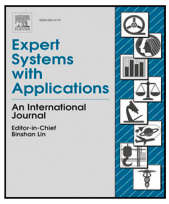
Expert Systems With Applications 263 (2025) 125785

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Multi-level Feature Attention Network for medical image segmentation

Yaning Zhang, Jianjian Yin, Yanhui Gu, Yi Chen *

School of Computer and Electronic Information/Artificial Intelligence, Nanjing Normal University, Nanjing 210023, China

ARTICLE INFO

Keywords:
Medical image segmentation
Swin Transformer
Cross-connection multi-level attention
Pyramid collaborative attention

ABSTRACT

Network architectures deriving from the Unet framework and its convolutional neural network variants have garnered significant attention for their impressive feats in computer vision. However, the shallow-level details and deep-level semantic information are underutilized in these methods, leading to the model's inability to adequately localize target regions. In this paper, we put forward a Multi-level Feature Attention Network, a novel method that cross-connects encoder and decoder features and focuses on multi-scale semantic features. Firstly, we extend UperNet using a hierarchical Swin Transformer with shifted windows, giving the network global modeling capabilities. Secondly, we introduce a Cross-connection Multi-level Attention module that connects encoder and decoder to refine the decoder's output features and supplement detailed information. Finally, we employ a Pyramid Collaborative Attention (PCA) module to mine the encoder's deepest semantic features across multiple scales. Our method establishes state-of-the-art performance on the ACDC, ISIC2017 and BUSI datasets, showcasing its exceptional capability in segmenting medical images.

1. Introduction

The advancement of deep learning (Dan, Lu, & Li, 2024b; Davidrajuh & Lin, 2011; Wu, Dai, Huang, Ma, & Xiao, 2024b, 2024c; Yin, Yan, Chen, Chen, & Yao, 2024) has revolutionized the domain of computer vision, particularly in medical image analysis, where it plays a pivotal role (Torres-Sospedra, Montoliu, Trilles, Belmonte, & Huerta, 2015). Image segmentation is an integral component of medical image analysis, and precise and stable image segmentation algorithms can provide important insights into the comprehensive analysis of anatomical regions, which is of paramount importance for lesions visualization, accurate diagnosis of diseases and the formulation of treatment plans in clinical settings (Chen, Lu, et al., 2021; Hatamizadeh et al., 2022).

A Fully Convolutional Neural Network (FCN) with a U-shaped structure have been widely utilized in various medical image segmentation tasks such as multi-organ segmentation and liver lesion segmentation (Hesamian, Jia, He, & Kennedy, 2019), and has demonstrated excellent segmentation capabilities. A typical U-shaped network, U-Net (Ronneberger, Fischer, & Brox, 2015), features an encoder–decoder structure, where the encoder extracts semantic information from the image using a series of convolutional layers and pooling layers, which is then transmitted to the decoder for up-sampling. Additionally, skip connections are used to connect corresponding levels of the encoder and decoder. Building on the foundational UNet architecture, a range of algorithms including Res-UNet (Xiao, Lian, Luo, & Li, 2018), U-Net++ (Zhou, Rahman Siddiquee, Tajbakhsh, & Liang, 2018) and Kiu-Net (Valanarasu, Sindagi, Hacihaliloglu, & Patel, 2020) have been developed and utilized across diverse medical imaging and volumetric segmentation tasks. These methods combine the powerful feature representation and local translational invariance of CNNs with the simplicity and high performance of the U-shaped structure, making them a hot topic in image segmentation discussions. However, the receptive field of each layer in these CNN architectural networks is restricted, thus failing to capture contextual information adequately and establish long-range dependencies (Li, Huang, Wang & Liu, 2024), especially when confronted with images that exhibit significant variations in texture, shape, and size, which further impedes their performance enhancement. To alleviate this problem, the feature pyramid-based UperNet (Xiao, Liu, Zhou, Jiang, & Sun, 2018) network has been proposed, which effectively integrates semantic features from different levels through its pyramid pooling module and lateral connections in the hierarchical structure. However, it remains a convolutional neural network-based approach, making it challenging to perform global information modeling or learn explicit long-range semantic information due to the intrinsic localization of convolutional operations.

In recent years, Transformer has been gradually applied to the field of Natural Language Processing (NLP) (Vaswani et al., 2017) to overcome the limitations of CNN in image processing, and Vision Transformer (ViT) proposed in literature (Dosovitskiy et al., 2020) has been applied to the task of image recognition for computer vision with excellent results. ViT primarily takes two-dimensional image patches with positional embeddings as input to the model, achieving

* Corresponding author.
E-mail addresses: 19210449@njnu.edu.cn (Y. Zhang), 212202015@njnu.edu.cn (J. Yin), gu@njnu.edu.cn (Y. Gu), cs_chenyi@njnu.edu.cn (Y. Chen).

https://doi.org/10.1016/j.eswa.2024.125785
Received 19 July 2024; Received in revised form 5 November 2024; Accepted 12 November 2024
Available online 20 November 2024

performance comparable to CNN-based methods. However, it also has obvious drawbacks: high computational cost, unsuitability for real-time applications, and inability to perform prediction tasks for high-resolution images. Therefore, the Swin Transformer (Liu et al., 2021) model has been proposed, which adopts a hierarchical ViT as its visual backbone, performing intra-window self-attention and inter-window communication with multi-head self-attention using shifted windows. Its superior capabilities have been proven in various applications such as object detection, image categorization, and semantic delineation. Despite its ability to capture global contextual representations, Swin Transformer still neglects shallow features. To mitigate this issue, methods combining the Transformer architecture with UNet model have been introduced to the stage of image segmentation, leading to innovative designs like SwinUNet (Cao et al., 2022), TransFuse (Zhang, Liu, & Hu, 2021), and TM-UNet (Tang et al., 2024). These designs effectively merge locally-focused CNNs with globally-focused transformers, which enhances the multi-scale representation of features and improves the segmentation capability (Huang, Li, Mao, Yuan, & Li, 2024). However, the aforementioned Transformer-based methods still simply concatenate the features generated by the encoder and decoder layers without considering the correlation between these features, hindering the network's ability to precisely identify target areas.

Since existing methods still fail to effectively combine low-level and high-level features and maintain feature consistency, the richness of the generated feature information is limited and the network is not precise enough to localize the target area. In this paper, we propose a Multi-level Feature Attention Network. The network leverages a hierarchical Swin Transformer architecture to enhance global contextual modeling. It supplements the rich detailed information generated by the encoder to the decoder with sibling cross-connection attention, refining the features generated by the decoder. Additionally, the network employs a multi-window based feature pyramid pooling module alongside a collaborative attention module to capture features at different scales from the deepest encoder layers for fusion, aiming to produce a more intricate feature map. Extensive experimental studies indicate that this approach achieves excellent segmentation accuracy and robust generalization capability with relatively few parameters. Specifically, our contributions are as follows:

- Multi-level Feature Attention Network makes deep use of features at different levels of the network, demonstrating precise localization capability on medical image segmentation tasks.
- The Network introduces the Cross-connection Multi-level Attention module and PCA module, which not only enhances the attention to detail information, but also extracts and fuses multi-scale semantic features.
- Extensive experiments validate the efficacy of our methodology.

The remainder of this paper is structured as follows. Section 2 offers a review of related work and analyzes the differences between these methods and our approach. Section 3 details the structure of our model and provides an in-depth explanation of our methods. The experimental results on various datasets and ablation studies are examined in Section 4. Ultimately, Section 5 provides a summary of our research.

2. Related works

2.1. CNN-based methods

Medical image segmentation necessitates detailed lesion feature learning. Therefore, initial approaches to medical image segmentation primarily focused on delineating the contours of pathological regions (Muthukrishnan & Radha, 2011; Ugarriza et al., 2009) and extensively employed traditional machine learning methods. The emergence of deep neural networks (Dan, Huang, Lu, & Li, 2024a; Wu, Dai, Chen, Huang, Ma, & Xiao, 2024a; Yin, Zheng, Gu, Zhou & and Chen, 2023; Yin, Zheng, Pan, Gu & and Chen, 2023) has markedly enhanced the precision of predictions in the field of medical image segmentation, especially the UNet (Ronneberger et al., 2015) network architecture, which has laid the cornerstone in medical image segmentation with the simplicity and superiority of its U-shaped structure. UNet consists of a symmetric encoder and decoder network with skip connections, which have demonstrated exceptional performance across various medical domains. In recent years, a multitude of UNet variants have surfaced, including R50 UNet (Xiao, Lian, et al., 2018), U-Net++ (Zhou et al., 2018), Kiu-Net (Valanarasu et al., 2020) and Dense-UNet (Cai et al., 2020), each tailored to address specific nuances of medical imaging tasks. For instance, R50 UNet (Xiao, Lian, et al., 2018) incorporates a symmetric depth structure in its encoder and decoder setup, integrating features from both at equivalent levels to bolster boundary delineation, which has markedly improved segmentation outcomes. Att-UNet (Oktay et al., 2018) enhances the UNet architecture with an attention mechanism that autonomously focuses on relevant structures of varying shapes and sizes within medical images. It skillfully suppresses the influence of irrelevant information while highlighting key features necessary for specific diagnostic tasks, significantly increasing the model's sensitivity and accuracy. MALUNet (Ruan, Xiang, Xie, Liu, & Fu, 2022) developed an ultra-lightweight UNet model that integrates advanced fusion techniques across multiple levels and scales. This model effectively integrates dilated convolutions with gated attention mechanisms, and uses external attention to strengthen inter-sample relationships, collectively reducing its dependency on a large number of input channels. Furthermore, MHorUNet (Wu et al., 2024e) for the first time integrated a higher-order spatial interaction mechanism (Rao et al., 2022) within the U-Net framework, demonstrating superior segmentation results in the treatment of skin lesions. Although CNN-based approaches have demonstrated impressive performance in medical image segmentation owing to their robust representation capabilities, they often exhibit a bias towards local details and struggle with holistic modeling because of the constrained receptive fields of convolutional kernels.

2.2. Transformer-based methods

Transformer models have exhibited exceptional performance across multiple tasks within the NLP domain (Devlin, Chang, Lee, & Toutanova, 2018). Consequently, researchers have explored their application in computer vision, culminating in the creation of the Vision Transformer (ViT) (Dosovitskiy et al., 2020). Employing self-attention mechanisms to process global information, ViT exhibits robust feature extraction capabilities and has shown outstanding performance in image recognition. However, ViT suffers from drawbacks such as requiring extensive data and the high computational overhead due to its highly complex model structure. The Swin Transformer model (Liu et al., 2021), based on ViT, effectively addresses these issues. Swin Transformer adopts the W-MSA attention mechanism to reduce computational overhead and utilizes SW-MSA to facilitate communication across different local windows, thereby enhancing its capability to understand spatial relationships. SwinUNet (Cao et al., 2022) integrates Swin Transformer blocks into the UNet architecture, resulting in a reduced parameter count compared to TransUNet (Chen, Lu, et al., 2021). MedT (Valanarasu, Oza, Hacihaliloglu, & Patel, 2021) enhances traditional architectures by incorporating gated axial transformer layers, designed specifically to address the challenges posed by the smaller datasets typical of medical imaging. This method introduces a sophisticated gated axial-attention mechanism, which dynamically focuses on essential features to enhance the precision of feature extraction. CrossViT (Chen, Fan, & Panda, 2021) develops a dual-branch vision transformer strategy that processes small and large patch tokens through separate branches, thereby efficiently capturing features at multiple scales for image classification. This design leverages a cross-attention module to facilitate effective communication between the branches, optimizing feature synthesis while ensuring computational and memory efficiency remains linear.

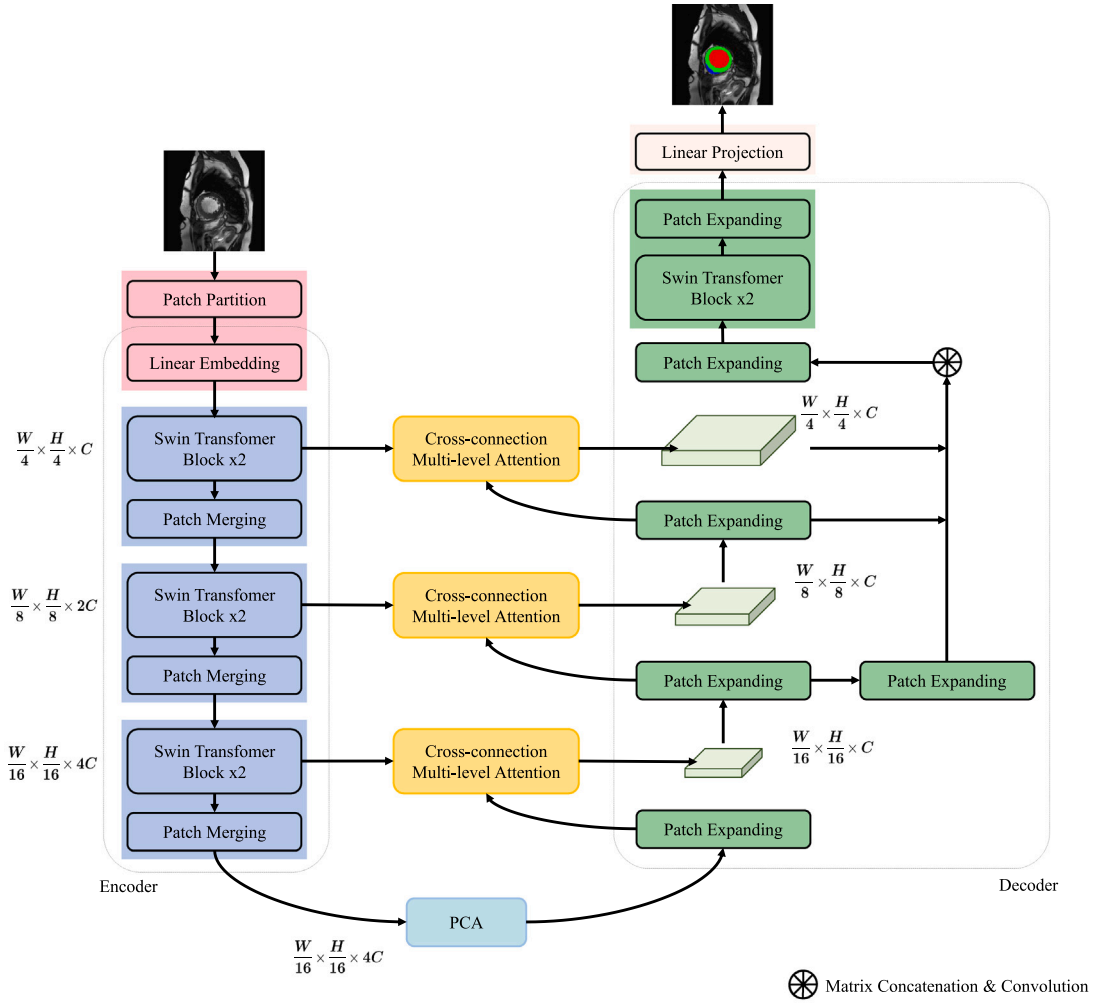


Fig. 1. The structure of our Multi-level Feature Attention Network. The network utilizes Swin Transformer blocks for feature extraction, Cross-connection Multi-level Attention module and PCA module to optimize the shallow and deep level features respectively. Finally, the decoder accumulates and fuses features from different layers and up-samples them to obtain the segmentation result.

In addition to purely transformer models, there are hybrid models that combine transformers and CNNs, such as TransUNet (Chen, Lu, et al., 2021), which better integrates the high-resolution spatial information from CNNs and the global contextual information from transformers. UNETR (Hatamizadeh et al., 2022) redefines the volumetric (3D) medical image segmentation task as a sequence-to-sequence prediction task, which consists of a transformer-based encoder for embedding the input 3D patches, coupled with a CNN-based decoder to generate the final 3D segmentation result. This architecture also adheres to the classic "U-shaped" design, connecting the transformer encoder directly to the decoder. Due to the high labor and time costs required from senior radiologists to annotate medical images, it is difficult to obtain sufficient labeled data, making semi-supervised and scribble-supervised segmentation methods popular (Han et al., 2024). ScribFormer (Li, Zheng, et al., 2024) excels in scribble-supervised segmentation by integrating a triple-branch structure that combines local CNN features with global Transformer representations and an attention-guided class activation map. HiFormer (Heidari et al., 2023) seeks to merge localized information obtained from CNNs with the extensive contextual interactions facilitated by Transformers, with the aim of capturing global dependencies. Although these hybrid methods show excellent performance, they still fall short in effectively utilizing both the shallow local information and the deepest global information within the network. The research in this paper utilizes a Cross-connection Multi-level Attention Module to comprehensively analyze the contour and texture details within the region of interest. In addition, we multi-scale mine and fuse the deeper features of the encoder to enhance the global modeling of spatial relationships.

3. Method

3.1. Architecture overview

The structural design of the proposed Multi-level Feature Attention Network is presented in Fig. 1. The network processes an image I: H×W×C, with H and W defining the spatial dimensions and C represents the channel count. It begins by dividing the image into a sequence of blocks using a 4 × 4 window size during the patch partition phase, aimed at retaining maximum detail. The tokenized input is subsequently routed through the encoder module, which consists of a series of three encoder blocks. Each one contains paired Swin Transformer blocks followed by a patch merging layer. Features derived from the deeper layers of the encoder are channeled into the PCA module to form a fine feature map, and the PCA module comprises a multi-window based Feature Pyramid Pooling Module and a Collaborative Attention Module that integrates spatial and channel information respectively. In the decoder, features from each patch expanding layer are merged with corresponding features from the encoder via Cross-connection Multi-level Attention, ensuring enhanced feature alignment. The resulting feature maps, varying in three different dimensions, are

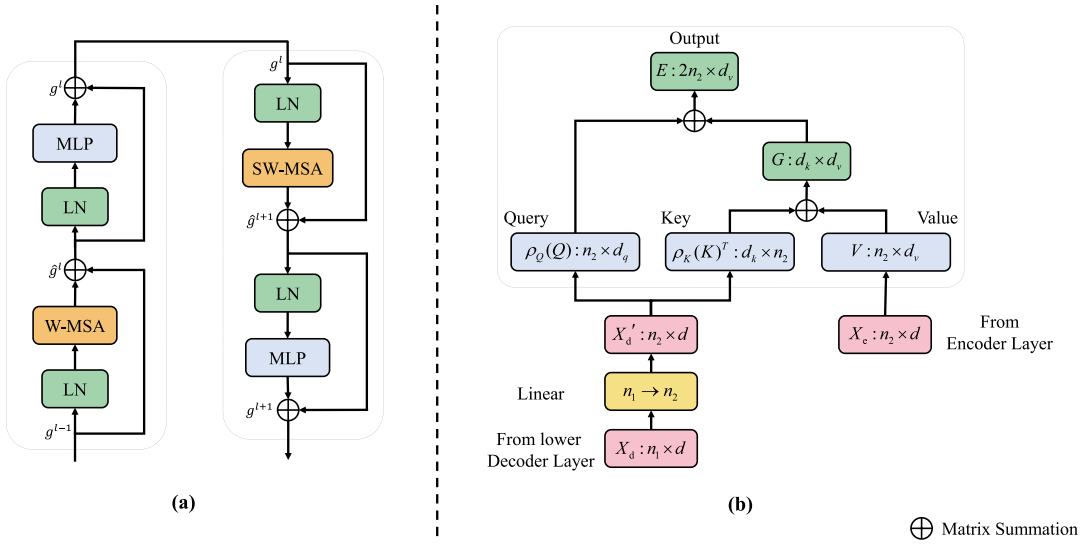


Fig. 2. (a) The Swin Transformer block. It comprises window multi-head self attention (W-MSA), shifted window multi-head self attention (SW-MSA), LayerNorm (LN) and MLP. (b) The Cross-connection Multi-level Attention module. It fuses encoder layer's features and lower decoder layer's, and efficiently computes self-attention.

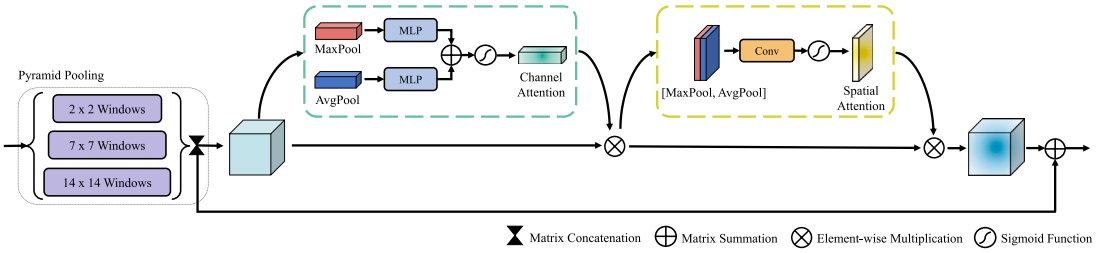


Fig. 3. The PCA module. Features traverse a pyramid pooling module featuring Swin Transformer blocks with three distinct window sizes, followed by processing through channel and spatial attention modules. MaxPool and AvgPool represent max pooling and average pooling, respectively. The Multi-Layer Perceptron (MLP) contains one hidden layer. Conv refers to a standard convolution layer.

sequentially added after dimension transformation, and finally upsampled to the original image size. In the next sections, we will introduce the Swin Transformer block, the PCA module and the Cross-connection Multi-level Attention module.

3.2. Swin Transformer block

Compared to ViT, the shifted window-based Swin Transformer incurs lower computational costs and is capable of executing prediction tasks on high-resolution images, such as image segmentation. Fig. 2(a). illustrates the internal configuration of the Swin Transformer module. The Swin Transformer module consists of two Multilayer Perceptrons (MLPs), four Layer Normalization (LN) layers, a Window Multi-Head Self-Attention module (W-MSA), and a Shifted Window Multi-Head Self-Attention module (SW-MSA). W-MSA segments the feature map into disjoint blocks according to the window size and conducts self-attention within each block, while SW-MSA enables communication between non-overlapping blocks. The equation governing the continuous operation of the Swin Transformer is presented below:

$$\hat{g}^{l} = F_{w}\left(G\left(g^{l-1}\right)\right) + g^{l-1}, \tag{1}$$

$$g^{l} = H\left(G\left(\hat{g}^{l}\right)\right) + \hat{g}^{l}, \tag{2}$$

$$\hat{g}^{l+1} = F_{sw}\left(G\left(g^{l}\right)\right) + g^{l}, \tag{3}$$

$$g^{l+1} = H\left(G\left(\hat{g}^{l+1}\right)\right) + \hat{g}^{l+1}, \tag{4}$$

Here, F_{sw} and F_{sw} represent the functions for W-MSA and SW-MSA, G is the function replacing LN, and H denotes the function previously represented by MLP.

As outlined in Hu, Gu, Zhang, Dai, and Wei (2018), Hu, Zhang, Xie, and Lin (2019), the calculation for self-attention is expressed as:

$$\mathrm{SA}(q,k,v) = \mathrm{SoftMax}\left(\frac{qk^{T}}{\sqrt{d}} + b\right)v, \tag{5}$$

where q, k, v are the matrices for the query, key and value respectively, d refers to the channel dimension, and b denotes the bias matrix.

3.3. Cross-connection Multi-level Attention

The Cross-connection Multi-level Attention module, as shown in Fig. 2(b), operates by performing sibling cross-connection attention between the corresponding features generated by the encoder and decoder. This operation refines the features produced by the decoder rather than simply concatenating them, thereby preserving richer detail information.

The input values are derived from the encoder X_e, while the inputs for queries and keys originate from the output of the lower decoder layer X_d. To facilitate feature fusion, X_d is scaled to match the embedding dimension with X_e using a linear layer.

$$\mathbf{X}'_{\mathbf{d}} = FC\left(\mathbf{X}_{\mathbf{d}}\right), \tag{6}$$

The keys, queries and values are then computed by projecting the features of the encoder and decoder, respectively.

$$\mathbf{K},\mathbf{Q} = Proj\left(\mathbf{X}'_{\mathbf{d}}\right), \tag{7}$$

$$\mathbf{V} = Proj\left(\mathbf{X}_{\mathbf{e}}\right), \tag{8}$$

Table 1
Results of comparison experiments on the ACDC dataset. All results are presented as the mean ±standard deviation. We report the DSC for each of the three categories (RV, Myo, and LV) along with the average.

Method	Avg (%)	RV (%)	Myo (%)	LV (%)
R50 UNet (Chen, Lu, et al., 2021)	87.60 ± 0.08	84.62 ± 0.10	84.52 ± 0.06	93.68 ± 0.08
R50 AttUNet (Chen, Lu, et al., 2021)	86.90 ± 0.09	83.27 ± 0.11	84.33 ± 0.05	93.53 ± 0.05
ViT-CUP (Chen, Lu, et al., 2021)	83.41 ± 0.07	80.93 ± 0.09	78.12 ± 0.03	91.17 ± 0.07
R50 ViT (Chen, Lu, et al., 2021)	86.19 ± 0.06	82.51 ± 0.07	83.01 ± 0.05	93.05 ± 0.08
TransUNet (Chen, Lu, et al., 2021)	89.71 ± 0.05	86.67 ± 0.08	87.27 ± 0.04	95.18 ± 0.06
UNETR (Hatamizadeh et al., 2022)	88.61 ± 0.04	85.29 ± 0.03	86.52 ± 0.04	94.02 ± 0.03
SwinUNet (Cao et al., 2022)	88.07 ± 0.07	85.77 ± 0.09	84.42 ± 0.04	94.03 ± 0.07
MISSFormer (Huang, Deng, Li, Yuan, & Fu, 2022)	89.47 ± 0.07	87.73 ± 0.10	87.51 ± 0.04	93.16 ± 0.06
HiFormer (Heidari et al., 2023)	89.68 ± 0.06	87.49 ± 0.08	86.55 ± 0.04	94.99 ± 0.06
CoST-UNet (Islam, Qaraqe, & Serpedin, 2024)	86.50 ± 0.03	87.30 ± 0.03	82.61 ± 0.03	89.50 ± 0.04
ScribFormer (Li, Zheng, et al., 2024)	88.53 ± 0.07	86.86 ± 0.10	87.08 ± 0.03	91.64 ± 0.06
MFAN	**90.79 ± 0.03**	**88.84 ± 0.08**	**88.12 ± 0.03**	**95.39 ± 0.02**

The module employs Efficient attention (Shen, Zhang, Zhao, Yi, & Li, 2021) to compute self-attention efficiently, which involves normalizing the keys and queries, multiplying the keys with the values, and subsequently combining the resultant global context vector with the queries to form a new representation. Efficient attention produces equivalent outputs to dot-product attention, but greatly reduces the computational complexity by not computing pairwise similarities between points first.

$$\mathbf{E} = \rho_{\mathbf{q}}(\mathbf{Q})\rho_{\mathbf{k}}(\mathbf{K})^{\mathbf{T}}\mathbf{V}, \tag{9}$$

Here, $\rho_{\mathbf{q}}$, $\rho_{\mathbf{k}}$ represent normalization functions, while *Proj* signifies a linear projection function.

3.4. Pyramid Collaborative Attention

To boost the global modeling capabilities and undertake the extraction of multi-scale features derived from the deep semantic outputs produced by the encoder, we utilize the Pyramid Collaborative Attention module (PCA), which consists of multi-window based feature pyramid pooling, channel attention and spatial attention. The architectural details of the PCA module are depicted in Fig. 3. Drawing inspiration from the convolution-based Pyramid Pooling Module (PPM) in the PSPNet (Zhao, Shi, Qi, Wang, & Jia, 2017), we build the feature pyramid pooling module using Swin Transformer blocks across three distinct window dimensions: 2 × 2, 7 × 7, and 14 × 14. Smaller windows focus on gathering local information, while larger windows are intended for capturing global information.

The multi-scale features generated are first aggregated through max pooling and average pooling to create two different spatial contexts. After passing through the same MLP, they are combined to form the channel attention map. Then, max pooling and average pooling operations are applied along the channel axis to generate two 2D feature maps. These are concatenated along the channel axis and convolved through a standard convolutional layer to produce the spatial attention map. Finally, the features with enhanced information is combined with the original features via a residual connection. It is noteworthy that the Collaborative Attention part requires only a very small number of parameters but markedly boosts the network's capabilities.

4. Experiments

4.1. Datasets and evaluation metrics

In our study, the model's efficacy was gauged using the Automated Cardiac Diagnosis Challenge dataset (ACDC) (Bernard et al., 2018), International Skin Imaging Collaboration (ISIC2017) (Codella et al., 2018), and Breast Ultrasound Images (BUSI) (Al-Dhabyani, Gomaa, Khaled, & Fahmy, 2020) datasets.

ACDC, a repository of cardiac MRI scans, is classified into three anatomical categories: left ventricle (LV), right ventricle (RV), and myocardium (Myo). This dataset is divided into 70 for training, 10 for validation, and 20 for testing samples. All images have been resized to a uniform dimension of 224 × 224.

ISIC2017, curated by the ISIC consortium, is an extensive compendium for the study of skin cancer segmentation. We adhere to the data partitioning guidelines set forth in Asadi-Aghbolaghi, Azad, Fathy, and Escalera (2020), Azad, Asadi-Aghbolaghi, Fathy, and Escalera (2019), with all ISIC2017 images uniformly resized to 224 × 224 pixels.

BUSI is a dataset tailored for breast cancer segmentation. This collection is partitioned into training and testing sets based on the data division methodology described in Chen, Lu, et al. (2021), Zhang et al. (2021). The settings are the same as ACDC and ISIC2017 datasets, with the BUSI image set to 224 × 224.

To qualitatively assess the model's performance, we apply specific metrics: (1) Dice Similarity Coefficient (DSC), (2) Sensitivity (SE), (3) Specificity (SP), and (4) Accuracy (ACC). It is important to note that for an equitable evaluation, we adopt the same metrics configuration used by other leading methods across the various datasets, rather than applying a universal metrics suite.

4.2. Implementation details

We developed our model using the PyTorch framework, conducting all experiments on an NVIDIA RTX 3090 GPU throughout both training and testing phases. For the ISIC2017 dataset, the batch size was configured at 24, and the training spanned 200 epochs with the use of the stochastic gradient descent (SGD) algorithm, starting at a learning rate of 0.05. In comparison, the training for the BUSI and ACDC datasets involved batch sizes of 8 and 24, respectively, with an initial learning rate set at 0.0001, and lasted for 100 epochs utilizing the Adam optimization algorithm. The network's training employed a composite loss function that integrated Dice loss and Cross-entropy loss in the following formulation:

$$\mathcal{L}_{\text{Loss}} = \alpha\mathcal{L}_{\text{CE}} + \beta\mathcal{L}_{\text{Dice}}, \tag{10}$$

where the coefficients α and β are configured to 0.3 and 0.7. We executed comprehensive ablation studies to examine the impact of these coefficients, with the results discussed in the ensuing sections.

4.3. Evaluation results

4.3.1. Evaluation on ACDC dataset

Table 1 displays the comparative results of our model against advanced methods on the ACDC dataset. Our method shows superior results in the DSC metrics, surpassing SwinUNet (Cao et al., 2022) and MISSFormer (Huang et al., 2022) by 2.72% and 1.32%. Additionally, the model outperforms MISSFormer by 1.11%, 0.61%, and 2.23% in RV, Myo, and LV categories, respectively. Moreover, our standard deviation is significantly smaller compared to other methods, such as only

Table 2
Results of comparison experiments on the ISIC2017 dataset. All results are presented as the mean ±standard deviation.

Method	DSC (%)	SE (%)	SP (%)	ACC (%)
UNet (Ronneberger et al., 2015)	81.59 ± 0.29	81.72 ± 0.20	96.80 ± 0.17	91.64 ± 0.18
Att-UNet (Oktay et al., 2018)	80.82 ± 0.27	79.98 ± 0.21	97.76 ± 0.19	91.45 ± 0.17
UNet++ (Zhou, Siddiquee, Tajbakhsh, & Liang, 2019)	85.80 ± 0.26	–	–	93.80 ± 0.11
DAGAN (Lei et al., 2020)	84.25 ± 0.20	83.63 ± 0.20	97.16 ± 0.12	93.04 ± 0.11
TransUNet (Chen, Lu, et al., 2021)	81.23 ± 0.16	82.63 ± 0.19	95.77 ± 0.08	92.07 ± 0.09
MedT (Valanarasu et al., 2021)	80.37 ± 0.21	80.64 ± 0.20	95.46 ± 0.13	90.90 ± 0.16
TransFuse (Zhang et al., 2021)	87.20 ± 0.19	–	–	94.40 ± 0.11
MSU-Net (Su, Zhang, Liu, & Cheng, 2021)	88.15 ± 0.13	82.66 ± 0.21	97.00 ± 0.06	92.90 ± 0.11
FAT-Net (Wu et al., 2022)	85.00 ± 0.17	83.92 ± 0.16	97.25 ± 0.10	93.26 ± 0.07
Ms-RED (Dai et al., 2022)	89.10 ± 0.15	84.37 ± 0.19	97.05 ± 0.07	93.40 ± 0.10
Swin-UNet (Cao et al., 2022)	88.45 ± 0.32	88.93 ± 0.46	97.78 ± 0.03	94.76 ± 0.11
PHCU-Net (Xu, Wang, Wang & Huang, 2023)	89.48 ± 0.15	87.72 ± 0.17	**98.21** ± 0.08	96.41 ± 0.06
MHorNet (Wu et al., 2024e)	89.44 ± **0.13**	88.35 ± **0.15**	97.82 ± 0.06	96.35 ± 0.08
MFAN	**90.42** ± 0.21	**91.11** ± 0.24	97.62 ± 0.04	**96.51** ± **0.05**

Table 3
Results of comparison experiments on the BUSI dataset.

Method	DSC (%)	ACC (%)	IoU (%)	Sensitivity (%)
UNet (Ronneberger et al., 2015)	76.35	95.20	62.66	75.06
UNet++ (Zhou et al., 2019)	77.54	95.64	63.55	70.99
DoubleU-Net (Jha, Riegler, Johansen, Halvorsen, & Johansen, 2020)	74.77	–	65.97	76.30
TransUNet (Chen, Lu, et al., 2021)	79.30	95.01	60.57	71.61
SSFormer (Wang et al., 2022)	79.27	96.53	64.75	77.06
DCSAU-Net (Xu, Ma, Na & Duan, 2023)	75.14	–	66.56	80.99
CMU-Net (Tang, Wang, Ning, Xian, & Ding, 2023)	78.29	95.71	65.74	75.70
NU-Net (Chen, Li, Zhang, & Dai, 2023)	79.42	–	70.35	82.46
MFAN	**79.73**	**98.07**	**78.14**	**88.13**

0.03% and 0.02% in the Myo and LV categories, respectively, and the lowest of 0.03% for the average DSC. This fully demonstrates the stability and reliability of our model. The data presented in the table clearly indicate that our method far exceeds the current leading techniques in DSC metrics, achieving excellent performance across all categories. This further confirms the effectiveness of the operations in refining the features generated by the decoder and extracting multi-scale semantic features in the Multi-level Feature Attention Network.

Furthermore, in Fig. 4, we provide a visualization of the segmentation output produced by our model alongside HiFormer and MISSFormer on the ACDC dataset. As depicted, our results most closely align with the ground truth, indicating the high efficacy of our approach. The Multi-level Feature Attention Network excels in integrating detail and global information, enabling effective segmentation at fine-edge details, which is particularly crucial for medical image segmentation tasks demanding high precision.

4.3.2. Evaluation on ISIC2017 dataset

The comparison results for the ISIC 2017 dataset are meticulously detailed in Table 2. Our method yields considerable enhancements relative to the current SOTA methods (PHCU-UNet (Xu, Wang, et al., 2023) and MHorNet (Wu et al., 2024e)) in DSC, SE and ACC metrics, with particularly notable improvements of 3.39% and 2.76% in the SE metric, respectively. It is important to note that although our method may not have attained the highest scores in SP metrics, higher SP values do not necessarily indicate superior model performance. Achieving optimal performance requires considering various trade-offs rather than solely relying on higher SP values. Furthermore, compared to some baselines, our method exhibits a slightly higher standard deviation in DSC and SE metrics, which is inherently due to the characteristics of the ISIC2017 dataset. This dataset contains significant variations in lighting conditions, pathological types and skin tones in skin cancer images. Such variations make the dataset more complex, resulting in higher standard deviations for most models. Nevertheless, MFAN consistently performs well on all key metrics, showcasing its robustness in handling such diverse image features. The slightly higher standard deviation reflects the natural variability of the dataset rather than a lack of consistency in our model's performance.

Fig. 5 shows an intuitive visualization illustration of the segmentation results generated by our model and TransFuse. As the figure demonstrates, the segmentation effect of our approach on the edges of the skin cancer pathology is outstanding, and clearly matches the ground truth masks better compared to TransFuse. This superior performance is attributed to the Cross-connection Multi-level Attention module of our method, which refines the decoder's output features using the encoder. This enhancement enables the model to capture contour information more accurately and obtain segmentation results with higher fine-grainedness.

4.3.3. Evaluation on BUSI dataset

We provide a detailed presentation of our findings on the BUSI dataset, as shown in Table 3. Compared to current SOTA methods, our model has achieved significant improvements in all four metrics, especially in IoU and Sensitivity metrics, where it outperforms NU-Net (Chen et al., 2023) by 7.79% and 5.67%, respectively. Additionally, MFAN surpasses the state-of-the-art methods (SSFormer (Wang et al., 2022) and CMU-Net (Tang et al., 2023)) in ACC metric by 1.54% and 2.36%, respectively. Meanwhile, in terms of DSC metric, our model further outperforms them by 0.46% and 1.44%, respectively. The experimental results from the above three datasets clearly illustrate the superior performance of our model relative to current leading methods. This is attributed to our effective integration of both fine-grained details and broad semantic insights within the network, thereby constructing features with stronger representational capabilities and improving the model's capacity to accurately localize target regions.

4.4. Ablation study

4.4.1. Effectiveness of PCA and CMA

On the ACDC dataset, we performed ablation studies to evaluate the performance of the PCA module and Cross-connection Multi-level Attention (CMA), and the results are shown in Table 4. Baseline refers to the model without the PCA module, where only convolutional operations are used to change the channel dimension and concatenate features at different scales. In addition, Baseline does not utilize CMA; instead, it simply concatenates features from both the encoder and decoder. The addition of PCA module improves the DSC metric by

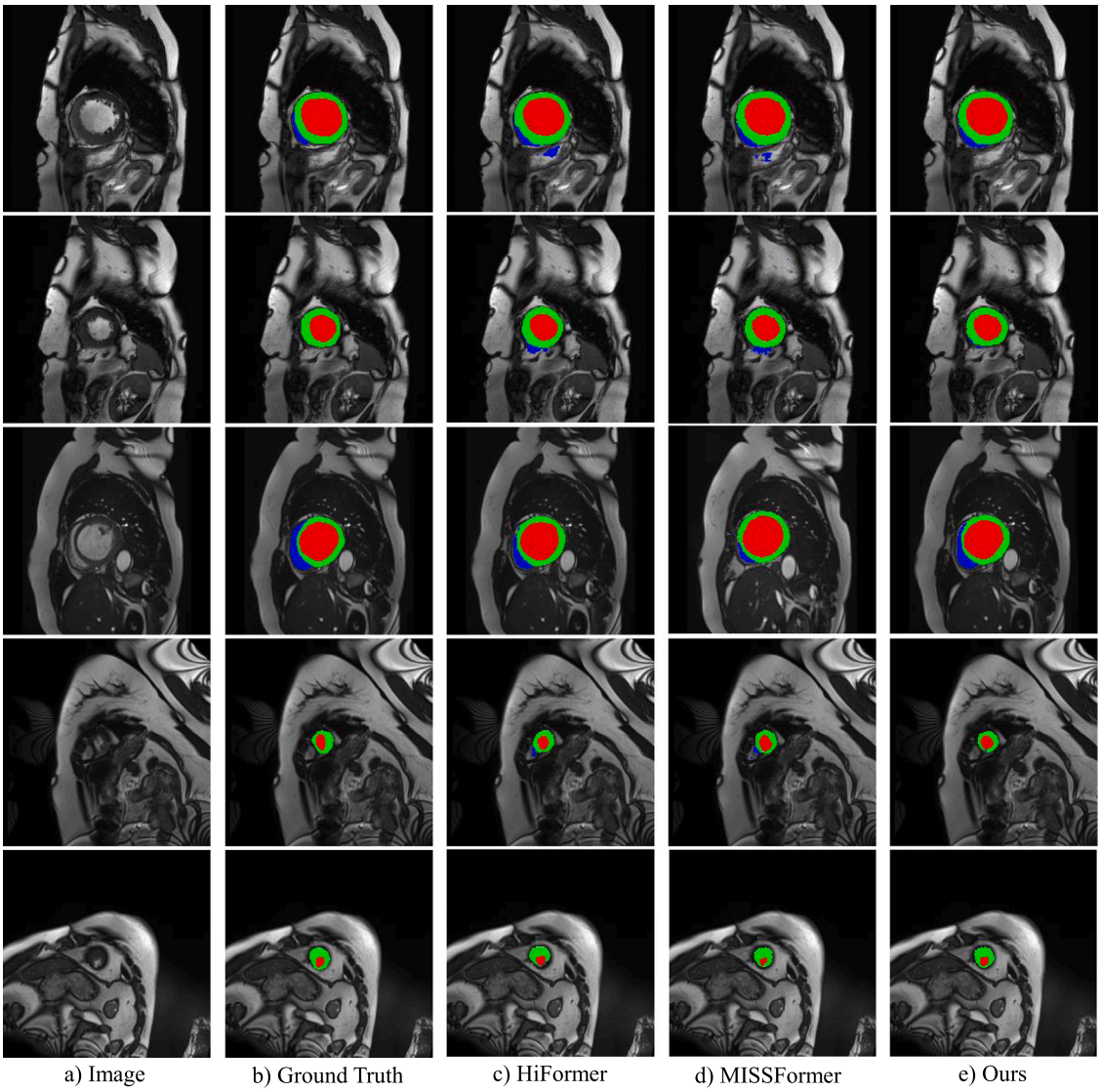


Fig. 4. Comparative visualization of segmentation results on the ACDC dataset.

Table 4
Ablation experiments were conducted on ACDC dataset to evaluate the efficacy of PCA and Cross-connection Multi-level Attention (CMA).

Baseline	PCA	CMA	Avg (%)	RV (%)	Myo (%)	LV (%)
✓			89.19	86.05	86.45	95.06
✓	✓		90.62	88.67	87.87	95.33
✓	✓	✓	**90.79**	**88.84**	**88.12**	**95.39**

Table 5
Ablation experiments on the ACDC dataset to evaluate the impact of PPM and its enhanced version PCA. The results highlight PCA's superior performance.

Method	Avg (%)	RV (%)	Myo (%)	LV (%)
Baseline	89.19	86.05	86.45	95.06
Baseline + PPM	89.78	86.71	87.18	95.16
Baseline + PCA	**90.62**	**88.67**	**87.87**	**95.33**

1.43%, and further incorporating CMA leads to an additional enhancement of 0.17% in DSC performance. Concurrent improvements are also observed in the RV, Myo, and LV categories. From the data presented, the PCA module effectively mines the deep semantic information from encoder, and the Cross-connection Multi-level Attention enhances the focus on edge information, which improves the model's segmentation capability.

4.4.2. Comparison between PCA and PPM

Inspired by the Pyramid Pooling Module (PPM) from PSPNet (Zhao et al., 2017), we designed the PCA module. Unlike PPM, which utilizes only convolutional operations for pyramid pooling and captures multi-scale features with convolution windows of different sizes, PCA introduces collaborative attention. This collaborative attention includes channel attention and spatial attention, which enhances the weight distribution among different feature channels and better captures the key features in the images. Table 5 shows the impact of PCA and PPM on model performance. PCA improves 1.96%, 0.69%, and 0.17% over PPM in terms of the DSC metric on the three categories of RV, Myo and LV, respectively. The experimental results demonstrate the superiority of PCA, which more effectively balances global information and local details, thus improving the network's performance.

4.4.3. The impact of α and β

Table 6 illustrates our study on the impact of α and β on ISIC2017 dataset. According to the experimental results, Cross-entropy loss and Dice loss complement and constrain each other in network training. When the settings for α and β are 0.3 and 0.7 respectively, the model

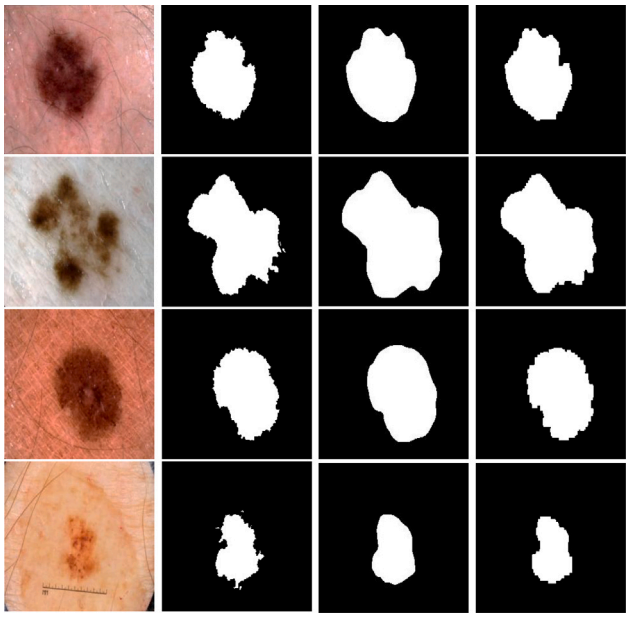


Fig. 5. Comparative visualization of segmentation results on the ISIC2017 dataset.

Table 6
The impact of α and β was evaluated through ablation experiments on the ISIC2017 dataset.

α	β	DSC (%)	SE (%)	SP (%)	ACC (%)
0.1	0.9	90.07	90.64	97.60	96.41
0.2	0.8	90.20	91.08	97.54	96.44
0.3	0.7	**90.42**	91.11	97.62	**96.51**
0.4	0.6	89.48	91.12	97.19	96.16
0.5	0.5	90.19	90.92	97.57	96.43
0.6	0.4	89.74	90.80	97.39	96.26
0.7	0.3	89.24	90.97	97.07	96.02
0.8	0.2	89.94	**91.63**	97.22	96.26
0.9	0.1	88.91	87.27	**97.94**	96.12

achieves the best performance on both DCS and ACC metrics. The best performance in SE metrics is obtained when α and β are adjusted to 0.8 and 0.2. When α is set to 0.9 and β to 0.1, the SP performance is optimal. Therefore, we default α and β values to 0.3 and 0.7 respectively in our studies.

4.4.4. Model complexity

Table 7 compares the number of parameters and complexity of different models. Time refers to the inference time required to process a batch. MFAN has the lowest FLOPs, compared to the 38.52G of TransUNet (Chen, Lu, et al., 2021) and 35.91G of ScribFormer (Li, Zheng, et al., 2024). At the same time, the model not only maintains lightweight but also demonstrates extremely very short inference time, nearly inferior to the state-of-the-art HiFormer (Heidari et al., 2023) model. It is worth noting that although HiFormer has lower computational complexity, its performance in medical image segmentation is not as good as ours. Considering the tendency of heavyweight networks to overfit when dealing with a small amount of medical image data, we lean towards a lightweight design while also ensuring performance improvement. Overall, MFAN achieves an advanced trade-off between performance, number of parameters and computational complexity.

Table 7
Comparison of parameters, FLOPs, and inference time with other state-of-the-art methods, using an input size of 224 × 224.

Method	Params (M)	FLOPs (G)	Time (ms)
Att-UNet (Oktay et al., 2018)	34.88	72.81	37.44
UNet++ (Zhou et al., 2019)	**9.16**	34.65	36.17
TransUNet (Chen, Lu, et al., 2021)	105.32	38.52	32.45
SSFormer (Wang et al., 2022)	66.22	13.56	29.49
HiFormer (Heidari et al., 2023)	29.52	**7.51**	**14.57**
ScribFormer (Li, Zheng, et al., 2024)	50.43	35.91	23.01
MFAN	29.44	31.88	19.29

5. Conclusion

In this paper, we introduce a Multi-level Feature Attention Network for medical image segmentation, mitigating the issue of insufficient localization of target regions. The network integrates a hierarchical Swin Transformer with shifted windows to model UperNet to capture global spatial correlation, while features from the encoder and decoder are fused using Cross-connection Multi-level Attention to supplement the detail information. In addition, the network introduces the Pyramid Collaborative Attention (PCA) module for multi-scale fusion of deep features to extract semantic information. Empirical analyses across three distinct datasets substantiate that our methodology sets a new benchmark in performance.

Although our method is effective, there are still some limitations: (1). In some cases, the model lacks effective utilization of detailed shallow features and has insufficient ability to handle fine textures and tiny regions. (2). The model is only suitable for 2D images and lacks the ability to process depth information, making it unable to utilize spatial information and thus cannot directly handle 3D image inputs. Future work will focus on these two aspects. The first is to introduce more refined attention mechanisms that can better integrate shallow features, especially in small target segmentation tasks that require high precision on pathology boundaries. The second is to broaden the application range of the model so that it can directly handle complex spatial relationships in 3D volumetric data. At the same time, we strive to maintain a low number of parameters and computational cost while achieving state-of-the-art performance.

CRediT authorship contribution statement

Yaning Zhang: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Investigation. **Jianjian Yin:** Data curation, Plotting graphs and charts. **Yanhui Gu:** Data curation, Resources. **Yi Chen:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work is supported by the Natural Science Foundation of China (Nos. 62377029, Nos. 92370127 and Nos. 22033002), and Frontier Technologies R&D Program of Jiangsu (BF2024076).

References

Al-Dhabyani, Walid, Gomaa, Mohammed, Khaled, Hussien, & Fahmy, Aly (2020). Dataset of breast ultrasound images. *Data in Brief, 28*, Article 104863.

Asadi-Aghbolaghi, Maryam, Azad, Reza, Fathy, Mahmood, & Escalera, Sergio (2020). Multi-level context gating of embedded collective knowledge for medical image segmentation. arXiv preprint arXiv:2003.05056.

Azad, Reza, Asadi-Aghbolaghi, Maryam, Fathy, Mahmood, & Escalera, Sergio (2019). Bi-directional convlstm U-net with densley connected convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

Bernard, Olivier, Lalande, Alain, Zotti, Clement, Cervenansky, Frederick, Yang, Xin, Heng, Pheng-Ann, et al. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, *37*(11), 2514–2525.

Cai, Sijing, Tian, Yunxian, Lui, Harvey, Zeng, Haishan, Wu, Yi, & Chen, Guannan (2020). Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quantitative Imaging in Medicine and Surgery*, *10*(6), 1275.

Cao, Hu, Wang, Yueyue, Chen, Joy, Jiang, Dongsheng, Zhang, Xiaopeng, Tian, Qi, et al. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision* (pp. 205–218). Springer.

Chen, Chun-Fu Richard, Fan, Quanfu, & Panda, Rameswar (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 357–366).

Chen, Gongping, Li, Lei, Zhang, Jianxun, & Dai, Yu (2023). Rethinking the unpretentious U-net for medical ultrasound image segmentation. *Pattern Recognition*, *142*, Article 109728.

Chen, Jieneng, Lu, Yongyi, Yu, Qihang, Luo, Xiangde, Adeli, Ehsan, Wang, Yan, et al. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Codella, Noel CF, Gutman, David, Celebi, M Emre, Helba, Brian, Marchetti, Michael A, Dusza, Stephen W, et al. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging* (pp. 168–172). IEEE.

Dai, Duwei, Dong, Caixia, Xu, Songhua, Yan, Qingsen, Li, Zongfang, Zhang, Chunyan, et al. (2022). Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Medical Image Analysis*, *75*, Article 102293.

Dan, Han-Cheng, Huang, Zhetao, Lu, Bingjie, & Li, Mengyu (2024a). Image-driven prediction system: automatic extraction of aggregate gradation of pavement core samples integrating deep learning and interactive image processing framework. *Construction and Building Materials*, *453*, 139056.

Dan, Han-Cheng, Lu, Bingjie, & Li, Mengyu (2024b). Evaluation of asphalt pavement texture using multiview stereo reconstruction based on deep learning. *Construction and Building Materials*, *412*, 134837.

Davidrajuh, Reggie, & Lin, Binshan (2011). Exploring airport traffic capability using Petri net based model. *Expert Systems with Applications*, *38*(9), 10923–10931.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Han, Kai, Sheng, Victor S, Song, Yuqing, Liu, Yi, Qiu, Chengjian, Ma, Siqi, et al. (2024). Deep semi-supervised learning for medical image segmentation: A review. *Expert Systems with Applications*, Article 123052.

Hatamizadeh, Ali, Tang, Yucheng, Nath, Vishwesh, Yang, Dong, Myronenko, Andriy, Landman, Bennett, et al. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574–584).

Heidari, Moein, Kazerouni, Amirhossein, Soltany, Milad, Azad, Reza, Aghdam, Ehsan Khodapanah, Cohen-Adad, Julien, et al. (2023). Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 6202–6212).

Hesamian, Mohammad Hesam, Jia, Wenjing, He, Xiangjian, & Kennedy, Paul (2019). Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of Digital Imaging*, *32*, 582–596.

Hu, Han, Gu, Jiayuan, Zhang, Zheng, Dai, Jifeng, & Wei, Yichen (2018). Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3588–3597).

Hu, Han, Zhang, Zheng, Xie, Zhenda, & Lin, Stephen (2019). Local relation networks for image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3464–3473).

Huang, Xiaohong, Deng, Zhifang, Li, Dandan, Yuan, Xueguang, & Fu, Ying (2022). MISSFormer: an effective transformer for 2D medical image segmentation. *IEEE Transactions on Medical Imaging*.

Huang, Zhenyang, Li, Jianjun, Mao, Ning, Yuan, Genji, & Li, Jinjiang (2024). DBEF-net: Diffusion-based boundary-enhanced fusion network for medical image segmentation. *Expert Systems with Applications*, Article 124467.

Islam, Md Rabiul, Qaraqe, Marwa, & Serpedin, Erchin (2024). CoST-UNet: Convolution and swin transformer based deep learning architecture for cardiac segmentation. *Biomedical Signal Processing and Control*, *96*, Article 106633.

Jha, Debesh, Riegler, Michael A, Johansen, Dag, Halvorsen, Pål, & Johansen, Håvard D (2020). Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd international symposium on computer-based medical systems* (pp. 558–564). IEEE.

Lei, Baiying, Xia, Zaimin, Jiang, Feng, Jiang, Xudong, Ge, Zongyuan, Xu, Yanwu, et al. (2020). Skin lesion segmentation via generative adversarial networks with dual discriminators. *Medical Image Analysis*, *64*, Article 101716.

Li, Guangju, Huang, Qinghua, Wang, Wei, & Liu, Longzhong (2024). Selective and multi-scale fusion mamba for medical image segmentation. *Expert Systems with Applications*, Article 125518.

Li, Zihan, Zheng, Yuan, Shan, Dandan, Yang, Shuzhou, Li, Qingde, Wang, Beizhan, et al. (2024). ScribFormer: Transformer makes CNN work better for scribble-based medical image segmentation. *IEEE Transactions on Medical Imaging*.

Liu, Ze, Lin, Yutong, Cao, Yue, Hu, Han, Wei, Yixuan, Zhang, Zheng, et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Muthukrishnan, Ranjan, & Radha, Miyilsamy (2011). Edge detection techniques for image segmentation. *International Journal of Computer Science & Information Technology*, *3*(6), 259.

Oktay, Ozan, Schlemper, Jo, Folgoc, Loic Le, Lee, Matthew, Heinrich, Mattias, Misawa, Kazunari, et al. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Rao, Yongming, Zhao, Wenliang, Tang, Yansong, Zhou, Jie, Lim, Ser Nam, & Lu, Jiwen (2022). Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Advances in Neural Information Processing Systems*, *35*, 10353–10366.

Ronneberger, Olaf, Fischer, Philipp, & Brox, Thomas (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* (pp. 234–241). Springer.

Ruan, Jiacheng, Xiang, Suncheng, Xie, Mingye, Liu, Ting, & Fu, Yuzhuo (2022). Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In *2022 IEEE international conference on bioinformatics and biomedicine* (pp. 1150–1156). IEEE.

Shen, Zhuoran, Zhang, Mingyuan, Zhao, Haiyu, Yi, Shuai, & Li, Hongsheng (2021). Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3531–3539).

Su, Run, Zhang, Deyun, Liu, Jinhuai, & Cheng, Chuandong (2021). Msu-net: Multi-scale u-net for 2d medical image segmentation. *Frontiers in Genetics*, *12*, Article 639930.

Tang, Hao, Cheng, Lianglun, Huang, Guoheng, Tan, Zhengguang, Lu, Junhao, & Wu, Kaihong (2024). Rotate to scan: Unet-like mamba with triplet SSM module for medical image segmentation. arXiv preprint arXiv:2403.17701.

Tang, Fenghe, Wang, Lingtao, Ning, Chunping, Xian, Min, & Ding, Jianrui (2023). Cmu-net: a strong convmixer-based medical ultrasound image segmentation network. In *2023 IEEE 20th international symposium on biomedical imaging* (pp. 1–5). IEEE.

Torres-Sospedra, Joaquín, Montoliu, Raúl, Trilles, Sergio, Belmonte, Óscar, & Huerta, Joaquín (2015). Comprehensive analysis of distance and similarity measures for Wi-Fi fingerprinting indoor positioning systems. *Expert Systems with Applications*, *42*(23), 9263–9278.

Ugarriza, Luis Garcia, Saber, Eli, Vantaram, Sreenath Rao, Amuso, Vincent, Shaw, Mark, & Bhaskar, Ranjit (2009). Automatic image segmentation by dynamic region growth and multiresolution merging. *IEEE Transactions on Image Processing*, *18*(10), 2275–2288.

Valanarasu, Jeya Maria Jose, Oza, Poojan, Hacihaliloglu, Ilker, & Patel, Vishal M (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *Medical image computing and computer assisted intervention–mICCAI 2021: 24th international conference, strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24* (pp. 36–46). Springer.

Valanarasu, Jeya Maria Jose, Sindagi, Vishwanath A, Hacihaliloglu, Ilker, & Patel, Vishal M (2020). Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations. In *Medical image computing and computer assisted intervention–mICCAI 2020: 23rd international conference, lima, peru, October 4–8, 2020, proceedings, Part IV 23* (pp. 363–373). Springer.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, et al. (2017). Attention is all you need. In *Advances in neural information processing systems: vol. 30*.

Wang, Jinfeng, Huang, Qiming, Tang, Feilong, Meng, Jia, Su, Jionglong, & Song, Sifan (2022). Stepwise feature fusion: Local guides global. In *International conference on medical image computing and computer-assisted intervention* (pp. 110–120). Springer.

Wu, Huisi, Chen, Shihuai, Chen, Guilian, Wang, Wei, Lei, Baiying, & Wen, Zhenkun (2022). FAT-net: Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis*, *76*, Article 102327.

Wu, Wangyu, Dai, Tianhong, Chen, Zhenhong, Huang, Xiaowei, Ma, Fei, & Xiao, Jimin (2024a). APC: Adaptive patch contrast for weakly supervised semantic segmentation. arXiv preprint arXiv:2407.10649.

Wu, Wangyu, Dai, Tianhong, Huang, Xiaowei, Ma, Fei, & Xiao, Jimin (2024b). Image augmentation with controlled diffusion for weakly-supervised semantic segmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6175–6179). IEEE.

Wu, Wangyu, Dai, Tianhong, Huang, Xiaowei, Ma, Fei, & Xiao, Jimin (2024c). Top-k pooling with patch contrastive learning for weakly-supervised semantic segmentation. *IEEE SMC*.

Wu, Renkai, Liang, Pengchen, Huang, Xuan, Shi, Liu, Gu, Yuandong, Zhu, Haiqin, et al. (2024e). MHorUNet: High-order spatial interaction unet for skin lesion segmentation. *Biomedical Signal Processing and Control*, *88*, Article 105517.

Xiao, Xiao, Lian, Shen, Luo, Zhiming, & Li, Shaozi (2018). Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th international conference on information technology in medicine and education* (pp. 327–331). IEEE.

Xiao, Tete, Liu, Yingcheng, Zhou, Bolei, Jiang, Yuning, & Sun, Jian (2018). Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision* (pp. 418–434).

Xu, Qing, Ma, Zhicheng, Na, H. E., & Duan, Wenting (2023). DCSAU-net: A deeper and more compact split-attention U-net for medical image segmentation. *Computers in Biology and Medicine*, *154*, Article 106626.

Xu, Jingchao, Wang, Xin, Wang, Wei, & Huang, Wendi (2023). PHCU-net: A parallel hierarchical cascade U-net for skin lesion segmentation. *Biomedical Signal Processing and Control*, *86*, Article 105262.

Yin, Jianjian, Yan, Shuai, Chen, Tao, Chen, Yi, & Yao, Yazhou (2024). Class probability space regularization for semi-supervised semantic segmentation. *Computer Vision and Image Understanding*, *249*, 104146.

Yin, Jianjian, Zheng, Zhichao, Gu, Yanhui, Zhou, Junsheng, & Chen, Yi (2023). Class-level multiple distributions representation are necessary for semantic segmentation. arXiv preprint arXiv:2303.08029.

Yin, Jianjian, Zheng, Zhichao, Pan, Yulu, Gu, Yanhui, & Chen, Yi (2023). Semi-supervised semantic segmentation with multi-reliability and multi-level feature augmentation. *Expert Systems with Applications*, *233*, Article 120973.

Zhang, Yundong, Liu, Huiye, & Hu, Qiang (2021). Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical image computing and computer assisted intervention–mICCAI 2021: 24th international conference, strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24* (pp. 14–24). Springer.

Zhao, Hengshuang, Shi, Jianping, Qi, Xiaojuan, Wang, Xiaogang, & Jia, Jiaya (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881–2890).

Zhou, Zongwei, Rahman Siddiquee, Md Mahfuzur, Tajbakhsh, Nima, & Liang, Jianming (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* (pp. 3–11). Springer.

Zhou, Zongwei, Siddiquee, Md Mahfuzur Rahman, Tajbakhsh, Nima, & Liang, Jianming (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, *39*(6), 1856–1867.