scientific reports

Check for updates

OPEN

Medical image segmentation by combining feature enhancement Swin Transformer and UperNet

Lin Zhang[1], Xiaochun Yin[1]✉, Xuqi Liu[1] & Zengguang Liu[2]

Medical image segmentation plays a crucial role in assisting clinical diagnosis, yet existing models often struggle with handling diverse and complex medical data, particularly when dealing with multi-scale organ and tissue structures. This paper proposes a novel medical image segmentation model, FE-SwinUper, designed to address these challenges by integrating the strengths of the Swin Transformer and UPerNet architectures. The objective is to enhance multi-scale feature extraction and improve the fusion of hierarchical organ and tissue representations through a feature enhancement Swin Transformer (FE-ST) backbone and an adaptive feature fusion (AFF) module. The FE-ST backbone utilizes self-attention mechanisms to efficiently extract rich spatial and contextual features across different scales, while the AFF module adapts to multi-scale feature fusion, mitigating the loss of contextual information. We evaluate the model on two publicly available medical image segmentation datasets: Synapse multi-organ segmentation dataset and the ACDC cardiac segmentation dataset. Our results show that FE-SwinUper outperforms existing state-of-the-art models in terms of Dice coefficient, pixel accuracy, and Hausdorff distance. The model achieves a Dice score of 91.58% on the Synapse dataset and 90.15% on the ACDC dataset. These results demonstrate the robustness and efficiency of the proposed model, indicating its potential for real-world clinical applications.

Keywords Medical image, Semantic segmentation, Deep learning, Feature enhancement

Medical image segmentation is a pivotal and challenging research problem that encompasses a variety of clinical applications, including polyp segmentation, lesion segmentation, and cell segmentation[1] As a fundamental yet complex step in medical image processing and analysis, it plays a critical role in computer-aided clinical diagnosis systems. By semi-automatically or automatically delineating anatomically or pathologically significant regions within medical images, segmentation enables the extraction of relevant features and provides a reliable basis for clinical diagnosis and pathological research, thus assisting physicians in making more accurate assessments[2]. However, medical images often contain artifacts, noise, and distortions introduced by imaging devices. In addition, organ or lesion boundaries may be blurred, and internal intensity distributions may be uneven, presenting significant challenges for accurate segmentation[3].

In traditional medical image segmentation methods, thresholding segmentation is a commonly used technique that has gained widespread attention due to its advantages such as real-time processing, effectiveness, automation, and wide applicability. The main idea behind this method is that different objects exhibit distinct characteristics, such as color, grayscale, or contour. By selecting a specific threshold based on these subtle differences, the method divides the objects from the background, achieving fast image segmentation. Typical thresholding algorithms include Otsu's Thresholding[4], Minimum Cross-Entropy Thresholding[5], K-means Thresholding[6], and Fuzzy Thresholding[7]. However, thresholding segmentation has some drawbacks. It performs poorly when the grayscale values or other feature differences between objects and the background are not obvious. Additionally, it does not account for the spatial information of the pixels, and noise or inhomogeneity can lead to artifacts, which may result in incorrect segmentation. In recent years, improved thresholding methods have significantly alleviated these issues. Jumiawi et al.[8] used heterogeneous mean filters to handle the poor quality of the images when estimating the mean value for Otsu's between-class variance. Then they proposed the utilization of a Maximum Likelihood Estimation of Gumbel's distribution or Extreme Value Type I distribution for the objective function of an MCET[9]. The goal is to introduce a dedicated image-thresholding model designed to enhance the accuracy and efficiency of image-segmentation tasks.

[1]College of Computer Science, Weifang University of Science and Technology, Weifang 262700, China. [2]School of Information Engineering, Shandong Vocational College of Science and Technology, Weifang 261053, China. ✉email: xiaochunyin@wfust.edu.cn

Since the introduction of U-Net[10], convolutional neural networks (CNNs) have become the predominant method for medical image segmentation[11–16]. Despite their popularity, CNNs inherently struggle to establish long-range dependencies and global contextual relationships due to the limited receptive-field imposed by convolutional operations. To alleviate this issue, numerous studies have attempted to expand the receptive-field and thereby improve the capacity for context modeling. For example, Gao et al.[17] employed dilated convolutions with adjustable dilation rates throughout the backbone, effectively tuning the field-of-view and achieving performance competitive with state-of-the-art methods. Peng et al.[18] introduced large kernels to capture richer global context, while Zhao et al.[19] aggregated multi-scale global information through pyramid pooling. Palash et al.[20] proposed an automatic and adaptive convolutional neural network architecture which has the capability to choose the local as well as the global features from the image data automatically using the interacting subpaths of different lengths available. Lei et al.[21] proposed the ConDSeg framework for medical image segmentation, featuring a Semantic Information Decoupling module to reduce uncertainty, a Contrast-Driven Feature Aggregation module for foreground-background differentiation, and a Size-Aware Decoder to accurately segment entities of varying sizes, improving overall segmentation performance. Additionally, Wang et al.[22] proposed non-local operations to capture remote dependencies, typically integrating them at the end of the encoder. Although these approaches enhance context modeling to some extent, they remain constrained by the CNN architecture, limiting their ability to fully establish global contextual relationships.

Transformer is a sequence-to-sequence prediction framework known for its robust modeling of long-range dependencies, making it highly effective in machine translation and natural language processing[23]. Its self-attention mechanism dynamically adjusts the receptive-field based on input content, thereby establishing global connections among sequential tokens more effectively than convolutional operations can for long-term dependencies. Recently, the Transformer has emerged as a compelling alternative to CNNs, achieving competitive performance in various computer vision tasks, including image recognition[24,25], semantic/instance segmentation[26,27], object detection[28,29], and image generation[30]. Vision Transformer (ViT) was the first image recognition model fully based on the Transformer architecture, demonstrating performance on par with state-of-the-art convolution-based methods. TransUNet[31] leverages CNNs for feature extraction and then employs a Transformer for long-range contextual modeling, while TransFuse[32] attempts to integrate features extracted by both Transformer and CNN. To address computational complexity and effectively integrate multi-scale, multi-level features with Transformer-based representations, Liu et al.[33] proposed the Swin Transformer. This hierarchical design utilizes window-based and shifted window-based multi-head attention mechanisms, surpassing previous state-of-the-art approaches in image classification and dense prediction tasks (such as object detection and semantic segmentation). Rohit et al.[34] proposed a novel biomedical image segmentation model using two parallel encoders and a dual-channel decoder. The decoder comprises a hierarchy of Attention-gated Swin Transformers with a fine-tuning strategy. The hierarchical Attention-gated Swin Transformers implements a multi-scale, multi-level feature embedding strategy that captures short and long-range dependencies and leverages the necessary features without increasing computational load. These advances highlight the tremendous potential of Transformer architectures in medical image segmentation. However, despite the promising performance of the Swin Transformer, the presence of self-attention still poses challenges when processing multi-scale, high-resolution feature maps-both crucial factors in achieving accurate image segmentation.

In this paper, we propose a medical image segmentation model based on feature enhancement Swin Transformer and UperNet (FE-SwinUper) to achieve an excellent segmentation performance in medical images. The proposed model employs a Swin Transformer network, which leverages self-attention, in combination with the UPerNet semantic segmentation framework as foundational architecture. Building upon this base, we introduce two key modules: Feature Enhancement Swin Transformer (FE-ST) module, designed to improve feature extraction, and Adaptive Feature Fusion (AFF) module, aimed at optimizing the feature pyramid. In the first stage, the FE-ST module serves as the backbone network, integrating a CNN to aggregate contextual information before and after perception within the Swin Transformer. By emphasizing inter-channel information interaction, supplemented by local spatial information exchange, the module exploits self-attention to achieve scale fusion based on visual correlations. Unlike existing feature enhancement techniques that typically rely solely on self-attention mechanisms or convolutional architectures, the FE-Swin module combines the advantages of both: it utilizes the self-attention mechanism to model long-range dependencies while leveraging CNNs to capture local and contextual consistency. This design enhances the model's robustness to noise and intensity distribution issues in medical images, while improving the spatial-to-channel feature representation, thereby making better use of organ-related information. Next, UPerNet, augmented with the AFF and Pyramid Pooling (PPM) modules, fuses multi-level features to enable effective semantic segmentation of organs and tissues in medical images. The AFF module proposes a selective multi-scale feature fusion strategy based on learnable weights, building upon existing adaptive fusion techniques. Unlike traditional feature fusion methods that typically use fixed or heuristic fusion weights, the AFF module dynamically adjusts the fusion weight of each feature layer based on contextual importance. This adaptive strategy ensures that high-level features in the feature pyramid are efficiently combined with their adjacent shallow features, enabling the model to focus on organ and tissue information across multiple scales in medical images. Furthermore, by integrating with PPM, the AFF module further enhances the segmentation capability for complex anatomical structures, effectively alleviating issues related to boundary blurring and intensity variations.

The proposed FE-ST module and AFF module combine state-of-the-art techniques in feature enhancement and adaptive fusion. While the concept of feature enhancement using attention mechanisms and multi-scale fusion has been explored in previous works, our FE-ST module integrates both convolutional and transformer architectures to capture both local and long-range dependencies more effectively. The AFF module introduces a novel dynamic weighting mechanism for multi-scale feature fusion, which differs from traditional static weight approaches in prior fusion techniques. Together, these modules represent a significant enhancement over

existing methods by improving segmentation accuracy in medical imaging tasks. The main contributions of our work are as follows:

1. We propose a FE-ST network as the backbone network to extract image feature information. This module not only has the excellent spatial feature information processing ability of Swin Transformer, but also uses CNN to enhance the correlation between feature mapping channels. It effectively suppresses the problem of insufficient feature extraction caused by artifacts and noise of medical images, obtains more significant feature information at different scales, and enhances the transmission ability of feature information.
2. We have constructed an AFF module, which allows the shallow feature information in the feature pyramid to selectively and adaptively fuse into the adjacent high-level feature information. The idea of learnable weights and near fusion reduces the huge information difference between the low-level and high-level features, and alleviates the problem of distraction in feature mapping.
3. We combine the FE-ST network with the UPerNet based on AFF and PPM to build the FE-SwinUper medical image segmentation model. Extensive experiments across two typical tasks for medical image segmentation show that the proposed FE-SwinUper consistently outperforms previous state-ofthe-art methods, which demonstrates the effectiveness of our method.

The rest of this paper is arranged as follows: "The proposed model" describes the proposed network. "Loss function" discusses the loss function of the proposed algorithm. "Experiment" analyses the experimental results of the network proposed in this paper and compares them with other algorithms. Finally, "Conclusion" gives conclusions about this paper.

The proposed model

This paper proposes a medical image segmentation model called FE-SwinUper based on feature enhancement Swin transformer and adaptive feature fusion UperNet. Firstly, combining the advantages of CNN structure and Swin Transformer, FE-ST is innovatively proposed as a new backbone network. Second, the FPN structure in UperNet is replaced with AFF to reconstruct the multi-scale feature pyramid. The overall framework is shown in Fig. 1.

In Fig. 1, the left side is the FE-ST backbone network, which consists of four stages. Each stage consists of a different number of Swin Transformer blocks and feature enhancement modules. The feature maps of different scales from each stage, as the input of the AFF module, ensure the effective global context prior representation, and at the same time, carry out the adaptive fusion of multi-scale hierarchical features. The PPM module is inserted between the output of the last stage of the backbone network and the first layer of the AFF module to ensure that the receptive field of the deep network is not missing, and to ensure effective global context prior

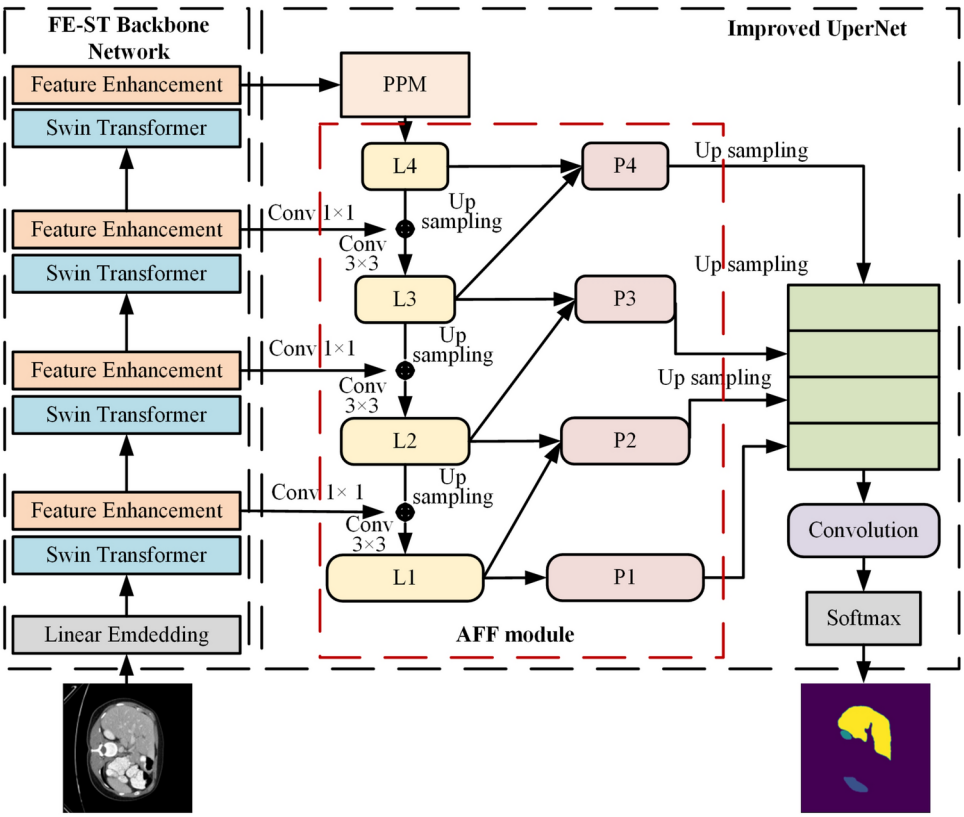

Figure 1. Feature enhancement Swin Transformer and UPerNet segmentation network.

representation. After further fusion of multi-scale hierarchical features from AFF, after size reduction of up sampling and convolution operation, a fusion feature map of the same size as the input image is generated, and the final semantic segmentation result is obtained after Softmax classifier.

Feature enhancement Swin transformer backbone network

In this paper, in order to obtain a feature map with more abundant organ information and transfer the feature information to the deeper level of the model, we propose a feature enhanced Swin transformer (FE-ST) network as the backbone network for feature extraction The structure of FE-ST is shown in Fig. 2. FE-ST establishes the association between image features based on Transformer idea, hierarchical structure and window multi-head self-attention mechanism. First, it uses skip-connections to fuse the context information before and after the Swin Transformer block at each stage. Secondly, it enhances the interaction between channels of fused information, optimizes the feature extraction ability, and has better organ and tissue detection accuracy in medical images.

In Fig. 2, the FE-ST backbone network consists of Swin Transformer blocks and feature enhancement modules. Based on the hierarchical design of Swin Transformer, the feature enhancement module is introduced in each feature extraction stage. The feature maps of each stage are further enhanced to generate more expressive and informative feature maps as output. At the same time, the output of each stage of FE-ST is used as the input of the next stage to obtain more advanced semantic information and enhance the whole backbone network. Since Swin Transformer performs self-attention operation at each stage and the information between each dimension is relatively independent, a feature enhancement module is added to the backbone network to aggregate the context information of different perceptions before and after the Swin Transformer block, and use CNN to enhance the interaction of channel information, further integrate channel and spatial information, and enhance the representation ability of the model.

The proposed feature enhancement module fuses the feature maps before and after feature extraction by using jump links. Then, they are fused in equal proportion with information interaction between channels as the main method and local spatial information interaction as the supplement. The convolution layer and activation function are used for local spatial information interaction to enhance local perception and obtain a larger receptive-field. Channel information interaction is performed through point-by-point convolution at each spatial location, and cross channel information aggregation is performed for each patch. Finally, the feature extraction ability is enhanced by fusing the two by weighting, so that the model has better expression ability. The structure of the feature enhancement module is shown in Fig. 3.

In Fig. 3, the feature maps before and after the Swin Transformer block are represented as X and Y, respectively. After the two feature maps are fused, channel information integration and spatial information integration are performed. In the channel information integration, pointwise convolution is used, allowing the interaction of the pointwise channel information at each spatial location. The output is denoted as $C(X) \in R^{C\times H\times W}$, as defined in Eq. (1). Besides channel information, for spatial information, each point in a single channel is integrated with its neighboring points simultaneously. The output is denoted as $S(X) \in R^{C\times H\times W}$, as defined in Eq. (2).

$$C(X) = LN(pwconv_2(\delta(LN(pwconv_1(X+Y))))), \tag{1}$$

$$S(X) = LN(conv_2(\delta(LN(conv_1(X+Y))))), \tag{2}$$

where, $pwconv_1(X) \in R^{C/4\times H\times W}$ (with 4 being the channel reduction ratio) denotes dimension reduction, and $pwconv_2(X) \in R^{C\times H\times W}$ denotes dimension expansion. The symbol δ represents the *ReLU* activation function, and *LN* denotes layer normalization. The kernel sizes of $Conv_1$ and $Conv_2$ are $(C/4)\times C\times 3\times 3$ and $C\times(C/4)\times 3\times 3$, respectively. The calculated $C(X)$ and $S(X)$ maintain the same dimensions as the input feature maps, both of which can preserve details from the original features to varying degrees. We denote by Z the resulting feature map after channel and spatial feature weighted fusion. The weights for this fusion are obtained via sigmoid activation on $(X+Y)$, denoted as $m(X+Y) \in R^{C\times H\times W}$. In the weighted fusion

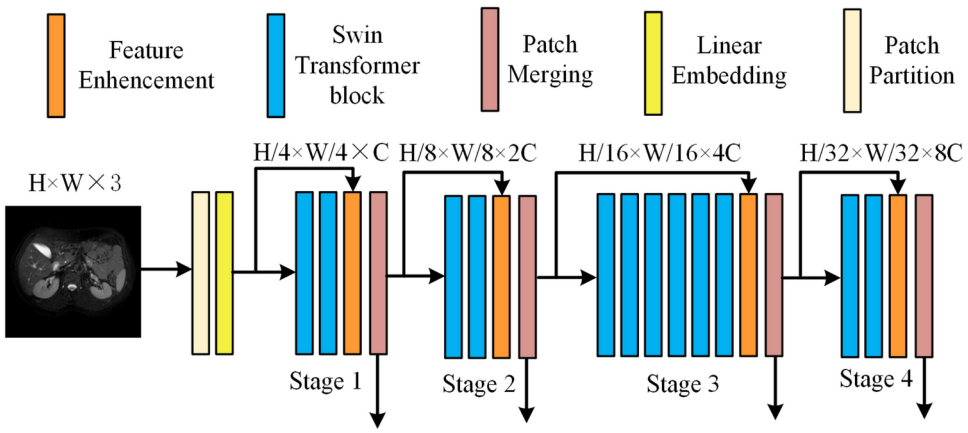

Figure 2. Feature enhanced Swin Transformer backbone network.

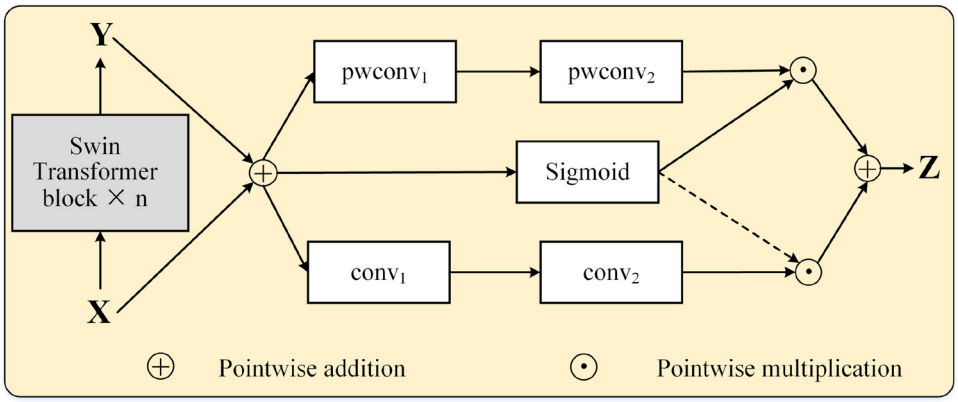

Figure 3. Feature enhancement module.

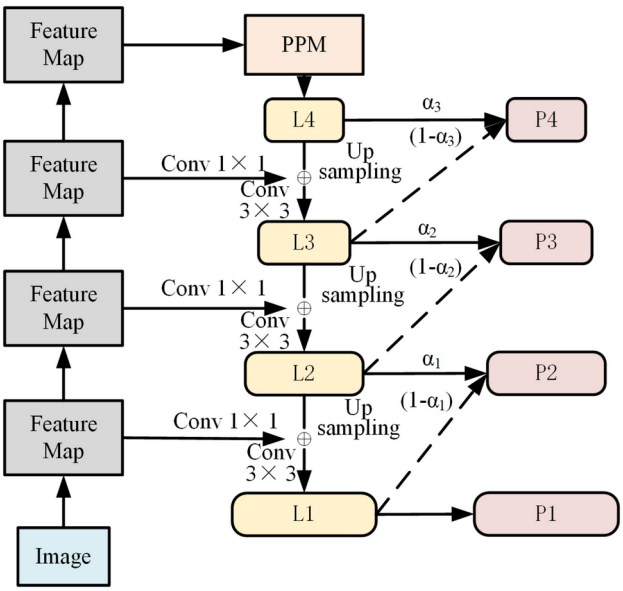

Figure 4. Adjacent feature fusion module.

scheme, the sum of the feature map weights is constrained to 1, and $\otimes$ represents element-wise multiplication. The above computation process is given by Eq. (3).

$$Z = m(X + Y) \otimes C(X) + (1 - m(X + Y)) \otimes S(X). \tag{3}$$

Adaptive feature fusion modual

In order to reduce the loss of feature information and the distraction of feature information attention, and make target information get attention in feature maps of different scales, the original feature pyramid network (FPN) module in UperNet is optimized in two aspects, and an adaptive feature fusion module is proposed. Its main structure is shown in Fig. 4. On the one hand, the features of adjacent layers are enhanced from bottom to top, while the bottom-up enhancement only combines the feature information fusion between the current layer and its adjacent shallow layers. There is no association outside the adjacent feature layers, and the fused content is relatively independent. On the other hand, the interlayer weighted fusion with learnable adaptive weights is used to select fusion, so as to obtain excellent fusion results. The AFF module combines the advantages of FPN and PANet, and the upward fusion of adjacent layers avoids the problem of large semantic information gap between multi-layers.

In Fig. 4, the results obtained from the initial feature pyramid can be expressed as {L1, L2, L3, L4}. For L2, L3 and L4, the high-level feature mapping is integrated with its adjacent feature mapping, while the shallow feature {L1, L2, L3} is expressed as {L1', L2', L3'} through the unified down sampling ratio. Then learnable weights α_1, α_2, and α_3 are generated using L2, L3, and L4 layers. Each weight is learned independently to form an adaptive fusion parameter for each feature mapping. Finally, the two adjacent feature mapping layers are multiplied by the

corresponding learnable weights, and the results are accumulated. The final feature mapping of the output result is represented by {P1, P2, P3, P4}, and the calculation process is as follows:

$$P1 = L1, \tag{4}$$

$$P2 = \alpha_1 \times L2 + (1 - \alpha_1) \times L1', \tag{5}$$

$$P3 = \alpha_2 \times L3 + (1 - \alpha_2) \times L2', \tag{6}$$

$$P4 = \alpha_3 \times L4 + (1 - \alpha_3) \times L3'. \tag{7}$$

It should be noted that the extension of adjacent fusion does not operate on L1. The sum of the weights used for the mapping of two adjacent features is controlled to 1 to ensure the stability of model training.

Pyramid pooling module

The pyramid pooling module is composed of a group of pooling blocks with different scales, which can better use the global image level prior knowledge to understand complex scenes, extract features with global context information to improve image recognition or segmentation results, and is an effective global context prior model. The structure of PPM is shown in Fig. 5.

In Fig. 5, PPM has four different pyramid scales. The input feature maps are first pooled to different target sizes, and the sizes of each layer are 1×1, 2×2, 3×3 and 6×6. The multi-scale pooling method can retain global context information at different scales. Then, in order to maintain the weight of the global characteristics, 1×1 convolution is performed on the pooled results, and the number of channels is reduced to the original $1/N$, where N is 4. Then, the low dimensional feature map is directly upsampled by bilinear interpolation to obtain the feature map with the same size as the original feature map, and then the original feature map and the feature map obtained by upsampling are spliced according to channel dimensions. The number of channels obtained is twice that of the original feature map. Finally, 1×1 convolution is used to reduce the number of channels to the original number of channels. Finally, a feature map with the same size and number of channels as the original feature map is obtained as the output of the pyramid pooling module.

As a hierarchical global prior structure, PPM can further reduce the loss of context information between different scales and different sub regions, and can construct global scene prior information on the final layer feature map of deep neural network.

The information processing flow of FE-SwinUper

The information processing flow of FE-SwinUper is as follows:

(1) **Input**

The image with the size of $H \times W \times 3$ is input into the FE-SwinUper. First, the image patches are segmented through the patch segmentation layer, and then through the mapping of the linear embedding layer, the dimension changes $(H/4) \times (W/4) \times C$.

(2) **FE-ST feature extraction**

The feature map with the size of $(H/4) \times (W/4) \times C$ undergoes four stages of feature extraction. Each stage contains a different number of Swin Transformer blocks, which are [2, 2, 6, 2] respectively. Each stage includes different numbers of Swin Transformer blocks with the same structure, a feature enhancement module, and a patch merge layer. Each Swin Transformer block contains two Transformer layers, the former is composed of LN layer, W-MSA module, residual connection and 2-layer MLP with GELU nonlinearity, and the latter is replaced by SW-MSA module. At the end of each phase, a feature enhancement module is cascaded, and the feature layers

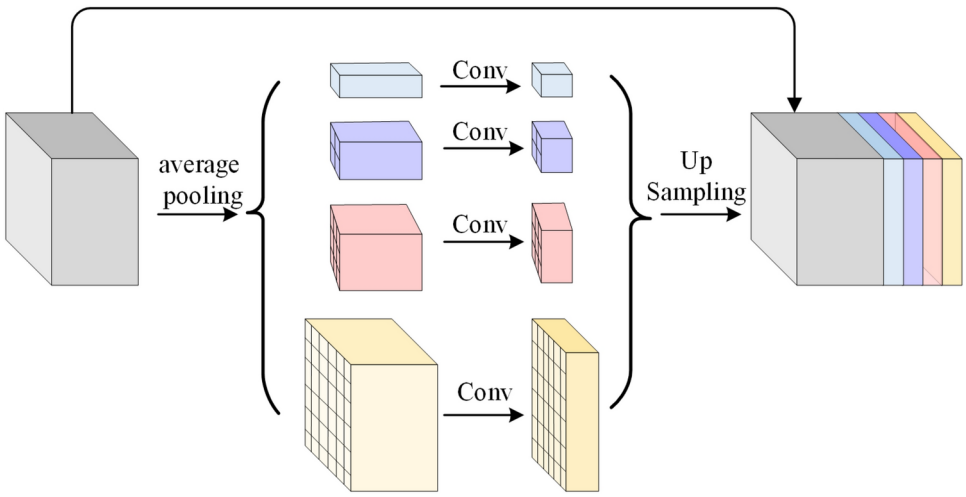

Figure 5. Pyramid pooling module.

of the first and last Swin Transformer blocks in this phase are fused with context information and enhanced with features.

After 4 stages, we obtained a multi-scale feature map with dimensions of $[(H/8) \times (W/8), (H/16) \times (W/16), (H/32) \times (W/32)]$ and channel dimensions of [2*C*, 4*C*, 8*C*]. The multi-scale feature output of each stage is used as the corresponding input of different levels of corresponding AFF.

(3) **Pyramid Pooling Module**

The last layer output from FE-ST is used as the input of PPM module. Firstly, the feature maps are pooled to different target scales on an average basis. The dimensions of each layer are $1 \times 1, 2 \times 2, 3 \times 3$ and 6×6 respectively. Then, the pooled results are convolved by 1×1, and the number of channels is reduced to the original 1/*N*, where *N* is 4. Then, the feature map with low-dimension is upsampled by bilinear interpolation to obtain the feature map with the same size as the original input, and the original feature map and the feature map obtained by upsampling are spliced according to channel dimensions. Finally, use 3×3 convolution to reduce the number of channels to 256 as the top-level feature map of the next AFF module.

(4) **AFF module**

The multi-scale feature outputs from each stage of the FE-ST network serve as inputs to the corresponding levels of the AFF module. Proceeding from top to bottom, each feature map in the AFF is first upsampled by a factor of two using bilinear interpolation. Subsequently, the corresponding feature map from the FE-ST at the same scale is processed via a 1×1 convolutional layer to maintain consistent channel dimensions before being added element-wise to the upsampled features. The fused feature map then undergoes a 3×3 convolution to produce the next-level feature map in the AFF, resulting in the multi-scale features {L1, L2, L3, L4}. Next, learnable weights α_1, α_2, and α_3 are derived from the L2, L3, and L4 feature layers. Finally, each pair of adjacent feature maps is multiplied by its corresponding learnable weight, and the results are summed. The final output feature maps are denoted as {P1, P2, P3, P4}.

(5) **Feature fusion**

The feature maps with different scales obtained from AFF are individually upsampled via bilinear interpolation to a unified spatial resolution. These upsampled feature maps are then concatenated along the channel dimension, followed by a 3×3 convolution to reduce their dimensionality to 256 channels. Finally, the resulting feature map is further upsampled via bilinear interpolation to the original input image size $(H \times W)$ and fed into the output layer.

(6) **Output**

At the output layer, the $(H \times W \times 256)$ feature map is passed through a Softmax classifier, yielding an $(H \times W)$ map of *n*-dimensional vectors, where each dimension corresponds to a segmentation class.

Loss function

In the training phase, the FE-SwinUper model is trained end-to-end using an objective function. Constructed by Sorensen-Dice loss and binary cross-entropy function, and the Softmax function is used on the final feature mapping to achieve pixel classification. The calculation formula is:

$$\mathcal{L}_{BCE} = \sum_{i=1}^{t} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)), \tag{8}$$

$$\mathcal{L}_{Dice} = 1 - \frac{\sum_{i=1}^{t} y_i p_i + \varepsilon}{\sum_{i=1}^{t} y_i + p_i + \varepsilon}, \tag{9}$$

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{BCE} + \beta \cdot \mathcal{L}_{Dice}, \tag{10}$$

where, *t* is the total number of pixels in each image, *y* represents the basic true value of the *i*-th pixel, and p_i represents the confidence score of the *i*-th pixel in the prediction result. In the experiment, $\alpha = \beta = 0.5$ and $\varepsilon = 10^{-6}$.

Experiment

In this paper, the method's effectiveness is verified by two datasets, Synapse and ACDC. Our approach is also compared with other state-of-the-art medical image segmentation algorithms.

Dataset

Synapse multi-organ segmentation dataset (Synapse)[31]: it is a widely used multi-organ segmentation dataset, specifically designed for abdominal CT scans. It contains a total of 30 cases, each featuring 3779 axial abdominal clinical CT images. The dataset is divided into 18 training samples and 12 testing samples. The goal of the dataset is to provide images of four main abdominal organs: the spleen, left kidney, right kidney, and liver. The images are provided with pixel-wise ground truth labels, which allow for evaluation of segmentation performance. The Synapse dataset is particularly challenging because it contains medical images with high intra-patient variability, including varying organ sizes, shapes, and different levels of contrast. Additionally, the presence of noise, artifacts, and varying slice thicknesses adds complexity to the segmentation task. Despite these challenges, the dataset is essential for evaluating methods targeting multi-organ segmentation in abdominal CT imaging, making it an ideal choice for testing FE-SwinUper's performance.

Automated cardiac diagnosis challenge dataset (ACDC)[35]: it is another prominent medical image segmentation dataset, specifically used for cardiac MRI segmentation. It includes left ventricle (LV), right

ventricle (RV), and myocardium (MYO) labels for each patient. The dataset consists of images from 70 patients, split into training, validation, and test sets (70 training, 10 validation, and 20 test). The images cover a variety of cardiovascular conditions such as ischemic, non-ischemic, myocardial infarction, and healthy cases, providing diversity in patient anatomy and pathology. ACDC is unique in that it captures longitudinal cardiac motion, and hence, segmentation models must not only deal with anatomical variability but also dynamic changes across the cardiac cycle. This dataset challenges segmentation methods to be both precise in structure delineation and robust against variations in the heart's geometry due to disease progression. The ACDC dataset also includes variations in image contrast, resolution, and patient-specific factors that further increase its difficulty. Despite these complexities, ACDC provides a benchmark for evaluating models in cardiac segmentation tasks, which are critical in clinical settings for diagnosing and monitoring cardiovascular diseases.

The evaluation metrics used for performance comparison on the two datasets include the Dice-Similarity coefficient (DSC), mean pixel accuracy (PA), and average Hausdorff Distance (HD), which provide a comprehensive assessment of segmentation accuracy, boundary precision, and shape similarity.

Implementation details

The experiments were implemented using Python 3.6 and PyTorch 1.7.0 on a Linux platform. The hardware configuration includes two Intel Xeon 4214R CPUs and three NVIDIA GeForce RTX 2080Ti (12GB) GPUs. During training, the batch_size was set to 12 and the model was trained for 100 epochs. We employed the AdamW optimizer with momentum terms $b_1 = 0.9$ and $b_2 = 0.95$. The weight decay was set to 0.05, and the base learning rate was initialized at $1.5e-4$. A linear scaling strategy[36] was applied to adjust the learning rate as $lr = base_lr/batch_size$. Additionally, a cosine decay schedule[37] was used for iterative learning rate reduction. To address issues related to limited data diversity, we augmented the original dataset with various data augmentation techniques, including random rotations, random translations, random scaling, and random flipping.

Experimental results and analysis

Experiment results on Synapse dataset

Figure 6 presents some representative visualization results of medical image segmentation using FE-SwinUper. In the figure, yellow areas represent the liver, green areas the spleen, blue areas the right kidney, and cyan areas the left kidney. The first row shows the original abdominal CT images, the second row displays the ground truth segmentation labels, and the third row illustrates the model's predicted segmentation results. As can be seen, the predicted segmentation results closely align with the ground truth.

To demonstrate the effectiveness of the proposed FE-SwinUper, we compare it against several mainstream semantic segmentation methods, including ResNet50+DeepLabV3+[38], ViT+UPerNet[39], TransUNet[40], SUnet[41] and VM-UNet[42]. All comparison models were implemented based on the authors' released code and trained on the same dataset using their default parameter settings. DeepLabV3+ employs a deep convolutional backbone (ResNet50) as the encoder and integrates multi-scale information through a spatial pyramid pooling module with dilated convolutions. A subsequent decoder module further fuses low-level and high-level features to improve boundary accuracy. The ViT+UPerNet model leverages a Vision Transformer (ViT) backbone for multi-level feature extraction, and utilizes a unified perceptual parsing framework built on a FPN and PPM to handle multi-scale contextual information, object detection, and semantic segmentation. TransUNet adopts a hybrid architecture that combines CNNs for extracting fine-grained, high-resolution spatial features with a Transformer network to model global contextual relationships. SUnet is a fully self-attention-based network that

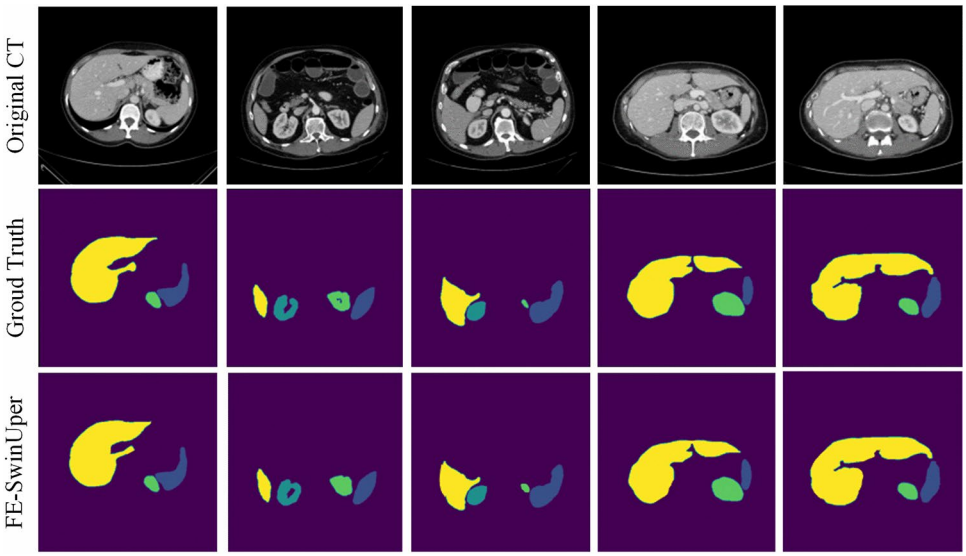

Figure 6. Visualization of segmentation results of FE-SwinUper on the Synapse dataset.

Model	PA (%)	Dice (%)	HD	Parameters(M)	Inference Time(s)
ResNet50+DeepLabV3+[38]	87.24	85.75	29.34	44.5	0.157
ViT+UPernet[39]	90.65	87.57	23.25	40.6	0.246
TransUNet[40]	91.67	89.13	20.58	78	0.309
SUnet[41]	86.23	84.29	13.22	20.9	0.122
VM-UNet[42]	85.58	81.08	19.21	44.27	0.136
FE-SwinUper	92.65	91.58	12.96	55	0.198

Table 1. Segmentation results of compare models on the Synapse dataset.

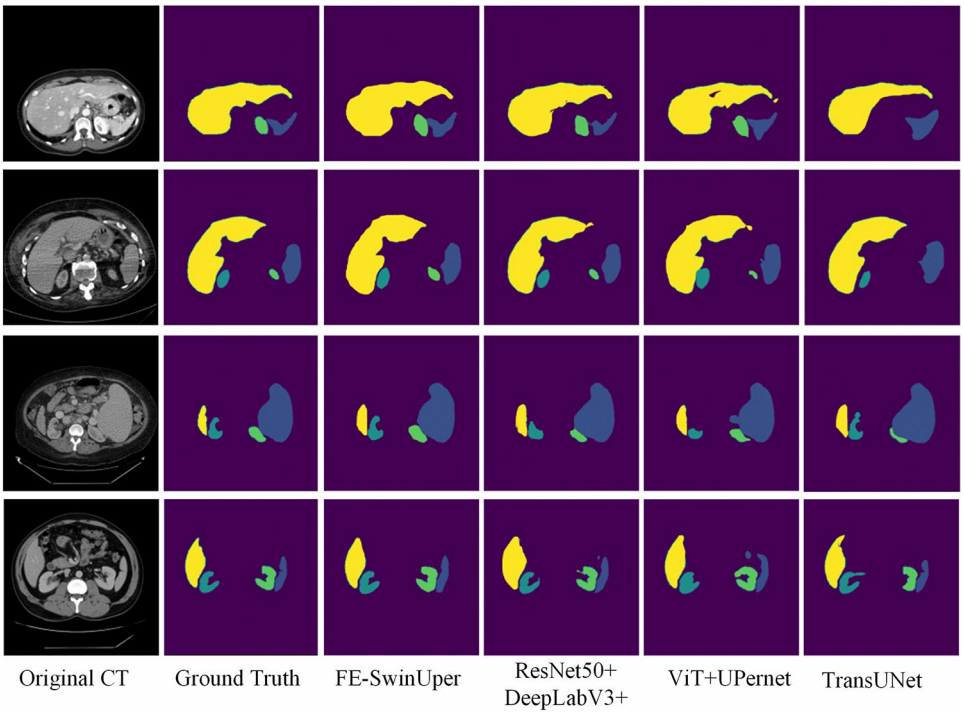

Figure 7. Qualitative comparison of different approaches by visualization on the Synapse dataset.

improves feature extraction and reduces overfitting with its efficient spatial reduction attention module. It also integrates cross-scale features through a multiple attention-based fusion module and enhances semantic features with an attention gate using grouped convolution and residual connections. VM-UNet introduce the Visual State Space block as the foundation block to capture extensive contextual information, and an asymmetrical encoder-decoder structure is constructed with fewer convolution layers to save calculation cost. The results are summarized in Table 1.

Table 1 shows that the proposed FE-SwinUper achieves the best performance metrics among all compared methods, with results of 92.65%, 91.58%, and 12.96, thus verifying the model's effectiveness. These outcomes validate that the hierarchical feature extraction mechanism based on self-attention and the feature enhancement module in FE-SwinUper significantly improve its multi-scale feature extraction capabilities. Additionally, the adjacent feature fusion module enhances the model's ability to integrate multi-scale features. The proposed model does not have an advantage in terms of parameter size and inference time due to the additional convolution-based feature enhancement module, which increases the number of parameters. Moreover, the adaptive fusion module incurs significant computational cost during inference. However, considering the performance improvements achieved, these trade-offs are acceptable. Figure 7 presents the visualization results of organ segmentation on abdominal CT images for different comparison models. From left to right are: the original MRI images, the ground-truth segmentation, the segmentation result of the FE-SwinUper model, the result of the ResNet50+DeepLabV3+ model, the result of the ViT+UPerNet model, and the result of the TransUNet model.

Figure 7 demonstrates that the proposed FE-SwinUper model produces segmentation results more closely aligned with the ground truth. It accurately captures all segmentation targets and yields clearer, smoother boundaries. In contrast, the purely convolution-based ResNet50+ DeepLabV3+ approach shows instances of both under-segmentation and mis-segmentation, particularly for smaller organs such as the kidneys and spleen. Compared to other self-attention-based models like ViT+UPerNet and TransUNet, FE-SwinUper achieves better

Model	PA (%)	Dice (%)	HD
ResNet50+DeepLabV3+[38]	85.35	82.48	30.13
ResNet50+UPernet[39]	88.21	85.87	28.32
TransUNet[40]	87.79	86.12	22.65
DCL-Net[43]	89.30	87.60	25.87
ASF-LKUNet[44]	88.75	89.45	19.24
FE-SwinUper	89.83	90.15	18.29

Table 2. Segmentation results of compare models on the ACDC dataset.

		Dice			
FE	AFF	Liver	Left kidney	Right kidney	Spleen
×	×	83.55	82.76	80.19	82.27
✓	×	88.25	87.51	86.38	90.19
×	✓	86.28	84.76	82.19	89.27
✓	✓	96.25	90.5	87.38	92.19

Table 3. Ablation experiment result on the Synapse dataset.

segmentation performance for smaller kidney organs that heavily overlap with the liver, resulting in smoother segmentation boundaries and showcasing superior robustness.

Experiment results on ACDC dataset

Similar to the Synapse dataset, the proposed Swin Unet is trained on the ACDC dataset to perform medical image segmentation, and two recent models are added for comparison. The experimental results are summarized in Table 2. By using the image data of MRI as input, FE-SwinUper is still able to achieve excellent performance with a Dice of 90.15%, which shows that our method has good generalization ability and robustness.

Ablation study

In order to explore the influence of different factors on the model performance, we conducted ablation studies on Synapse dataset. Specifically, feature enhancement module and AFF module are discussed below. The comparison results are shown in Table 3.

Table 3 shows that for FE-SwinUper, the separate addition of feature enhancement (FE) module or AFF module can significantly improve Dice-score. After adding the FE module separately, the recognition Dice-scores of all four organs increased by more than 4.7%, especially for the small and heavily occluded right kidney and spleen, the improvement in indicators was more significant, reaching 6.19% and 7.92% respectively, proving its key role in improving feature extraction ability. The FE module significantly improves the model's ability to capture richer, more informative feature maps. By performing context aggregation between feature maps before and after the Swin Transformer block, the FE module enhances both spatial and channel-wise interactions. This leads to improved localization and delineation of organs, particularly those with complex boundaries, such as kidneys or the spleen.

The separate introduction of the AFF module can also comprehensively improve the segmentation performance of the model on cardiac MRI images, with a Dice-score improvement of over 2%, especially for small target spleens, which increased by 7%, demonstrating its ability to enhance multi-scale feature fusion. The AFF module is indispensable in managing multi-scale feature map fusion, especially in addressing challenges related to hierarchical integration of deep and shallow feature representations. It enables the model to effectively combine features at different levels, ensuring the preservation of shallow fine details while integrating deeper, more abstract representations. Table 3 shows that the AFF module improves segmentation quality by providing a more balanced fusion of low-level and high-level features, especially when segmenting smaller organs or structures.

The joint introduction of FE module and AFF module has greatly improved the performance of the model, especially for the liver with large volume, long boundary and fuzzy, the Dice score has increased by 12.7%, which fully proves the rationality of introducing these two modules. The feature enhancement module can capture more scale features to ensure that low-level features are not lost. The AFF module can effectively fuse features of different scales and levels, improving the model's ability to distinguish and reconstruct fuzzy features.

Discussion

The FE-SwinUper method has shown substantial improvements in segmentation accuracy across two distinct datasets (Synapse and ACDC), suggesting that the model can handle variability in image quality and modality. However, further validation across a broader set of imaging modalities (such as ultrasound or PET scans) is required to ensure the robustness of the model in clinical environments. Additionally, while the proposed method demonstrates excellent segmentation performance, its computational demands may be a limiting factor for real-time deployment in resource-constrained clinical settings. Future work will focus on optimizing the

model for faster inference and exploring data augmentation strategies to mitigate the reliance on large annotated datasets.

Limitations and future work

Although FE-SwinUper shows promising results on CT and MRI datasets (Synapse and ACDC), its performance on other imaging modalities, such as ultrasound, PET, or X-ray, remains untested. The model may face challenges when applied to such modalities due to differences in image characteristics, such as noise levels, resolution, and contrast. The robustness of FE-SwinUper across diverse imaging technologies requires further exploration. The FE-SwinUper model incorporates several complex modules, such as the Swin Transformer backbone and the multi-scale adaptive feature fusion (AFF) module, which may lead to high computational demands in terms of memory usage and processing power. This could pose challenges in real-time clinical applications, especially in settings with limited hardware resources. While the model achieves impressive accuracy, its inference speed may not meet the requirements for real-time applications, such as during surgeries or in emergency departments where rapid decisions are crucial. Future work should focus on optimizing the model to balance segmentation accuracy with fast processing times.

Future work will focus on evaluating FE-SwinUper on a broader range of imaging modalities, including ultrasound, PET scans, and X-ray images. This will allow for a more comprehensive understanding of the model's adaptability and robustness in clinical practice. Further optimization of the FE-SwinUper model, including pruning techniques and quantization, can help reduce its computational requirements and inference time. This would make the model more suitable for clinical environments where real-time predictions are necessary.

Conclusion

In this paper, we propose FE-SwinUper for medical image segmentation. FE-ST and a UperNet enhanced with AFF and PPM serve as the key components of FE-SwinUper. The FE-ST module undertakes multi-scale feature extraction to acquire richer and more informative representations, while the AFF and PPM modules facilitate the adaptive fusion of multi-scale organ and tissue features. Experimental results on the Synapse and ACDC datasets demonstrate that FE-SwinUper outperforms existing methods in terms of Dice coefficient, pixel accuracy, and Hausdorff distance, achieving state-of-the-art performance. These results highlight the potential of FE-SwinUper in improving medical image segmentation tasks, particularly for complex organ structures. The results of the ablation experiment show that the introduction of FE module and AFF module has significantly improved the segmentation performance of the model, with Dice score improving by more than 7.19%, proving the effectiveness of the proposed and combined effects of the two modules. Although the model shows strong segmentation accuracy, its computational cost and inference speed may be limiting factors in time-sensitive clinical settings. Future work will focus on enhancing the model's efficiency, addressing data scarcity issues through techniques like semi-supervised learning, and further validating its performance across additional clinical datasets. The proposed method lays a strong foundation for more robust and accurate medical image segmentation in clinical practice.

Data availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Received: 13 December 2024; Accepted: 7 April 2025
Published online: 25 April 2025

References

1. Qureshi, I. et al. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Inf. Fusion* **90**, 316–352 (2023).
2. Xiao, H., Li, L., Liu, Q., Zhu, X. & Zhang, Q. Transformers in medical image segmentation: A review. *Biomed. Signal Process. Control* **84**, 104791 (2023).
3. Zhang, Y., Shen, Z. & Jiao, R. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine* 108238 (2024).
4. Otsu, N. et al. A threshold selection method from gray-level histograms. *Automatica* **11**, 23–27 (1975).
5. Li, C. H. & Lee, C. Minimum cross entropy thresholding. *Pattern Recogn.* **26**, 617–625 (1993).
6. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press* (1967).
7. Aja-Fernández, S., Curiale, A. H. & Vegas-Sánchez-Ferrero, G. A local fuzzy thresholding methodology for multiregion image segmentation. *Knowl.-Based Syst.* **83**, 1–12 (2015).
8. Jumiawi, W. A. H. & El-Zaart, A. Otsu thresholding model using heterogeneous mean filters for precise images segmentation. In *2022 International Conference of Advanced Technology in Electronic and Electrical Engineering (ICATEEE)*, 1–6 (IEEE, 2022).
9. Jumiawi, W. & El-Zaart, A. Gumbel (evi)-based minimum cross-entropy thresholding for the segmentation of images with skewed histograms. *Appl. Syst. Innov.* **6**, 87 (2023).
10. Azad, R. et al. Medical image segmentation review: The success of u-net. *IEEE Trans. Pattern Anal. Mach. Intell.* (2024).
11. Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P. & Johansen, H. D. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*, 558–564 (IEEE, 2020).
12. Du, G., Cao, X., Liang, J., Chen, X. & Zhan, Y. Medical image segmentation based on u-net: A review. *J. Imaging Sci. Technol.* **64** (2020).
13. Patel, S. An overview and application of deep convolutional neural networks for medical image segmentation. In *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, 722–728 (IEEE, 2023).
14. Xie, L. et al. Deep label fusion: A generalizable hybrid multi-atlas and deep convolutional neural network for medical image segmentation. *Med. Image Anal.* **83**, 102683 (2023).
15. Huang, H. et al. Channel prior convolutional attention for medical image segmentation. *Comput. Biol. Med.* **178**, 108784 (2024).

16. Tang, F. et al. Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5 (IEEE, 2024).
17. Gao, R. Rethinking dilated convolution for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4675–4684 (2023).
18. Peng, C., Zhang, X., Yu, G., Luo, G. & Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4353–4361 (2017).
19. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890 (2017).
20. Ghosal, P. et al. A deep adaptive convolutional network for brain tumor segmentation from multimodal mr images. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, 1065–1070 (IEEE, 2019).
21. Lei, M., Wu, H., Lv, X. & Wang, X. Condseg: A general medical image segmentation framework via contrast-driven feature enhancement. *arXiv preprint* arXiv:2412.08345 (2024).
22. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803 (2018).
23. Gillioz, A., Casas, J., Mugellini, E. & Abou Khaled, O. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on computer science and information systems (FedCSIS)*, 179–183 (IEEE, 2020).
24. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929 (2020).
25. Touvron, H. et al. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357 (PMLR, 2021).
26. Wang, Y. et al. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8741–8750 (2021).
27. Zheng, S. et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890 (2021).
28. Carion, N. et al. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229 (Springer, 2020).
29. Zhu, X. et al. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint* arXiv:2010.04159 (2020).
30. Jiang, Y., Chang, S. & Wang, Z. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Adv. Neural. Inf. Process. Syst.* **34**, 14745–14758 (2021).
31. Chen, J. et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint* arXiv:2102.04306 (2021).
32. Zhang, Y., Liu, H. & Hu, Q. 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*, 14–24 (Springer, 2021).
33. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).
34. Agarwal, R., Ghosal, P., Sadhu, A. K., Murmu, N. & Nandi, D. Multi-scale dual-channel feature embedding decoder for biomedical image segmentation. *Comput. Methods Programs Biomed.* **257**, 108464 (2024).
35. Bernard, O. et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?. *IEEE Trans. Med. Imaging* **37**, 2514–2525 (2018).
36. He, T. et al. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 558–567 (2019).
37. Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V. & Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 113–123 (2019).
38. Chen, H., Qin, Y., Liu, X., Wang, H. & Zhao, J. An improved deeplabv3+ lightweight network for remote-sensing image semantic segmentation. *Complex Intell. Syst.* **10**, 2839–2849 (2024).
39. Ruiping, Y., Kun, L., Shaohua, X., Jian, Y. & Zhen, Z. Vit-upernet: a hybrid vision transformer with unified-perceptual-parsing network for medical image segmentation. *Complex Intell. Syst.* 1–13 (2024).
40. Chen, J. et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **97**, 103280 (2024).
41. Li, X. et al. Sunet: A multi-organ segmentation network based on multiple attention. *Comput. Biol. Med.* **167**, 107596 (2023).
42. Ruan, J. & Xiang, S. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint* arXiv:2402.02491 (2024).
43. Wen, L. et al. Dcl-net: Dual contrastive learning network for semi-supervised multi-organ segmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1876–1880 (IEEE, 2024).
44. Wang, R. et al. Asf-lkunet: Adjacent-scale fusion u-net with large kernel for multi-organ segmentation. *Comput. Biol. Med.* **181**, 109050 (2024).

Acknowledgements

This research was supported in part by the Natural Science Foundation of Shandong Province (Grant number: ZR2021MF086), in part by the Scientific Talent Fund Project of Weifang University of Science & Technology (Grant number: KJRC2021002), and in part by the Key Technologies R&D Program of Weifang (Grant number: 2023GX063 and 2024RKX078).

Author contributions

Lin Zhang conceived the methodology and experiments, and wrote the manuscript. Xiaochun Yin and Zengguang Liu conducted the experiments. Xiaochun Yin and Xuqi Liu analyzed the results. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.Y.