shortcut of residual network to increase the depth of the network. It solves the problem of excessive parameters and high computational costs in traditional 3D convolutional neural networks as the network depth increases. Moreover, dense connections are added in the residual block to fuse high-level and low-level features, maximizing the flow of feature information. The

method	Accuracy (%)	
	CK+	KTH
3DCNN	87.93	89.87
3DCNN_ResNet	91.28	92.56
D3DCNN_ResNet	97.94	98.86

FIGURE 19
The average recognition rate (%) of clip frames and clip frames + original frames.

TABLE 2 Comparison results of different algorithms on KTH.

Method	Feature	Recognition rate
Jia et al. (2018)	LSDA	95.6
Wang et al. (2011)	HOG-HOF-MBH	94.2
Tong et al. (2013)	3D-DenseNet	93.5
Wang et al. (2013)	Interest point detection based on flow vorticity	95.3
Zhang et al. (2019)	3DCNN	96.2
This article	D3DCNN ResNet	98.9

TABLE 3 Accuracy of different algorithms on CK+ and KTH.

Method	Accuracy (%)	
	CK+	KTH
3DCNN	87.93	89.87
3DCNN_ResNet	91.28	92.56
D3DCNN_ResNet	97.94	98.86

combination of residual blocks and dense connections not only reduces feature redundancy and improves the gradient correlation of the network, but also reduces the amount of network parameters, making the model have a certain anti over fitting effect. Through data preprocessing and data enhancement, the robustness and generalization of the model are enhanced, and it is not easy to be disturbed by external environmental factors. The clip frame and the original frame of the recognition region obtained from SSD target detection are trained respectively. Then multi-level features are extracted in parallel, and the classification is fused. Compared with a single network model, the parallel network can extract the local or even the overall spatial feature information of the image sequence more completely and effectively so as to improve the recognition rate. In the conventional teaching classroom, the students' feedback on facial expression is an important way for teachers to know whether the student is suitable for his own teaching. However, the teacher will not always pay attention to the student's expression and analyze of it, nor can he fully take into account the expression changes of all the students in the class. In this case, it is very meaningful to use computer technology as an assistant teacher to identify and record the expressions of students, analyze the quality of the class, and adjust the teaching progress to improve the teaching method. Through the application of the D3DCNN_ResNet, it is of great significance to effectively improve the recognition rate of classroom expressions. At the same time, the traditional English classroom oral speech training purely relies on speech recognition, which has a flaw that it is impossible to maintain a high recognition accuracy in a noisy environment. The visual recognition method is not affected by the ambient sound, and the accurate path of speech recognition can be improved through multimodal recognition. The new model structure of the D3DCNN_ResNet mentioned in this paper can also assist in speech interaction and image recognition, which is widely used in the field of English spoken speech teaching. However, given the complexity of the language environment, it will take time to truly put into practice, and it is still necessary to further strengthen the integration of research in areas such as big data visual analysis and artificial intelligence technology.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

XL: Writing – original draft. TY: Writing – original draft. MT: Methodology, Writing – original draft. PX: Data curation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and publication of this article. This work was supported by Teaching Reform Research Project of Shangrao Normal University, 'Exploration of the teaching mode of English and American literature in normal colleges from the perspective of curriculum ideology and politics' in 2022, and also support by the funding NSFC with award number 62163024.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5, 157–166. doi: 10.1109/72.279181

Fan, X., and Tjahjadi, T. (2015). A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. *Pattern Recogn.* 48, 3407–3416. doi: 10.1016/j.patcog.2015.04.025

Gers, F. A., Schraudolph, N. N., and Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* 3, 115–143. doi: 10.1162/153244303768966139

Graves, A., Fernández, S., and Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification and recognition. Artificial Neural Networks: Formal Models and Their Applications—ICANN 2005, 15th International Conference. Warsaw, Poland, September 11–15, 2005. 799–804.

Gu, W., Xiang, C., Venkatesh, Y. V., Huang, D., and Lin, H. (2012). Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recogn.* 45, 80–91. doi: 10.1016/j.patcog.2011.05.006

Hong, K. (2022). Facial expression recognition based on anomaly feature. *Opt. Rev.* 29, 178–187. doi: 10.1007/s10043-022-00734-3

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille. IEEE. 448–456.

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 221–231. doi: 10.1109/TPAMI.2012.59

Jia, X., Yu, F., and Chen, Y. (2018). Improved explicit shape regression for face alignment algorithm. *J. Compu. Appl.* 38:1289. doi: 10.1109/CCDC.2018.8407305

Kacem, A., Daoudi, M., Amor, B. B., and Alvarez-Paiva, J. C. (2017). A novel space-time representation on the positive semidefinite con for facial expression recognition. 2017 IEEE International Conference on Computer Vision (ICCV). Venice. IEEE

Laptev, I., and Lindeberg, T. (2005). Space-time interest points. *Int. J. Comput. Vis.* 64, 107–123. doi: 10.1007/s11263-005-1838-7

Liu, M., Li, S., Shan, S., Wang, R., and Chen, X. (2014). Deeply learning deformable facial action parts model for dynamic expression analysis. Asian Conference on Computer Vision. Springer. Cham.

Liu, S.-S., Tian, Y.-T., and Wan, C. (2011). Facial expression recognition method based on Gabor multi-orientation features fusion and block histogram. *Acta Automat. Sin.* 37, 1455–1463. doi: 10.3724/SP.J.1004.2011.01455 (in Chinese)

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA. IEEE.

Ma, Z., Teschendorff, A. E., Leijon, A., Qiao, Y., Zhang, H., and Guo, J. (2015). Variational bayesian matrix factorization for bounded support data. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 876–889. doi: 10.1109/TPAMI.2014.2353639

Niu, L., Huang, S., Zhao, X., Kang, L., Zhang, Y., and Zhang, L. (2022). Hallucinating uncertain motion and future for static image action recognition. *Comput. Vis. Image Underst.* 215:103337. doi: 10.1016/j.cviu.2021.103337

Qiu, Z., Yao, T., and Mei, T. (2017). Learning spatio-temporal representation with pseudo-3D residual networks. 2017 IEEE International Conference on Computer Vision (ICCV). Venice. IEEE

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Tong, L., Song, Q., Ge, Y., and Liu, M. (2013). HMM-based human fall detection and prediction method using tri-axial accelerometer. *IEEE Sensors J.* 13, 1849–1856. doi: 10.1109/JSEN.2013.2245231

Wang, H., Alexander, K., Schmid, C., and Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.* 103, 60–79. doi: 10.1007/s11263-012-0594-8

Wang, H., Klaser, A., Schmid, C., and Liu, C.-L. (2011). Action recognition by dense trajectories. IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO. IEEE.

Wang, X., Peng, M., Hu, M., Jin, C., and Ren, F. (2018). Combination of valence-sensitive loss with restrictive center loss for facial expression recognition. 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI). Xiamen. IEEE. 528–533.

Wanshu, L., and Bin, N. (2022). High-dynamic dance motion recognition method based on video visual analysis. *Sci. Program.* 2022, 1–9. doi: 10.1155/2022/6724892

Wu, H., Lu, Z., and Zhang, J. (2023). A privacy-preserving student status monitoring system. *Complex Intell. Syst.* 9, 597–608. doi: 10.1007/s40747-022-00796-5

Xun, Z. (2022). Monitoring and analysis of English classroom teaching quality based on big data. *Secur. Commun. Netw.* 2022, 1–9. doi: 10.1155/2022/5365807

Yang, Y., Peng, Y., and Han, S. (2017). Video segmentation based on patch matching and enhanced onecut. 2017 2nd International Conference on Image, Vision and Computing (ICIVC). IEEE. Chengdu. 346–350.

Yu, Q., Jieyu, Z., and Yanfang, W. (2016). Facial expression recognition using temporal relations among facial movements. *Acta Electron. Sin.* 44, 1307–1313. doi: 10.3969/j.issn.0372-2112.2016.06.007

Zhang, X. J., Li, C. Z., Sun, L. Y., and Zhang, M. L. (2019). Behavior recognition method based on improved 3D convolutional neural network. *Comput. Integr. Manuf. Syst.* 25, 2000–2006. doi: 10.13196/j.cims.2019.08.014

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 11 October 2024
DOI 10.3389/fnins.2024.1441791

OPEN ACCESS

EDITED BY
Yanqiu Huang,
University of Twente, Netherlands

REVIEWED BY
Guangyu Dan,
University of Illinois Chicago, United States
Zihong Zhu,
Harbin Medical University, China

*CORRESPONDENCE
Xin Xie
✉ xxhedf@163.com
Xin Ding
✉ dingxin81@163.com

RECEIVED 03 July 2024
ACCEPTED 26 September 2024
PUBLISHED 11 October 2024

CITATION
Li L, Lu Z, Jiang A, Sha G, Luo Z, Xie X and Ding X (2024) Swin Transformer-based automatic delineation of the hippocampus by MRI in hippocampus-sparing whole-brain radiotherapy.
Front. Neurosci. 18:1441791.
doi: 10.3389/fnins.2024.1441791

Swin Transformer-based automatic delineation of the hippocampus by MRI in hippocampus-sparing whole-brain radiotherapy

Liang Li[1], Zhennan Lu[2], Aijun Jiang[1], Guanchen Sha[3], Zhaoyang Luo[4], Xin Xie[1]* and Xin Ding[1]*

[1]Department of Radiotherapy, The Affiliated Hospital of Xuzhou Medical University, Xuzhou, China, [2]Department of Equipment, Affiliated Hospital of Nanjing University of Chinese Medicine (Jiangsu Province Hospital of Chinese Medicine), Nanjing, China, [3]Department of Radiation Oncology, Xuzhou Central Hospital, Xuzhou, China, [4]HaiChuang Future Medical Technology Co., Ltd., Zhejiang, China

Objective: This study aims to develop and validate SwinHS, a deep learning-based automatic segmentation model designed for precise hippocampus delineation in patients receiving hippocampus-protected whole-brain radiotherapy. By streamlining this process, we seek to significantly improve workflow efficiency for clinicians.

Methods: A total of 100 three-dimensional T1-weighted MR images were collected, with 70 patients allocated for training and 30 for testing. Manual delineation of the hippocampus was performed according to RTOG0933 guidelines. The SwinHS model, which incorporates a 3D ELSA Transformer module and an sSE CNN decoder, was trained and tested on these datasets. To prove the effectiveness of SwinHS, this study compared the segmentation performance of SwinHS with that of V-Net, U-Net, ResNet and VIT. Evaluation metrics included the Dice similarity coefficient (DSC), Jaccard similarity coefficient (JSC), and Hausdorff distance (HD). Dosimetric evaluation compared radiotherapy plans generated using automatic segmentation (plan AD) versus manual hippocampus segmentation (plan MD).

Results: SwinHS outperformed four advanced deep learning-based models, achieving an average DSC of 0.894, a JSC of 0.817, and an HD of 3.430 mm. Dosimetric evaluation revealed that both plan (AD) and plan (MD) met treatment plan constraints for the target volume (PTV). However, the hippocampal D_{max} in plan (AD) was significantly greater than that in plan (MD), approaching the 17 Gy constraint limit. Nonetheless, there were no significant differences in $D_{100\%}$ or maximum doses to other critical structures between the two plans.

Conclusion: Compared with manual delineation, SwinHS demonstrated superior segmentation performance and a significantly shorter delineation time. While plan (AD) met clinical requirements, caution should be exercised regarding hippocampal D_{max}. SwinHS offers a promising tool to enhance workflow efficiency and facilitate hippocampal protection in radiotherapy planning for patients with brain metastases.

KEYWORDS

hippocampus, whole brain radiotherapy, automatic segmentation, Swin Transformer, MRI

1 Introduction

Whole-brain radiotherapy (WBRT) is an effective treatment for patients with brain metastases (Berghoff and Preusser, 2018). Prophylactic cranial irradiation (PCI) can significantly reduce the probability of brain metastasis and improve the overall survival rate of patients (Gondi et al., 2010). However, WBRT can cause hippocampal damage and cognitive disorders, with an incidence ranging from 50 to 90%. This often manifests as short-term memory impairment, decreased attention, and problem-solving abilities, seriously affecting the patient's quality of life (Fike et al., 2009; Peters et al., 2016). With advancements in radiotherapy and growing emphasis on post-radiotherapy quality of life, hippocampal avoidance whole-brain radiotherapy (HA-WBRT) has been shown to significantly improve cognitive function in patients post-treatment. The Radiation Therapy Oncology Group (RTOG) 0933 phase II trial demonstrated that protecting the hippocampus could reduce the incidence of cognitive dysfunction to 7% (Gondi et al., 2014). Subsequently, the NRG Oncology CC001 phase III trial confirmed these findings. Notably, the results showed that WBRT combined with memantine for hippocampal protection resulted in superior cognitive preservation in adult patients with brain metastases, compared to WBRT with memantine but without hippocampal protection. Importantly, there was no significant difference in intracranial progression-free survival (PFS) or overall survival (OS) (Brown et al., 2020). Therefore, protecting the hippocampus during midbrain radiotherapy for brain tumor patients can mitigate memory and cognitive impairment, consequently enhancing the overall quality of life.

According to RTOG 0933, outlining the hippocampus on axial T1-weighted MR images is essential (Gondi et al., 2014). However, the hippocampus is a complex structure, and accurate delineation is crucial for effective radiation treatment planning and minimizing radiation-related side effects (Mukesh et al., 2012; Walker et al., 2014). Additionally, the hippocampus is situated between the thalamus and the medial temporal lobe of the brain. In magnetic resonance imaging, the gray matter intensity of the hippocampus is very similar to that of surrounding structures like the amygdala, caudate nucleus, and thalamus, with no distinct boundary, making delineation difficult. Currently, the method of hippocampal delineation is mainly based on the anatomical expertise of the doctor, who refers to the patient's MR images to outline the CT images. The accuracy of this approach depends on the registration precision between MR and CT images, as well as the physician's proficiency and anatomical knowledge. Significant variability exists between the delineation results of different doctors. Therefore, improving the accuracy, efficiency, and standardization of hippocampal delineation is a key step in reducing the risk of radiation-induced brain injury. Automatic segmentation of the hippocampus from MR images remains a challenging task.

Deep learning approaches based on convolutional neural networks (CNNs) have been widely used due to their efficiency and accuracy (Erickson, 2021). In 2015, U-Net was first proposed, constructing a U-shaped deep network using encoders and symmetric decoders, achieving commendable performance in segmenting image edges (Ronneberger et al., 2015). Specifically, U-Net employs an encoder to extract low-level details and high-level semantic features from the image and utilizes a decoder to map the features back to the original size, thereby generating the segmented image. By establishing connections between the encoder and decoder, the features from corresponding layers in both components can be merged, which enhances the preservation of detailed information in the input image and improves segmentation outcomes. Given that CT and MR images are typically three-dimensional, a 3D U-Net (Çiçek et al., 2016) was designed. Building upon this framework, V-Net (Milletarì et al., 2016) integrates encoder information filtered by the decoder and adds a ResNet (He et al., 2016), which prevents gradient vanishing, accelerates network convergence, and achieves superior performance. Subsequently, several U-Net variants have been developed.

The CNN method typically utilizes deep convolution layers with an encoder-decoder architecture to capture global information. However, this process often relies on skip connections to compensate for the loss of shallow feature information. The convolution operation is inherently local due to the receptive field size, which limits its effectiveness, particularly in segmenting small objects (Fei et al., 2023).

In recent years, the Transformer has been widely adopted in medical image segmentation as an alternative architecture featuring a global self-attention mechanism. Models like TransFuse (Zhang Y. et al., 2021), MFSuse (Basak et al., 2022), TFormer (Zhang et al., 2023), and TransCeption (Azad et al., 2023) effectively capture edge information, enhance segmentation accuracy, and optimize network performance. However, these advancements come with increased parameters, computational complexity, and longer inference times. Despite their enhanced localization capabilities, these models still struggle to capture low-level details.

As a transformer-based model, the Vision Transformer (VIT) (Dosovitskiy et al., 2020) surpasses CNNs due to its global and long-range modeling capabilities. However, VIT's computational efficiency is relatively low because it depends on a self-attention mechanism for feature extraction. Swin Transformer (Liu et al., 2021), a new variant of VIT, introduces a sliding window approach to constrain self-attention. This model integrates locality into multihead self-attention (MHSA) through local self-attention (LSA), embedding local details in the earlier layers. However, LSA's performance is comparable to that of convolution and is inferior to dynamic filters. To improve this, an enhanced LSA module (ELSA) (Zhou et al., 2021) has been introduced to better capture local information. SwinBTS (Jiang et al., 2022), the first model to incorporate the ELSA Transformer module in brain tumor segmentation tasks, brings forward innovative approaches.

Building upon the success of the Swin Transformer and the detailed feature extraction capabilities of enhanced local self-attention (ELSA), we propose SwinHS, a novel neural network designed for the automatic segmentation of hippocampal MR images. SwinHS improves local detail extraction by incorporating a 3D ELSA Transformer module. Additionally, we introduce the spatial squeeze excitation (sSE) block, which allows feature maps to be more informative both spatially and across channels. The primary goal of this study was to develop an AI tool for automated hippocampal delineation, with a focus on validating the segmentation's accuracy and clinical applicability, ultimately aiming to enhance workflow efficiency for clinicians.

2 Materials and methods

2.1 Data collection

The Ethics Committee (No. XYFY2023-KL155-01) approved the retrospective collection of 100 three-dimensional T1-weighted

(3D-T1) MR images from patients who underwent hippocampus-protected whole-brain radiotherapy at the Department of Oncology and Radiology of Xuzhou Medical University between 2018 and 2023. The patient cohort included 61 males and 39 females, aged between 30 and 83 years, with a median age of 60 years. Any images depicting hippocampal tumor invasion were excluded prior to MRI. The images were obtained using a GE Discovery MR750 3.0 T (GE Healthcare, Milwaukee, WI, United States) magnetic resonance imaging system. The scanning protocol employed a slice thickness of 0.8 mm, a spatial resolution of $(0.8 \times 0.496 \times 0.496)$ mm^3. The sequence used was 3D BRAVO, with repetition time (TR) = 7 ms, echo time (TE) = 3 ms, flip angle (FA) = 12°, and the resulting images were exported and saved in DICOM files. For the study, 70 patients were allocated to the training set and 30 to the test set.

2.2 Manual delineation

In accordance with the hippocampus atlas contouring guidelines proposed by RTOG0933 (Gondi et al., 2014), a tumor radiotherapist who was thoroughly trained and experienced in hippocampal delineation manually outlined the hippocampus on 100 axial MR images using the Varian Eclipse 13.6 planning system (Varian Medical Systems, Palo Alto, CA, United States). To ensure accuracy, the delineation results were subsequently reviewed and, where necessary, adjusted by another expert in tumor radiotherapy. The two radiologists involved in this study have 13 and 14 years of experience respectively, ensuring a high level of expertise in interpreting the imaging data.

2.3 Model training and testing

The overall architecture is illustrated in Figure 1. The input consists of a multimodal MR medical image $X \in R^{H\times W\times D\times C}$, where the image size is $H \times W \times D$ and C is the number of channels. These images are divided into non-overlapping patches, which are then passed to the transformer-encoder. The encoded features are subsequently processed through the ELSA module and the Swin Transformer module. Next, the feature representations are transmitted to the sSE CNN-decoder via skip connections at multiple resolutions, generating the final segmentation output. Each component of the proposed architecture is detailed in the following sections.

2.3.1 Transformer encoder

Initially, we employ a 3D patch partition layer to segment medical images into nonoverlapping 3D patches with a volume of $\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}$. Subsequently, these patches are projected into an embedding space with a dimensionality of C, enabling us to generate a feature map of size $\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}\times C$.

2.3.2 ELSA Transformer module

The ELSA Transformer module is employed to enhance local detailed feature extraction. ELSA introduces a novel local self-attention mechanism that outperforms both LSA and dynamic filters in the Swin Transformer. A key element of ELSA is Hadamard attention, which applies the Hadamard product to improve attention across neighboring elements while maintaining high-order mapping. In deep learning, it is commonly assumed that higher-order mappings offer stronger fitting capabilities. The low accuracy of some attention mechanisms may stem from their lower mapping order (Gondi et al., 2010), as the attention mechanism typically performs second-order mapping of the input, as described in Supplementary Formula 1.

As illustrated in Figure 2, the ELSA Transformer module is derived by incorporating an identical MLP module subsequent to the attention structure in conjunction with the Transformer architecture, as depicted in Supplementary Formula 3.

2.3.3 Swin Transformer module

The Swin Transformer is a hierarchical VIT that performs self-attention computations through an efficient shifted window partitioning scheme. This approach significantly reduces the number of parameters while enabling multiscale feature extraction with improved feature learnability. As shown in Figure 1, the Swin Transformer block in the architecture consists of a normalization layer (LN), window-based multihead self-attention module (MHSA), and multilayer perceptron (MLP).

2.3.4 SSE CNN decoder

The decoder has the same depth as the encoder and is used to decode the feature representation of the extracted encoder. A skip connection is used between the encoder and decoder at each resolution. The output characteristics are reshaped to the size $\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}$ at each stage i ($i \in 0, 1, 2, 3, 4$) of the encoder and the bottom, and then the residual block composed of two $3\times3\times3$ normalized convolutional layers is input. Then, the sSE block is applied to the extracted features so that the feature map can provide more information both spatially and channelwise for image segmentation (Hu et al., 2018).

A linear transformation of the feature map is performed to enhance the dimension $(\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}\times C)$, which subsequently allows obtaining an output resolution similar to that of the image input, i.e., $H\times W\times D$ resolution output. The final segmentation output is calculated by using a $1\times1\times1$ convolutional layer and a sigmoid activation function.

For the model, the Adam optimizer is used for training, the initial learning rate is 1×10^{-4}, and the weight is attenuated to 1×10^{-5}. The default batch size is 50, and the default number of training iterations is 150. All experiments were performed using an Nvidia RTX2080Ti GPU.

2.4 Model evaluation

We use the Dice similarity coefficient (DSC), Jaccard similarity coefficient (JSC) and Hausdorff distance (HD) to evaluate the performance of our automatic delineation tool in the test set. The formulas are provided in the Supplementary material.

The DSC is the most commonly used metric for measuring the overlap between two contours, and its value is between 0 and 1. The larger the DSC value is, the greater the similarity between the two contour lines (Taha and Hanbury, 2015). Similarly, JSC compares the similarities and differences between finite sets (Eelbode et al., 2020).

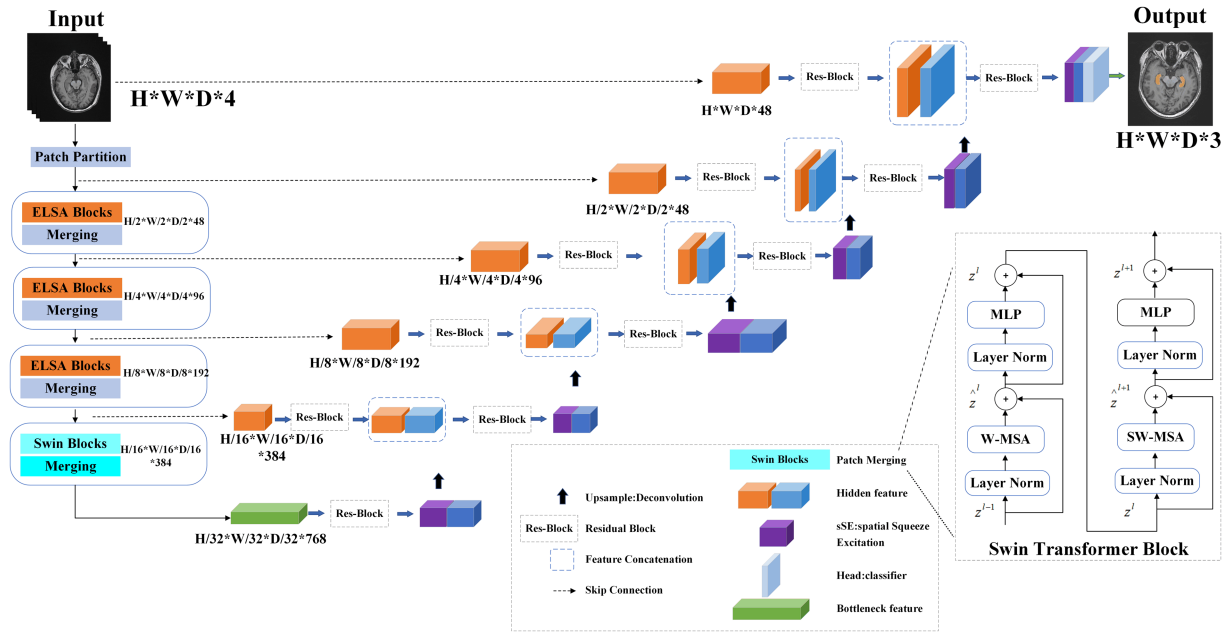


FIGURE 1
Overview of the model training process. The Transformer encoder includes three levels: each level includes two ELSA Transformer modules, and the next level includes two Swin Transformer modules. In the first level, the linear embedding layer is used to create a three-dimensional feature map. In the first, second and third stages, the ELSA module is used to extract very detailed feature information. Stage 4 uses the Swin Transformer to extract multiresolution features and the shift window mechanism to calculate self-attention. In these four stages, a slice merging layer is used to reduce the resolution of the feature by 2 times. On the right, the encoder is used to decode the extracted feature representations of the encoder through skip connections.

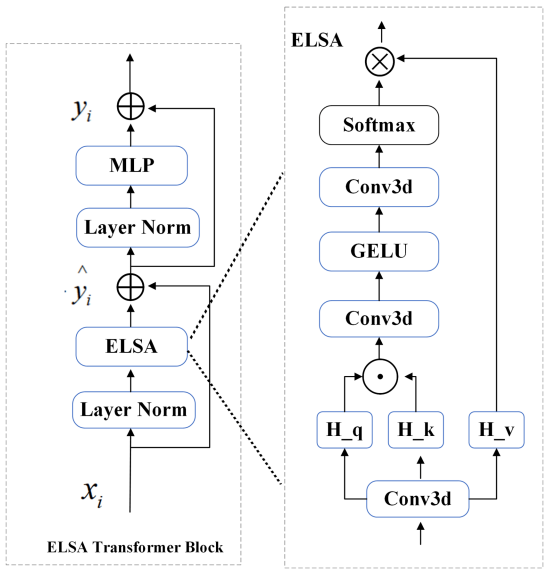


FIGURE 2
Structure of the ELSA Transformer block.

The larger the JSC is, the greater the sample similarity. HD describes the boundary similarity of 2 point sets by measuring the maximum distance of the closest pair of points. The smaller the HD is, the greater the coincidence degree between A and B, and the better the segmentation effect (Taha and Hanbury, 2015). In the special dosimetric evaluation reported by Gondi et al. (2014), deviations greater than 7 mm are considered unacceptable.

2.5 Comparison of model performance

To demonstrate the effectiveness of our proposed SwinHS, we compared its segmentation performance with that of four advanced deep learning-based methods: V-Net (Milletari et al., 2016), U-Net (Ronneberger et al., 2015), ResNet (He et al., 2016), and VIT (Chen et al., 2022). Then, each method was trained and tested on the same dataset using their respective frameworks. We evaluated the performance of SwinHS and the four other methods using DSC, JSC, and HD.

2.6 Radiotherapy planning and dosimetric evaluation

To evaluate the feasibility of applying hippocampal delineation via the SwinHS model in clinical practice, we conducted a study comparing simulated whole-brain radiotherapy plans for 10 randomly selected patients. We compared the differences in dosimetric distribution between two sets of radiotherapy plans: one using the manually delineated hippocampi and the other using the automatically delineated hippocampi by the SwinHS model. A large aperture CT simulator (Philips, Cleveland, OH, United States) was used to collect CT localization images of the patients' head area, with a slice thickness of 1.5 mm. According to the RTOG 0933 report, the hippocampus is a low-signal gray matter structure that begins medially from the inferior

horn of the lateral ventricle's temporal horn and is bounded externally by the cerebrospinal fluid, typically forming a crescent shape. On the MR image, the contoured hippocampal tissue is expanded by 5 mm to create the hippocampus planning risk volume (HC-PRV). We registered the patient positioning CT image with the 3DT1 MR image on the MIM workstation (MIM Software Inc., Beachwood, OH). The delineated hippocampus and HC-PRV on the MR image were then mapped to the positioning CT, and the CT image was subsequently imported back to the treatment planning system (TPS) for the creation of the simulated radiotherapy plan. The target area was defined as follows: the patient's whole brain tissue was the CTV, and the CTV was expanded by 3 mm and subtracted from the hippocampus to generate the planned target volume (PTV). The prescribed radiotherapy dose was 30 Gy, administered in 10 fractions. The goal of all plans was to cover at least 95% of the PTV with a 100% prescription dose.

We created two radiotherapy plans, plan (AD) and plan (MD), using the automatically contoured hippocampus (AD) and manually contoured hippocampus (MD) respectively, where the optimization parameters are identical. The radiotherapy plan was designed using the Varian VitalBeam (Varian Medical Systems, Palo Alto, CA, United States) 6 MV X-ray in FFF mode, utilizing the dynamic intensity-modulated radiotherapy (sIMRT) technique. A dose rate of 1,200 MU/min was applied across 9 noncoplanar irradiation fields. Dose calculations were performed with the Varian Eclipse 13.6 planning system, employing an anisotropic analytical algorithm with a spatial resolution of 2.5 mm.

Given that we used the manually delineated hippocampus as the reference standard for evaluating the accuracy of the automatically delineated hippocampus, our plan (AD) and plan (MD) evaluations were based on the manually delineated hippocampus to accurately reflect the hippocampal dose during radiotherapy. We compared the dose distribution differences between the radiotherapy plan (AD) and the radiotherapy plan (MD) using a dose-volume histogram (DVH) and assessed whether the relevant indicators in the radiotherapy plans met the dose limits outlined in the RTOG-0933 protocol (Gondi et al., 2015) and the NRG Oncology CC001 phase III trial (Brown et al., 2020). When administering whole-brain radiotherapy at 30 Gy/10 F, the indices included the following: (1) PTV: $D_{2\%}$ ≤37.5 Gy ($D_{2\%}$: the dose received by 2% of the PTV), $D_{98\%}$ ≥25 Gy ($D_{98\%}$: the dose received by 98% of the PTV), V30 Gy ≥90% ($V_{30\,Gy}$: the percentage of the PTV volume receiving 30 Gy). (2) Hippocampus: D_{max} ≤17 Gy (maximum dose), $D_{100\%}$ ≤10 Gy ($D_{100\%}$: the minimum dose received by the entire hippocampus). All treatment plans were designed by the same medical physicist with 5 years of experience in radiotherapy planning, and subsequently reviewed by other experts to ensure quality and adherence to clinical standards.

2.7 Statistical analysis

Paired t tests were conducted to compare the hippocampal volume, DSC, JSC and HD between the manual delineation group (MD) and the automatic delineation group (AD), as well as assess the differences in dosimetric parameters between the MD and AD plans. All the statistical analyses were performed using SPSS v22.0 software. A significance level of $p < 0.05$ was considered considered statistically significant.

3 Results

3.1 Patient characteristics

The characteristics of the patients in the training dataset and the test dataset are presented in Table 1. In the test dataset, a significant difference was observed between the hippocampus volumes in the manual delineation (MD) and automatic delineation (AD) groups, with a *p*-value of 0.019. Specifically, the hippocampus volume in the AD group was smaller than that in the MD group.

3.2 Performance comparison of the SwinHS models

We compared the segmentation results of five different models in the test dataset, as presented in Table 2, using DSC, JSC, and HD as evaluation metrics. The table demonstrates that our proposed model outperforms the other four models across all indicators. Specifically, the average DSC is 0.894 ± 0.017, the average JSC is 0.817 ± 0.020, and the average HD is 3.430 ± 0.245 mm.

The segmentation results of the hippocampus at different levels between the proposed model and other models are visually compared, in Figure 3. The first column displays the actual manual segmentation of the hippocampus. From the second column onwards, it becomes evident that the proposed method shows greater consistency with the manual delineation of the hippocampus contour. In the third column, the contour delineated by VIT appears smooth but shows slight deviations from the actual delineation. The fourth and fifth columns reveal rough hippocampal contours delineated by 3D ResNet and 3D U-Net, respectively. Finally, in the sixth column, the hippocampal contour delineated by V-Net is depicted inaccurately and incompletely.

TABLE 1 Basic characteristics of the 100 patients.

	Total subjects (*n* = 100)	Training cohort (*n* = 70)	Testing cohort (*n* = 30)	
			MD	AD
Number of male patients (%)	61 (61)	43 (61.4)	18 (60)	
Number of female patients (%)	39 (39)	27 (38.6)	12 (40)	
Median age in years (range)	60 (30–83)	62.5 (33–81)	57.5 (30–83)	
Volume of hippocampus (±SD cm^3)	4.02 ± 0.83	3.89 ± 0.85	4.32 ± 0.70	4.15 ± 0.67
p-value	-	-	0.019	

MD, manual hippocampus segmentation; AD, automatic segmentation; SD, standard deviation.

3.3 Dosimetric evaluation of the SwinHS model

According to the requirements of the RTOG0933 phase II trial (Gondi et al., 2015) and the NRG Oncology CC001 phase III trial (Brown et al., 2020), specific criteria must be met for radiotherapy planning. When administering whole-brain radiotherapy at 30 Gy/10 F, it is essential to ensure that the dose received by 2% of the planning target volume (PTV $D_{2\%}$) is ≤40 Gy and that the dose received by 98% of the PTV ($D_{98\%}$) is ≥25 Gy. Additionally, it is considered unacceptable if the volume of the PTV receiving 30 Gy ($V_{30\ Gy}$) exceeds 90%. Furthermore, for the hippocampus, it is imperative that the minimum dose ($D_{100\%}$) does not exceed 10 Gy, and the maximum dose (D_{max}) remains under 17 Gy. In this study, radiotherapy plans were generated using both AD and MD hippocampus, denoted as plan (AD) and plan (MD), respectively. We utilized the MD hippocampus as the reference standard to assess the accuracy of the AD hippocampus. Subsequently, we compared the dose indicators and distribution differences between plan (AD) and plan (MD) based on MD hippocampus delineation.

TABLE 2 Results of different models.

	DSC	JSC	HD (mm)	*p*-value
SwinHS	0.894 ± 0.017	0.817 ± 0.020	3.430 ± 0.245	
VIT	0.891 ± 0.016	0.803 ± 0.016	3.959 ± 0.328	0.002
3D ResNet	0.871 ± 0.024	0.783 ± 0.022	4.730 ± 0.262	0.016
3D U-Net	0.845 ± 0.025	0.759 ± 0.019	6.895 ± 0.268	2.5×10^{-4}
V-Net	0.778 ± 0.020	0.674 ± 0.023	7.785 ± 0.277	0.008

DSC, Dice similarity coefficient, JSC, Jaccard similarity coefficient; HD, Hausdorff distance.

The representative patient dose distributions comparing automatic and manual hippocampus segmentation plans are shown in Figure 4. Plan (AD) was generated using automatically delineated hippocampus, while plan (MD) was based on hippocampus manually contoured by experienced clinicians. The contours in both plan (AD) and plan (MD) are the same; however, the manually contoured hippocampus serves as the reference standard for evaluating both plans. The volume of the automatically segmented hippocampus was smaller than that of the manually delineated hippocampus, resulting in the 17 Gy dose color brush being closer to the actual hippocampus in the automatic segmentation plan. As shown in Table 3, the dose indicators for PTV in both plan (AD) and plan (MD) met the treatment plan constraints recommended by the RTOG 0933 trial, with no significant differences observed between the two groups of plans. Regarding hippocampus dosimetry, although both plan (AD) and plan (MD) met acceptable variations, the hippocampus D_{max} in plan (AD) was significantly greater than that in plan (MD), with a notable difference ($p < 0.001$) at 1697.03 ± 11.02 cGy, approaching the limit of the 17 Gy constraint. Moreover, there was no significant difference in $D_{100\%}$ between the two groups ($p = 0.236$).

3.4 Delineation time analysis

The median time required for automatic hippocampal delineation in the test group of 30 patients was 13.3 s (range: 11.7–14.9 s). This result was significantly shorter than the time required for manual delineation (MD) ($p < 0.001$), which was 786 s (range: 635–905 s).

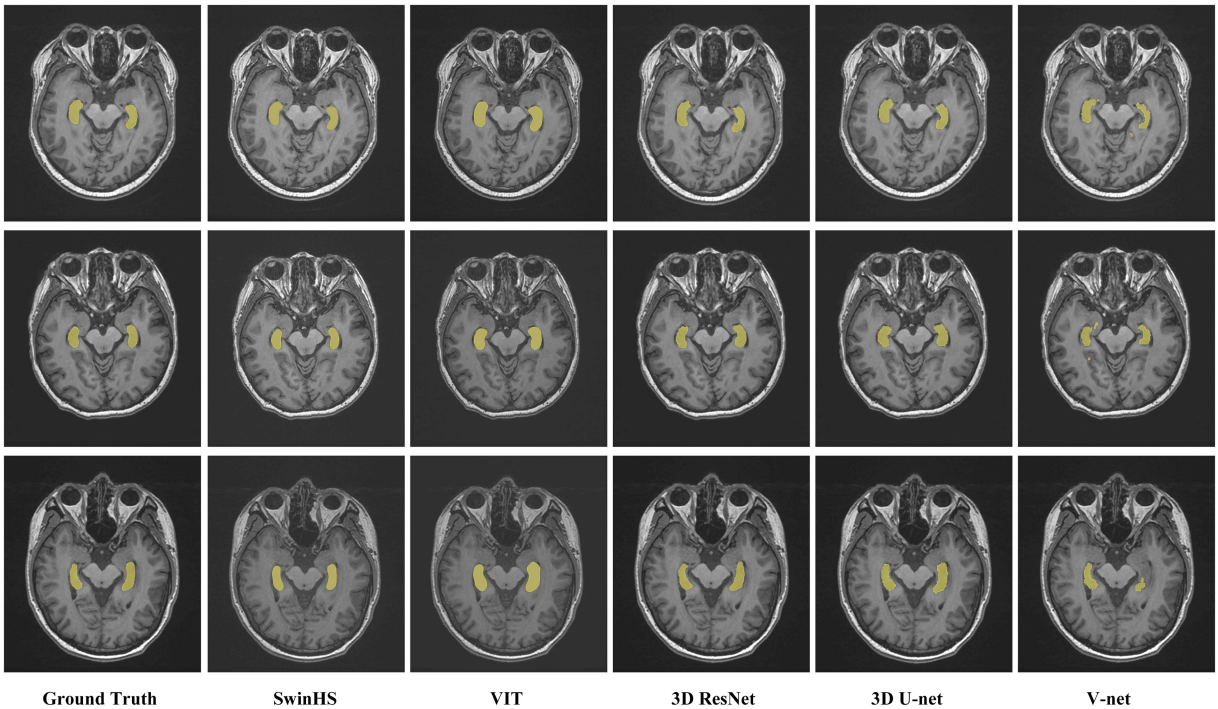


FIGURE 3
Representative cases of the proposed model and other model methods. The first column shows the contour of the hippocampus in the real MR image. Columns 2 to 6 show the results of the proposed model, VIT, 3D ResNet, 3D U-Net, and V-Net, respectively.

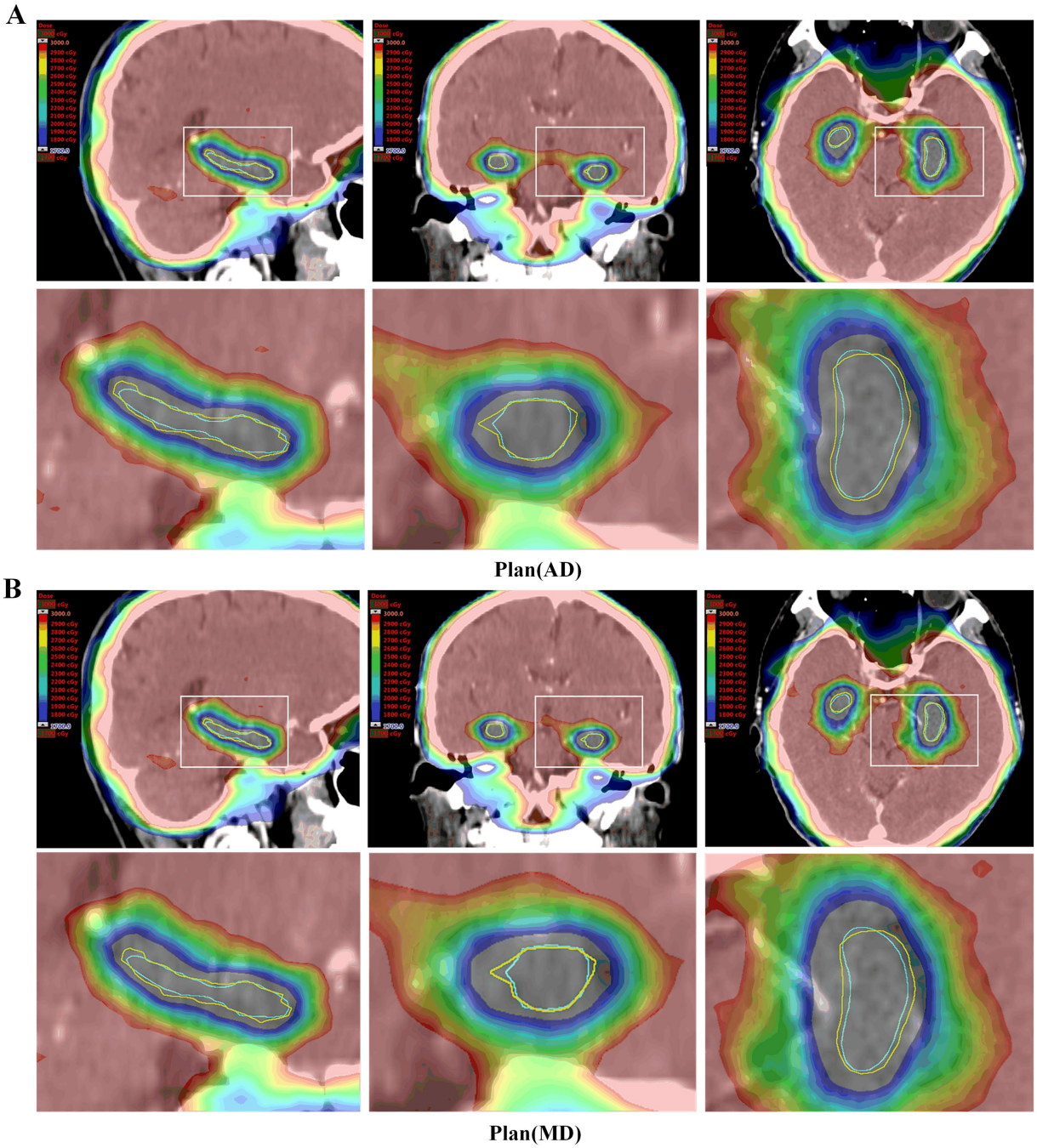


FIGURE 4
Dose distribution of representative patient plans (MD) and plans (AD). **(A)** Plan (AD) automatic delineation of the hippocampus-generated radiotherapy plan. **(B)** Plan (MD) manually outlines the radiotherapy plan generated by the hippocampus. The horse body was manually sketched (yellow line), and the hippocampus was automatically depicted (blue). Both plans were evaluated using manual delineation of the hippocampus. In the Plan (AD), a small portion of the manually delineated hippocampus was closer to a dose of 1700 cGy.

4 Discussion

In this study, we employed a Swin Transformer-based neural network, SwinHS, to automatically segment hippocampal MR images. This network incorporated a 3D ELSA Transformer module to enhance local detailed feature extraction and a spatial squeeze excitation module (sSE) to integrate spatial and channel information. Four deep learning models, namely, V-Net, U-Net, ResNet, VIT, and the SwinHS network developed in this study, were trained and tested on the same dataset. Performance was evaluated using DSC, JSC, and HD metrics, and the dosimetric parameters of plan (AD) and plan (MD) were compared. The results demonstrated that the proposed model outperformed the other four models across all indicators, achieving a contouring effect more consistent with manual hippocampal delineation. The PTV of both the AD and MD plans met the constraints outlined in the RTOG 0933 treatment plan. However, the D_{max} of the

TABLE 3 Dosimetric comparison between plan (AD) and plan (MD).

		Plan (AD) (±SD)	Plan (MD) (±SD)	*p*-value
PTV	$D_{2\%}$ (cGy)	3353.25 ± 29.31	3352.71 ± 28.66	0.111
	$D_{98\%}$ (cGy)	2795.22 ± 21.75	2793.43 ± 22.32	0.172
	V_{30Gy} (%)	94.30 ± 0.73	94.30 ± 0.74	0.545
Hippocampus	D_{max} (cGy)	1697.03 ± 11.02	1474.25 ± 35.51	0.000
	$D_{100\%}$ (cGy)	997.22 ± 5.88	994.54 ± 7.13	0.236

D_{max}, the maximum dose; $V_{30\,Gy}$, the volume of PTV getting 30 Gy; $D_{98\%}$, the dose received at 98% of PTV; $D_{2\%}$, the dose received at 2% of PTV; $D_{100\%}$, the dose to 100% of hippocampus.

hippocampus in the AD plans was significantly greater than that in the MD plans ($p < 0.001$), while the $D_{100\%}$ remained below 10 Gy, with no significant difference observed. These findings suggest that the automatic hippocampal segmentation method proposed in this study effectively extracts global features, accurately outlines hippocampal contours, and enhances hippocampal segmentation accuracy.

The RTOG-0933 protocol requires that the hippocampus be delineated on the patient's high-resolution 3DT1-weighted MR image before HA-WBRT and then registered with the positioning CT for planning design (Gondi et al., 2015). In practice, manual segmentation by clinicians is time-consuming and labor-intensive, often leading to large segmentation errors. According to the RTOG-0933 test, nearly 7% of the hippocampus delineations by a doctor were deemed unqualified (Gondi et al., 2015). Additionally, manual segmentation of organs is highly subjective, with significant variation among doctors (Zhang J. et al., 2021). To address these challenges, scholars have conducted extensive research on accurate automatic hippocampal segmentation. Feng et al. (2020) used NeuroQuant software approved by the U.S. Food and Drug Administration to perform hippocampal segmentation on T1 MR in patients undergoing whole-brain radiotherapy. Among 100 patients, 99 underwent acceptable automatic hippocampal segmentation without manual intervention, with all plans meeting the PTV dose-volume target set by the NRG CC001 protocol. However, the segmentation technology of NeuroQuant is based on atlas-based registration (Fischl et al., 2002). Although this method provides accurate results and reduces manual effort, it requires significant computation and depends heavily on the choice of atlas, resulting in unstable segmentation performance. Wang et al. (2022) found that the deep learning (DL) model demonstrated superior segmentation performance, especially for smaller OARs, by comparing the differences between the multiatlas segmentation method and the deep learning method in the automatic segmentation (OARs) scheme of nasopharyngeal carcinoma risk organs. In recent years, deep learning methods based on convolutional neural network CNNs have been widely used in the field of medical images (Li and Shen, 2022). Among them, 3D U-Net-based models are widely used in medical image segmentation tasks (Li and Shen, 2022) (deep learning-based methods have been proposed, in which 3D U-Net was employed because it is widely used in medical image segmentation tasks). In addition, scholars have carried out in-depth research on automatic segmentation of the hippocampus. Lin et al. (2023) developed an improved 3D U-Net segmentation model. For CT images of 10 patients in the independent test set, the overall average DSC and 95% HD of the hippocampal contour were greater than 0.8 mm and less than 7 mm, respectively. All the plans met the RTOG 0933 standard. Porter et al. (2020) proposed the attention-gated 3D ResNet (proposed Attention-Gated 3D ResNet) network model to study the segmentation of the hippocampus on patients' noncontrast CT, with Dice coefficients of 0.738/0.737 (left/right). However, these studies require strict registration of MR and CT images. The automatic segmentation tool for the hippocampus based on CT has made progress, but MRI is still the most reliable method for excluding the metastasis of the hippocampus. Hänsch et al. (2020) compared the automatic segmentation of the hippocampus based on a convolutional neural network (CNN) for MR and CT images and found that high-quality and anatomically accurate training contours can be generated on MR images and propagated to CT images to obtain optimal results. Therefore, Qiu et al. (2021) proposed a 3D U-Net model of multitask edge-aware learning for segmenting T1-weighted MR images of patients and obtained a Dice coefficient of 0.8483 ± 0.0036, an HD of 7.5706 ± 1.2330 mm, and an AVD of 0.1522 ± 0.0165 mm. In addition, Pan et al. (2021) proposed a CNN network structure based on 3D U-Net to segment the hippocampus on 3D-T1 MR images, with average DSC and AVD values of 0.86 and 1.8 mm, respectively. Encouraged and inspired by previous research, we propose a new automatic hippocampal segmentation model for 3DT1 MRI called SwinHS, which is based on the Swin Transformer. This model is designed to address the limitations of conventional CNN models and traditional Vision Transformers (VIT). Unlike these models, the Swin Transformer leverages a self-attention mechanism to capture long-range dependencies and context information across the entire input, significantly enhancing the model's ability to understand complex spatial relationships in the hippocampal region. This global attention mechanism enables the model to accurately capture the spatial positioning of the hippocampus in MR images. Additionally, the network incorporates an enhanced version of local self-attention (ELSA) instead of LSA. The introduction of the Hadamard product in ELSA facilitates more efficient attention generation while preserving high-order mapping relationships (Ghazouani et al., 2024), thereby enhancing the extraction of local detailed features. Finally, the feature representation extracted by the decoder is passed through a multiresolution skip connection to the sSE CNN decoder, resulting in the final output segmentation map.

In traditional Vision Transformer (ViT) models, the input tokens have a fixed size, and the model operates at a fixed sampling rate of 16, which is effective for image classification tasks. However, for dense prediction tasks on high-resolution images, the computational complexity scales quadratically with image size, leading to significant computational costs (Fang et al., 2023). To address this limitation, we introduced a hybrid model that combines the strengths of Transformer architectures, which excel at capturing long-range dependencies, with the hierarchical structure of convolutional neural networks (CNNs), thereby reducing computational complexity while retaining the model's ability to capture both global and local features.

When evaluating hippocampus segmentation performance, SwinHS demonstrated exceptional results across key performance metrics, including Dice similarity coefficient (DSC), Jaccard similarity coefficient (JSC), and Hausdorff distance (HD). Compared to other models, SwinHS achieved a high DSC of 0.894 and significantly reduced HD values. This improvement can be attributed to the Transformer architecture's ability to capture global information while preserving local detail, making it highly effective for segmenting small and intricate structures like the hippocampus.

In terms of data processing efficiency, the innovative design of the SwinHS architecture significantly accelerates processing speed. While manual segmentation typically requires an average of 786 s, SwinHS completes the same task in just 13.3 s, drastically reducing the workload for clinical practitioners. Compared to other models, SwinHS combines the efficiency of deep learning with the adaptability of Transformers, speeding up the computation process without compromising accuracy. In hippocampus segmentation tasks, this translates to faster and more precise outcomes.

In conclusion, the high accuracy and efficiency of the Swin Transformer model are expected to positively impact the development of hippocampus avoidance whole-brain radiotherapy treatment plans in clinical practice. Accurate hippocampus segmentation is also expected to assist in the early detection and monitoring of diseases related to hippocampus atrophy, such as Alzheimer's disease. Additionally, the model's rapid processing capabilities can shorten the time patients wait for diagnostic results, improving the overall responsiveness of healthcare services.

On the other hand, our proposed model operates as a supervised learning model, which requires sufficient data and precise manual contours as training labels. Hence, to optimize the automatic segmentation performance of the hippocampus in hippocampus-shielded whole-brain radiotherapy, we deliberately excluded data from healthy adults and individuals with mental disorders. Instead, we compiled training datasets from relevant patient cohorts, a strategy also supported by Lei et al. (2023). Experimental findings indicate that this approach effectively leverages hippocampus guidance information from MR images of patients undergoing whole-brain radiotherapy, leading to improved hippocampal segmentation accuracy compared to traditional deep learning methods.

We employed the dynamic IMRT technique to compare the dosimetric differences between plan (AD) and plan (MD) in order to evaluate the clinical feasibility of the SwinHS model for automatic hippocampal segmentation. According to previous studies, VMAT technology provides excellent treatment plan quality for hippocampus-protected whole-brain radiotherapy (Lin et al., 2023) and is superior to IMRT in terms of efficiency (Soydemir et al., 2021). In a study by Jiang et al. (2019), conducted by our team, all treatment plans, including static IMRT, dynamic IMRT, VMAT, and TomoTherapy, met the RTOG 0933 dose standards for hippocampus protection in patients with limited brain metastases undergoing hippocampus-sparing whole-brain radiotherapy. However, compared to VMAT and TOMO, the average maximum doses delivered to the hippocampus using sIMRT and dIMRT were significantly lower. Despite this, the differences in the mean hippocampal dose among the sIMRT, dIMRT, VMAT, and TOMO groups were not statistically significant.

Additionally, studies have shown that flattening filter free (FFF) beams not only provide higher dose rates and reduce field scatter and electron contamination but also minimize normal tissue exposure outside the target area (Hrbacek et al., 2011; Ghemiş and Marcu, 2021; Ji et al., 2022). To enhance treatment effectiveness and reduce hippocampal dose, we opted for 9-field noncoplanar FFF-dynamic IMRT. According to our experimental results, the automatically delineated hippocampus had a smaller volume than the manually delineated one, as shown in Table 1. In most cases, the contours of the automatically delineated hippocampus closely matched the manual delineations, as illustrated in Figure 1. Additionally, since the manually delineated hippocampus was used to evaluate plan (MD), while the automatically delineated hippocampus was used to generate plan (AD), there was a notable difference in the average maximum dose (D_{max}) to the hippocampus. Specifically, the average D_{max} in the manual plan (MD) was 1474.25 ± 35.51 cGy, while in the automatic plan (AD), it was 1697.03 ± 11.02 cGy. Despite this difference, both values remained within the permissible limits specified by RTOG 0933 for hippocampal doses. In terms of $D_{100\%}$, there was no statistically significant difference between the automatic and manual plans, with both remaining below 10 Gy. Additionally, no dosimetric differences were observed in the PTV between plan (AD) and plan (MD). This is consistent with our expectations, as the volume variation of the hippocampus is negligible compared to the overall PTV. According to the RTOG 0933 study, these findings are considered clinically acceptable.

Our model has some limitations and presents opportunities for future improvements. First, we focused solely on hippocampal segmentation, so future research should explore automatic segmentation of other normal tissues in MR images, such as the crystalline lens, eyeballs, and brainstem, to further enhance treatment efficiency. Second, in the RTOG 0933 trial, 15.85% of participants failed the centralized review due to fusion or hippocampal segmentation errors (Gondi et al., 2015). This highlights the need to explore accurate automatic MR and CT registration as a critical area for future development. Moreover, our model's training process was limited by the use of a relatively small dataset. Given the variability in hippocampal shapes among patients, future research should involve larger, multicenter datasets to improve the model's robustness and generalizability. While our model assists physicians in hippocampal segmentation, the importance of the hippocampus in whole-brain radiotherapy means it is not yet capable of fully automating the segmentation process. Post-segmentation review and calibration by clinicians remain essential.

5 Conclusion

In this paper, we propose a hippocampus segmentation method based on the Swin Transformer, which effectively captures global features and enhances segmentation accuracy. We believe this approach has the potential to significantly improve clinical treatment efficacy for patients undergoing whole-brain radiotherapy (WBRT), leading to better prognoses by reducing treatment-associated cognitive decline and improving overall outcomes.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Affiliated Hospital of Xuzhou Medical University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

LL: Data curation, Writing – review & editing, Resources, Supervision. ZnL: Formal analysis, Investigation, Writing – original draft. AJ: Data curation, Writing – original draft. GS: Formal analysis, Writing – original draft. ZgL: Software, Validation, Writing – original draft. XX: Visualization, Writing – review & editing. XD: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was supported by the Project of Xuzhou Medical University Affiliated Hospital (2022ZL19), and the Innovation and Entrepreneurship Project of Xuzhou Medical University Science and Technology Park (CXCYZX2022003).

Acknowledgments

We would like to extend our heartfelt gratitude to all the patients who participated in this study. Additionally, we sincerely appreciate the invaluable contributions of our research team, whose dedication and efforts were essential to the success of this research.

Conflict of interest

ZL was employed by HaiChuang Future Medical Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2024.1441791/full#supplementary-material

References

Azad, R., Jia, Y., Aghdam, E. K., Cohen-Adad, J., and Merhof, D. (2023) Enhancing medical image segmentation with TransCeption: a multi-scale feature fusion approach. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2301.10847. [Epub ahead of preprint]

Basak, H., Kundu, R., and Sarkar, R. J. P. R. (2022). MFSNet: a multi focus segmentation network for skin lesion segmentation 128, 108673. doi: 10.1016/j.patcog.2022.108673

Berghoff, A. S., and Preusser, M. (2018). New developments in brain metastases. *Ther. Adv. Neurol. Disord.* 11:1756286418785502. doi: 10.1177/1756286418785502

Brown, P. D., Gondi, V., Pugh, S., Tome, W. A., Wefel, J. S., Armstrong, T. S., et al. (2020). Hippocampal avoidance during whole-brain radiotherapy plus memantine for patients with brain metastases: phase III trial NRG oncology CC001. *J. Clin. Oncol.* 38, 1019–1029. doi: 10.1200/jco.19.02767

Chen, J., Frey, E. C., He, Y., Segars, W. P., Li, Y., and Du, Y. (2022). TransMorph: transformer for unsupervised medical image registration. *Med. Image Anal.* 82:102615. doi: 10.1016/j.media.2022.102615

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016. Springer International Publishing. 424–432.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16 × 16 words: transformers for image recognition at scale. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2010.11929. [Epub ahead of preprint].

Eelbode, T., Bertels, J., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., et al. (2020). Optimization for medical image segmentation: theory and practice when evaluating with dice score or Jaccard index. *IEEE Trans. Med. Imaging* 39, 3679–3690. doi: 10.1109/tmi.2020.3002417

Erickson, B. J. (2021). Basic artificial intelligence techniques: machine learning and deep learning. *Radiol. Clin. North Am.* 59, 933–940. doi: 10.1016/j.rcl.2021.06.004

Fang, Y., Wang, X., Wu, R., and Liu, W. (2023). What makes for hierarchical vision transformer? *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 12714–12720. doi: 10.1109/tpami.2023.3282019

Fei, X., Li, X., Shi, C., Ren, H., Mumtaz, I., Guo, J., et al. (2023). Dual-feature fusion attention network for small object segmentation. *Comput. Biol. Med.* 160:106985. doi: 10.1016/j.compbiomed.2023.106985

Feng, C. H., Cornell, M., Moore, K. L., Karunamuni, R., and Seibert, T. M. (2020). Automated contouring and planning pipeline for hippocampal-avoidant whole-brain radiotherapy. *Radiat. Oncol.* 15:251. doi: 10.1186/s13014-020-01689-y

Fike, J. R., Rosi, S., and Limoli, C. L. (2009). Neural precursor cells and central nervous system radiation sensitivity. *Semin. Radiat. Oncol.* 19, 122–132. doi: 10.1016/j.semradonc.2008.12.003

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355. doi: 10.1016/s0896-6273(02)00569-x

Ghazouani, F., Vera, P., and Ruan, S. (2024). Efficient brain tumor segmentation using Swin Transformer and enhanced local self-attention. *Int. J. Comput. Assist. Radiol. Surg.* 19, 273–281. doi: 10.1007/s11548-023-03024-8

Ghemiş, D. M., and Marcu, L. G. (2021). Progress and prospects of flattening filter free beam technology in radiosurgery and stereotactic body radiotherapy. *Crit. Rev. Oncol. Hematol.* 163:103396. doi: 10.1016/j.critrevonc.2021.103396

Gondi, V., Cui, Y., Mehta, M. P., Manfredi, D., Xiao, Y., Galvin, J. M., et al. (2015). Real-time pretreatment review limits unacceptable deviations on a cooperative group radiation therapy technique trial: quality assurance results of RTOG 0933. *Int. J. Radiat. Oncol. Biol. Phys.* 91, 564–570. doi: 10.1016/j.ijrobp.2014.10.054

Gondi, V., Pugh, S. L., Tome, W. A., Caine, C., Corn, B., Kanner, A., et al. (2014). Preservation of memory with conformal avoidance of the hippocampal neural stem-cell compartment during whole-brain radiotherapy for brain metastases (RTOG 0933): a phase II multi-institutional trial. *J. Clin. Oncol.* 32, 3810–3816. doi: 10.1200/jco.2014.57.2909

Gondi, V., Tolakanahalli, R., Mehta, M. P., Tewatia, D., Rowley, H., Kuo, J. S., et al. (2010). Hippocampal-sparing whole-brain radiotherapy: a "how-to" technique using helical tomotherapy and linear accelerator-based intensity-modulated radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.* 78, 1244–1252. doi: 10.1016/j.ijrobp.2010.01.039

Hänsch, A., Hendrik Moltz, J., Geisler, B., Engel, C., Klein, J., Genghi, A., et al. (2020). Hippocampus segmentation in CT using deep learning: impact of MR versus CT-based training contours. *J. Med. Imaging* 7:064001. doi: 10.1117/1.Jmi.7.6.064001

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

Hrbacek, J., Lang, S., and Klöck, S. (2011). Commissioning of photon beams of a flattening filter-free linear accelerator and the accuracy of beam modeling using an anisotropic analytical algorithm. *Int. J. Radiat. Oncol. Biol. Phys.* 80, 1228–1237. doi: 10.1016/j.ijrobp.2010.09.050

Hu, J., Shen, L., and Sun, G.. (2018). Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7132–7141.

Ji, T., Sun, L., Cai, F., and Li, G. (2022). Comparison between flattening filter-free (FFF) and flattened photon beam VMAT plans for the whole brain radiotherapy (WBRT) with hippocampus sparing. *Asia Pac. J. Clin. Oncol.* 18, e263–e267. doi: 10.1111/ajco.13624

Jiang, A., Sun, W., Zhao, F., Wu, Z., Shang, D., Yu, Q., et al. (2019). Dosimetric evaluation of four whole brain radiation therapy approaches with hippocampus and inner ear avoidance and simultaneous integrated boost for limited brain metastases. *Radiat. Oncol.* 14:46. doi: 10.1186/s13014-019-1255-7

Jiang, Y., Zhang, Y., Lin, X., Dong, J., Cheng, T., and Liang, J. (2022). SwinBTS: a method for 3D multimodal brain tumor segmentation using Swin Transformer. *Brain Sci.* 12. doi: 10.3390/brainsci12060797

Lei, Y., Ding, Y., Qiu, R. L. J., Wang, T., Roper, J., Fu, Y., et al. (2023). Hippocampus substructure segmentation using morphological vision transformer learning. *Phys. Med. Biol.* 68:235013. doi: 10.1088/1361-6560/ad0d45

Li, Q., and Shen, L. (2022). Neuron segmentation using 3D wavelet integrated encoder-decoder network. *Bioinformatics* 38, 809–817. doi: 10.1093/bioinformatics/btab716

Lin, C. Y., Chou, L. S., Wu, Y. H., Kuo, J. S., Mehta, M. P., Shiau, A. C., et al. (2023). Developing an AI-assisted planning pipeline for hippocampal avoidance whole brain radiotherapy. *Radiother. Oncol.* 181:109528. doi: 10.1016/j.radonc.2023.109528

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin Transformer: hierarchical vision transformer using shifted windows. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2103.14030. [Epub ahead of preprint].

Milletarì, F., Navab, N., and Ahmadi, S.-A. (2016). V-Net: fully convolutional neural networks for volumetric medical image segmentation. *arXiv*. Available at: https://doi.org/10.48550/arXiv.1606.04797. [Epub ahead of preprint]. 565–571.

Milletari, F., Navab, N., and Ahmadi, S. A.. (2016). V-Net: fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV)). 565–571.

Mukesh, M., Benson, R., Jena, R., Hoole, A., Roques, T., Scrase, C., et al. (2012). Interobserver variation in clinical target volume and organs at risk segmentation in post-parotidectomy radiotherapy: can segmentation protocols help? *Br. J. Radiol.* 85, e530–e536. doi: 10.1259/bjr/66693547

Pan, K., Zhao, L., Gu, S., Tang, Y., Wang, J., Yu, W., et al. (2021). Deep learning-based automatic delineation of the hippocampus by MRI: geometric and dosimetric evaluation. *Radiat. Oncol.* 16:12. doi: 10.1186/s13014-020-01724-y

Peters, S., Bexelius, C., Munk, V., and Leighl, N. (2016). The impact of brain metastasis on quality of life, resource utilization and survival in patients with non-small-cell lung cancer. *Cancer Treat. Rev.* 45, 139–162. doi: 10.1016/j.ctrv.2016.03.009

Porter, E., Fuentes, P., Siddiqui, Z., Thompson, A., Levitin, R., Solis, D., et al. (2020). Hippocampus segmentation on noncontrast CT using deep learning. *Med. Phys.* 47, 2950–2961. doi: 10.1002/mp.14098

Qiu, Q., Yang, Z., Wu, S., Qian, D., Wei, J., Gong, G., et al. (2021). Automatic segmentation of hippocampus in hippocampal sparing whole brain radiotherapy: a multitask edge-aware learning. *Med. Phys.* 48, 1771–1780. doi: 10.1002/mp.14760

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Springer International Publishing. 234–241.

Soydemir, G. P., Bilici, N., Tiken, E. E., Balkanay, A. Y., Sisman, A. F., and Karacetin, D. (2021). Hippocampal sparing for brain tumor radiotherapy: a retrospective study comparing intensity-modulated radiotherapy and volumetric-modulated arc therapy. *J. Cancer Res. Ther.* 17, 99–105. doi: 10.4103/jcrt.JCRT_32_19

Taha, A. A., and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* 15:29. doi: 10.1186/s12880-015-0068-x

Walker, G. V., Awan, M., Tao, R., Koay, E. J., Boehling, N. S., Grant, J. D., et al. (2014). Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiother. Oncol.* 112, 321–325. doi: 10.1016/j.radonc.2014.08.028

Wang, J., Chen, Z., Yang, C., Qu, B., Ma, L., Fan, W., et al. (2022). Evaluation exploration of atlas-based and deep learning-based automatic contouring for nasopharyngeal carcinoma. *Front. Oncol.* 12:833816. doi: 10.3389/fonc.2022.833816

Zhang, J., Gu, L., Han, G., and Liu, X. (2021). AttR2U-Net: a fully automated model for MRI nasopharyngeal carcinoma segmentation based on spatial attention and residual recurrent convolution. *Front. Oncol.* 11:816672. doi: 10.3389/fonc.2021.816672

Zhang, Y., Liu, H., and Hu, Q. J. A. (2021). TransFuse: fusing transformers and CNNs for medical image segmentation. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2102.08005. [Epub ahead of preprint].

Zhang, Y., Xie, F., and Chen, J. (2023). TFormer: a throughout fusion transformer for multi-modal skin lesion diagnosis. *Comput. Biol. Med.* 157:106712. doi: 10.1016/j.compbiomed.2023.106712

Zhou, J., Wang, P., Wang, F., Liu, Q., Li, H., and Jin, R. J. A. (2021). ELSA: enhanced local self-attention for vision transformer. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2112.12786. [Epub ahead of preprint]

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 11 November 2024
DOI 10.3389/fnins.2024.1502499

OPEN ACCESS

EDITED BY
Lu Tang,
Xuzhou Medical University, China

REVIEWED BY
Huibing Wang,
Dalian Maritime University, China
Zhenhua Wang,
Northwest A&F University, China

*CORRESPONDENCE
Daokuan Qu
qudaokuan_cumt@163.com

RECEIVED 27 September 2024
ACCEPTED 21 October 2024
PUBLISHED 11 November 2024

CITATION
Qu D and Ke Y (2024) Asymmetric Large Kernel Distillation Network for efficient single image super-resolution.
Front. Neurosci. 18:1502499.
doi: 10.3389/fnins.2024.1502499

Asymmetric Large Kernel Distillation Network for efficient single image super-resolution

Daokuan Qu[1,2]* and Yuyao Ke[3]

[1]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, Jiangsu, China, [2]School of Energy and Materials Engineering, Shandong Polytechnic College, Jining, Shandong, China, [3]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu, China

Recently, significant advancements have been made in the field of efficient single-image super-resolution, primarily driven by the innovative concept of information distillation. This method adeptly leverages multi-level features to facilitate high-resolution image reconstruction, allowing for enhanced detail and clarity. However, many existing approaches predominantly emphasize the enhancement of distilled features, often overlooking the critical aspect of improving the feature extraction capabilities of the distillation module itself. In this paper, we address this limitation by introducing an asymmetric large-kernel convolution design. By increasing the size of the convolution kernel, we expand the receptive field, which enables the model to more effectively capture long-range dependencies among image pixels. This enhancement significantly improves the model's perceptual ability, leading to more accurate reconstructions. To maintain a manageable level of model complexity, we adopt a lightweight architecture that employs asymmetric convolution techniques. Building on this foundation, we propose the Lightweight Asymmetric Large Kernel Distillation Network (ALKDNet). Comprehensive experiments conducted on five widely recognized benchmark datasets-Set5, Set14, BSD100, Urban100, and Manga109-indicate that ALKDNet not only preserves efficiency but also demonstrates performance enhancements relative to existing super-resolution methods. The average PSNR and SSIM values show improvements of 0.10 dB and 0.0013, respectively, thereby achieving state-of-the art performance.

KEYWORDS

single image super-resolution, efficient method, asymmetric large kernel convolution, information distillation, convolutional neural network

1 Introduction

Single image super-resolution (SISR) is a fundamental task in low-level computer vision, aimed at recovering fine details lost during image degradation and reconstructing a high-resolution (HR) image from a given low-resolution (LR) input. In recent years, the advancement of deep learning has led to numerous methods leveraging deep neural networks to address the challenges of image SR.

Dong et al. (2014) were the first to apply convolutional neural networks to image SR. Their method involved upsampling the low-resolution image to match the high-resolution size using bicubic interpolation, followed by the use of a Super-Resolution Convolutional Neural Network (SRCNN) to learn the mapping from the upsampled image to the high-resolution counterpart. Although SRCNN consisted of only three convolutional layers, it achieved remarkable performance. Kim et al. (2016a) introduced residual connections in their Very Deep Super-Resolution (VDSR) network, which enabled deeper networks (up to 20 layers) and significantly improved reconstruction performance. In response to

the limitations of residual networks for low-level vision tasks, Lim et al. (2017) proposed the Enhanced Deep Super-Resolution (EDSR) network, which utilized simplified residual blocks by removing redundant batch normalization layers. Their findings demonstrated that batch normalization was unnecessary for SR tasks, leading to fewer reconstruction artifacts and reducing the computational complexity of the model. Nevertheless, the reliance of these super-resolution methods on intricate deep convolutional neural networks poses significant challenges for practical deployment, particularly in resource-constrained settings such as real-time processing, mobile platforms, or embedded devices.

Various methods have been introduced to address lightweight SR task, including recurrent learning (Kim et al., 2016b), neural network pruning (Zhang et al., 2021a,b; Wang et al., 2023), knowledge distillation (Gao et al., 2018; He et al., 2020), neural architecture search (Chu et al., 2021), etc. Recently, information distillation (Hui et al., 2018) has emerged as a preferred strategy for designing lightweight networks for super-resolution. This technique involves stacking distillation blocks, which incorporate feature enhancement and compression units, to extract features at different depths for image reconstruction. IMDN (Hui et al., 2019) expands on the concept of information distillation by employing a distillation module and a fusion module within each Information Multi-Distillation Block (IMDB) to extract and integrate hierarchical features. Building on this foundation, RFDN (Liu et al., 2020) introduces a shallow residual block that enhances performance without increasing the number of parameters. BSRN (Li et al., 2022) employs Blueprint Separable Convolutions (BSConv) (Haase and Amthor, 2020) to optimize the Super Resolution Block (SRB) and integrates enhanced spatial attention for feature refinement, achieving state-of-the-art results. BSConv operates on the premise that a blueprint serves as a template for the convolutional weights, allowing all convolution kernels within a model to be derived through linear transformations of this blueprint. Specifically, BSConv first performs a weighted combination of depth features, followed by channel-wise convolutions to regulate the interdependencies within the learned convolution kernels. However, this regulation inadvertently limits the potential for further enhancement in feature extraction capacity.

To address this issue, we present an Asymmetric Large Kernel Distillation Network (ALKDNet), designed to enhance the quality of reconstructed images while maintaining efficient super-resolution performance. The proposed method incorporates large kernel convolutions to better extract and refine features. Increasing the kernel size effectively expands the receptive field, allowing the model to leverage more contextual information for improved task completion. However, directly enlarging the kernel size leads to a dramatic increase in parameters and computational cost. To mitigate this, we propose an asymmetric large kernel convolution, which replicates the effects of a large kernel by utilizing two asymmetric rectangular convolutions and a smaller square convolution. Additionally, we introduced an Anchor-Based Residual Learning (ABRL) (Du et al., 2021) method, built upon the conventional feature space residual learning, to further enhance the visual quality of the reconstructed images. This method establishes anchor points for each pixel in the high-resolution image using the corresponding low-resolution pixels, providing richer detail for image reconstruction.

Our contributions in this paper can be summarized as follows:

- We propose a novel Asymmetric Large Kernel Distillation Network (ALKDNet) aimed at addressing the challenge of lightweight super-resolution. Experiments on benchmark datasets demonstrate that the proposed ALKDNet achieves state-of-the-art performance.
- We design a novel Asymmetric Large Kernel Convolution (ALKConv), which enhances model performance while preserving computational efficiency and manageable complexity.
- We incorporate an anchor-based residual learning method into our ALKDNet alongside the conventional feature space residual learning, which results in improved performance compared to using either residual learning method in isolation.

The remainder of the paper is organized as follows: Section 2 shows an overview of the related work, Section 3 details the proposed model, Section 4 presents the empirical research results, and Section 5 shows the conclusion.

2 Related work

2.1 Efficient SR methods

As previously mentioned, Dong et al. (2014) were the first to apply CNNs to the SR problem, though their initial method was highly inefficient. In response, they introduced FSRCNN (Dong et al., 2016), which utilized a deconvolution layer as the upsampling module placed at the end of the network. This significantly accelerated the model and established a new paradigm for network design in SR tasks. Subsequently, ESPCN (Shi et al., 2016) proposed a sub-pixel convolutional upsampling method that delivered superior performance, making it the go-to upsampling strategy for SR tasks. Kim et al. (2016b) introduced recursive learning in DRCN, reducing the model size without sacrificing effectiveness. Subsequently, Tai et al. (2017) enhanced DRCN by proposing the Deep Recurrent Residual Network (DRRN), which achieved superior performance with fewer parameters while maintaining the same network depth. Building upon the Laplacian pyramid framework, Lai et al. (2017) developed a deep laplacian pyramid network (LapSRN), which leverages low-resolution feature maps at each pyramid layer to predict high-frequency details, achieving notable performance improvements. Ahn et al. (2018) advanced this by proposing CARN, which incorporated a cascading mechanism into the residual network. Hui et al. (2018) were the first to apply the information distillation mechanism for efficient SR in their IDN. Later, Hui et al. (2019) extended this concept with IMDN, introducing information multi–distillation, which considerably boosted model performance. RFDN (Liu et al., 2020) further lightened the model while improving its performance by designing shallow residual blocks and incorporating extensive feature distillation connections. Finally, BSRN (Li et al., 2022)

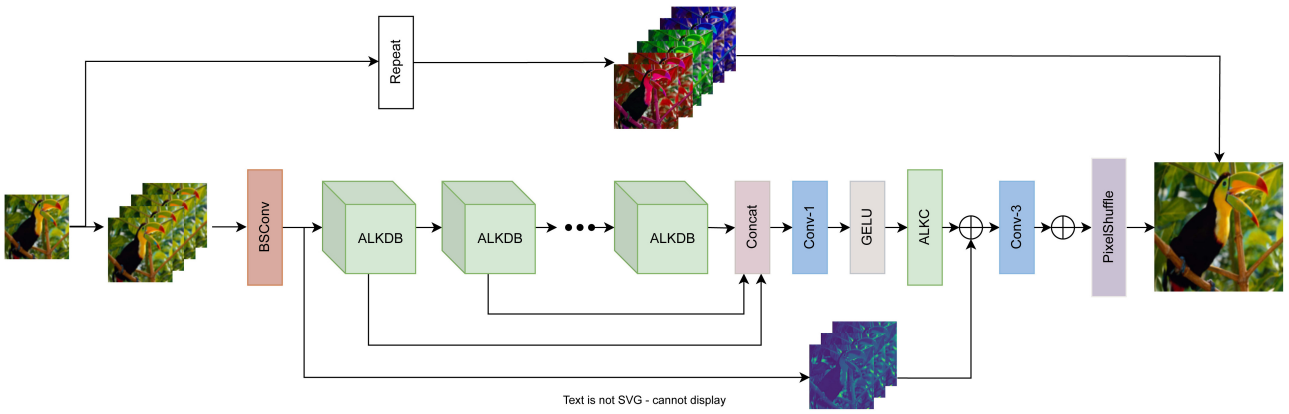


FIGURE 1
The architecture of ALKDNet. The ALKDNet replicates the LR image *m* times as the input, maps it to the feature space via BSConv, and progressively refines the features through a series of ALKDB blocks. After feature fusion stage, the fused features are sent to the image reconstruction module to obtain the reconstructed image.

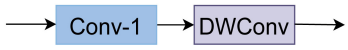


FIGURE 2
The architecture of BSConv. BSConv initially applies a 1 × 1 pointwise convolution, which is subsequently followed by independent depthwise convolutions executed for each individual channel.

achieved state-of-the-art results by replacing standard convolutions with blueprint separable convolutions and enhancing feature extraction through enhanced spatial attention, further reducing model complexity. Furthermore, Hui et al. (2020) integrated non-local operations into the residual block architecture, introducing a lightweight Feature Enhancement Residual Network (FERN). This design significantly strengthened the model's capacity to capture long-range dependencies. Moreover, Wang et al. (2021) developed a Sparse Masked Super-Resolution (SMSR) model that utilizes sparse masks. This method employs spatial masks to identify salient regions and channel masks to filter out unnecessary channels, thereby reducing redundant computations and enhancing super-resolution performance. Kong et al. (2022) streamlined the feature aggregation process by employing three convolutional layers for local feature learning, and introduced a Residual Local Feature Network (RLFN), achieving a balance between model performance and inference time. Additionally, Gendy et al. (2023) further advanced the SISR task by proposing a Mixer-based Local Residual Network (MLRN), which utilizes convolutional mixer blocks to blend channel and spatial features, achieving favorable performance.

2.2 Large kernel convolution

Since VGG (Simonyan and Zisserman, 2014) popularized the method of replacing large convolution kernels with stacked smaller convolutions, it has been widely adopted for its lightweight and efficient characteristics. With the advent of Transformer (Vaswani, 2017), many researchers sought to understand the source of their superior performance. Some attributed this to the extensive receptive field provided by the attention mechanism and aimed to enhance CNNs by expanding their receptive fields. According to the theory of effective receptive fields (ERF) (Luo et al., 2016), the ERF is proportional to $O(K\sqrt{L})$, where K represents the kernel size and L the network depth. This shows that increasing the kernel size is a more effective way to expand the ERF than merely stacking smaller convolutions. ConvNeXt (Liu Z. et al., 2022) expands the convolution kernel size to enhance the receptive field, ultimately achieving performance comparable to that of the Swin Transformer (Liu et al., 2021). RepLKNet (Ding et al., 2022) leveraged reparameterization technique and depth-wise convolution to scale the kernel size up to 31 × 31, achieving results that are comparable to, and in some cases surpass, those of the Swin Transformer across various tasks. Guo et al. (2023) integrated large kernel convolution with an attention mechanism, introducing a novel Large Kernel Attention (LKA) module in their VAN architecture, which demonstrated significant effectiveness across various tasks. LargeKernel3D (Chen et al., 2023) applied the concept of large kernel design to 3D networks, expanding the kernel size to 17 × 17 × 17. SLaK (Liu S. et al., 2022) simulated large kernel convolutions with two rectangular convolutions and integrated dynamic sparsity, pushing the kernel size to 51 × 51. Meanwhile, PeLK (Chen et al., 2024) further extended the kernel to 101 × 101 using a parameter-sharing mechanism and kernel-based position embedding, achieving impressive results across various computer vision tasks.

2.3 Asymmetric convolution

Szegedy et al. (2016) first introduced the concept of asymmetric convolution decomposition in Inception-v3, wherein the 7 × 7 convolution kernel is split into two smaller kernels of 7 × 1 and 1 × 7 to reduce the parameters for image recognition. This

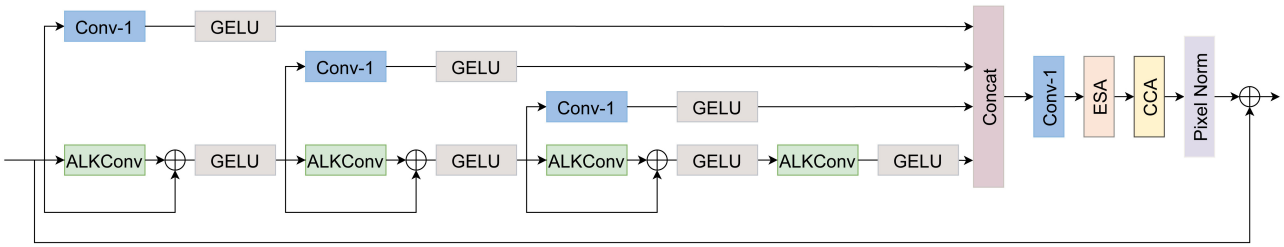


FIGURE 3
The architecture of ALKDB. ALKDB extracts and refines features using a sequence of ALKConvs, while simultaneously employing 1 × 1 convolutions for feature compression. Following the concatenation and fusion of the features, the ESA and CCA modules are utilized for further feature enhancement.

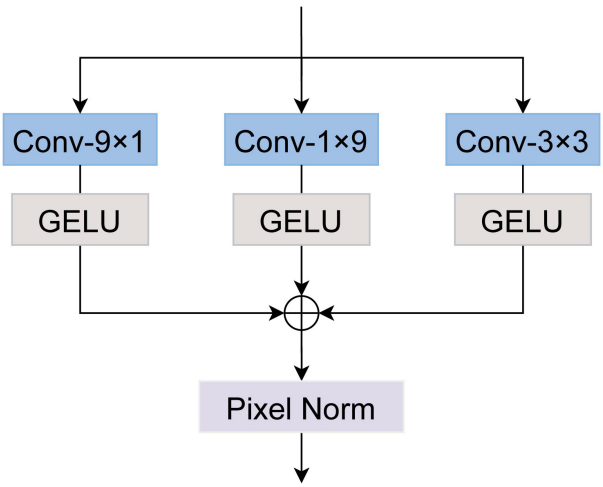


FIGURE 4
The architecture of ALKConv. We decompose a 9 × 9 convolution into 9 × 1 and 1 × 9 convolutions. In line with prior research, we also introduce a 3 × 3 convolution to operate in parallel with the large kernel convolution, subsequently summing their outputs.

technique was adopted in Global Convolutional Network (GCN) (Peng et al., 2017) to increase the kernel size to 15 × 15, enhancing performance in semantic segmentation tasks. However, it has been reported that this method may lead to a decrease in performance on ImageNet. EDANet (Lo et al., 2019) also employed this strategy by substituting 3 × 3 convolutions with 3 × 1 and 1 × 3 convolutions to reduce computational cost, albeit at the expense of performance. Nevertheless, it experienced a decline in performance when applied to semantic segmentation tasks. In contrast, Ding et al. (2019) utilized asymmetric convolution for structural reparameterization in ACNet, where asymmetric convolutions were employed to strengthen horizontal and vertical information, which was then aggregated on a square convolution kernel, leading to significant performance improvements. Furthermore, Tian et al. (2021) were the first to apply asymmetric convolution in the realm of image super-resolution, achieving notable results. Building on this foundation, SLaK (Liu S. et al., 2022) integrates convolution decomposition with dynamic sparsity, expanding the kernel size to 51 × 51 and thereby significantly improving model performance.

3 Proposed method

In this section, we firstly introduce the overall network architecture of ALKDNet and the loss function, then we give a detailed introduction to the designed asymmetric large kernel distillation block. Next, we introduce the proposed asymmetric large kernel convolution in detail.

3.1 Network architecture

The proposed method adopts the structural design of BSRN (Li et al., 2022), as illustrated in Figure 1. The complete model consists of four main components: a shallow feature extraction module, a deep feature extraction module, a deep feature fusion module, and a high-resolution image reconstruction module.

Initially, the input image I_{LR} is duplicated m times and concatenated along the channel dimension to form I_{LR}^{m}. This process is described as follows:

$$I_{LR}^{m} = Concat_{m}(I_{LR}), \tag{1}$$

where $Concat(\cdot)$ represents the concatenation operation along the channel dimension, where m indicates the number of times the input image I_{LR} is replicated and concatenated. Subsequently, higher-dimensional shallow features are extracted through the shallow feature extraction module:

$$F_0 = H_{SFE}(I_{LR}^{m}), \tag{2}$$

where $H_{SFE}(\cdot)$ represents the shallow feature extraction module, implemented as a 3 × 3 BSConv, with F_0 denoting the extracted shallow features. The structure of BSConv, illustrated in Figure 2, consists of both a channel convolution and a depthwise convolution. Following this, a series of asymmetric large kernel distillation blocks (ALKDB) are employed to progressively extract and refine deep features. This process can be expressed as follows:

$$F_k = H_k(F_{k-1}), k = 1, 2, ..., n, \tag{3}$$

where H_k represents the i-th ALKDB, while F_k and F_{k-1} refer to the output and input of the i-th ALKDB, respectively.

After the progressive extraction and refinement of ALKDBs, all intermediate features are concatenated via a 1 × 1 convolution,

TABLE 1 Ablation study on large kernel convolution.

Method	Params	Multi-adds	Set5		Set14		BSD100		Urban100		Manga109	
	(K)	(G)	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
BSRN	332	73.0	38.09	0.9609	33.74	0.9193	32.24	0.9007	32.36	0.9301	39.11	0.9780
ALKConv5 × 5	354	78.3	38.09	0.9607	**33.81**	0.9197	32.24	0.9005	32.44	0.9312	39.15	0.9781
ALKConv7 × 7	361	79.9	38.11	0.9608	33.77	0.9191	32.25	0.9008	32.41	0.9307	39.20	**0.9782**
ALKConv9 × 9	368	81.6	**38.13**	**0.9610**	33.78	0.9191	**32.27**	0.9009	**32.51**	**0.9318**	**39.21**	**0.9782**
ALKConv11 × 11	375	83.2	38.08	0.9609	33.80	**0.9198**	**32.27**	**0.9010**	32.50	0.9316	39.20	0.9781

The best and second-best results are **highlighted** and underlined, respectively.

TABLE 2 Ablation study on residual learning.

Method	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
FSRL	38.09	0.9609	33.74	0.9193	32.24	0.9007	32.36	0.9301	39.11	0.9780
ABRL	38.12	0.9609	33.76	0.9194	32.24	0.9006	32.45	0.9310	39.19	0.9782
FSRL+ABRL	38.09	0.9608	33.68	0.9190	32.24	0.9006	32.40	0.9308	39.14	0.9780
ALKConv+FSRL	38.13	**0.9610**	33.78	0.9191	32.27	0.9009	32.51	0.9318	39.21	0.9782
ALKConv+ABRL	**38.14**	0.9609	**33.81**	**0.9197**	**32.28**	**0.9010**	32.49	0.9316	39.17	0.9782
ALKConv+FSRL+ABRL	38.13	0.9609	33.76	0.9192	**32.28**	**0.9010**	**32.61**	**0.9327**	**39.26**	**0.9783**

The best and second-best results are **highlighted** and underlined, respectively.

followed by GELU activation for feature fusion and activation. Finally, asymmetric Large Kernel Convolution (ALKConv) is applied to smooth the features. This deep feature fusion process can be described as follows:

$$F_{fused} = H_{fusion}(Concat(F_1, ..., F_k)), \tag{4}$$

where F_{fused} represents the aggregated deep features, while H_{fusion} refers to the feature fusion module as described above.

In the final stage, the image reconstruction module of BSRN employs a long-range skip connection for residual learning. While maintaining this residual learning in the feature space, we introduce an anchor-based residual learning method. This method repeats the squared upscaling factor for each pixel in the LR space, using it as an anchor point for the corresponding pixel in the HR space. Subsequently, the pixel shuffle operation is applied to generate the reconstructed image. This process can be formulated as follows:

$$I_{SR} = H_{PS}(Conv_{up}(F_{fused} + f_0) + H_{repeat}(I_{LR})), \tag{5}$$

where $H_{PS}(\cdot)$ denotes the pixel shuffle operation, while $H_{repeat}(\cdot)$ refers to repeating the squared upscaling factor of the LR images, organizing them by color channels, and concatenating them along the channel dimension. The $Conv_{up}(\cdot)$ operation is a 3 × 3 convolution, used to expand the fused features learned through residual learning in the feature space, ensuring that their channels are aligned with the output of $H_{repeat}(\cdot)$.

Our model is optimized using the L1 loss function, which is formulated as:

$$L_1 = \|I_{SR} - I_{HR}\|_1. \tag{6}$$

3.2 Asymmetric large kernel distillation block

Drawing inspiration from the ESDB structure in BSRN (Li et al., 2022), we designed a asymmetric large kernel distillation block (ALKDB) with a similar architecture. The ALKDB is composed of three key components: feature distillation, feature condensation, and feature enhancement. The overall structure of ALKDB is illustrated in Figure 3. Given an input feature F_{in}, the feature distillation process in the initial stage can be formulated as follows:

$$\begin{aligned} F_{d1}, F_{r1} &= D_1(F_{in}), R_1(F_{in}), \\ F_{d2}, F_{r2} &= D_2(F_{r1}), R_2(F_{r1}), \\ F_{d3}, F_{r3} &= D_3(F_{r2}), R_3(F_{r2}), \\ F_{d4} &= D_4(F_{r3}), \end{aligned} \tag{7}$$

where D_i represents the i-th distillation layer, responsible for extracting the distilled feature F_{di}, while R_i denotes the i-th refinement layer, used to iteratively refine the feature F_{ri}. Specifically, the distillation layer is composed of a 1 × 1 convolution followed by GELU activation, while the refinement layer consists of a asymmetric large kernel convolution with skip connections, also followed by GELU activation. In the feature condensation stage, the four distilled features are concatenated along the channel dimension, followed by a 1 × 1 convolution for feature fusion. This process can be described as follows:

$$F_{condensed} = Conv\_1(Concat(F_{d1}, ..., F_{d4})), \tag{8}$$

where $F_{condensed}$ represents the condensed feature obtained from the fusion process. In the subsequent feature enhancement stage,

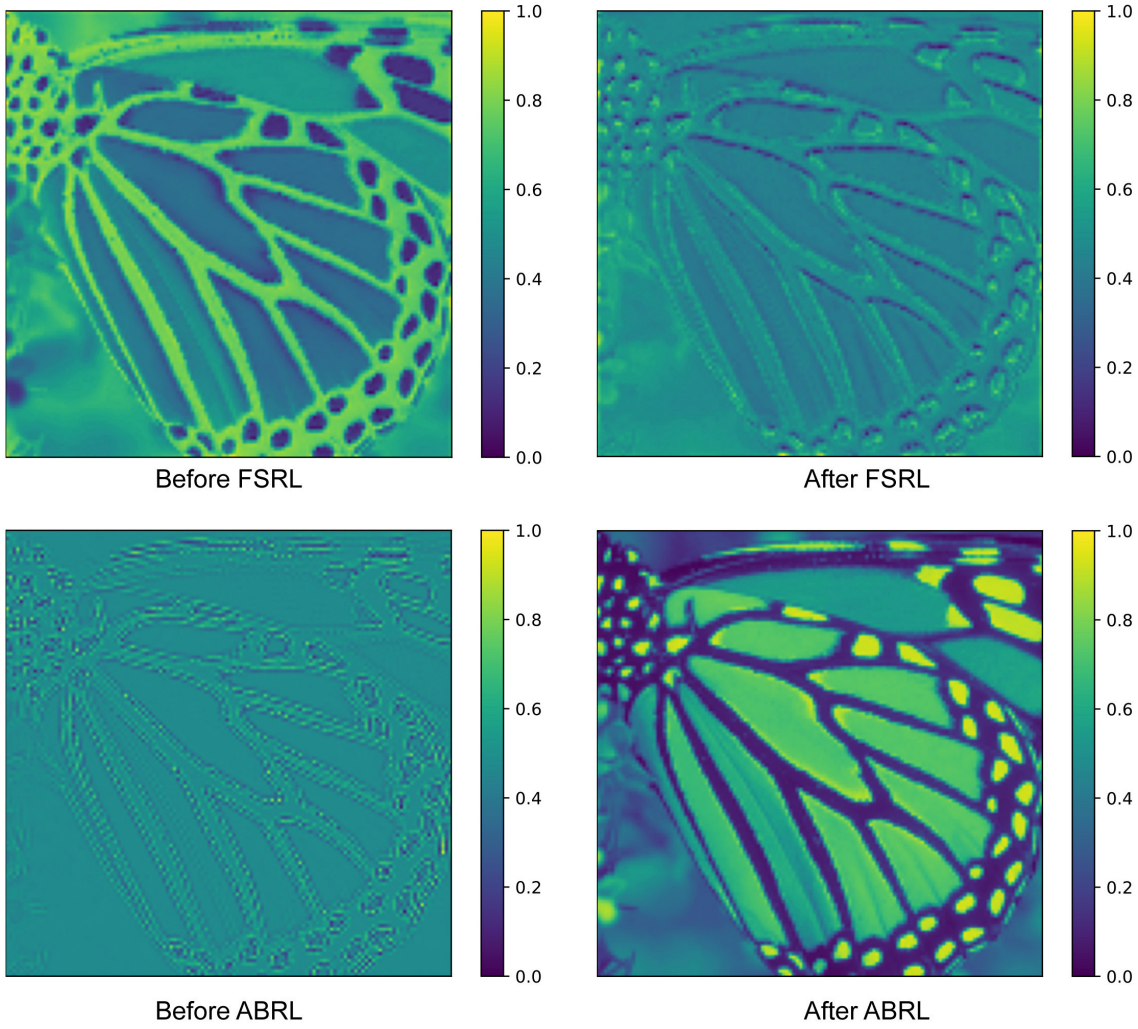


FIGURE 5
To further explore the impact of the two residual learning methods, we visualize the average feature maps obtained before and after applying FSRL and ABRL. The feature map following FSRL exhibits enhanced activation of high-frequency textures, while the feature map after ABRL contains richer detailed information.

we employ both a Enhanced Spatial Attention (ESA) block and a Contrast-aware Channel Attention (CCA) block to further enhance the features. Simultaneously, the pixel normalization module is applied to ensure stability during the model's training process:

$$F_{enhanced} = Norm_{pixel}(H_{CCA}(H_{ESA}(F_{condensed}))), \quad (9)$$

where $H_{CCA}(\cdot)$ and $H_{ESA}(\cdot)$ represent the CCA and ESA modules, respectively, while $Norm_{pixel}(\cdot)$ denotes the pixel-level normalization module. The output, $F_{enhanced}$, is the enhanced feature. Ultimately, the input features F_{in} are employed for long-range residual learning to derive the final output features F_{out}:

$$F_{out} = F_{enhanced} + F_{in}. \quad (10)$$

3.3 Asymmetric large kernel convolution

Liu S. et al. (2022) proposed the decomposition of a large 51 × 51 convolutional kernel into three smaller kernels of size 51 × 5, 5 × 51, and 5 × 5 in their SLaK model, enhancing performance while keeping computational complexity manageable. Drawing inspiration from this method, we adopt a similar strategy to construct a 9 × 9 large kernel convolution, as illustrated in Figure 4.

Specifically, for the input feature F_{in}, we apply three convolution operations with kernel sizes of 9 × 1, 1 × 9, and 3 × 3, respectively. Feature activation is performed using the GELU function. The resulting three feature maps are then summed together, followed by a pixel normalization operation to enhance the stability of the training process. This procedure can be formulated as follows:

$$F_{out} = Norm_{pixel}(H_{act}(Conv_{\_9\times1}(F_{in})) + H_{act}(Conv_{\_1\times9}(F_{in})) + H_{act}(Conv_{\_3}(F_{in}))), \quad (11)$$

where F_{out} represents the output feature after processing with the large kernel convolution, and H_{act} denotes the GELU activation function.

TABLE 3 Ablation study on pixel normalization.

Method	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
BSRN	38.09	0.9609	33.74	0.9193	32.24	0.9007	32.36	0.9301	39.11	0.9780
BSRN+norm	38.09	0.9608	33.69	0.9189	32.25	0.9006	32.42	0.9308	39.15	0.9780
BSRN+ABRL	38.09	0.9608	33.68	0.9190	32.24	0.9006	32.40	0.9308	39.14	0.9780
BSRN+ABRL+norm	38.06	0.9607	33.72	0.9190	32.25	0.9006	32.50	0.9316	39.16	0.9781
BSRN+ALKConv	38.13	**0.9610**	33.78	0.9191	32.27	0.9009	32.51	0.9318	39.21	0.9782
BSRN+ALKConv+norm	38.13	0.9609	**33.89**	**0.9198**	32.27	0.9010	32.45	0.9313	39.21	0.9782
BSRN+ALKConv+ABRL	38.13	0.9609	33.76	0.9192	32.28	0.9010	32.61	0.9327	39.26	**0.9783**
BSRN+ALKConv+ABRL+norm	**38.14**	0.9609	33.81	0.9193	**32.29**	**0.9011**	**32.71**	**0.9332**	**39.28**	**0.9783**

The best and second-best results are **highlighted** and underlined, respectively.

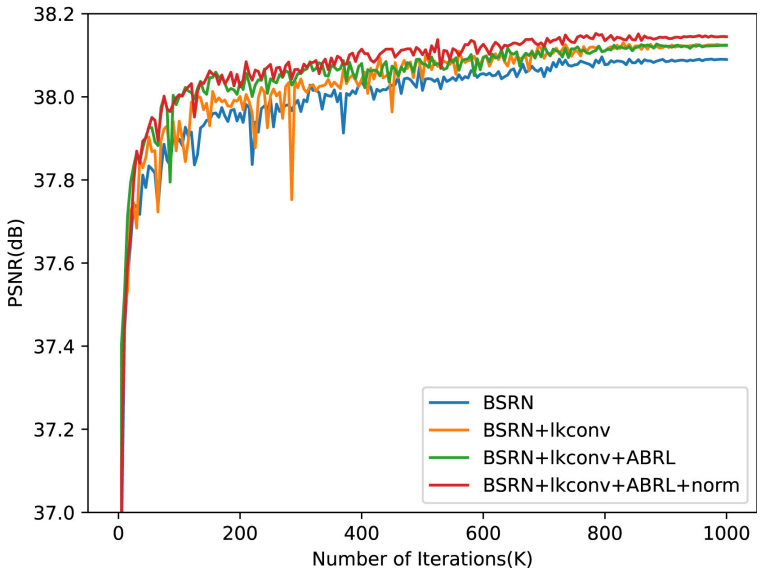


FIGURE 6
The PSNR test results on ×2 scale benchmark dataset Set5 during training. The proposed final model, as indicated by the red line, demonstrated superior performance and convergence speed, while also exhibiting the highest stability throughout the training process.

4 Experiments

In this section, the datasets, evaluation metrics and implementation details are firstly introduced in detail, and then a series of ablation experiments on ALKDNet are conducted to verify the efficiency. Next, we compare our ALKDNet with many other state-of-the art lightweight SR methods quantitatively and visually.

4.1 Datasets and evaluation metrics

We follow the method in previous work (Li et al., 2022) for model training and testing. DIV2K (Timofte et al., 2017) and Flickr2K (Lim et al., 2017) datasets were used for model training, and five benchmark datasets Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2012), BSD100 (Arbelaez et al., 2010), Urban100 (Huang et al., 2015) and Manga109 (Matsui et al., 2017) were used for testing. LR images were generated from HR images through bicubic degradation. The evaluation of super-resolution reconstruction results is to convert the image to YCbCr format, and only calculate the PSNR and SSIM (Wang et al., 2004) of the Y component. The Multi-Adds of the evaluation method is based on the acquisition of output image with a spatial resolution of 1280 × 720 pixels.

4.2 Implementation details

The proposed method consists of 8 blocks and the number of channels is set to 64. The size of all convolution kernels is set to 3 unless otherwise noted. Data augmentation was performed by random rotations of 90°, 180°, 270° and horizontal flipping. The minibatch size is set to 64 and the patch size of each LR input is

TABLE 4 Quantitative results of state-of-the-art lightweight SR methods on benchmark datasets.

Method	Scale	Params	Multi-adds	Set5	Set14	BSD100	Urban100	Manga109
		(K)	(G)	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Bicubic	×2	-	-	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403	30.80/0.9339
SRCNN	×2	8	52.7	36.66/0.9542	32.45/0.9067	31.36/0.8879	29.50/0.8946	35.60/0.9663
FSRCNN	×2	13	6.0	37.00/0.9558	32.63/0.9088	31.53/0.8920	29.88/0.9020	36.67/0.9710
VDSR	×2	666	612.6	37.53/0.9587	33.03/0.9124	31.90/0.8960	30.76/0.9140	37.22/0.9750
DRRN	×2	298	6796.9	37.74/0.9591	33.23/0.9136	32.05/0.8973	31.23/0.9188	37.88/0.9749
IDN	×2	553	124.6	37.83/0.9600	33.30/0.9148	32.08/0.8985	31.27/0.9196	38.01/0.9749
IMDN	×2	694	158.8	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
RFDN	×2	534	95.0	38.05/0.9606	33.68/0.9184	32.16/0.8994	32.12/0.9278	38.88/0.9773
FMEN	×2	748	172.0	38.10/0.9609	33.75/0.9192	32.26/0.9007	32.41/0.9311	38.95/0.9778
BSRN	×2	332	73.0	38.10/**0.9610**	33.74/**0.9193**	32.24/0.9006	32.34/0.9303	39.14/0.9782
SAFMN	×2	228	52.0	38.00/0.9605	33.54/0.9177	32.16/0.8995	31.84/0.9256	38.71/0.9771
MLRN	×2	488	90.4	38.07/0.9607	33.59/0.9180	32.21/0.9000	32.28/0.9297	38.76/0.9773
HSNet	×2	302	81	38.07/0.9607	33.65/0.9185	33.22/0.9002	32.27/0.9295	39.00/0.9778
CFSR	×2	291	62.6	38.07/0.9607	33.74/0.9192	32.24/0.9005	32.28/0.9300	39.00/0.9778
ALKDNet(Ours)	×2	373	83.7	**38.14**/0.9609	**33.81/0.9193**	**32.29/0.9011**	**32.71/0.9332**	**39.28/0.9783**
Bicubic	×3	-	-	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349	26.95/0.8556
SRCNN	×3	8	52.7	32.75/0.9090	29.30/0.8215	28.41/0.7863	26.24/0.7989	30.48/0.9117
FSRCNN	×3	13	5.0	33.18/0.9140	29.37/0.8240	28.53/0.7910	26.43/0.8080	31.10/0.9210
VDSR	×3	666	612.6	33.66/0.9213	29.77/0.8314	28.82/0.7976	27.14/0.8279	32.01/0.9340
DRRN	×3	298	6796.9	34.03/0.9244	29.96/0.8349	28.95/0.8004	27.53/0.8378	32.71/0.9379
IDN	×3	553	56.3	34.11/0.9253	29.99/0.8354	28.95/0.8013	27.42/0.8359	32.71/0.9381
IMDN	×3	703	71.5	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
RFDN	×3	541	42.2	34.41/0.9273	30.34/0.8420	29.09/0.8042	28.21/0.8525	33.67/0.9449
FMEN	×3	757	77.2	34.45/0.9275	30.40/0.8435	29.17 0.8063	28.33/0.8562	33.86/0.9462
BSRN	×3	340	33.3	34.46/0.9277	30.47/0.8449	29.18/0.8068	28.39/0.8567	34.05/0.9471
SAFMN	×3	233	23.0	34.34/0.9267	30.33/0.8418	29.08/0.8048	27.95/0.8474	33.52/0.9437
MLRN	×3	496	40.9	34.46/0.9267	30.35/0.8426	29.10/0.8054	28.20/0.8533	33.66/0.9450
HSNet	×3	302	36	34.49/0.9278	30.44/0.8434	29.15/0.8063	28.36/0.8555	33.95/0.9466
CFSR	×3	298	28.5	34.50/0.9279	30.44/0.8437	29.16/0.8066	28.29/0.8553	33.85/0.9462
ALKDNet(Ours)	×3	381	37.3	**34.56/0.9284**	**30.50/0.8457**	**29.22/0.8079**	**28.58/0.8608**	**34.18/0.9478**
Bicubic	×4	-	-	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN	×4	8	52.7	30.48/0.8626	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
FSRCNN	×4	13	4.6	30.72/0.8660	27.61/0.7550	26.98/0.7150	24.62/0.7280	27.90/0.8610
VDSR	×4	666	612.6	31.35/0.8838	28.01/0.7674	27.29/0.7251	25.18/0.7524	28.83/0.8870
DRRN	×4	298	6796.9	31.68/0.8888	28.21/0.7720	27.38/0.7284	25.44/0.7638	29.45/0.8946
IDN	×4	553	32.3	31.82/0.8903	28.25/0.7730	27.41/0.7297	25.41/0.7632	29.41/0.8942
IMDN	×4	715	40.9	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
RFDN	×4	550	23.9	32.24/0.8952	28.61/0.7819	27.57/0.7360	26.11/0.7858	30.58/0.9089
FMEN	×4	769	44.2	32.24/0.8955	28.70/0.7839	27.63/0.7379	26.28/0.7908	30.70/0.9107
BSRN	×4	352	19.4	32.35/0.8966	28.73/0.7847	27.65/0.7387	26.27/0.7908	30.84/0.9123

(Continued)

TABLE 4 (Continued)

Method	Scale	Params (K)	Multi-adds (G)	Set5 PSNR/SSIM	Set14 PSNR/SSIM	BSD100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
SAFMN	×4	240	14.0	32.18/0.8948	28.60/0.7813	27.58/0.7359	25.97/0.7809	30.43/0.9063
MLRN	×4	507	23.5	32.30/0.8956	28.62/0.7824	27.57/0.7365	26.10/0.7867	30.56/0.9092
HSNet	×4	313	30	32.32/0.8970	28.65/0.7838	27.63/0.7393	26.29/0.7918	30.72/0.9124
CFSR	×4	307	17.5	32.33/0.8964	28.73/0.7842	27.63/0.7381	26.21/0.7897	30.72/0.9111
ALKDNet(Ours)	×4	393	21.6	**32.37/0.8976**	**28.80/0.7860**	**27.69/0.7399**	**26.46/0.7970**	**30.97/0.9137**

The best and second-best results are **highlighted** and underlined, respectively.

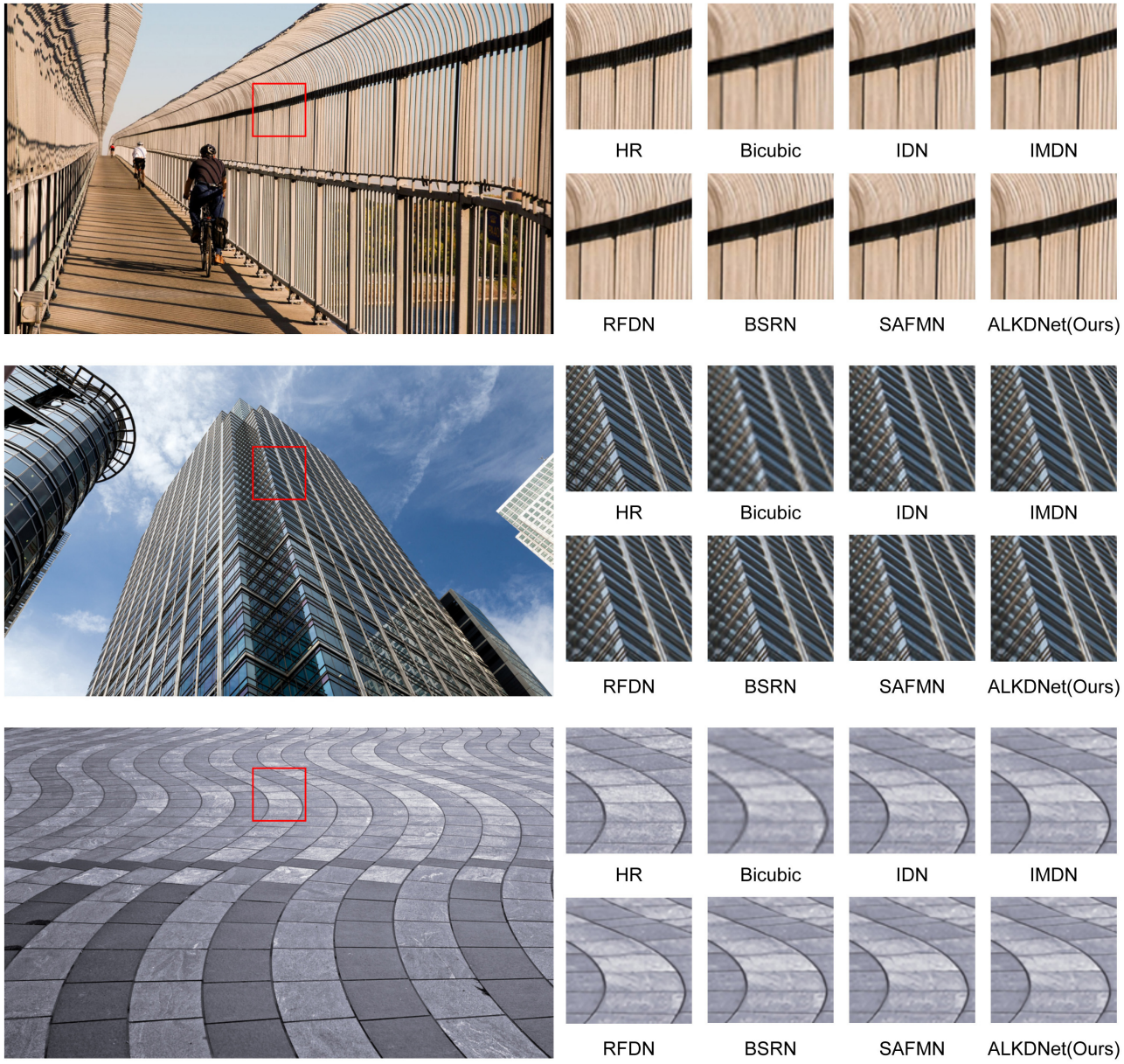


FIGURE 7
Qualitative comparison of ours model with the state-of-the-art methods for ×2 SR. We selected three images from the Urban100 dataset to evaluate and compare the visual effects of the reconstructed images. The images, presented in order from top to bottom, are img024, img047 and img071.

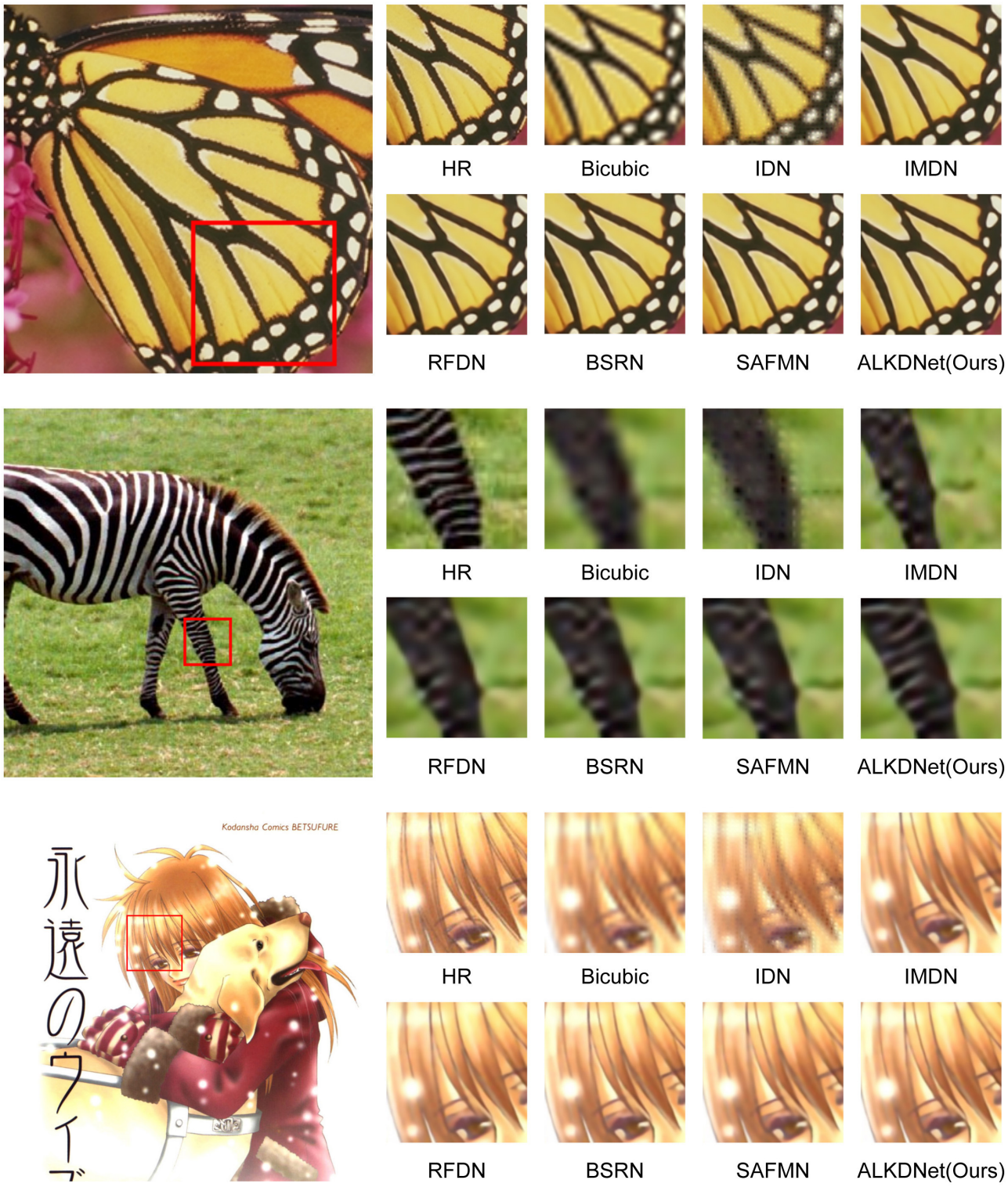


FIGURE 8
Qualitative comparison of ours model with the state-of-the-art methods for ×4 SR. We selected three images from the Set5 dataset, Set14 dataset and Manga109 dataset to evaluate and compare the visual effects of the reconstructed images. The images, presented in order from top to bottom, are butterfly in Set5, zebra in Set14 and EienNoWith in Manga109.

set to 48 × 48. We trained our model using the Adam optimizer (Kingma, 2014) with the initial learning rate set to 1×10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and adjusted the learning rate using cosine learning rate decay.L_1 loss is used to optimize the model for total 1×10^6 iterations. We use Pytorch 2.2.0 to implement our model on a single GeForce RTX 3090 GPU.

4.3 Ablation study

In this section, we demonstrate the effectiveness of the proposed method. All experiments presented here are conducted at the ×2 scaling factor.

4.3.1 Impact of asymmetric large kernel convolution

We conduct ablation experiments to verify the effectiveness of the proposed large kernel convolution. We simply replaced the BSConvs used in ESDB of BSRN with the ALKConvs we designed, and explored the impact of the size of the convolution kernel on the performance. The results are shown in the Table 1. It can be found that the performance of the model has been improved when the convolution kernel size is only 5, and the comprehensive performance on each benchmark dataset has reached the best when the convolution kernel size is 9. Specifically, when the convolution kernel size is expanded to 5, the model demonstrates improved performance on all benchmark datasets except Set5 and BSD100, with an average PSNR increase of 0.04 dB and an average SSIM increase of 0.0002. Expanding the kernel size further to 9 results in an average PSNR improvement of 0.07 dB and an SSIM increase of 0.0004. We speculated that continuing to expand the convolution kernel would help further improve the performance of the model, but we decided to set the size of the convolution kernel to 9 as a trade-off between model performance and efficiency.

4.3.2 Impact of residual learning method

In this section, we explored the impact of two residual learning methods on model performance, and the results are presented in the Table 2. Among them, FSRL is the original BSRN, ABRL is to replace the FSRL method in BSRN with ABRL, FSRL+ABRL is to add the ABRL method on the basis of the original BSRN, and with lkconv means that we replace the BSConvs in the ESDB of BSRN with our ALKConvs. It can be seen from the data in the table that the model performance has been improved after replacing the FSRL method with ABRL, but the performance decreases after applying the two residual learning methods on BSRN at the same time. However, it is interesting to see that the performance of the model is significantly improved after using large kernel convolution and two kinds of residual learning at the same time. Except for the slightly worse performance on Set5 and Set14, the best results are obtained on the other Benchmark datasets. Specifically, replacing the FSRL method with ABRL leads to an average improvement of 0.04 dB in PSNR and 0.0002 in SSIM. The highest performance is obtained when ALKConv is combined with both residual learning methods, resulting in an average gain of 0.10 dB in PSNR and 0.0006 in SSIM. On the Urban100 dataset, this method achieves a significant increase of 0.25 dB in PSNR and 0.0026 in SSIM.

We visualized the average feature maps before and after residual learning in Figure 5 to demonstrate the impact of residual learning. As observed, the high-frequency texture details in the feature map are effectively activated after applying FSRL. This can be attributed to FSRL's utilization of shallow features extracted by the convolutional layer for feature fusion. The convolutional layer possesses a strong capability to capture local high-frequency features, which contributes to this activation. Furthermore, after applying ABRL, the feature map exhibits a significant enhancement in image detail richness. This is primarily due to ABRL's direct utilization of information from the low-resolution image, allowing it to effectively enrich the detail representation.

4.3.3 Impact of pixel normalization

In this section, we evaluate the effect of pixel normalization on model performance, as shown in Table 3. The term +norm indicates the application of pixel normalization at the end of the original ESDB. The addition of pixel normalization results in minimal impact on overall model performance, with only slight improvements observed on certain benchmarks. Specifically, incorporating the pixel normalization layer yields the greatest performance improvement on the Urban100 dataset, with an average increase of 0.05 dB in PSNR and 0.0004 in SSIM.

Figure 6 presents the PSNR test results during training after integrating our proposed method. The inclusion of ALKConv leads to a notable improvement in model performance, though the PSNR exhibits significant fluctuations in the early stages, suggesting instability in the training process. When ABRL is further incorporated, while the performance gain is modest, the convergence speed is notably accelerated in the initial training phase, and the overall training process becomes more stable. Finally, with the addition of pixel normalization, model performance continues to improve, and PSNR fluctuations are further reduced, indicating enhanced training stability.

4.4 Comparison with the state-of-the-art methods

In this section, we contrast our model with 13 other state of the art methods in lightweight SR, including SRCNN (Dong et al., 2014), FSRCNN (Dong et al., 2016), VDSR (Kim et al., 2016a), DRRN (Kim et al., 2016b), IDN (Hui et al., 2018), IMDN (Hui et al., 2019), RFDN (Liu et al., 2020), FMEN (Du et al., 2022), BSRN (Li et al., 2022), SAFMN (Sun et al., 2023), MLRN (Gendy et al., 2023), HSNet (Cui et al., 2024), and CFSR (Wu et al., 2024). Table 4 shows quantitative comparisons for ×2, ×3, and ×4 SR. It is easy to find that our model performs slightly worse on set5 of ×2 and the SSIM result is 0.0001 lower than that of BSRN, and the other test results are better than the compared advanced methods.

Specifically, the performance of our model is improved compared with the suboptimal method at all three scales, for the ×2 scale, our model achieves an average improvement of 0.11 dB in PSNR and 0.0005 in SSIM. At the ×3 scale, the PSNR shows an average increase of 0.09 dB, while the SSIM improves by 0.0014. For the ×4 scale, the model delivers an average gain of 0.09 dB in PSNR and 0.0019 in SSIM. Among them, the gain of our model is the most obvious on Urban100, and the performance increases at ×2, ×3, and ×4 scales are 0.30dB/0.0021, 0.19dB/0.0041, and 0.17dB/0.0052, respectively.

To demonstrate the visual effects of our model's reconstructed images, we use six images from the benchmark dataset to conduct

a qualitative evaluation of the model. Figures 7, 8 displays the reconstruction results of our model compared to other state-of-the-art methods. It can be seen that our reconstruction results are still better even in the state-of-the-art methods. For example, in the image captured from img024, the images obtained by other methods have obvious artifacts at the top left continuous curved to the left texture, and the images obtained by other methods are very blurred at the bottom middle continuous vertical texture. In contrast, the image reconstructed by the proposed method is free from prominent artifacts and demonstrates the highest clarity, closely resembling the HR reference in terms of visual quality. Furthermore, within the zebra from the Set14 dataset, our method was the only one to reconstruct the high-resolution image without introducing any erroneous textures.

5 Conclusion

In this paper, we introduced the Asymmetric Large Kernel Distillation Network (ALKDNet), designed for lightweight super-resolution based on the BSRN architecture. The proposed method combines Asymmetric Large Kernel Convolution (ALKConv) in the distillation block, effectively balancing efficiency and performance to enhance model capability while maintaining acceptable complexity. Additionally, we introduced an anchor-point-based residual learning method in the image reconstruction module, which establishes anchor points for each corresponding pixel in the HR image using pixels from the LR image, thereby improving the quality of the reconstruction output. Results from five widely used benchmark datasets demonstrate that the proposed method achieves state-of-the-art performance.

Despite the contributions of our research, certain limitations remain. The low-resolution images used in the paper's experiments were generated through bicubic downsampling. However, in real-world scenarios, low-resolution images may be affected by various complex factors, such as limitations of acquisition devices, noise interference, and data compression. Therefore, further research is needed to effectively apply the proposed method in practical environments.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

DQ: Conceptualization, Data curation, Methodology, Project administration, Resources, Validation, Writing – original draft. YK: Data curation, Formal analysis, Methodology, Resources, Software, Visualization, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahn, N., Kang, B., and Sohn, K. A. (2018). "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Berlin: Springer Verlag), 252–268. doi: 10.1007/978-3-030-01249-6_16

Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J. (2010). Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Analy. Machine Intellig.* 33, 898–916. doi: 10.1109/TPAMI.2010.161

Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L. (2012). *Low-Complexity Single-Image Super-Resolution Based on Nonnegative Neighbor Embedding.* Durham: BMVA Press.

Chen, H., Chu, X., Ren, Y., Zhao, X., and Huang, K. (2024). "PeLK: Parameter-efficient large kernel convnets with peripheral convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ: IEEE), 5557–5567.

Chen, Y., Liu, J., Zhang, X., Qi, X., and Jia, J. (2023). "LargeKernel3D: Scaling up kernels in 3D sparse CNNS," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 13488–13498.

Chu, X., Zhang, B., Ma, H., Xu, R., and Li, Q. (2021). "Fast, accurate and lightweight super-resolution with neural architecture search," in *2020 25th International Conference On Pattern Recognition (ICPR)* (Milan: *IEEE)* 59–64.

Cui, Z., Yao, Y., Li, S., Zhao, Y., and Xin, M. (2024). A lightweight hash-directed global perception and self-calibrated multiscale fusion network for image super-resolution. *Image Vision Comp.* 151:105255. doi: 10.1016/j.imavis.2024.105255

Ding, X., Guo, Y., Ding, G., and Han, J. (2019). "ACNet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul: IEEE), 1911–1920. doi: 10.1109/ICCV.2019.00200

Ding, X., Zhang, X., Han, J., and Ding, G. (2022). "Scaling up your kernels to 31x31: revisiting large kernel design in CNNS," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 11963–11975.

Dong, C., Loy, C. C., He, K., and Tang, X. (2014). "Learning a deep convolutional network for image super-resolution," in *Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13* (Zurich: Springer), 184–199. doi: 10.1007/978-3-319-10593-2_13

Dong, C., Loy, C. C., and Tang, X. (2016). "Accelerating the super-resolution convolutional neural network," in *Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14* (Amsterdam: Springer), 391–407. doi: 10.1007/978-3-319-46475-6_25

Du, Z., Liu, D., Liu, J., Tang, J., Wu, G., and Fu, L. (2022). "Fast and memory-efficient network towards efficient image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 853–862. doi: 10.1109/CVPRW56347.2022.00101

Du, Z., Liu, J., Tang, J., and Wu, G. (2021). "Anchor-based plain net for mobile image super-resolution," in *Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 2494–2502. doi: 10.1109/CVPRW53098.2021.00283

Gao, Q., Zhao, Y., Li, G., and Tong, T. (2018). "Image super-resolution using knowledge distillation," in *Asian Conference on Computer Vision* (Cham: Springer), 527–541.

Gendy, G., Sabor, N., Hou, J., and He, G. (2023). "Mixer-based local residual network for lightweight image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 1593–1602.

Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M., and Hu, S.-M. (2023). Visual attention network. *Comp. Visual Media* 9, 733–752. doi: 10.1007/s41095-023-0364-2

Haase, D., and Amthor, M. (2020). "Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved MobileNets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 14600–14609.

He, Z., Dai, T., Lu, J., Jiang, Y., and Xia, S.-T. (2020). "Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution," in *2020 IEEE International Conference on Image Processing (ICIP)* (Abu Dhabi: IEEE), 518–522. doi: 10.1109/ICIP40778.2020.9190917

Huang, J.-B., Singh, A., and Ahuja, N. (2015). "*Single image super-resolution from transformed self-exemplars*," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Boston, MA: IEEE), 5197–5206. doi: 10.1109/CVPR.2015.7299156

Hui, Z., Gao, X., and Wang, X. (2020). Lightweight image super-resolution with feature enhancement residual network. *Neurocomputing* 404, 50–60. doi: 10.1016/j.neucom.2020.05.008

Hui, Z., Gao, X., Yang, Y., and Wang, X. (2019). "Lightweight image super-resolution with information multi-distillation network," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2024–2032. doi: 10.1145/3343031.3351084

Hui, Z., Wang, X., and Gao, X. (2018). "Fast and accurate single image super-resolution via information distillation network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 723–731.

Kim, J., Lee, J. K., and Lee, K. M. (2016a). "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 1646–1654.

Kim, J., Lee, J. K., and Lee, K. M. (2016b). "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 1637–1645. doi: 10.1109/CVPR.2016.181

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv* [Preprint]. arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980

Kong, F., Li, M., Liu, S., Liu, D., He, J., Bai, Y., et al. (2022). "Residual local feature network for efficient super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 766–776.

Lai, W.-S., Huang, J.-B., Ahuja, N., and Yang, M.-H. (2017). "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 624–632.

Li, Z., Liu, Y., Chen, X., Cai, H., Gu, J., Qiao, Y., et al. (2022). "Blueprint separable residual network for efficient image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 833–843. doi: 10.1109/CVPRW56347.2022.00099

Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. (2017). "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Honolulu, HI: IEEE), 136–144.

Liu, J., Tang, J., and Wu, G. (2020). "Residual feature distillation network for lightweight image super-resolution," in *Computer vision-ECCV 2020 workshops: Glasgow, UK, August 23-28, 2020, proceedings, part III 16* (Cham: Springer), 41–55.

Liu, S., Chen, T., Chen, X., Chen, X., Xiao, Q., Wu, B., et al. (2022). More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv* [Preprint]. arXiv:2207.03620. doi: 10.48550/arXiv.2207.03620

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC: IEEE), 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 11976–11986.

Lo, S.-Y., Hang, H.-M., Chan, S.-W., and Lin, J.-J. (2019). "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proceedings of the 1st ACM International Conference on Multimedia in Asia* (New York, NY: ACM), 1–6.

Luo, W., Li, Y., Urtasun, R., and Zemel, R. (2016). "Understanding the effective receptive field in deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (Cambridge, MA: MIT Press), 29.

Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., et al. (2017). Sketch-based manga retrieval using manga109 dataset. *Multimed. Tools Appl.* 76, 21811–21838. doi: 10.1007/s11042-016-4020-z

Peng, C., Zhang, X., Yu, G., Luo, G., and Sun, J. (2017). "Large kernel matters-improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 4353–4361.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., et al. (2016). "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 1874–1883. doi: 10.1109/CVPR.2016.207

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* [Preprint]. arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556

Sun, L., Dong, J., Tang, J., and Pan, J. (2023). "Spatially-adaptive feature modulation for efficient image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Paris: IEEE), 13190–13199.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 2818–2826.

Tai, Y., Yang, J., and Liu, X. (2017). "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 3147–3155. doi: 10.1109/CVPR.2017.298

Tian, C., Xu, Y., Zuo, W., Lin, C.-W., and Zhang, D. (2021). Asymmetric cnn for image superresolution. *IEEE Trans. Syst. Man, Cybernet.: Syst.* 52, 3718–3730. doi: 10.1109/TSMC.2021.3069265

Timofte, R., Agustsson, E., Van Gool, L., Yang, M.-H., and Zhang, L. (2017). "NTIRE 2017 challenge on single image super-resolution: methods and results," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Honolulu, HI: IEEE), 114–125. doi: 10.1109/CVPRW.2017.149

Vaswani, A. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Cambridge, MA: MIT Press).

Wang, J., Wang, H., Zhang, Y., Fu, Y., and Tao, Z. (2023). "Iterative soft shrinkage learning for efficient image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Paris: IEEE), 12590–12599.

Wang, L., Dong, X., Wang, Y., Ying, X., Lin, Z., An, W., et al. (2021). "Exploring sparsity in image super-resolution for efficient inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE), 4917–4926.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Proc.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Wu, G., Jiang, J., Jiang, J., and Liu, X. (2024). "Transforming image super-resolution: a convformer-based efficient approach," in *arXiv* [Preprint]. arXiv:2401.05633. doi: 10.1109/TIP.2024.3477350

Zeyde, R., Elad, M., and Protter, M. (2012). "On single image scale-up using sparse-representations," in *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7* (Cham: Springer), 711–730.

Zhang, Y., Wang, H., Qin, C., and Fu, Y. (2021a). Aligned structured sparsity learning for efficient image super-resolution. *Adv. Neural Inform. Proc. Syst.* 34, 2695–2706. doi: 10.5555/3540261.3540467

Zhang, Y., Wang, H., Qin, C., and Fu, Y. (2021b). "Learning efficient image super-resolution networks via structure-regularized pruning," in *International Conference on Learning Representations* (Washington, DC: ICLR).

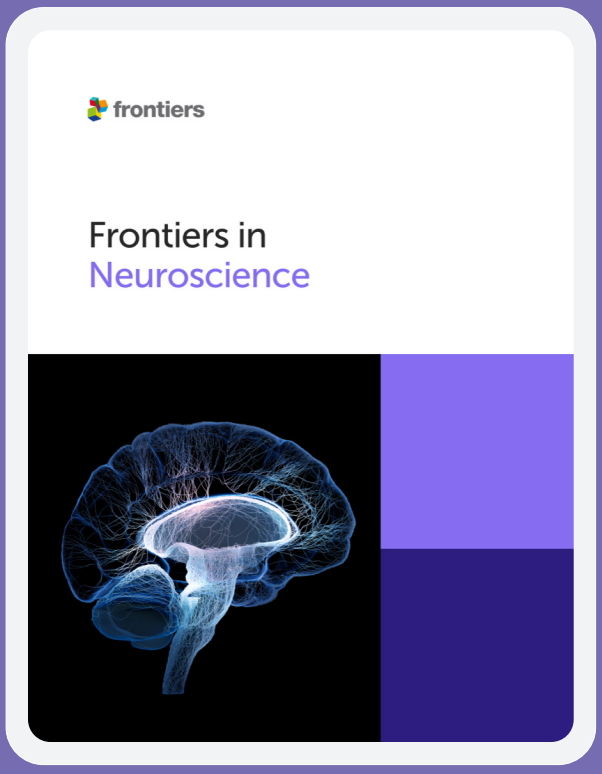
Frontiers in Neuroscience

Provides a holistic understanding of brain function from genes to behavior

Part of the most cited neuroscience journal series which explores the brain - from the new eras of causation and anatomical neurosciences to neuroeconomics and neuroenergetics.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact