MLSE-Net: Multi-level Semantic Enriched Network for Medical Image Segmentation

Di Gai[1], Heng Luo[3], Jing He[3], Pengxiang Su[3], Zheng Huang[1], Song Zhang[3], and Zhijun Tu[2*]

[1] School of Mathematics and Computer Sciences, Nanchang University, Nanchang, China

[2] School of Information Engineering, Nanchang University, Nanchang, China

[3] School of Software, Nanchang University, Nanchang, China

[e-mail: tuzhijun@ncu.edu.cn]

*Corresponding author: Zhijun Tu

Received January 4, 2023; revised June 28, 2023; accepted August 17, 2023; published September 30, 2023

Abstract

Medical image segmentation techniques based on convolution neural networks indulge in feature extraction triggering redundancy of parameters and unsatisfactory target localization, which outcomes in less accurate segmentation results to assist doctors in diagnosis. In this paper, we propose a multi-level semantic-rich encoding-decoding network, which consists of a Pooling-Conv-Former (PCFormer) module and a Cbam-Dilated-Transformer (CDT) module. In the PCFormer module, it is used to tackle the issue of parameter explosion in the conservative transformer and to compensate for the feature loss in the down-sampling process. In the CDT module, the Cbam attention module is adopted to highlight the feature regions by blending the intersection of attention mechanisms implicitly, and the Dilated convolution-Concat (DCC) module is designed as a parallel concatenation of multiple atrous convolution blocks to display the expanded perceptual field explicitly. In addition, Multi-Head Attention-DwConv-Transformer (MDTransformer) module is utilized to evidently distinguish the target region from the background region. Extensive experiments on medical image segmentation from Glas, SIIM-ACR, ISIC and LGG demonstrated that our proposed network outperforms existing advanced methods in terms of both objective evaluation and subjective visual performance.

Keywords: Transformer, Attention mechanism, Semantic enriched, Medical image segmentation.

1. Introduction

Medical image segmentation is a key step in medical image processing and analysis, which is also a core component of other advanced medical image analysis and interpretation systems. Segmentation of medical images provides the basis and prerequisites for target separation, feature extraction and quantitative measurement of parameters, making higher-level medical image understanding and diagnosis possible. Medical image segmentation has a wide range of applications and research values in medical research, clinical diagnosis, pathological analysis, surgical planning, image information processing, computer-aided surgery and other medical research and practice fields. Deep learning-based medical image segmentation networks can greatly assist doctors in medical image segmentation [1-5].

Traditional manual feature extraction methods are usually based on features such as grayscale values, shapes and textures in the image. However, these features are designed manually using expert knowledge to achieve automatic segmentation of the target region. Manual methods frequently require a large amount of prior knowledge to extract manual features for segmentation, which are valid for segmentation tasks on specific datasets and the segmentation performance is not stable. Fortunately, deep learning-based segmentation methods use the idea of pixel classification, which can automatically extract the semantic information and learn the representation of the data, effectively overcoming the limitations of traditional manual feature segmentation.

The previous medical image segmentation networks are convolution neural networks, which use repeatedly stacked convolution operations for monotonic down sampling. Nevertheless, the inconsistent contextual information leads to coarser segmentation results. To solve this problem, U-shape networks with context-skip connections have been proposed and are constantly being changed and developed. For instance, U-Net [6] is the basis of U-shape networks, which enables some improvement in segmentation accuracy. Attention U-Net [7] achieves a groundbreaking update to U-shape's structure, which combines the attention layer and U-shape networks. U-Net ++ [8] integrates different levels of features and uses a flexible network structure with deeply supervised U-shape networks, and ET-Net [9] embeds edge-attentive U-shape networks. However, these networks are deficient in learning local information and global feature information. In many cases, these networks cannot correctly distinguish between background and target regions, lacking the analysis of local pixel information extraction.

The U-shaped structure is difficult to learn explicit global and remote semantic information interactions in a primordial CNN approach due to the inherent limitations of convolution operations [10]. To overcome such limitations, existing studies have suggested a self-attention mechanism based on CNN features [11, 12]. The transformer is powerful in modeling global contexts [10]. However, the primitive transformer approach also suffers from feature loss. As a result, approaches based on CNN architectures combined with transformer modules flourished. For example, Bello *et al.* [13] used a self-attention mechanism to augment convolution operators by connecting convolution feature mappings that emphasize localization with self-attention feature mappings that are capable of modeling the global. DETR [14] uses a conventional CNN skeleton network to learn a two-dimensional representation of the input image, where both the encoder and decoder are composed of a transformer. Later researchers started to combine transformer and U-shape networks. Swin-UNet [15] is a U-shaped medical image segmentation network based on a pristine transformer architecture, which feeds tokenized image blocks through jump connections into a transformer-based En-Decoder architecture for local and global semantic feature learning. Nevertheless, the original transformer approach has the problem of feature

loss. ViT [16] can process image blocks or CNN outputs directly through a self-attention mechanism just like the transformer. In addition, Ramachandran *et al.* [17] gave a local self-attention module that can completely replace the convolution in the ResNet architecture. However, these designs ignore the rich context between neighboring keys. A novel transformer-style module called CoTNet has emerged, which makes full use of the contextual information between input keys to guide the learning of dynamic attention matrices [18]. The arrival of the transformer continues to improve the segmentation accuracy of images, but it makes network deployment difficult because of the enormous parameters it uses in the network. making the required parameters for the network not easy for model deployment.

In this paper, a multi-level semantic enriched neural network approach for medical image segmentation is proposed. To improve the differentiation of target and background regions, as well as to optimize the learning and extraction of global and local feature information, the CDT module is used for feature enhancement in the decoder. In the CDT module, the Cbam attention module is first used, which uses its unique hybrid attention mechanism to weigh the effective features while suppressing the invalid features or noise. The DCC module uses its dilated convolution layers to increase the perceptual field and thus obtain more semantic information, optimizing the learning ability of local and global features. Finally, the MDTransformer module is adopted to solve the problem of distinguishing the background from the target area. In order to ameliorate the problem of a large number of parameters, we use Pooling-Conv-Former (PCFormer) module in the process of down sampling feature extraction. PCFormer module adopts pooling as the token mixer structure in the transformer and thus significantly reduces the computation of parameters. The main contributions of this work are as follows:

(1) We present the PCFormer module as the encoder of the network to avoid the parameter explosion caused by the traditional transformer in the process of down sampling.

(2) We design the CDT module as the decoder of the network. In order to optimize the network for the inadequacy of global and local feature region extraction, the Cbam attention module is utilized to weigh the target region, and the DCC module is designed to increase the perceptual field. In addition, the MDTransformer module is employed to distinctly distinguish the target region from the background region.

(3) On four medical image segmentation benchmark datasets, including the ISIC, Glas, SIIM-ACR and LGG datasets, our method achieves state-of-the-art performance in both subjective segmentation results and objective evaluation metrics.

2. Related work

2.1 Medical image segmentation

With the manipulation of deep learning in clinical medical image segmentation, numerous segmentation networks for medical images have flourished. In the field of retinal vessel segmentation [19], Fu *et al.* [20] introduced conditional random fields into convolution neural networks to optimize segmentation results with vessel segmentation and put forward a network that was based on U-Net improved retinal vessel segmentation method. Gu *et al.* [21] derived a U-shaped network through an experiment that has residual structure and dilated convolution and improved segmentation results on vessel segmentation. Li *et al.* [22] adopted the end-to-end structure of 3D output to 2D giving an image projection network (IPN) for effective feature selection and dimensionality reduction for vessel segmentation. In the field of liver tumor segmentation, Liu *et al.* [23] proposed a liver tumor segmentation

method that is based on deep Unet and graph-cut abdominal CT sequences. Li and Tso *et al.* [24] suggested a bottleneck supervised Unet model, which is segmented by making full use of the information between the layers of the network. Additionally, Schlempr *et al.* applied the A-unet model to incorporate attention mechanisms into the Unet segmentation framework, which can suppress problems such as irrelevant features and segmentation inaccuracies, and difficult detection of tiny tumors. Many other fields, such as brain tumors, heart tumors, etc., have developed a large number of appealing segmentation networks, which can successfully assist doctors in clinical diagnosis.

2.2 Medical image segmentation based on encoder-decoder architecture

The encoder-decoder architecture has been greatly treasured and largely developed in medical image segmentation. Ronneberger *et al.* designed the first deep learning model U-Net for biomedical image segmentation based on FCN, which has the satisfactory performance of U-Net on medical images. Consequently, many researchers have given various improvements that are based on the encoder-decoder structure. Nabil *et al.* [25] applied a variant of U-Net on a multimodal medical image segmentation task after optimizing the encoder of U-Net, which achieved excellent performance. The authors of D-Net [26] proposed a multiscale information fusion module that uses parallel convolution layers with different expansion rates to better capture information about retinal vessels of different sizes. Chen *et al.* [27] presented a dual-stream architecture that contains a scale-context-selected attention module to enhance multiscale processing. Zhang *et al.* [28] introduced a boundary-enhanced structure to combine spatial information through dilated convolution. Kushnure *et al.* [29] proposed a multi-scale approach to capture broader and deeper contextual features. The authors of BA-Net [30] introduce a jump connection from the boundary decoder to the segmentation decoder and define a consistency loss to drive both decoders to produce the same result. In addition, the CPF-Net [31] is used by combining two pyramid modules that incorporate global and multi-scale contextual information. These networks are based on the encoder-decoder, with modular innovations and structural adjustments that allow a breakthrough in the extraction of semantic information.

2.3 Transformer-based medical image segmentation

The self-attention mechanism in a transformer is able to globally compute pairwise relationships between patches, thus enabling feature interaction over a longer range. Beyer *et al.* [16] used a primitive transformer framework for vision tasks, treating images as a collection of spatial patches. The same self-attention mechanism is combined in DGFAU-Net [32]. In medical image semantic segmentation, transformer combinatorial architectures can be divided into two categories. One mainly utilizes self-attention mechanisms to complement convolution neural networks [33]. The other exploits primordial transformers to build encoder-decoder architectures to capture the depth of information [34]. Two types of transformers are combined in the network mentioned in this paper, a pristine transformer is employed to build the encoder with pooling instead of the self-attention mechanism, while a transformer with the self-attention mechanism is adopted in the decoder. In fact, the self-attention mechanism in the transformer is permutation equivalent [35], which omits the order of blocks in the input sequence. Nevertheless, since medical image segmentation results are highly correlated with location, the nature of substitution equivalence may be detrimental to medical image segmentation. Previous work usually used absolute position encoding (APE) [35] or relative position encoding (RPE) [36] to complement the position information. However, APE requires a pre-given fixed number of patches and therefore cannot be

generalized to different image sizes. RPE ignores absolute position information, which is precisely important information for a pixel-level task like medical image segmentation.

3. Methodology

In this section, the proposed model is described in detail. The proposed method adopts a four-layer down sampling architecture, using the PCFormer module as the encoder module for each layer. In the decoder construction, the CDT module is utilized as a semantic enrichment module. Specifically, the entire network is introduced in Section 3.1. The encoder structure based on the PCFormer module is presented in Section 3.2. The CDT module is designed in Section 3.3, and the loss function is described in Section 3.4.

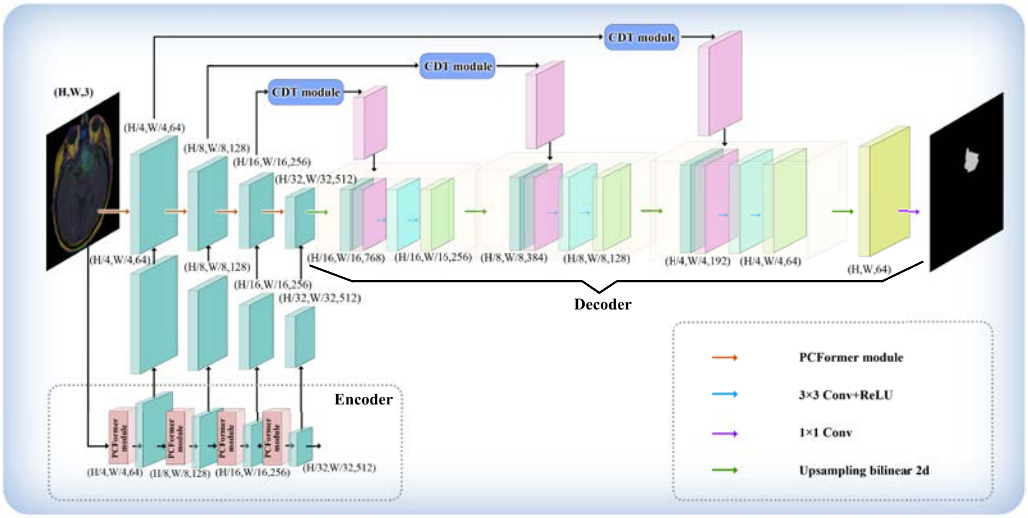


Fig. 1. The architecture of the MLSE-Net

3.1 Overview

Fig. 1. illustrates the pipeline of the proposed method in this paper. The encoder process is divided into four layers in steps. In particular, the input image $\mathcal{I}_{in}$ is processed by the PCFormer module in the first layer to obtain the output image $\mathcal{I}_1$. $\mathcal{I}_1$ is used as the input image in the second layer and is processed to obtain the output image $\mathcal{I}_2$. $\mathcal{I}_2$ is adopted as the input image of the third layer, and the output image $\mathcal{I}_3$ is derived after processing. $\mathcal{I}_3$ is selected as the input image of the fourth layer, and the output image $\mathcal{I}_4$ is obtained after processing. In the decoder process, the feature enhancement step before feature fusion is first performed by the CDT module, and $\mathcal{I}_1$ is processed by the CDT module to obtain the feature enhancement image $\mathcal{Z}_1$. $\mathcal{I}_2$ is subjected to the CDT module to obtain the feature enhancement image $\mathcal{Z}_2$. $\mathcal{I}_3$ is applied to the CDT module to obtain the feature-enhanced image $\mathcal{Z}_3$. $\mathcal{I}_4$ is concatenated with $\mathcal{Z}_3$ in the channel dimension, and then the image channel dimension is changed by the convolution operation to get $\mathcal{H}_1$. $\mathcal{H}_1$ is concatenated with $\mathcal{Z}_2$ in the channel dimension, and then the image channel dimension is modified by the convolution operation to get $\mathcal{H}_2$. After concatenating $\mathcal{H}_2$ with $\mathcal{Z}_1$ in the channel dimension,

the image channel dimension size is adjusted using the convolution operation to obtain $\mathcal{H}_3$. Finally, the binary map $\mathcal{F}$ is generated by bilinear interpolation and convolution operation.

3.2 Coding structure based on PCFormer module

The transformer is prone to parameter explosion in the network due to using self-attention as the token mixer module, which makes the model difficult to deploy [35]. Fortunately, the computational complexity of pooling is linear in sequence length and learnable parameters are unnecessary, which can compensate for the shortcomings of the parameter explosion. As shown in **Fig. 2 (a),** the PCFormer module contains a convolution layer, a feature capture layer composed of two pooling transformers (PT), and a closing convolution layer (CCL).

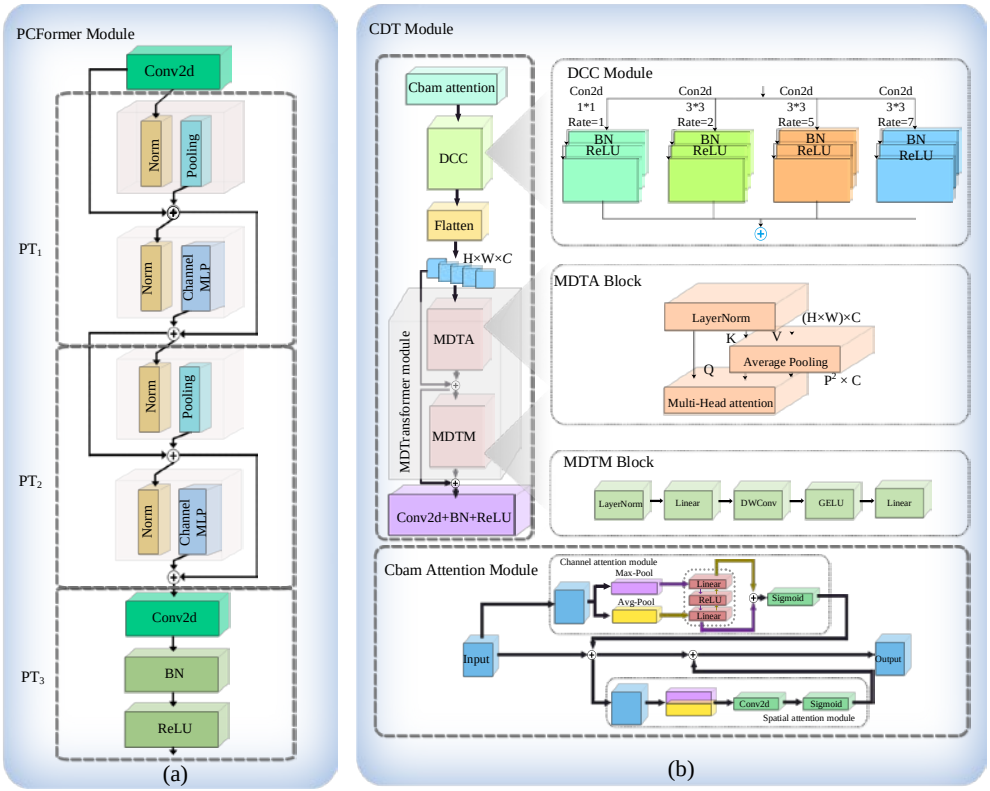


Fig. 2. PCFormer module & CDT module

Fig. 2 (a) shows the specific structure of the PCFormer module. The feature map to be down-sampled is processed by the first convolution layer and then feature capture is performed by two PT layers. The first PT layer focuses on the overall structure of the feature image for generalized feature extraction. The feature map obtained by using the output of the first convolution layer is complemented with the feature map obtained after the pooling operation in the first PT layer to prevent the feature loss caused by pooling. The overall can be expressed as follows:

$$\mathrm{PT}_1 = \begin{cases} X_{\mathrm{in}} \xrightarrow[\text{Convolution}]{3\times 3} \delta_1 \\ \delta_1 \oplus \delta_1 \xrightarrow[\text{Pooling}]{\text{Norm}} \delta_2 \\ \delta_2 \oplus \delta_2 \xrightarrow[\text{ChannelMLP}]{\text{Norm}} \delta_3 \end{cases} \tag{1}$$

Where δ_1 is the result obtained after convolutional layer processing, X_{in} is the image to be processed in each layer input, δ_2 is the result obtained by adding the feature map obtained by δ_1 after the norm and pooling operations in the first PT layer with δ_1, and δ_3 is the feature map obtained in the first PT layer. Where channel MLP can be expressed as follows:

$$\psi_{in} \xrightarrow[\text{GELU}]{\text{1×1 Convolution}} \xrightarrow[\text{Convolution}]{\text{1×1}} \psi_{out} \tag{2}$$

ψ_{in} is the input feature map and ψ_{out} is the output feature map. The featured image processed by the first PT layer has more weight on the outer contour of the feature region than the other regions, so the trailing PT_1 layer can focus on the local detail regions and complete the detail feature capture. Finally, the closing convolution layer (CCL) performs information integration as well as gives linear structure to the feature map. The overall can be expressed as follows:

$$\text{PCFormer\_module} = \begin{cases} \delta_3 \oplus \delta_3 \xrightarrow[\text{Pooling}]{\text{Norm}} \delta_4 \\ \underbrace{\delta_4 \oplus \delta_4 \xrightarrow[\text{ChannelMLP}]{\text{Norm}} \delta_5}_{PT_2} \\ \underbrace{\delta_5 \xrightarrow[\text{BN+ReLU}]{\text{Convolution}} \delta_6}_{PT_3} \end{cases} \tag{3}$$

Where δ_3 is the resultant feature map generated by the PT_1 layer, δ_4 is the feature map generated by the Norm and Pooling layers of the PT_2 layer, δ_5 is the feature map generated by the PT_2 layer, and δ_6 is the feature map finally generated by the PCFormer module.

3.3 Decoding structure based on CDT module

In the traditional U-Net model, the directness skip connection is unable to correctly distinguish the target and background regions due to insufficient feature capture ability. For suppressing sample noise, U-Net fails to make effective initiatives. The CDT module in the decoder structure can improve the differentiation of target and background regions, suppress sample noise, as well as optimize the learning and extraction ability of global and local feature information. The CDT module contains a Cbam attention module, a DCC module, and an MDTransformer module.

3.3.1 Cbam attention module

The Cbam attention module is a simple and effective attention module for convolutional neural networks (CNNs). Given an arbitrary intermediate feature map in a CNNs, the Cbam attention module injects the attention mapping along two independent dimensions of the channel and space of the feature map. Then, it multiplies the attention by the input feature map to perform adaptive feature refinement on the input feature map. Due to the Cbam attention module is an end-to-end generic module, it can be seamlessly integrated into any CNNs architecture and can be trained end-to-end with basic CNNs. The structure of channel attention and spatial attention in the Cbam attention module is shown in **Fig. 2 (b)**. Given an intermediate feature map $F \in \mathbb{R}^{C \times H \times W}$ as input, the operation process of the Cbam attention module is generally divided into two stages: Firstly, MaxPooling and AvgPooling operations are performed on the input by channel, and the two one- dimensional vectors after pooling are sent to the fully connected layer operation and then summed to generate one-

dimensional channel attention $M_C \in \mathbb{R}^{C\times1\times1}$, and then the channel attention is multiplied with the input by element to obtain the channel attention-adjusted feature map F^1; Secondly, F^1 is subjected to global MaxPooling and AvgPooling operations by space, and the two two-dimensional vectors generated by pooling are concated together and subjected to convolution operations to finally generate two-dimensional spatial attention $M_S \in \mathbb{R}^{C\times1\times1}$, and then the spatial attention is multiplication dot with F^1. The specific operational flow is shown in **Fig. 2 (b),** and the overall attention generation process of the Cbam attention module can be described as follows:

$$F_{\text{out}} = S_a(C_a(F)\odot F)\odot C_a(F)\odot F \tag{4}$$

Where F_{out} denotes the output feature map of the Cbam attention module and $\odot$ denotes the corresponding element multiplication. F is the input feature map. $S_a(\cdot)$ is the Spacial attention operation. $C_a(\cdot)$ is a Channel attention operation. Before the multiplication operation, channel attention and spatial attention need to be broadcasted in spatial dimension and channel dimension, respectively.

3.3.2 DCC module

The DCC module is composed of four dilated convolution blocks with expansion coefficients of 1, 2, 5, 7 respectively. Each dilated module consists of one dilated convolution, one BN and one RelU layer. The convolution kernel size is set to expand the field of perception for the expansion coefficients of 2, 5, 7, and the kernel size of 1 × 1 is set to supplement the expansion coefficient of 1 to suppress the loss of information of important feature regions by the expansion convolution. Finally, the output feature maps of the four expansion convolution modules are stitched together. In the field of image segmentation, such as FCN [37], the convolution operation is done before the pooling operation to reduce the image size and increase the image field of perception at the same time, but it will cause the lack of accuracy of the image when up sampling the reduced image size. Dilated convolution increases the receptive field of the neural network without decreasing the image size so that each convolution output contains a larger range of information. The different perceptual fields obtained with four 3 × 3 ordinary convolutions and four 3 × 3 dilated convolutions, respectively, are clearly shown in **Fig. 3.** Assuming that the size of the convolution kernel is $K \times K$ and the dilation rates are d, its equivalent convolution kernel size K is calculated by the following formula:

$$K' = K + (K-1)\times(d-1) \tag{5}$$

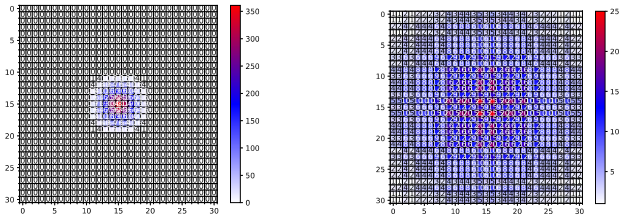
After the processing of four inflated convolution blocks, the output feature map is finally stitched in the channel dimension, which can be expressed in general as

$$X_{\text{out}} = [\varphi^1_{1\times1}(X_{in})\otimes\varphi^2_{3\times3}(X_{in})\otimes\varphi^5_{3\times3}(X_{in})\otimes\varphi^7_{3\times3}(X_{in})] \tag{6}$$

Where X_{out} is the output result, X_{in} is the input feature map, $\otimes$ is the concat operation, and $\varphi^1{}_{1\times1}(\cdot)$ is a dilated convolution block with an expansion factor of 1 and a convolution kernel size of 1 × 1. $\varphi^2{}_{3\times3}(\cdot)$ is a dilated convolution block with an expansion factor of 2 and a convolution kernel size of 3 × 3. $\varphi^5{}_{3\times3}(\cdot)$ is a dilated convolution block with an expansion factor of 5 and a convolution kernel size of 3 × 3. $\varphi^7{}_{3\times3}(\cdot)$ is a dilated convolution block with an expansion factor of 7 and a convolution kernel size of 3 × 3.

3.3.3 MDTransformer module

The MFTransformer is composed of an MDTransformer-Attention block (MDTA block) and an MDTransformer-MLP block (MDTM block). The MDTA block consists of a LayerNorm



(a) (b)

Fig. 3. the different perceptual fields of the two convolutions without changing the image size, (a) the perceptual field of the normal convolution, and (b) the perceptual field of the dilated convolution.

layer, an Average Pooling layer, and a Multi-Head Attention layer. The MDTM block is composed of a LayerNorm layer, a first linear processing layer, a DWConv layer, a GELU layer, and a tail linear processing layer. MDTransformer module greatly reduces the use of parameters while enhancing the feature capture capability of the CDT module. The input feature map is processed by Flatten linear expansion to obtain a linear feature vector of size $(H \times W) \times C$. The vector is input to the MDTA block for feature capture processing, and then fed to MDTM block for processing, and the output linear feature vector is reshaped to get the feature map of size $H \times W \times C$. In summary, it can be described as:

$$X_{\text{out}} = \text{reshape}(\text{MDTA}(X_{\text{in}}) + X_{\text{in}} + \text{MDTM}(\text{MDTA}(X_{\text{in}}) + X_{\text{in}})) \tag{7}$$

Where X_{out} is the output of the MDTransformer module. reshape $(\cdot)$ is the reorganization function. MDTA $(\cdot)$ is the MDTA block operation. X_{in} is the input feature map. MDTM $(\cdot)$ is the MDTM block operation.

MDTA block. Multi-Head Attention in the MDTA block enhances the feature capture capability for the CDT module and assists the CDT module in feature enhancement. Unlike Vit's [16] Multi-Head Attention, the Multi-Head Attention in this paper reduces the computational effort to a great extent. If the input feature map of $H \times W \times C$ is given, the complexity of Vit's Multi-Head Attention is:

$$\Omega = 3 \times H^2 \times W^2 \times C \tag{8}$$

Where Ω denotes the parameter complexity.

If the input feature map of $H \times W \times C$ is given, the complexity of this MDTA block is:

$$\Omega = 2 \times H \times W \times P^2 \times C \tag{9}$$

Where P is the linear pool size. The final formula for the self-attention calculation is:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{\text{head}}}}\right) \tag{10}$$

Where Attention (Q, K, V) is the self-attentive mechanism. Q, K, and V are the matrices of multiple linear vectors combined from the deformation of the original input feature map. The results obtained from each set of Q_i, K_i, and V_i vector operations are stitched together to obtain d_{head}. K^T is the transpose matrix of the K. Softmax $(\cdot)$ is an activation function. But before K, V is passed in the input size changes from $H \times W \times C$ to a fixed size of $P^2 \times C$, where P is the linear pool size. This greatly reduces computational complexity and memory consumption. Therefore, larger feature maps can be processed with limited resources.

MDTM block. DW convolution is used in the MDTM block to remove the position coding and reduce the required parameters of the network. DW convolution can learn the contour information of the feature map based on the outer circle of the feature map. In other words, DW can learn some absolute position information to model the position information.

MDTM block can be represented as:

$$X_{\text{out}} = \text{Drop}(\text{Linear}(\text{GELU}(\text{DWConv}(\text{Linear}(\text{LayerNorm}(X_{\text{in}})))))) \quad (11)$$

Where X_{out} is the output feature map of the MDTM block. Drop (·) is a Dropout operation. Linear (·) is a linear activation function. GELU (·) is an activation function that incorporates the idea of stochastic regularity. DWConv (·) is a DWConv operation. LayerNorm (·) is a normalized processing layer.

3.4 Hybrid loss

Different loss functions have different characteristics in different aspects. To take advantage of the different loss functions in backpropagation, the method in this paper designs a Hybrid loss, which is composed of CrossEntropy loss and Dice loss. Setting Dice loss to L_1, and CrossEntropy loss to L_2, the Hybrid loss can be expressed as:

$$L = \alpha L_1 + \beta L_2 \quad (12)$$

3.4.1 Dice loss

The Dice loss function, which essentially measures the overlap of two samples, addresses the case of category imbalance. Dice loss can be expressed as:

$$L_1 = 1 - \frac{2\sum_{i=1}^{N}\sum_{c=1}^{c} g_i^c s_i^c}{\sum_{i=1}^{N}\sum_{c=1}^{c} g_i^{c^2} + \sum_{i=1}^{N}\sum_{c=1}^{c} s_i^{c^2}} \quad (13)$$

Where L_1 refers to Dice loss, i denotes each pixel point, c denotes classification, g_i^c denotes whether the classification is correct, and s_i^c denotes the probability of being classified into a certain class.

3.4.2 CrossEntropy loss

CrossEntropy loss is able to measure subtle differences. It is calculated as follows:

$$L_2 = -\sum_{i=1}^{m}\sum_{j=1}^{n} P(y_{ij}) \log(Q(y_{ij})) \quad (14)$$

Where L_2 refers to the CrossEntropy loss function, n is the number of categories to be classified, $P(y_{ij})$ represents the true labels corresponding to the y_i categories, i.e., the predicted probability of the n categories, and $Q(y_{ij})$ represents the predicted value of the y_i categories.

4. Experiment

4.1 Experimental Settings

Dataset To demonstrate the generalization of the network, the experimental design is based on four different categories of datasets, including the Glas dataset, SIIM-ACR dataset, ISIC dataset, and LGG dataset.

1) Glas dataset: The Glas dataset is a publicly available dataset from the MICCAI 2015 Glandular Segmentation Challenge and consists of 165 images from 16 hematoxylins and eosin-stained slides of colorectal cancer tissue sections. The original images are set to 775 × 522. The dataset is divided into a training set and a test set, in which 100 images are used as the training set, and 65 images are applied as the test set.

2) SIIM-ACR dataset: The SIIM-ACR dataset is a public dataset for the SIIM-ACR Lung Image Segmentation Kaggle competition hosted by the Society for Imaging Informatics in

Medicine (SIIM) 2019. After slicing and processing its 3D images, 101 chest radiographs are obtained for the experiments, 88 of which are used as the training set and 13 for the validation set.

3) ISIC dataset: The ISIC dataset contains a total of 5423 images of skin lesion areas with different scale sizes, different shapes, and different colors, of which 3461 images are selected as the training set and 2002 images serve as the test set.

4) LGG dataset: The LGG dataset contains 110 MR image slices of the brains with low-grade gliomas, which are obtained from the Cancer Genome Atlas and the Cancer Imaging Archive. After processing the dataset, a total of 1311 images are collected for the experiments, of which 1049 images are randomly selected for training and 262 for testing.

Implementation Details The experiments are conducted on Ubuntu 16. 04 LTS 64-bit operating system with 64GB RAM and NVIDIA Tesla K80 GPU. The initial learning rate is 1e-4 using Adam optimizer, and the learning rate is adjusted using CosineAnnealingLR algorithm for each epoch completed during the training process. The variation interval of the learning rate is between (1e-4,1e-6). α set to 0.3 and β set to 0.7. The batch size is set to 4, and the size of the training image is 256*256. The epochs of the four datasets LGG, SIIM-ACR, ISIC, and Glas are set to 100 times during training.

Evaluation metrics In order to evaluate the usefulness of the algorithm fairly, we analyze it from different perspectives and use different evaluation metrics including Dice, Intersection over Union (IoU), Weighted F-measure (wFm), Structure-based Metric (Sm), Enhanced-alignment Metric (Em) and Sensitive (Sen). Specifically, the Dice index focuses on the similarity information of pixel points inside the region and is used to measure the similarity between two samples. The IoU criteria calculates the similarity or overlap between two samples by the ratio of predicted borders to true borders. We use the wFm metric to assign different weights to the errors generated at different locations based on the adjacency information. The Sm metric is a reconciliation index, which can effectively reflect the structural similarity between two image collections. The Em index can effectively reflect the local pixel-matching information between two image collections. The Sen criteria indicates the proportion of all positive examples that are judged to be correct, and it measures the classifier's ability to recognize positive examples.

4.2 Comparison experiments

To verify the effectiveness of the method in this paper, comparison experiments are conducted on four different types of datasets with Unet [6], NestedUNet [8], BiSeNetV1 [38], BiSeNetV2 [39], KiUnet [40], SSformer [41], TransUnet [10], Uctransnet [42], ScaleFormer [43]. The following are the results of the experiments with data sets respectively.

4.2.1 SIIM-ACR Dataset

Quantitative Evaluation: Unet, NestedUNet, BiSeNetV1, BiSeNetV2, KiUnet, SSformer, TransUnet, and the proposed method are utilized to segment the same test set, and the mean values of the five networks are calculated for the evaluation metrics Dice, IoU, wFm, Sm, Em, and Sen, respectively. The evaluation results are shown in **Table 1**. Comparing the experimental results, it can be seen that SSformer has the lowest Dic and IoU, indicating that the segmentation effect of the network has the largest gap compared with the original label. BiSeNetV1 and BiSeNetV2 outperform SSformer in terms of Dic and IoU, but the indexes are not as good as the proposed method. This is because the proposed network structure in the decoder part enhances the feature capture capability and develops the field of view perception, which results in higher

accuracy of segmentation. The wFm, Sm, and Em of Unet and NestedUNet achieve equal metrics, which indicates that the dense residual edges used in NestedUNet are less advantageous on lightweight datasets. In this paper, the proposed network advances the Sm to 95.9% while ensuring an insignificant improvement in wFm and Em, achieving a balance between the accuracy of local segmentation and overall structural segmentation. Consequently, the proposed method also has significant enhancement in meanDic, meanIoU, and meanSen. Combining the six evaluation results, the overall performance of the proposed method surpasses other methods.

Table 1. The quantitative result on the SIIM-ACR dataset
(Bold numbers indicate the best performance)

Network module (SIIM-ACR)	Dic	IoU	wFm	Sm	Em	Sen
Unet	0.971	0.945	0.976	0.957	0.985	0.970
NestedUNet	0.973	0.949	0.976	0.957	0.985	0.978
BiSeNetV1	0.963	0.931	0.968	0.948	0.981	0.961
BiSeNetV2	0.972	0.947	0.976	0.957	0.985	0.978
Kiunet	0.973	0.949	0.977	0.958	0.985	0.976
SSformer	0.951	0.916	0.958	0.940	0.971	0.943
Tranunet	0.969	0.943	0.973	0.955	0.984	0.974
ScaleFormer	0.971	0.946	0.973	0.953	0.982	0.977
Uctransnet	0.967	0.948	0.970	0.949	0.974	0.973
MLSE-Net(Ours)	**0.974**	**0.951**	**0.978**	**0.959**	**0.986**	**0.981**

Quality Evaluation: The segmentation results of all networks are shown in **Fig. 4.** On the one hand, the segmentation results of each network cannot effectively suppress the sample noise, as shown in the circled part of case1 in **Fig. 4**. In case 1, there has some misclassification phenomenon due to the inability to suppress the sample noise in the segmentation results, while there are relatively few cases of misclassification in this method. Meanwhile, TransUnet and BiSeNetV1 which use the attention module in the encoder part have more serious misclassification than Unet and SSformer, which proves that the attention module in the encoder cannot effectively suppress the sample noise in lightweight data sets. The significant reason why SSormer only produces smaller segmentation errors is that it designs the LE and SFA modules in the decoder part to effectively suppress the sample noise. The proposed method has an eminent decoder module and therefore the perfect segmentation results in this case1. On the other hand, in the segmentation results of each network, the left lung end is lost to different degrees as shown in the circled part in case2. Case2 has an obvious loss of the left lung end in the segmentation results of Transnet, SSormer and BiSeNetV1, while the segmentation results of NestedUNet and Kiunet are revised, which proves that in the lightweight datasets in encoder using the attention module cannot effectively suppress the sample noise. The PCFormer module that is proposed in this paper in encoder structure circumvents this structure very well. In general, the method in this paper can effectively solve the above two types of issues, and the segmentation results of this method are closer to the real results compared with other networks.

4.2.2 Glas Dataset

Quantitative Evaluation: As can be seen from **Table 2**, compared to Unet, the KiUnet network with a dual-network mechanism drops to 87.8% and 78.9% on Dic and IoU, respectively. Although the TransUnet and SSformer networks optimized the encoder and decoder structures, the scores on Dic and IoU were still lower than those of Unet. NestedUNet optimizes the feature fusion process so that its scores on Dic and IoU arrive to

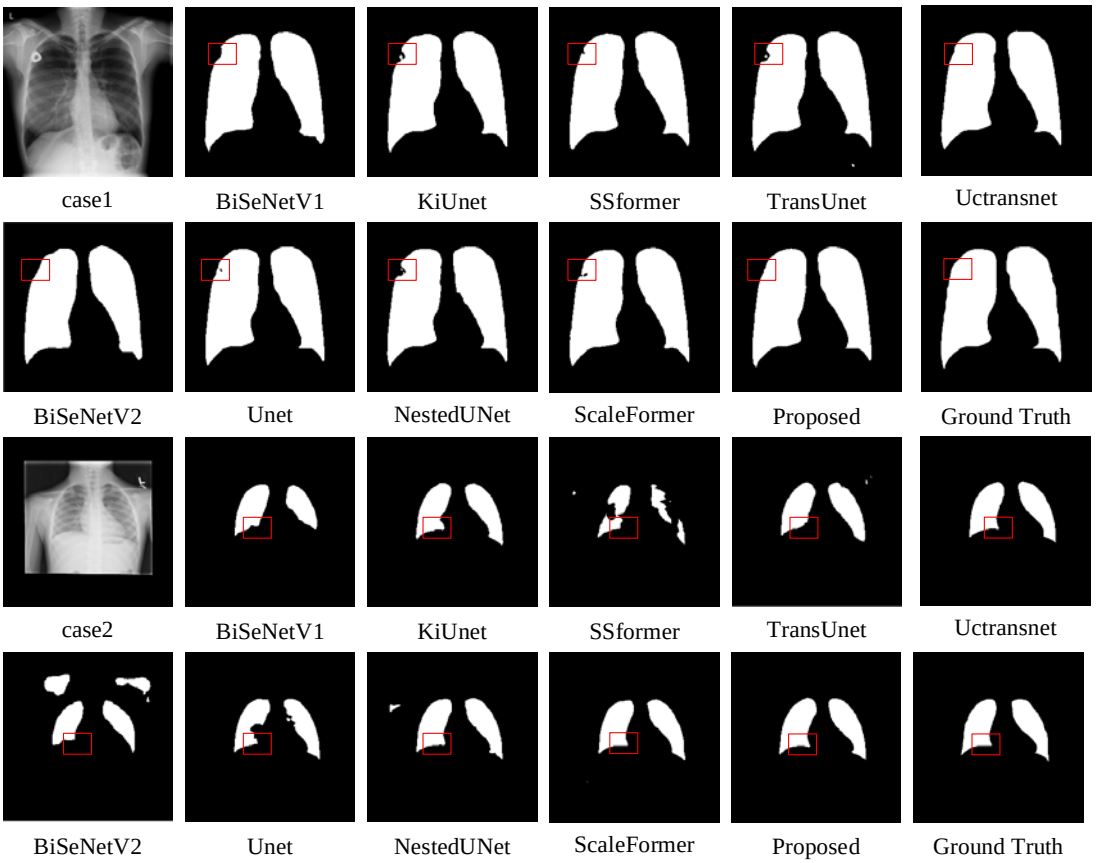


Fig. 4. Quality results on the SIIM-ACR Dataset

92.6% and 86.5%, respectively. The scores of the proposed method on Dic and IoU, after using the PCFormer module in the backbone and CDT module in a decoder, are increased to 93.0% and 87.2%, respectively, which proves that the two modules proposed in this paper's network have excellent improvement for segmentation accuracy. BiSeNetV1 and TransUnet are close in wFm, Sm, Em and Sen. From **Table 2**, we can see that TransUnet outperforms BiSeNetV1 in terms of the overall structure, and its Sm reaches 82.6%. This indicates that the transformer structure used in the encoder part of TransUnet is more excellent than the ARM module proposed in BiSeNetV1 in terms of the global structure of the segmentation. In addition, the proposed method achieves 85.3% in Sm since the CDT module used expands

Table 2. The quantitative result on the Glas dataset (Bold numbers indicate the best performance)

Network module (Glas)	Dic	IoU	wFm	Sm	Em	Sen
Unet	0.919	0.852	0.908	0.830	0.914	0.935
NestedUNet	0.926	0.865	0.920	0.843	0.919	0.937
BiSeNetV1	0.915	0.845	0.908	0.821	0.907	0.924
BiSeNetV2	0.897	0.816	0.884	0.798	0.886	0.910
KiUnet	0.878	0.789	0.861	0.780	0.874	0.895
SSformer	0.890	0.804	0.867	0.766	0.849	0.929
TransUnet	0.917	0.849	0.907	0.826	0.905	0.930
ScaleFormer	0.924	0.864	0.920	0.841	0.917	0.932
Uctransnet	0.925	0.863	0.912	0.821	0.902	0.927
MLSE-Net(Ours)	**0.930**	**0.872**	**0.925**	**0.853**	**0.921**	**0.941**

the perception field of the network segmentation in each layer. Combining all the metrics data, the proposed method has a positive effect on the medical image segmentation task.

Quality Evaluation: The segmentation results of all networks are shown in **Fig. 5**. After comparison, it can be seen that each network has segmentation errors in different degrees, and the conflict between the local segmentation effect and the global segmentation effect is highlighted. In case1, BiSeNetV2 is less effective than BiSeNetV1 segmentation because BiSeNetV2 eliminates the fast-downsampling strategy, which reduces the range of sensory field acquisition and leads to a lack of overall segmentation accuracy. However, in case2 BiSeNetV1 is less effective than BiSeNetV2 segmentation due to the fact that the detailed branch used in BiSeNetV2 is more prominent in its ability to capture details. The same phenomenon is reflected in the segmented images of Unet and NestedUNet. NestedUNet, which uses the residual structure in the feature fusion process, outperforms Unet in the overall segmentation in case1, but its segmentation error in case2 is more prominent than that of Unet, due to the enhanced noise effect factor in the residual fusion process. Case1 shows that the global segmentation of TransUnet and SSformer outperforms the other networks, while in case2 SSformer shows more effective local segmentation than TransUnet, due to the enhanced local details of SSformer using LE and SFA modules. This shows that using transformer structure in an encoder structure can effectively help the network to perform global segmentation of images, while a good decoder module is more

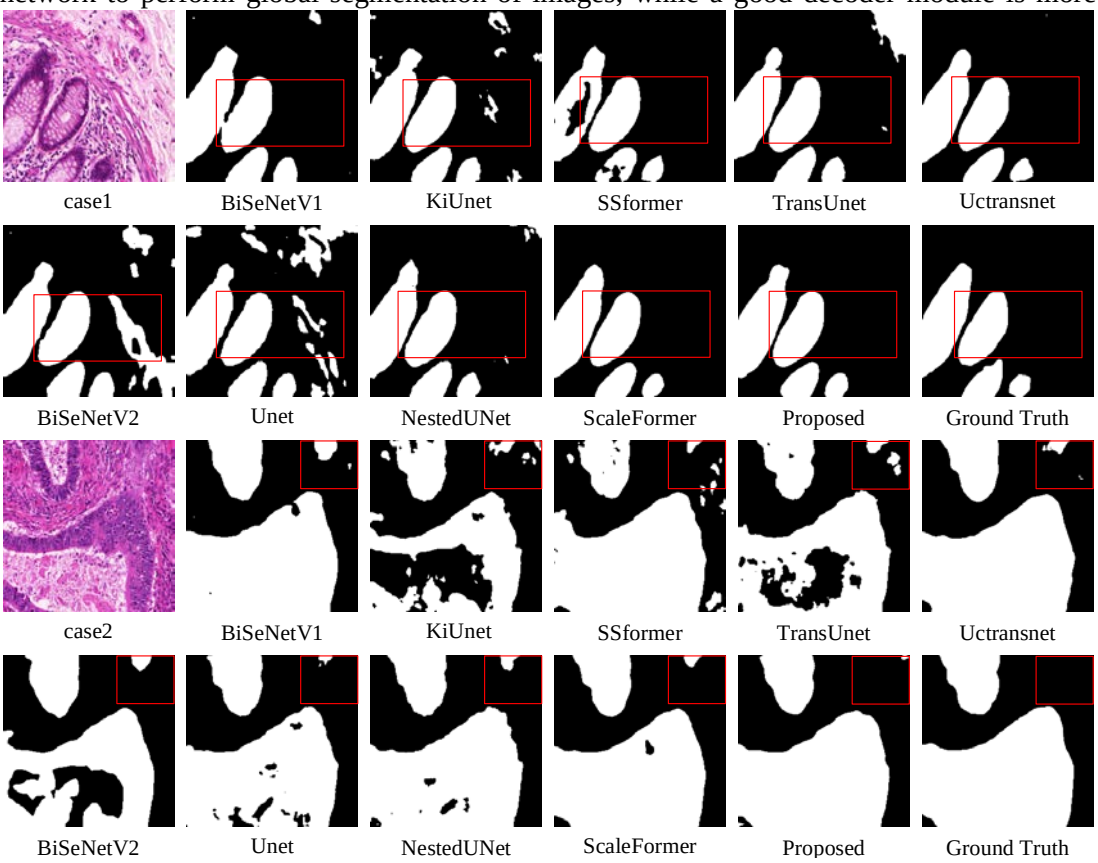


Fig. 5. Quality results on the Glas Dataset. The Glas dataset is a cell image set, and the main requirement is that the network has good boundary segmentation ability and environment differentiation effect. From the segmentation effect, MLSE-Net is better than other network models both in terms of environment differentiation effect and boundary segmentation ability.

important for local segmentation. The reason why this network can get the glorious segmentation results is in case1 and case2. It is because this network uses PCFormer the encoder part to enhance the feature capture ability and the CDT module in the decoder part to enhance the semantic details, which can effectively solve the conflict between the local segmentation effect and the global segmentation effect.

4.2.3 ISIC Dataset

Quantitative Evaluation: As can be seen from **Table 3**, for the U-Net network, NestedUNet's scores on Dic and IoU did not change significantly, while KiUnet with a dual-network mechanism decreased to 88.1% and 81.0% on Dic and IoU instead. The TransUnetnetwork combined with a transformer in the backbone decreased to 87.4% and 79.9% for Dic and IoU, while SSformer with optimized decoder structure increased to 91.0% and 85.1% for Dic and IoU. In contrast, adding a transformer to the backbone weakens the overall segmentation accuracy of medical images to some extent, while using an effective feature enhancement module can effectively suppress this phenomenon. Among them, the PCFormer module and CDT module proposed in this paper enhance the scores of Dic and IoU to 91.5% and 85.8%, respectively. Combined with the analysis of the comprehensive indexes of Dic, IoU, wFm, Sm, and Em, the proposed method achieves the best score, which demonstrates that the proposed method is advanced both from the overall segmentation effect and from the local segmentation details. Sm and Sen have certain correlations in the case of data unification when one of them has a higher value, the other index will be affected to some extent. From **Table 3**, it can be seen that Unet, NestedUNet, and BiSeNetV1 have slightly higher Sen scores than the proposed method, which shows that the expanded convolution layer used in the CDT module incorrectly learns some information while expanding the perceptual field. Nevertheless, the proposed method is higher than the other networks in all other metrics, which verifies that the proposed method has a positive effect on the medical image segmentation task.

Table 3. The quantitative result on the ISIC dataset (Bold numbers indicate the best performance)

Network module (ISIC)	Dic	IoU	wFm	Sm	Em	Sen
Unet	0.905	0.842	0.887	0.888	0.933	0.955
NestedUNet	0.905	0.842	0.886	0.887	0.931	0.958
BiSeNetV1	0.906	0.843	0.889	0.888	0.932	**0.959**
BiSeNetV2	0.912	0.853	0.900	0.896	0.938	0.944
KiUnet	0.881	0.810	0.856	0.867	0.913	0.944
SSformer	0.910	0.851	0.899	0.895	0.936	0.941
TransUnet	0.874	0.799	0.852	0.861	0.909	0.928
ScaleFormer	0.903	0.843	0.885	0.883	0.929	0.947
Uctransnet	0.901	0.842	0.890	0.873	0.925	0.949
MLSE-Net(Ours)	**0.915**	**0.858**	**0.902**	**0.898**	**0.940**	0.954

Quality Evaluation: **Fig. 6** shows the visualization results of the other methods compared with the proposed method. As shown by the comparison results, BiSeNetV1, BiSeNetV2, KiUnet, SSformer, TransUnet, Unet, and NestedUNet networks in case1 display significant bias in segmenting the target lesions and cannot effectively suppress the sample noise to ensure the certainty of the segmentation results. In case2, KiUnet and TransUnet cannot accurately locate the feature region, resulting in serious segmentation errors. In addition, NestedUNet can identify the target region, but cannot guarantee the integrity of the segmentation results resulting in large overall segmentation errors. Unet refines the issue of

target region segmentation error that exists in NestedUNet, but the problem of inability to effectively suppress sample noise with local segmentation error appears. BiSeNetV1 and BiSeNetV2 address the problem of local segmentation error, but still had the problem of not being able to suppress sample noise. The Segformer used by SSformer effectively suppressed sample noise, but the problem of local segmentation error appeared. In comparison, the proposed method can precisely predict the shape and location of lesions and tackle the issue of local segmentation error. Besides, the method in this paper is closer to the manual segmentation effect of doctors and has significant advantages in segmentation completeness and accuracy.

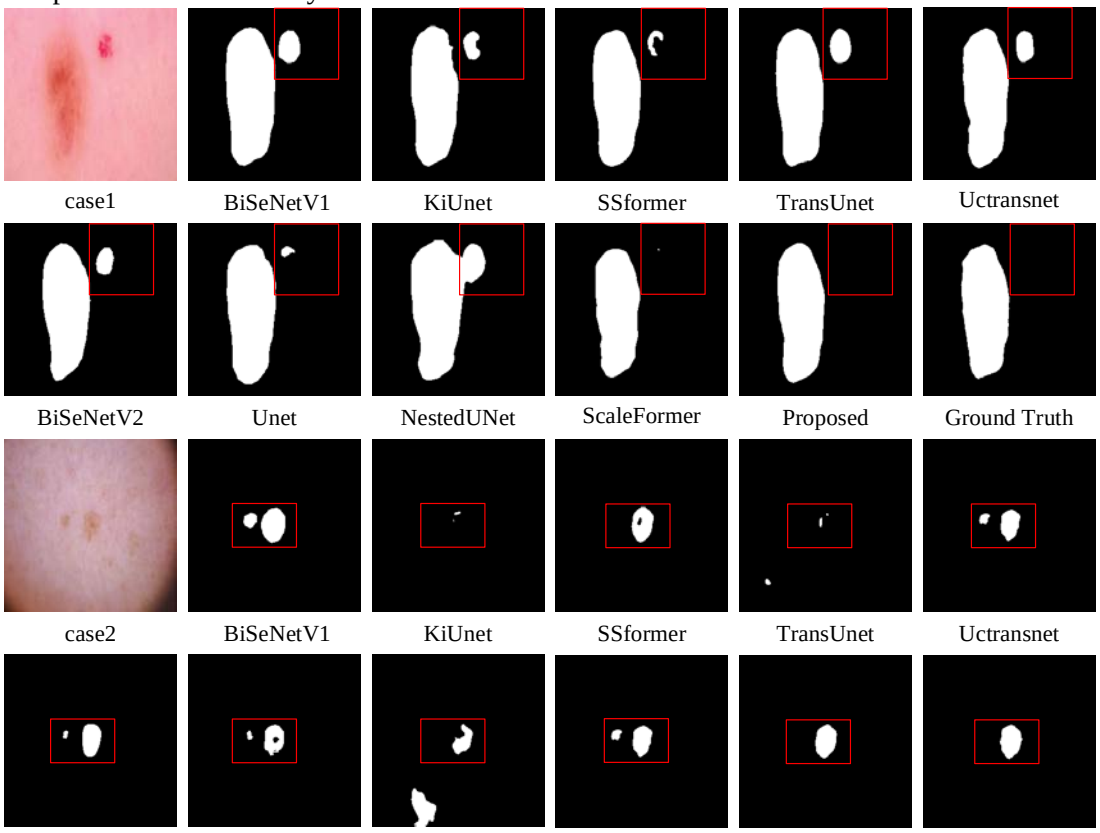


Fig. 6. Quality results on the ISIC Dataset. The ISIC dataset is highly valued for the function of suppressing noise interference, and non-lesion interference regions appear around the skin lesion regions. Comparing with other network segmentation effects MLSE-Net can effectively suppress the noise region and get the segmented image correctly.

4.2.4 LGG Dataset

Quantitative Evaluation: From **Table 4**, we can see that for the Unet network, NestedUNet's Dic and IoU scores are slightly reduced after adopting dense residual edges in the backbone, which is due to the fact that the network adopts the residual structure to fuse the underlying semantic information and the high-level semantic information, which makes it difficult for the network to deal with the errors caused by the sample noise in the high-Level semantic information and cannot correctly distinguish the background information from the target region. BiSeNetV1, SSformer, TransUnet compared to Unet in both Dic and IoU scores decreased by 2% and Sm score decreased by 1%, which indicates that the use of

Table 4. The quantitative result on the LGG dataset (Bold numbers indicate the best performance)

Network module (LGG)	Dic	IoU	wFm	Sm	Em	Sen
Unet	0.924	0.865	0.924	0.939	0.984	0.941
NestedUNet	0.923	0.863	0.918	0.938	0.983	**0.957**
BiSeNetV1	0.901	0.828	0.903	0.922	0.977	0.908
BiSeNetV2	0.918	0.855	0.921	0.936	0.985	0.929
KiUnet	0.902	0.832	0.904	0.924	0.975	0.911
SSformer	0.905	0.836	0.910	0.926	0.975	0.906
TransUnet	0.904	0.835	0.909	0.926	0.974	0.903
ScaleFormer	0.921	0.869	0.923	0.938	0.972	0.951
Uctransnet	0.927	0.871	0.929	0.940	0.977	0.940
MLSE-Net(Ours)	**0.930**	**0.874**	**0.934**	**0.944**	**0.987**	0.942

the attention-based transformer structure in the down sampling process leads to an inferior overall segmentation effect compared to Unet. The reason why the method in this paper can enhance the scores of Dic, IoU, and Sm to 93.0%, 87.4%, and 94.4% is that the Pooling operation is used instead of the attention structure for down sampling. This method of deleting the unnecessary attention mechanism in the down sampling process and instead using the attention mechanism in the feature fusion process can supplement the local semantic information while enhancing the overall segmentation effect of the network. Therefore, the proposed method has a significant improvement in the local comprehensive evaluation index. However, the score of NestedUNet on Sen is higher than the proposed method on Sen, which is due to the higher classification ability of its adopted residual structure, but the proposed method is the highest in all other indexes, which integrally reflects the robustness of our method.

Quality Evaluation: **Fig. 7** shows the visualization comparison results of the other methods in this paper. As seen from the comparison results, in case1, Unet, NestedUNet, BiSeNetV1, KiUnet, SSformer, and TransUnet exhibit significant deviations in segmenting the target lesions due to sampling noise interference, and the overall segmentation effect has a large gap with the labels. BiSeNetV2 adopts the Bilateral Guided Aggregation Layer to enhance the interconnection of feature regions, so its segmentation effect is outstanding, but the segmentation effect at the target lesion boundary is poor, and it cannot distinguish the lesion boundary and background pixels well. In this paper, the proposed method adopts a non-attention transformer structure as an encoder and cooperates with the CDT module to fuse multi-level semantic information, which enriches local semantic information while suppressing sample noise. In comparison, the method in this paper can exactly predict the shape and location of lesions, effectively raise the segmentation of lesion boundaries, and effectively suppress sample noise. In summary, the method has significant advantages in segmentation completeness and accuracy.

4.3 Ablations Experiments and Analysis

4.3.1 Effectiveness of Hybrid loss

Different combinations are designed for the hybrid loss function in the experiments. Specifically, to further analyze the impact of the use of mixed loss on the experiments. Different combinations of CELoss and DiceLoss used in the experiments are performed, which included using CELoss without DiceLoss, using DiceLoss without CELoss, and using

both CELoss and DiceLoss. The experimental results are shown in **Table 5**.

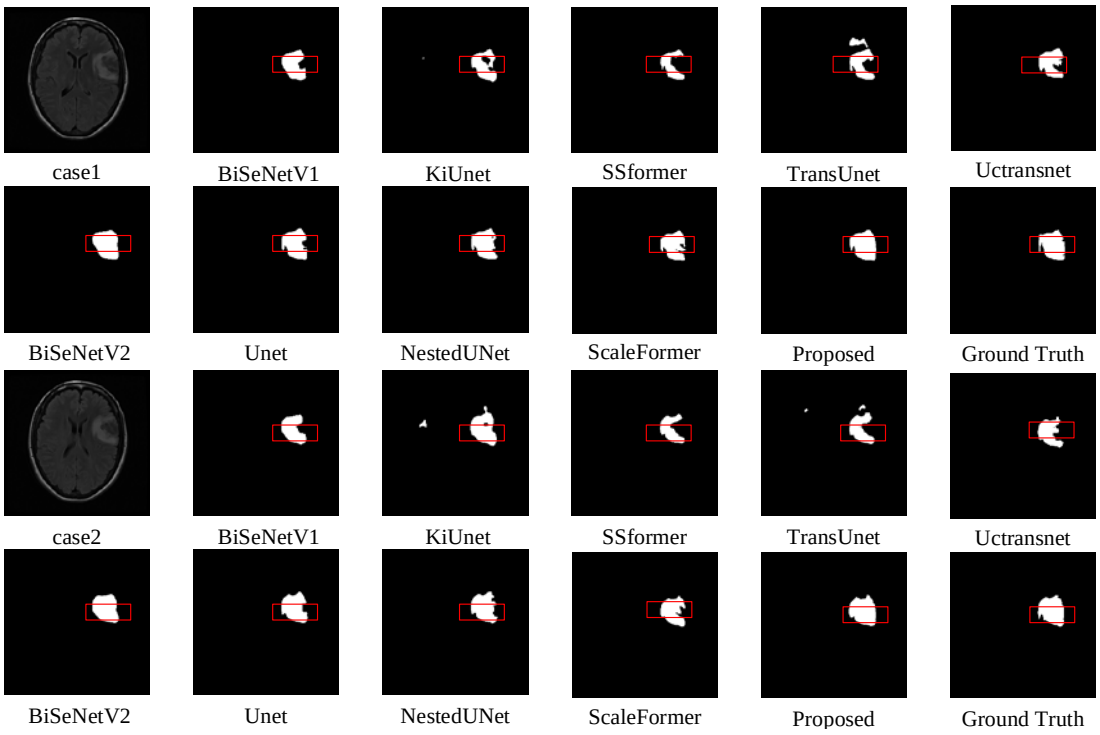


Fig. 7. Quality results on the LGG Dataset

Table 5. Result of the effectiveness of Hybrid loss (Bold numbers indicate the best performance)

Loss		**Dic**			**IoU**			**Sm**			**Em**		
CELoss	DiceLoss	Glas	ISIC	LGG	Glas	ISIC	LGG	Glas	ISIC	LGG	Glas	ISIC	LGG
√		0.863	0.910	0.926	0.762	0.850	0.867	0.746	0.894	0.942	0.850	0.936	**0.987**
	√	0.259	0.804	0.914	0.152	0.701	0.849	0.247	0.795	0.934	0.382	0.852	0.982
√	√	**0.930**	**0.915**	**0.930**	**0.872**	**0.858**	**0.874**	**0.853**	**0.898**	**0.944**	**0.921**	**0.940**	**0.987**

From **Table 5**, it can be found that the method using CELoss alone has higher Dic, IoU, Sm, and Em scores on all three data sets than the method using DiceLoss alone. When the experiments are conducted on the Glas dataset, the DiceLoss-only approach produced large biases in the experiments with low scores on all four metrics. This is because DiceLoss essentially measures the overlap of two samples, which is more generalized than CELoss, so it has poor backpropagation on Glas datasets that require abundantly detailed semantic information. The analysis of the data shows that the back-propagation process of the loss function has a vital role in the segmentation task. As the number of iterations increases, the learnable parameters of the network are continuously updated, which has a positive effect on the segmentation results. For this paper, the proposed method achieves the best results for medical image segmentation when two loss functions are used simultaneously.

4.3.2 Effectiveness of PCFormer module

To verify the effect of PCFormer module usage on network detection, ablation experiments

are conducted on Glas, ISIC, and LGG datasets. VGG16 is used as the baseline model of the network backbone, and different combinations of PCFormer modules are utilized to replace the corresponding down sampling layers. The experimental results are shown in **Table 6**.

Table 6. Result of the effectiveness of the PCFormer module (Bold numbers indicate the best performance)

Usage of PCFormer module				**Dic**			**IoU**		
First layer	Second layer	Third layer	Fourth layer	Glas	ISIC	LGG	Glas	ISIC	LGG
				0.851	0.883	0.907	0.756	0.823	0.826
√				0.879	0.912	0.927	0.786	0.852	0.869
	√			0.872	0.904	0.921	0.776	0.841	0.860
		√		0.874	0.914	0.926	0.779	0.855	0.867
			√	0.871	0.914	0.924	0.774	0.856	0.864
√	√			0.866	0.913	0.925	0.767	0.854	0.867
√		√		0.882	0.912	0.923	0.791	0.853	0.863
√			√	0.893	0.912	0.924	0.809	0.853	0.865
	√	√		0.874	0.913	0.926	0.779	0.855	0.868
	√		√	0.891	0.912	0.926	0.805	0.853	0.867
		√	√	0.870	0.912	0.925	0.772	0.852	0.866
√	√	√		0.905	0.913	0.925	0.829	0.855	0.866
√	√		√	0.878	0.910	0.925	0.786	0.850	0.867
√		√	√	0.877	0.913	0.927	0.783	0.854	0.869
	√	√	√	0.868	0.910	0.925	0.770	0.851	0.866
√	√	√	√	**0.930**	**0.915**	**0.930**	**0.872**	**0.858**	**0.874**

From **Table 6**, we can see that the segmentation performance of the network changes to different degrees after applying different numbers and combinations of PCFormer modules in the network. Using the PCFormer module only in the first layer is the best combination to using only one PCFormer module. Its average scores for Dic, IoU in the three datasets are 90.6% and 83.6%. Using the PCFormer module only in the first and fourth layers is the best combination of using only two PCFormer modules. Its average scores of Dic, IoU in the three datasets are 91.0% and 84.2%. Using the PCFormer module only in the first, second and third layers is the best combination of using only three PCFormer modules. The average scores of Dic, IoU in the three datasets are 91.4% and 85.0%. The average scores of Dic, IoU in the three datasets using a combination of four PCFormer modules at the same time were 92.5% and 86.8%. Collectively, the segmentation performance of the network gradually increases as the number of PCFormer modules used increases, and the best results for medical image segmentation are achieved when PCFormer modules are used in each layer.

4.3.3 Effectiveness of CDT module

The CDT module is the feature enhancement module of the method in this paper. In order to synthetically evaluate the impact of the CDT module on this paper, ablation experiments are conducted on Glas, ISIC, and LGG datasets. The experiments use different combinations of CDT modules to replace the feature fusion process of direct splicing. The experimental results are shown in **Table 7**.

From **Table 7**, it can be found that when only CDT module is used, the segmentation performance of the network on different datasets changes as the usage level changes. When the network uses the CDT module in the first layer of the feature fusion process, the Dic and IoU metrics scores of our method on the LGG dataset are the highest among those also using the CDT module in only one layer, reaching 92.4% and 86.4%, respectively. When the

network uses the CDT module in the second layer of the feature fusion process, the Dic and IoU metrics scores of our method on the Glas dataset are the highest among those also using the CDT module in only one layer, reaching 90.8% and 83.4%, respectively. When the network uses the CDT module in the feature fusion process in the third layer, the Dic and IoU metrics scores of the proposed method on the ISIC dataset are the highest among those also using the CDT module in only one layer, reaching 91.0% and 85.5%. Analyzing the cases of using the CDT module in both layers in **Table 7**, it can be found that the case of including the CDT module in the first layer feature fusion process has superior segmentation results on the LGG dataset than the case of not including the CDT module in the first layer feature fusion process. The segmentation results on the Glas dataset with the CDT module included in second-layer feature fusion were superior to those without the CDT module included in the second-layer feature fusion. The segmentation results on the ISIC dataset with the CDT module included in the third-layer feature fusion process are stronger overall than those without the CDT module included in the third-layer feature fusion process. When the CDT module is used simultaneously in the three-layer feature fusion process, the Dic and IoU scores of the network on the three datasets are greatly increased, and the average scores of Dic and IoU reach 92.5% and 86.8%, respectively. Comprehensively, the best results and generalization performance of medical image segmentation are achieved when the three-layer feature fusion process uses the CDT module at the same time.

Table 7. The quantitative result on the CDT module
(Bold numbers indicate the best performance)

Usage of CDT module			**Dic**			**IoU**		
First layer	Second layer	Third layer	Glas	ISIC	LGG	Glas	ISIC	LGG
			0.831	0.893	0.897	0.738	0.823	0.829
√			0.863	0.910	0.924	0.763	0.850	0.864
	√		0.908	0.910	0.924	0.834	0.850	0.864
		√	0.899	0.914	0.923	0.820	0.855	0.863
√	√		0.893	0.910	0.927	0.809	0.851	0.869
√		√	0.866	0.909	0.926	0.767	0.849	0.867
	√	√	0.899	0.912	0.920	0.820	0.852	0.858
√	√	√	**0.930**	**0.915**	**0.930**	**0.872**	**0.858**	**0.874**

4.3.4 The impact of backbone and feature enhancement modules

In this paper, we change the downsampling method and feature fusion method based on U-Net. The PCFormer module is used to downsample layer by layer, and then a feature enhancement module, the CDT module, is used to enhance the feature images to be fused with features. To evaluate the effect of these two structures on the network segmentation performance, ablation experiments were performed on Glas, ISIC, and LGG datasets. The experimental results are shown in **Table 8**.

Table 8. Result of the backbone and feature enhancement modules
(Bold numbers indicate the best performance)

Whether to use PCFormer/CDT module		Dic			IoU		
PCFormer module	CDT module	Glas	ISIC	LGG	Glas	ISIC	LGG
		0.823	0.912	0.927	0.703	0.854	0.869
	√	0.871	0.913	0.925	0.774	0.854	0.867
√		0.887	0.903	0.922	0.800	0.839	0.861
√	√	**0.930**	**0.915**	**0.930**	**0.872**	**0.858**	**0.874**

From **Table 8**, it can be found that the mean values of Dic and IoU on the three datasets for the network structure without the CDT module and PCFormer module are 88.7% and 80.9%, respectively, while its segmentation effect on the LGG dataset is slightly effective than that of the network structure using only CDT module and PCFormer module network structure. The mean values of Dic and IoU on the three datasets for the network structure using only the CDT module are 90.3% and 83.2%, respectively, while its segmentation effect on the ISIC dataset is better than that of the network structure with the out CDT module and PCFormer module and the network structure using only PCFormer module network structure without CDT module and PCFormer module and with PCFormer module only. The average scores of Dic and IoU on the three datasets for the network structure using the only PCFormer module are 90.4% and 83.3%, respectively, while its segmentation effect on the Glas dataset is better than that of the network structure with the out CDT module and PCFormer module and the network structure using only CDT module network structure without CDT module and PCFormer module and with CDT module only. Comparing the experimental data, we found that the average segmentation effect of the network structure using the PCFormer module alone and the network structure using the CDT module alone is improved more than that of the network structure without the CDT module and PCFormer module, and the different structures have the optimal performance on different data sets. The mean values of Dic and IoU for the network structure using both the PCFormer module and CDT module are 92.5% and 86.8%, respectively, while their segmentation results are optimal on all three datasets. Taken together, for the U-Net network, the best results and generalization of medical image segmentation are obtained when both the PCFormer module and CDT module are used.

5. Conclusion

In this paper, we propose a multi-level semantic enriched neural network approach for medical image segmentation. In the encoder structure, we present the PCFormer module. The PCFormer module uses pooling as the structure of the token mixer in the transformer to avoid unnecessary parameter operations in the down sampling process, thus avoiding the parameter explosion problem caused by using the transformer in the down sampling process. Among the decoder structures, the CDT module is presented, in which Cbam attention module utilizes its unique hybrid attention mechanism with the broad perceptual field generated by the DCC module, which can effectively optimize the problem of insufficient extraction of global and local feature regions by traditional networks. The MDTransformer module in the CDT module is employed to address the problem that traditional networks cannot effectively distinguish between background and target regions. The results on Glas, SIIM-ACR, ISIC and LGG datasets show that the method in this paper can significantly improve the segmentation accuracy of medical images. At present, the method in this paper is only suitable for segmentation on two-dimensional medical image slices, and in future work will be devoted to the task of segmenting medical images in higher dimensions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62076117 and 62166026, the Jiangxi Science and Technology Program under Grant No. 20232ABC03A32 and the Jiangxi Provincial Natural Science Foundation under Grant No. 20224BAB212011, 20232BAB212008 and 20232BAB202051.

References

[1] F. Shamshad, S. Khan, S. Zamir, M. Khan, M. Hayat, F. Khan and H. Fu, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, p. 102802, Jun. 2023. Article (CrossRef Link).

[2] Z. Mirikharaji, K. Abhishek, A. Bissoto, C. Barata, S. Avila, E. Valle, M. Celebi and G. Hamarneh, "A survey on deep learning for skin lesion segmentation," *Med. Image Anal.*, vol. 88, p. 102863, Jun. 2023. Article (CrossRef Link).

[3] D. Gai, J. Zhang, Y. Xiao, W. Dong, Y. Zhong, and Y. Zhong, "RMTF-Net: Residual Mix Transformer Fusion Net for 2D Brain Tumor Segmentation," *Brain Sci.*, vol. 12, no. 9, p. 1145, Sep. 2022. Article (CrossRef Link).

[4] D. Gai, X. Shen, H. Chen and P. Su, "Multi-focus image fusion method based on two stage of convolutional neural network," *Signal Process.*, vol. 176, p. 107681, Nov. 2020. Article (CrossRef Link).

[5] Q. Wang, W. Min, Q. Han, Q. Liu, C. Zha, H. Zhao and Z. Wei, "Inter-Domain Adaptation Label for Data Augmentation in Vehicle Re-Identification," *IEEE Trans. Multimedia*, vol. 24, pp. 1031-1041, 2021. Article (CrossRef Link).

[6] O. Ronneberger, P. Fischer, T. Brox, N. Navab, J. Hornegger, W. Wells and A. Frangi, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. of the 18th International Conference on Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, Munich, Germany, pp. 234-241, Oct. 2015. Article (CrossRef Link).

[7] O. Oktay, J. Schlemper, L. Folgoc, M. Lee and H. Mattias "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv Preprint arXiv:1804.03999*, 2018. Article (CrossRef Link).

[8] Z. Zhou, M. M. R. Siddiquee, and N. Tajbakhsh, "UNet plus plus: A Nested U-Net Architecture for Medical Image Segmentation," in *Proc. of the 4th International Workshop on Deep Learn. Med. Image Anal. (DLMIA*), Granada, Spain, pp. 3-11, Sep. 2018. Article (CrossRef Link).

[9] Z. Zhang, H. Fu and H. Dai, "ET-Net: A Generic Edge-aTtention Guidance Network for Medical Image Segmentation," in *Proc. of the 22nd International Conference on Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, Shenzhen, China, pp. 442-450, Oct. 2019. Article (CrossRef Link).

[10] J. Chen, Y. Lu, Q. Yu, X. Luo, A. Ehsan and W. Yan, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv Preprint arXiv:2102.04306*, 2021. Article (CrossRef Link).

[11] J. Schlemper, O. Oktay, M. Schaa, M. Heinrich and B. Kainz, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197-207, Apr. 2019. Article (CrossRef Link).

[12] X. Wang, R. Girshick, A. Gupta and K. He, "Non-local Neural Networks," in *Proc. of the 31st IEEE/CVF Conf. on Comput. Vis. Pattern Recog.*, Salt Lake City, UT, USA, pp. 7794–7803, Jun. 2018. Article (CrossRef Link).

[13] I. Bello, B. Zoph, Q. Le, A. Vaswani and J. Shlens, "Attention Augmented Convolutional Networks," in *Proc of the 2019 IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)*, Seoul, Korea (South), pp. 3286–3295, Oct. 2019. Article (CrossRef Link).

[14] N. Carion, F. Massa, G. Synnaeve, U. Nicolas and K. Alexander, "End-to-End Object Detection with Transformers," in *Proc. of Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, US, pp. 213–229, Nov. 2020. Article (CrossRef Link).

[15] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian and M. Wang, "Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation," in *Proc. of the 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, pp. 205-218, Oct. 2021. Article (CrossRef Link).

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov and D. Weissenborn, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv Preprint arXiv:2010.11929*, 2020. Article (CrossRef Link).

[17] P. Ramachandran, N. Parmar, A. Vaswani and I. Bello, “Stand-Alone Self-Attention in Vision Models,” in *Proc. of the 33rd Conference on Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, pp. 68–80, Dec. 2019. Article (CrossRef Link).
[18] Y. Li, T. Yao, Y. Pan and T. Mei, "Contextual Transformer Networks for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1489-1500, Feb. 2023. Article (CrossRef Link).
[19] M. Fraz, P. Remagnino, A. Hopp, “Blood vessel segmentation methodologies in retinal images–a survey,” *Comput. Meth. Prog. Bio.*, vol. 108, no. 1, pp. 407-433, Oct. 2012. Article (CrossRef Link).
[20] H. Fu, Y. Xu, S. Li, D. Wong and J. Liu, “DeepVessel: Retinal Vessel Segmentation via Deep Learning and Conditional Random Field,” in *Proc. of the 19th International Conference on Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, Istanbul, Turkey, pp. 132-139, Oct. 2016. Article (CrossRef Link).
[21] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao and T. Zhang “CE-Net: Context Encoder Network for 2D Medical Image Segmentation,” *IEEE Trans. Med. Imaging*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019. Article (CrossRef Link).
[22] M. Li, Y. Chen, Z. Ji, K. Xie, S. Yuan, Q. Chen and S. Li, “Image Projection Network: 3D to 2D Image Segmentation in OCTA Images,” *IEEE Trans. Med. Imaging*, vol. 39, no. 11, pp. 3343-3354, Nov. 2020. Article (CrossRef Link).
[23] Z. Liu, Y. Song, V. Sheng, L. Wang and R. Jiang, “Liver CT sequence segmentation based with improved u-net and graph cut,” *Expert Syst. Appl.*, vol. 126, pp. 54-63, Jul. 2019. Article (CrossRef Link).
[24] S. Li, G. Tso, and H. Kaijian, “Bottleneck feature supervised u-net for pixel-wise liver and tumor segmentation,” *Expert Syst. Appl.*, vol 145, p. 113131, May. 2020. Article (CrossRef Link).
[25] N. Ibtehaz and M. S. Rahman, “MultiResUNet: Rethinking the u-net architecture for multimodal biomedical image segmentation,” *Neural Networks*, vol. 121, pp. 74-87, Jan. 2020. Article (CrossRef Link).
[26] Y. Jiang, N. Tan, T. Peng and H. Zhang, “Retinal Vessels Segmentation Based on Dilated Multi-Scale Convolutional Neural Network,” *IEEE Access*, vol. 7, pp. 76342-76352, Jul. 2019. Article (CrossRef Link).
[27] Y. Liu, C. Xu, Z. Chen, C. Chen, H. Zhao and X. Jin, “Deep Dual-Stream Network with Scale Context Selection Attention Module for Semantic Segmentation,” *Neural Process. Lett.*, vol. 51, no. 3, pp. 2281-2299, Jun. 2020. Article (CrossRef Link).
[28] H. Zhu, B. Wang, X. Zhang, and J. Liu, “Semantic image segmentation with shared decomposition convolution and boundary reinforcement structure,” *Appl. Intell.*, vol. 50, no. 9, pp. 2676–2689, Sep. 2020. Article (CrossRef Link).
[29] D. Kushnure and N. Talbar, “MS-UNet: A multi-scale UNet with feature recalibration approach for automatic liver and tumor segmentation in CT images,” *Comput. Med. Imaging Grap.*, vol. 89, p. 101885, Mar. 2021. Article (CrossRef Link).
[30] S. Hu, J. Zhang, and Y. Xia, “Boundary-aware network for kidney tumor segmentation,” in *Proc. of the Int. Workshop on Mach. Learn. Med. Imaging (MLMI)*, Lima, Peru, pp. 189-198, Sep. 2020. Article (CrossRef Link)
[31] S. Feng, H. Zhao, F. Shi, X. Chen, M. Wang and Y. Ma, “CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation,” *IEEE Trans. Med. Imaging*, vol. 39, no. 10, pp. 3008-3018, Oct. 2020. Article (CrossRef Link)
[32] D. Peng, X. Yu, W. Peng and J. Lu, “DGFAU-Net: Global feature attention upsampling network for medical image segmentation,” *Neural Comput. Appl.*, vol. 33, no. 18, pp. 12023-12037, Sep. 2021. Article (CrossRef Link)
[33] Y. Zhang, H. Liu, and Q. Hu, “TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation,” in *Proc. of the 24th International Conference on Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, pp. 14–24, Sep. 2021. Article (CrossRef Link)

[34] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu and D. Zhang, "DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, May 2022. Article (CrossRef Link)
[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," in *Proc. of the 31st Conference on Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017. Article (CrossRef Link)
[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc of the 18th IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)*, Montreal, QC, Canada, pp. 9992–10002, Oct. 2021. Article (CrossRef Link)
[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017. Article (CrossRef Link)
[38] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu and N. Sang, "BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation," in *Proc. of the 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, pp. 334-349, Oct. 2018. Article (CrossRef Link)
[39] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen and N. Sang, "BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051-3068, Nov. 2021. Article (CrossRef Link)
[40] J. Valanarasu, V. Sindagi, I. Hacihaliloglu, and M. Patel, "KiU-Net: Towards Accurate Segmentation of Biomedical Images Using Over-Complete Representations," in *Proc. of the 23rd International Conference on Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, Lima, Peru, pp. 363–373, Oct. 2020. Article (CrossRef Link).
[41] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, S. Song, "Stepwise Feature Fusion: Local Guides Global," in *Proc. of the 25th International Conference on Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, Singapore, pp. 110-120, Sep. 2022. Article (CrossRef Link).
[42] H. Wang, P. Cao, J. Wang, and O. Zaiane, "UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer," in *Proc. of the 36th AAAI Conf. Artif. Intell.(AAAI)*, vol. 36(3), pp. 2441-2449, Jun. 2022. Article (CrossRef Link).
[43] H. Huang, S. Xie, L. Lin, Y. Iwamoto, X. Han, W. Chen and R. Tong, "ScaleFormer: Revisiting the Transformer-based Backbones from a Scale-wise Perspective for Medical Image Segmentation," in *Proc. of the 31st Int. Joint Conf. Artif. Intel. (IJCAI)*, Vienna, Austria, pp. 964-971, Jul. 2022. Article (CrossRef Link).

Di Gai received the M.E. and Ph.D. degrees in College of Computer Science and Technology from Jilin University, China, in 2018 and 2021, respectively. He is currently a lecturer, School of Mathematics and Computer Sciences, Nanchang University, China. He also is an assistant researcher in Jiangxi Key Laboratory of Smart City, China. His research interests include medical image processing and pattern recognition, especially on image fusion.

Heng Luo is enrolled in the School of Software at Nanchang University, pursuing a B.S. degree. His research interests include computer vision and deep learning.

Jing He is enrolled in the School of Software at Nanchang University, pursuing a B.S. degree. Her research interests include computer vision and deep learning.

Pengxiang Su received the Ph.D. degree in College of Computer Science and Technology from Jilin University, China in 2022. He is a lecturer at School of Software, Nanchang University, China. His research interests include computer vision, human motion analysis, and image recognition.

Zheng Huang received the B.E. degree in School of Civil Engineering and Communication, North China University of Water Resources and Electric Power, China, in 2021. He is currently pursuing the master's degree in computer technology at Nanchang University, China. His research interests include computer vision and deep learning.

Song Zhang is enrolled in the School of Software at Nanchang University, pursuing a B.S. degree. His research interests include computer vision and deep learning.

Zhijun Tu graduated from the School of Computer Science, Zhengzhou University. He is now working as an experimenter in the School of Information Engineering, Nanchang University. His research interests include computer vision and deep learning.