



FTUNet: A Feature-Enhanced Network for Medical Image Segmentation Based on the Combination of U-Shaped Network and Vision Transformer

Yuefei Wang¹ · Xi Yu² · Yixi Yang³ · Shijie Zeng¹ · Yuquan Xu² · Ronghui Feng²

Accepted: 8 January 2024 / Published online: 4 March 2024
© The Author(s) 2024

Abstract

Semantic Segmentation has been widely used in a variety of clinical images, which greatly assists medical diagnosis and other work. To address the challenge of reduced semantic inference accuracy caused by feature weakening, a pioneering network called FTUNet (Feature-enhanced Transformer UNet) was introduced, leveraging the classical Encoder-Decoder architecture. Firstly, a dual-branch Encoder is proposed based on the U-shaped structure. In addition to employing convolution for feature extraction, a Layer Transformer structure (LTrans) is established to capture long-range dependencies and global context information. Then, an Inception structural module focusing on local features is proposed at the Bottleneck, which adopts the dilated convolution to amplify the receptive field to achieve deeper semantic mining based on the comprehensive information brought by the dual Encoder. Finally, in order to amplify feature differences, a lightweight attention mechanism of feature polarization is proposed at Skip Connection, which can strengthen or suppress feature channels by reallocating weights. The experiment is conducted on 3 different medical datasets. A comprehensive and detailed comparison was conducted with 6 non-U-shaped models, 5 U-shaped models, and 3 Transformer models in 8 categories of indicators. Meanwhile, 9 kinds of layer-by-layer ablation and 4 kinds of other embedding attempts are implemented to demonstrate the optimal structure of the current FTUNet.

Keywords Semantic segmentation · Vision transformer · U-shaped network · Attention mechanism

✉ Yuefei Wang
wangyuefei@cdu.edu.cn
<https://github.com/YF-W/FTUNet>

¹ College of Computer Science, Chengdu University, 2025 Chengluo Rd., Chengdu 610106, Sichuan, China

² Stirling College, Chengdu University, 2025 Chengluo Rd., Chengdu 610106, Sichuan, China

³ Institute of Cancer Biology and Drug Discovery, Chengdu University, 2025 Chengluo Rd., Chengdu 610106, Sichuan, China

1 Introduction

With the extensive and in-depth study of deep learning, Computer Vision (CV) has been fully and significantly developed with rich technologies and extensive applications [1]. As a kind of computer core and advanced vision technology, Semantic Segmentation defines the image processing mechanism through the network structure, completes the object segmentation, and realizes the scene understanding [2–4]. In essence, this kind of technology belongs to pixel-level classification model. It uses Convolutional Neural Network (CNN) to capture rich information, allowing applications to infer knowledge from images [5]. In addition, Semantic Segmentation plays a central role in medical image diagnosis [6–9]. In a variety of important scenarios such as image-guided interactions, radiological diagnostics, and radiology [8], it can help medical personnel effectively extract the lesion area, greatly reduce the work intensity, and accurately carry out disease analysis. Since 1992, Semantic Segmentation in medical scenes involves dividing an image into non-overlapping regions, forming the whole image [10]. Since then, the convenience brought by the subsequent technological development has shown a unique role in the medical environment, and has achieved certain results in a series of multi-modal tasks such as brain lesion segmentation [11, 12], skin cancer tissue segmentation [13, 14], and tongue segmentation [15].

Semantic Segmentation has rich technical support. With the development and influence of deep learning, a large number of different types of models have been born. Such as region-based [16], weakly supervised methods [17, 18], and fully convolutional networks [19]. Among them, Semantic Segmentation based on convolutional networks is the most popular method in recent years. This kind of framework divides the model into two parts: encoding and decoding. The coding region is used to extract feature information to infer knowledge, and the decoding region recovers image size to determine position. In recent years, the field of medical image segmentation has seen the emergence of several pivotal network architectures, such as SegNet [20], Fully Convolutional Networks (FCN) [19], and U-Net [21], among others. These models have significantly contributed to the technical foundation of semantic segmentation. Moreover, a multitude of variant network models has surfaced subsequently, adding further depth and richness to research in this domain. While the CNN has achieved remarkable progress over the past decade, its limitations are becoming increasingly conspicuous. The feature extraction method, characterized by strong local constraints, gives rise to a notable challenge known as the perception limitation problem. This issue substantially hampers the mining of semantic features in image processing, even with the application of techniques like dilated convolution, which, to a certain extent, alleviate the problem [22].

In the past two years, a new global information aware CV structure Transformer has aroused the attention and discussion of scholars once it was proposed, and has developed rapidly in subsequent research, keeping pace with CNN structure in the CV field [23, 24]. In 2017, the Google team adopted attention structures including Q, K, and V to achieve translation tasks in the field of NLP (Natural Language Processing) with high quality, and demonstrated the substitutability of this mechanism for convolution [25]. Meanwhile, another kind of ViT (Vision Transformer) model which is applied well in Visual recognition is born [26]. As a new structure in the field of CV, the ViT is a framework that directly applies the Transformer to the sequence of image patches. In this model, the image is segmented into several 2D patches, and the Transformer is directly embedded in the image patches sequence in the classification task to enhance the self-attention of classification. This mechanism of embedding pure transformer directly into the model has aroused extensive discussion, and has spawned a large number of application research in different fields such as object detection

and video understanding [27, 28]. At present, some work still embeds the ViT structure directly into the model, and combines the convolution operation to realize the semantic encoding and decoding [29]. Compared with other Semantic Segmentation models, the effect is significantly improved. Recent studies have shown that different variants of ViT, such as DeiT (Data-efficient image Transformers) [30] and Swin [31], have achieved remarkable results in CV and are emerging in the subfield of Semantic Segmentation research. In a word, Transformer has good features in the field of Semantic Segmentation by virtue of its excellent features of attention mechanism, and has been favored by more and more scholars.

1.1 Motivation

1) *Rough processing of global context information* After the Transformer structure implements 1D conversion and embedding on segmented image patches, multiple Transformer layers are adopted to extract information about the Global context. While serialized extraction systematically deepens the impact of image semantic mining, the fusion with U-shaped networks frequently adopts a simplistic and rudimentary mode. Regrettably, this strategy lacks the differentiated integration of deep and shallow semantic information, leading to an insufficient extraction of valuable image information. 2) *Inadequacy of Feature Mining at the Bottleneck* The Bottleneck is a crucial link in U-shaped processing, housing rich semantic information. The original U-Net model adopts double convolution processing, while the existing optimization structures often do not pay too much attention to this structure part [32]. Therefore, the lack of multi-scale convolution information extraction results in a waste of semantic information. 3) *Limitations of differences in channel attentional features* Channel attention is an information-enhanced structure that can effectively express the importance of different channels. The traditional method uses the maximum pooling or average pooling of channel pixels to represent the importance of the channel and acts on the corresponding tensor. However, this method cannot amplify the difference between different tensors, and the association of channel representation values is quite accidental and coincidental. Therefore, the existing attention mechanism urgently needs a method to enlarge the difference of features to enhance or suppress the corresponding channel features.

In view of the above motivations, based on the combination of ViT and U-Net model as the Baseline, this study aims at the coarse-grained problem of Transformer layer serialization, the problem of Bottleneck feature loss, and the problem of small channel feature difference. The main work is as follows:

- (1) A mechanism Layer Transformer (LTrans) for multilevel semantic fusion between Transformer and U-Net is proposed. The continuity of the internal recurrent structure of ViT is not intervened by traditional networks. In contrast to conventional methods, nuanced semantic details are extracted from Transformer layers 3, 6, 9, and 12 in our approach. These details are ingeniously integrated into U-type input layers 1–4, leading to a sophisticated fusion of the Vision Transformer (ViT) and U-Net architectures. This process achieves a deep amalgamation of their respective structures.
- (2) A semantic feature enhancement model, ASPP-Inception (AI), is proposed. Inspired by the Google Inception module, a parallel ASPP Inception module is designed with multi-microscale convolution kernels at the key point of coding and decoding connection. This facilitates the improved learning and strengthening of semantic feature information at the object's edge. To mitigate local limitations, the module incorporates the multi-dilated convolution structure ASPP, accompanied by the utilization of a pyramid-like sub-module to impact image segmentation within a larger receptive field

- (3) Rethinking of channel weights, and a Polarized Attention (PA) mechanism is proposed. The application of the transformation relationship between tensor and list is employed to magnify and differentiate the importance gap of different channels from the perspective of mathematical statistics. Consequently, important features are fortified, and irrelevant features are subdued, thereby optimizing channel attention.

The structure of this paper is as follows. The second section summarizes the research work related to this paper. The third section discusses the overall model and describes the basic composition of the design structure. The fourth section carries out the comparative demonstration of the models and integrates the structure ablation analysis. The last section is the conclusion.

2 Related Work

2.1 Medical Image Segmentation

With the development of deep learning networks, researchers began to try to use image segmentation technology to assist clinical medical diagnosis. Early researchers tried to apply CNN [5] to the field of semantic inference, and proposed the technology of adopting convolutional neural network to infer image semantics [33]. This innovation has made great achievements in medical treatment. Since then, a large number of classic networks continue to emerge. Long et al. [19] proposed FCN based on the fully convolutional neural network, which replaces the fully connected part of CNN with a convolutional block, so that the model can flexibly handle pixel-level classification problems. Besides, Pyramid Scene Parsing Network (PSPNet) [34] applies a pyramid module to extract context information by fusing features from four pooling scales, which improves the feature extraction mechanism. In addition, the DeepLab family [35–38] attempted to employ ResNet to increase the depth of the network with the support of atrous convolution, and realized the fusion of more contextual information while expanding the receptive field. As an equally excellent model, the embedding of attention mechanism enables Attention Deeplabv3 + to learn more feature information [39].

Due to the increasing influence of FCN network in the field of Semantic Segmentation, Olaf Ronneberger et al. [21] proposed the U-Net model in 2015. This model adopts the framework of encoding and decoding, which standardizes the network structure of feature extraction and has positive segmentation effect. Since the structure of Encoder and Decoder can effectively complete the task of image partition, quite a few models based on Encoder and Decoder have been generated. For example, Refine Net [40] combines Residential Conv Unit (RCU), Multi resolution Fusion, and Chained Residential Pooling to add them to the Decoder, which can more effectively generate high-resolution visual features. Besides, SegNet [20] employs the location information to recover the feature map instead of deconvolution, reducing the learning of up-sampling during training. Furthermore, more scholars regard U-Net as Baseline to achieve updates and breakthroughs. For instance, inspired by the shape of the model, a large number of structures that can inherit encoding and decoding types have been created, such as W-net [41], X-net [42], etc. Because of the combination of down-sampling and up-sampling, the semantic information can be basically captured. In addition to model shape optimization, a multitude of scholars choose to study from the direction of model structure. As an example, Context Encoder Network (CE-net) [43] added the Dense Atrous Convolution module (DAC) and Residual Multi-kernel pooling (RMP) to the Bottleneck layer of U-Net.

Multi-scale atrous convolution and pooling operations are used to obtain more contextual features while retaining more spatial information. Likewise, Outlined Attention U-net (OAU-net) replaced the double convolution of the third and fourth layers with Res2 Conv2d, and adopts multi-scale parallelized convolution to improve the fusion of multi-scale features [44]. Moreover, Skip Connection is also an important part of U-net, which recovers the edge information lost in the down-sampling process and retains more dimension information by splicing. Small Attention-UNet (SmaAt-UNet) added a spatial attention mechanism CBAM module at the Skip Connection to capture the semantic information in the high-level feature map [45]; DC-net [46] increased maximum pooling at Skip Connection.

2.2 Transformer

2.2.1 Variant Networks of Transformer

With the introduction of the Transformer prototype by Vaswani et al. [25] it has been widely used not only in Natural Language Processing (NLP) field but also has contributed to the research in CV. For example, Dosovitskiy et al. introduced the Vision Transformer model [26] that merges image patches and Transformer, initiating the ViT variant network research boom in the CV field. Since then, many researchers have applied ViT and its variant networks to a variety of tasks in the CV field, such as Pose Estimation [47–49], Object Detection [27, 50, 51], and Classification [52]. He et al. integrated the mask idea from BERT [53] into ViT, proposing masked autoencoders [54]. This enables the network to learn existing features while predicting on masked features, enhancing robustness and accuracy. In addition, a new Transformer variant, Swin-Transformer [55], uses the shifted windows technique to reduce the training cost of the model, which is favored by contemporary researchers and has inspired many scientific works. All of these models play an important role in various tasks of the CV field, and it is foreseeable that the boom started by Transformer will continue in the near future.

2.2.2 Integration of Transformer with Other Networks

Since the Transformer has a good track record in CV, many researchers have combined it with other networks, including but not limited to combining the Transformer with CNN [56–58], Transformer with DeepLab series [59, 60], and Transformer with FCN[61]. Among the more prominent ones, He et al. [62] proposed Fully Transformer Network model to learn long-range contextual information, making their model beyond the contemporaneous State-Of-The-Art (SOTA) CNNs. Similarly, Xie et al. [63] proposed the Convolutional neural network and Transformer (CoTr) model to capture features at key locations with a deformable Transformer instead of focusing on global features, thus reducing computational costs and enabling improved accuracy and robustness. Furthermore, Wang et al. [64] put forward the Max-DeepLab model using Mask Transformer to predict masks and classes directly. Also inspired by masked Transformer, Yu et al. combined it with clustering ideas and proposed a Clustering Mask Transformer (CMT-DeepLab), which became the new SOTA on the MS COCO (Microsoft Common Objects in Context) dataset [65] at that time. From what has been discussed above, we can see that the Transformer can be better combined with other networks to improve the accuracy and stability of the model, and thus better accomplish the tasks in the CV field.

2.2.3 Transformer + U-Net

The Transformer plays a pivotal role in Semantic Segmentation tasks, particularly in the domain of medical image segmentation. Consequently, a substantial body of research has emerged, combining the Transformer architecture with U-Net. These hybrid models have found application across various domains. As can be seen from Table 1, Hatamizadeh and his team applied Transformer to 3D medical datasets by proposing the U-Net Transformers (UNETR) model [66], where the model inverts U-Net. To be more specific, the model firstly upsamples the deep features of the Transformer, then continuously splices the deep features with the shallow features, and then fuses this context information by convolution and other means as ways to capture more 3D information. In addition, many models apply Transformer to 2D datasets, for example, Wu et al. [29] introduced Feature Adaptive Transformers (FAT-Net), merging features from the 12th Transformer layer and U-Net coding layer, improving accuracy in capturing skin lesion sites. Moreover, inspired by U-Net, Hu Cao et al. designed an Encoder-Decode architecture based on symmetric transformer, and used the patch extension layer to perform up-sampling to reshape the feature map of adjacent dimensions to restore the spatial resolution of the feature map [31]. Besides, SUNet adopted Swin Transformer as the basic module to apply it to U-Net for image noise reduction, and reduces parameters through shifting the pane to achieve advanced performance in a large number of pixel visual tasks [67]. In contrast, our work is not only carried out on the 2D Skin dataset but also adds the Cell and Lung datasets to validate the robustness of the model from multiple perspectives.

Besides, the combination of Transformer and U-Net is also very abundant. For example, H Wang et al. [68] applied Transformer in the different positions of the codec body and adopted a Transformer in the Encoder and Decoder to extract features before performing convolution or deconvolution. Furthermore, Gao et al. applied the self-attentive module [69] in both Encoder and Decoder blocks, which can better capture the long-range dependency, but it will increase the computational resources. In order to save computational resources, Teams, like in the Medical Transformer [70], have added the multi-headed attention mechanism to the left side of U-Net's Encoder, integrating it into each layer. Similarly, the TransU-Net is proposed in [71], which firstly extracts features through CNN blocks, and then passes these feature information into Transformer blocks to extract deeper context information. What's more, The Levit-U-Net proposed in [72] is similar to the TransU-Net in that both pass through the CNN block at first, and then combine with the Transformer block to extract deeper features, but the difference is that the Levit-U-Net will process downsample and reshape after the Transformer block, to make the subsequent features can be better integrated. In addition, some other teams apply the Transformer to other locations of the U-shaped network. For example, the U-Net Transformer proposed in [73] applies the attention mechanism Multi-Head Self-Attention (MHSA) and Multi-Head Cross-Attention (MHCA) to the Bottleneck and decoding area, respectively. Likewise, the paper [74] constructs a Twisted Pattern and implements Twisted Information-sharing Pattern for Multi-branched Network, which improves the feature information fusion effect and alleviates the problem of semantic isolation. Moreover, Wang et al. introduced DBUNet (Dual-Decoding Branch U-shaped Network), incorporating the ViT Encoder into the Decoder branches to enhance the representation of shallow features during image upsampling [75]. The results show that DBUNet has excellent lesion extraction effects. Different from these ideas, considering the computational resources and the stability of the model, Our approach applies Transformer to Skip Connection and Bottleneck, blending U-Net and Transformer features via a residual network, and subsequent experiments confirm the model's excellent performance.

Table 1 Recent work on the combination of transformer and U-Net

Model	Transformer position	Description	Year & reference
UNETR	Encoder	Taking features from Transformer 3rd, 6th, 9th, and 12th layer, deconvoluting them before merging with other features, and processing convolution after that	2022, [66]
FAT-Net	Bottleneck	Taking the Transformer's 12th feature and splicing it with the features extracted from the 4th Encoder layer	2022, [29]
Mixed transformer	Encoder; bottleneck; decoder	In Encoder, use Transformer to extract features before convolution. Add Transformer to Bottleneck. In Decoder, processing deconvolution at first and then using Transformer to extract the features	2022, [68]
UTNet	Encoder; decoder	Transformer Encoder is proposed to be used after the Residual Basic Block, and Transformer Decoder is proposed as well to replace the original upsampling part	2021, [69]
Medical transformer	Encoder; decoder	With the Transformer as the main body, the attention mechanism is integrated into each layer of U-Net	2021, [70]
TransUNet	Encoder	The CNN block and Transformer block do not intersect, and the CNN block will pass the extracted features to the Transformer block	2021, [71]
LeViT-UNet	Encoder	The features are first extracted by CNN, and then the extracted features are passed to Transformer, and then downsampling and reshaping are processed	2021, [72]
U-Net transformer	Bottleneck; decoder	Adding Multi-Head Self-Attention to the Bottleneck and applying Multi-Head Cross-Attention to the Decoder	2021 [73],
TP-MNet	Encoder	The network addresses the primary issue of semantic isolation by simultaneously implementing inter-layer connections and cross-branch connections in both the convolutional and ViT branches	2023, [74]
DBUNet	Decoder	Is a class of multi-branch codecs. The attention mechanism is added inside ViT to realize the optimization of ViT	2023, [75]
Ours	Skip connection	Taking features from the 3rd, 6th, 9th, and 12th layers of the Transformer and splice them together with the results from PA	–

3 Model

This section begins by presenting the overall architecture of the proposed model, FTUNet. Subsequently, it sequentially introduces three key component modules within this architecture, providing a detailed discussion of the principles and effects of each component.

3.1 Overall Architecture

In this study, a “Double Encoding—Single Decoding” network is proposed, which is U-shaped and asymmetric. The dual-pathway coding region of the model is composed of the optimized ViT structure and the Contracting Path. In addition, combined with the created attention mechanism and embedded with the designed ASPP-Inception module, multi-components jointly construct a deep network for optimizing the Semantic Segmentation effect. Figure 3 shows the model construction architecture.

Figure 1 demonstrates the FTUNet model proposed in this study, which shows our optimal design of the model from the overall perspective. Observably, the model’s input path is comprised of Transformer ViT and convolution layers. The convolutional layer contributes semantic information derived from the feature channels, whereas ViT incorporates contextual information from a global perspective. As a result, these dual pathways collaboratively contribute to the extraction of comprehensive image features. Additionally, to enhance macro and micro semantics at Bottleneck, we introduce an attention mechanism at the Skip Connection to modulate feature channel weights. In short, the network model structure shows that FTUNet enriches the extracted semantic features, distinguishes the importance of channels, and amplifies the microscopic details of semantics.

The main module design includes:

- 1) *Layer Transformer* The ViT mechanism is introduced to extract the image information of Transformer Layer at different levels and sent to the Decoder to realize semantic fusion.

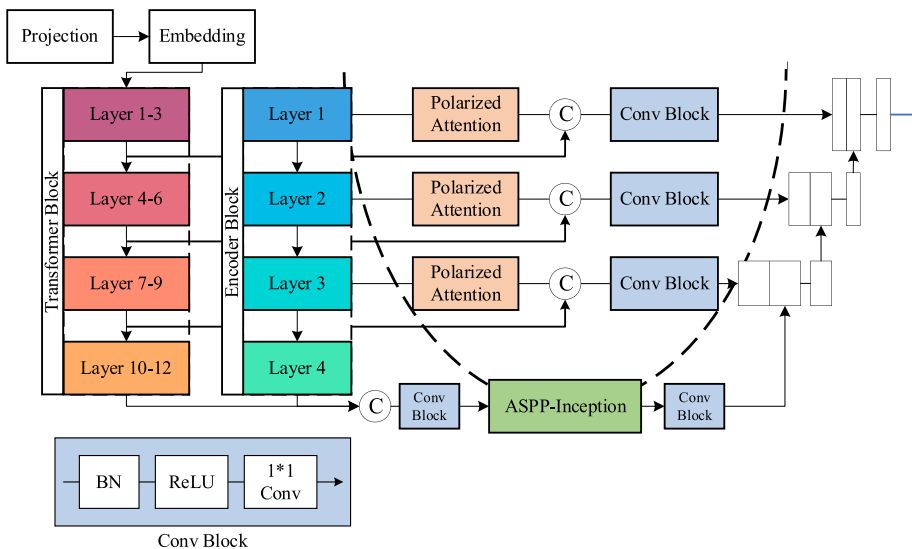


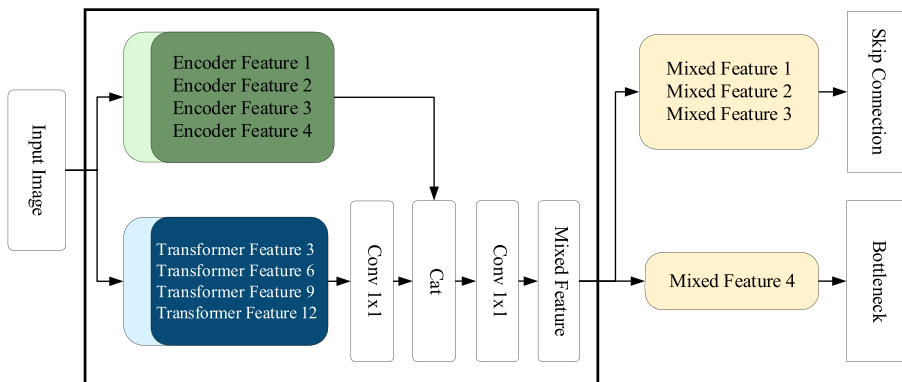
Fig. 1 Overview of the proposed FTUNet framework

- 2) *ASPP-Inception* The Inception structure is constructed, on the one hand, different dilation rates are adopted to enhance the receptive field, and overcome the small convolution kernel defect; meanwhile, the small-core parallel network is employed to deeply mine the edge features of the object and infer the microscopic results of the object edge.
- 3) *Polarized Attention* Enlarge important feature information, capture important channels through mathematical statistics technology and increase their weights, while reducing the weight of secondary feature channels.

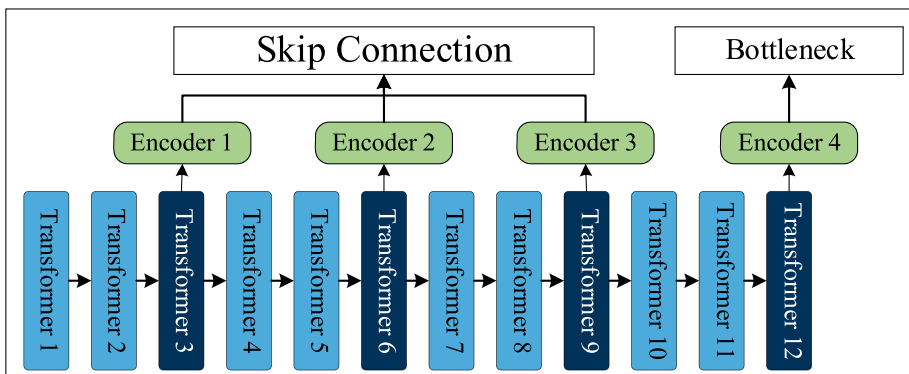
3.2 Layer Transformer

Medical images pose challenges like irregular graphics and fuzzy edge semantics. Depending solely on U-Net for feature extraction may be insufficient for these complexities. Therefore, Transformer is introduced into the model to combine the features extracted by U-Net with those extracted by Transformer.

Figure 2 shows our proposed Layer Transformer, which is a method that extracts context information at different levels in the Transformer and implements the same layer matching



(a) Overall structure of Layer Transformer



(b) The Internal Connection Structure of LTrans and U-shaped network

Fig. 2 Overview of proposed LTrans structure

with different depths of the U-shaped network. In our experiment, we believe that the primary core of this mechanism is constituted by a significant number of Transformer blocks following the embedding structure. Sequential processing can gradually transition from the shallow expression of the image to the deep semantic information with the goal of inferring semantics. However, only combining deep information with U-shaped network will waste shallow semantic expression. Accordingly, we demonstrate through experiments that using the hierarchical structure can mine semantic information more reasonably. In fact, the input images will be sent to the Encoder Block and Transformer Block respectively. The Encoder Block refers to a continuous sequence of convolution operations composed of convolutional encoders, spanning layers 1 to 4. Meanwhile, the Transformer Block refers to a continuous sequence of attention extraction operations composed of the ViT, encompassing layers 1 to 12 in the encoder. The Encoder Feature 1–4 will be obtained from the Encoder Block, and the Transformer Features 1–12 will be collected from the Transformer Block. As shown in Fig. 2a, the semantic fusion of different network depths is achieved by extracting attention information and convolution feature in different levels of two branches. Figure 2b shows the combination of two coding branches with the deepening of the Transformer in a process-oriented manner.

Before merging Encoder Features and Transformer Features, resize the Transformer Features to match Encoder Features using the View function. Subsequently, the number of channels in the Transformer Features is adjusted to align with the Encoder Features through the application of a 1×1 convolution. This ensures that both components carry equal weight, facilitating mutual complementation. Next, the Encoder Features and the Transformer Features need to be merged by using the Cat function. Finally, the number of channels needs to be restored to the same size as the initial Encoder Features by 1×1 convolution, and then Mixed Feature 1–4 can be obtained. Mixed Features 1, 2, 3 are sent to Skip Connection to combine with upsampling features, while Mixed Feature 4 goes to Bottleneck for deeper processing. In short, the Eq. (1) denotes the structure of LTrans.

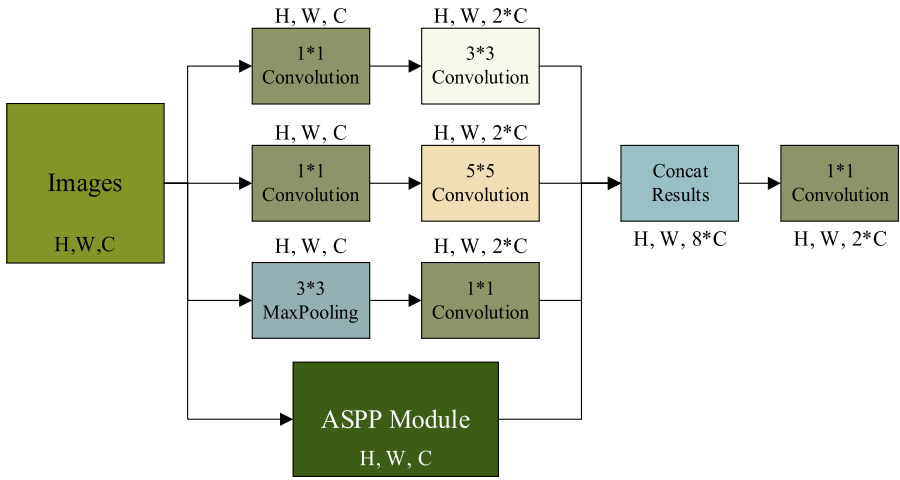
$$F(x) = Conv_{1 \times 1}(Cat(Conv_{1 \times 1}(View(x)), Encoder\ Feature)) \quad (1)$$

3.3 ASPP-Inception

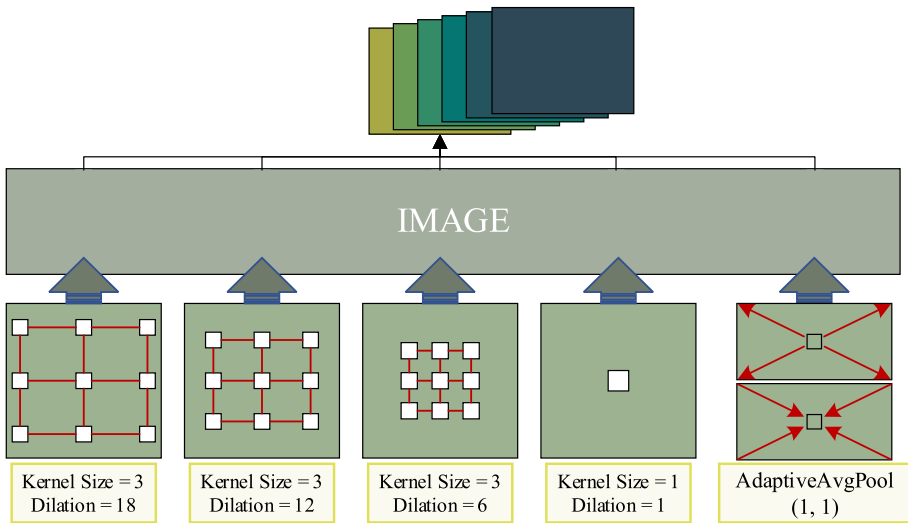
We believe that the most important reason for the loss of accuracy is the limitation of the receptive field. Although pooling can effectively avoid this problem, it also brings a lot of information loss. Based on this, we use the excellent characteristics of dilated convolution to improve the receptive field during convolution, and also pay attention to the context information of different scales. Earlier, He et al. proposed the Spatial Pyramid Pooling module [76], which is a pyramid structure. Subsequently, Chen constructed an ASPP module in the proposed DeepLab v3 network [37], and organized and built a pyramid structure with multiple dilation rates in a cascading manner. This structure consists of a 1×1 convolution and three types of 3×3 convolution with different dilation rates. Furthermore, it gives full play to the advantages of dilated convolution to a greater extent.

Although ASPP has certain advantages in obtaining context semantics, the attention of modules to micro details is often unsatisfactory. According to the Inception structure, this research proposes an ASPP-Inception module, which is used to capture the feature information of different receptive fields while focusing on the local features of the object.

Figure 3 shows the network module unit for ASPP-Inception, where Fig. 3a shows the



(a) Parallel structure of ASPP-Inception



(b) ASPP structure

Fig. 3 Network structure of ASPP-Inception

semantic mining module of the designed Inception structure, and Fig. 3b shows the ASPP structure.

Inspired by the Inception module, the network has constructed three basic convolution branches, namely 3*3, 5*5, Pooling, and ASPP Module. Different convolutions not only take into account that they can extract features of different scales and enrich receptive fields, but also because the convergence speed can be accelerated after the dense matrix is decomposed

into sparse matrixes. In addition, 1*1 convolution can be used to achieve dimension specification. In general, a 1*1 convolution helps the pixel to achieve a fully connected calculation on all features; and through the series of two convolutions, more nonlinear feature results can also be discovered.

On the basis of the above branches, the parallel mechanism is used to implement feature processing. This model effectively organizes the convolution feature tensors of different branches, pastes all the feature results together, and reduces the feature dimension with 1*1 convolution. In short, the ASPP-Inception module has a good effect in overcoming local limitations and mining features under different receptive fields, and has a strong sensitivity to the edges of objects in Semantic Segmentation.

3.4 Polarized Attention

Influenced by style transfer, Lee et al. [77] proposed a style-based recalibration module (SRM), which is a style-based attention mechanism for recalibrating channels. The proposed model not only enhances the representation ability of CNN but also offers the advantages of being lightweight and easily embeddable. Leveraging this inspiration, this paper introduces a channel attention mechanism to heighten the differentiation between image cut regions and edge features. The channel attention mechanism utilizes the median pixel value of each channel as a reference and employs mathematical statistics blocks (MS) to modulate the feature channel weights, thereby amplifying the discrepancy in pixel values across different regions. Simultaneously, it enables the model to capture important features more accurately during training while suppressing the influence of weakly correlated channel weights. The key advancements are as follows:

- 1) Introducing Parallel Global Maximum Pooling (GMP) in addition to Style Pooling.
- 2) Incorporate the method of mathematical statistics, and calculate the tensor value in each channel using Eq. (2) to analyze image feature weights. This process aims to capture key features of the image and enhance the processing ability of image edges and details.

$$f(x) = (y - median)^3 \tag{2}$$

Figure 4 shows the basic structure of the model. First, the original image feature dimensions

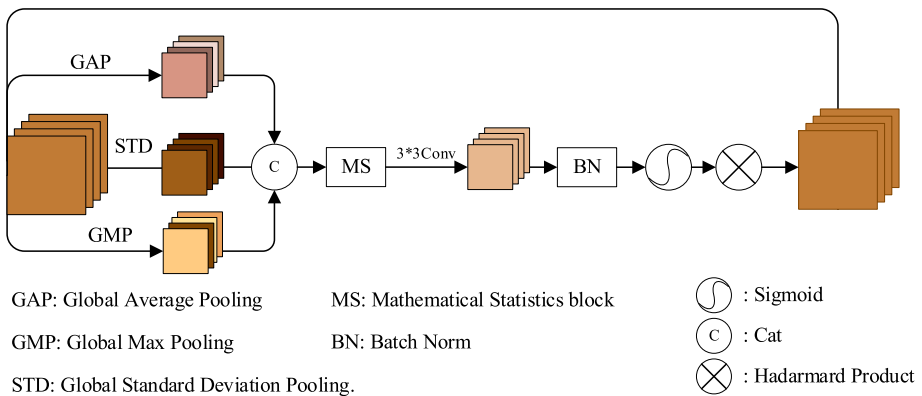


Fig. 4 Channel attention mechanism model: polarized attention

are (b, c, h, w), which are implemented through three parallel pooling processes. Next, the data is combined to obtain the image format of (b, c, 3). Then, during the process of mathematical statistics, calculate the median pixel value for each channel of the combined image. Subtract each pixel value from the respective channel's median to obtain an absolute value, and raise each absolute value to the power of three to obtain a new pixel value. Subsequently, the features of the new image are extracted again through convolution, and the weights of each channel are recalibrated. Apply image processing techniques to mitigate overfitting and activate the image using the Sigmoid function. Finally, multiply the image in the format (b, c, 1, 1) by the original image (b, c, h, w) to obtain the result. Equation (3) represents the core content of this mechanism.

4 Experimental Analysis

This chapter conducts comprehensive experimental exploration of FTUNet, including metric validation, image comparison, training process analysis, and model ablation study.

4.1 Dataset

In this experiment, different data sets are used to achieve comprehensive verification, and medical image segmentation is implemented in the SKIN (<https://www.kaggle.com/code/hashbanger/skin-lesion-segmentation-using-segnet/notebook>), LUNG (<https://www.kaggle.com/datasets/kmader/finding-lungs-in-ct-data>), and the DRIVE (<https://drive.grand-challenge.org/>), respectively.

4.2 Evaluation Metric

In order to show the effect of the model from multiple perspectives, this experiment adopts 8 different evaluation indicators, which are Accuracy, Dice, Jaccard, Precision, Recall, and Specificity. In the composition of factors in their formulas, TP (True Positive) Indicates the pathological pixels that are judged successfully, TN (True Negative) means the background pixels that are judged successfully, FP (False Positive) indicates the background pixels that are incorrectly judged as lesion pixels, while FN (False Negative) denotes the lesion pixels that are wrongly judged as background pixels. The corresponding formula of each index is shown in Eqs. (3–9).

$$Acc = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} \times 100\% \tag{3}$$

$$Dice = \frac{2 \sum TP}{\sum FP + 2 \sum TP + \sum FN} \times 100\% \tag{4}$$

$$Jaccard = \frac{\sum TP}{\sum FP + \sum TP + \sum FN} \times 100\% \tag{5}$$

$$Precision = \frac{\sum TP}{\sum TP + \sum FP} \times 100\% \tag{6}$$

$$Recall = \frac{\sum TP}{\sum TP + \sum FN} \times 100\% \tag{7}$$

$$F1 = \frac{2(Precision \cdot Recall)}{(Precision + Recall)} \times 100\% \tag{8}$$

Table 2 Brief introduction of comparison networks

Order	Network	Year	References	Type
1	FCN	2015	[19]	Non U-shaped network
2	PSPNet	2017	[34]	
3	DeepLab v1	2014	[35]	
4	DeepLab v2	2017	[36]	
5	DeepLab v3	2017	[37]	
6	DeepLab v3 +	2018	[38]	
7	U-Net	2015	[21]	U-shaped network
8	SmaAt-UNet	2021	[45]	
9	OAU-net	2022	[44]	
10	SegNet	2017	[20]	
11	Refine Net	2017	[40]	Combined with transformer
12	FAT-Net	2022	[29]	
13	SUNet	2022	[67]	
14	Swin-Unet	2021	[31]	

$$Specificity = \frac{\sum TN}{\sum TN + \sum FP} \times 100\% \quad (9)$$

4.3 Contrast Models

As shown in Table 2, we introduce 14 comparison models to objectively verify the Semantic Segmentation effect of FTUNet, the networks we introduced have the following implications. From the perspective of network structure, these models are constructed by traditional segmentation networks (FCN, PSPNet, DeepLab v1, DeepLab v2, DeepLab v3, DeepLab v3 +), U-Net and its variants (U-Net, SmaAt-UNet, OAU-net, SegNet, Refine Net), and the Transformer structures (FAT-Net, SUNet, Swin-Unet). Therefore, we adopt these representative networks as comparison models to verify the segmentation effect of FTUNet.

4.4 Implementation Detail

The 5-Fold cross-validation is implemented for the image samples to be trained, which includes 4 training parts and 1 test part. Validation occurred 5 times, with 100 rounds each, resulting in 500 epochs per model. The average of 8 indicators was then computed. Other training details include: *Batch Size* is 4, *Learning Rate* is 1e-4, the *Loss Function* is BCE-WithLogitsLoss. For this Loss Function, its formula is shown as Eq. (10), where the *pix* represents the number of pixels, $pred_i$ is prediction, gt_i is Ground Truth, sig is Sigmoid activation function.

$$\begin{aligned} & BCEWithLogitsLoss(pred, gt) \\ &= \frac{1}{pix} \sum_{i=1}^{pix} [gt_i \cdot \log(\text{sig}(pred_i)) + (1 - gt_i) \cdot \log(1 - \text{sig}(pred_i))] \end{aligned} \quad (10)$$

4.5 Comparative Study

4.5.1 Comparison of Indicator Values

We first perform a comparative analysis of multi-indexing with U-shaped and non-U-shaped networks, which include both the classical model and the newly proposed SOTA models in recent two years.

Table 3 shows FTUNet outperforming 11 models in 5 of 8 indicators, notably in ACC (98.8236%), Dice (97.4635%), and Jaccard (95.1379%). Specificity and Precision ranked 2nd and 3rd, showcasing superior pathological figure segmentation. Furthermore, FTUNet achieved 98.0188% in Recall and 97.4635% in F1, indicating increased sensitivity to pathological pixels at the expense of some Precision (97.0171%, ranking 3rd), enhancing segmentation accuracy. Given FTUNet's superiority, these differences signify its significant outperformance. Apart from FTUNet, the DeepLab family, notably DeepLab v3 +, displayed outstanding performance, achieving the highest Precision (97.8074%) and AUC (99.6031%). U-Net, SmaAT-UNet, SegNet, FCN, and PSPNet also excelled with Dice consistently above 95%, reflecting high pixel-level similarity. Their ACC exceeded 98%, highlighting accuracy superiority. Specifically, SegNet and FCN exhibited Jaccard values above 94% (94.5945% and 94.0644%, respectively), indicating substantial overlap between segmented and labeled pixels. SmaAT-UNet stood out with the highest Specificity (99.8270%) but a relatively lower AUC (83.3388%), suggesting a tendency to classify all pixels as background. In contrast, OAU-net performed relatively well, especially in Specificity (99.7675%), while Refine Net lagged behind with 87.8590%, 84.8032%, and 96.3426% for Dice, Jaccard, and Acc. Notably, the Wilcoxon Test *P*-values for ACC differences between FTUNet and other models in the Lung dataset were all below 0.05, indicating statistical significance. UNet was an exception with a *P*-value of 6.51E-01, suggesting no significant difference from FTUNet. Overall, these *P*-values underscore the substantial distinctions in accuracy between FTUNet and the other models.

In Table 4, FTUNet excelled in 7 out of 8 evaluation indicators, achieving outstanding scores in Dice (95.0725%), Jaccard (90.9817%), Precision (96.4227%), F1 (95.0725%), Specificity (99.6840%), AUC (97.0795%), and Acc (97.5609%). Notably, FTUNet outperformed DeepLab v3 + in key metrics—Acc, Dice, and Jaccard—by 1.1121%, 0.5637%, and 1.1970%, affirming its superiority over the DeepLab family. DeepLab v3 + showed enhanced Recall at 94.5349%, slightly surpassing FTUNet and indicating heightened sensitivity in segmenting pathological pixels. Other notable models include FCN and OAU-net, both exceeding 90% in Dice and 82.5% in Jaccard, demonstrating effective differentiation of pathological pixels from skin pixels, though slightly below FTUNet and DeepLab v3 +. SmaAT-UNet ranked 3rd in Acc at 94.5913%, surpassing OAU-net and FCN but trailing FTUNet by 2.9695%. Similarly, SegNet ranked 3rd in Precision at 98.4664%, showcasing proficiency in identifying pathological pixels, though FTUNet outperformed by 1.2176%. FCN, ranking 3rd in F1 at 90.6123%, demonstrated effective segmentation while being surpassed by FTUNet by 4.4602%. Underperforming models in this dataset include Refine Net, DeepLab v1-3, and PSPNet, all with lower Acc than U-Net. Refine Net exhibited the lowest performance across all indicators, attributed to sensitivity to learning rate, as confirmed by subsequent image analysis. The Wilcoxon Test, applied to the Skin dataset, revealed significant differences between FTUNet and other algorithms, including UNet, with *P*-values below 0.05. Notably, the lowest *P*-value corresponded to DeepLab v1, indicating substantial differences in ACC data distribution between this model and FTUNet.

Table 3 Comparison results of different models in eight categories of indicators and *P*-value (Wilcoxon Test *P*-value) in LUNG dataset

Model	Dice (%)	Jaccard (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)	AUC (%)	Acc (%)	<i>P</i> -value
UNET	96.6593	94.2108	96.8988	97.1587	96.6025	99.5955	97.3221	98.5840	6.51E - 01
SmaAT-UNet	96.0271	93.2061	97.3708	95.5969	96.0271	99.8270	83.3388	98.5997	2.71E - 42
OAU-net	91.0611	86.5992	96.3607	89.8421	91.0611	99.7675	80.4857	96.9809	7.53E - 69
SegNet	97.1735	94.5945	96.6820	97.9383	97.2592	99.5651	97.3844	98.6621	2.33E - 29
FCN	96.8021	94.0644	96.8437	97.2002	96.8630	99.5881	97.2544	98.5422	1.25E - 36
DeepLab v1	94.4088	89.6818	94.9117	94.0741	94.4267	99.4733	96.2992	97.3756	1.51E - 76
DeepLab v2	91.9255	85.5843	92.7374	91.6451	92.0155	99.2148	94.6579	96.2143	8.97E - 80
DeepLab v3	95.4359	91.7641	96.5862	96.9499	96.7334	99.6031	99.6031	97.5958	7.03E - 73
DeepLab v3 +	97.3353	95.0209	97.8074	96.9283	97.3293	99.7659	98.1209	98.7238	3.53E - 04
PSPNet	96.4123	93.3503	96.7119	96.2687	96.4142	99.6242	97.7587	98.2623	5.02E - 46
Refine Net	87.8590	84.8032	89.1213	87.7230	88.0109	99.6062	91.9298	96.3426	3.62E - 74
FTUNet(Ours)	97.4635	95.1379	97.0171(3rd)	98.0188	97.4635	99.7878(2nd)	98.1640	98.8236	-

Bolded font means that the data for the indicator is ranked first in all comparative models

Table 4 Comparison results of different models in eight categories of indicators and *P*-value (Wilcoxon Test *P*-value) in SKIN dataset

Model	Dice (%)	Jaccard (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)	AUC (%)	Acc (%)	<i>P</i> -value
U-Net	87.8256	79.0820	88.4518	88.1383	87.9490	97.5255	90.9575	92.0855	2.51E – 82
SmaAT-UNet	89.4118	82.0488	89.8450	90.1655	89.4118	99.0863	96.0252	94.5913	7.19E – 82
OAU-net	90.1664	82.6990	90.0888	91.6244	90.1664	98.9396	96.6879	94.5452	2.07E – 80
SegNet	89.3189	81.3840	92.2174	87.4434	89.4348	98.4664	92.0721	93.2778	3.74E – 82
FCN	90.5480	83.3961	91.3181	90.6006	90.6123	98.1576	93.3318	93.9590	2.70E – 81
DeepLab v1	87.8611	79.3213	87.8644	89.1851	88.0179	97.2409	91.4769	91.9308	1.28E – 83
DeepLab v2	84.7690	75.0274	75.0274	86.7744	85.0496	96.4875	89.1721	90.0076	3.82E – 83
DeepLab v3	82.0111	71.8141	78.0204	89.1730	82.0603	96.1983	88.1126	89.3183	8.02E – 83
DeepLab v3 +	94.5088	89.7847	94.7553	94.5349	94.5639	98.8933	95.9210	96.4488	1.61E – 65
PSPNet	87.2458	78.6190	87.7046	88.2722	87.3264	97.2023	90.8753	91.7733	1.35E – 83
Refine Net	65.2284	56.7123	61.5762	73.4122	66.2747	95.6030	84.8812	85.7370	5.91E – 83
FTUNet(Ours)	95.0725	90.9817	96.4227	94.3093 (2nd)	95.0725	99.6840	97.0795	97.5609	–

Bolded font means that the data for the indicator is ranked first in all comparative models

In examining the Drive dataset (Table 5), FTUNet experiences a slight decline, yet still secures the top position in Jaccard (67.7485%) and Accuracy (97.0639%), while ranking second in the remaining six indicators (Dice, Precision, Recall, F1, Specificity, AUC). Despite this dip, FTUNet's overall stability remains significant. The algorithm encounters challenges due to the fine segmentation granularity of veins in eye blood vessels, resulting in initial rounds with indices at 0 in every 100 rounds within 500 iterations. However, subsequent rounds witness a surge, surpassing most comparison models. In general, FTUNet exhibits the best average value among the eight indicators at 82.6418%, outperforming U-Net (2nd) at 82.1904%. This underscores FTUNet's strong segmentation performance, especially on fine-grained objects.

Combining the above data, we discuss the reasons for the excellent results of FTUNet model on 8 indicators in the following 3 aspects. (1) ASPP-Inception Module. The FTUNet's prowess is notably anchored in the deployment of the ASPP-Inception module, a sophisticated fusion of ASPP and Inception structures. (2) Dual Encoder Mechanism. The FTUNet's Encoder structure stands out due to the strategic integration of a dual Encoder mechanism, which harmoniously combines local and global information. (3) Receptive Field Widening with PA Module. Inspired by established models like FCN and PSPNet, the FTUNet employs a sophisticated approach to widen its receptive field. Overall, the average Acc of DeepLab families, U-Net series, and receptive field widening series models are 94.1430%, 94.8394%, and 94.4737%, respectively, while the average Acc of FTUNet is 97.8161%, which is 3.6732%, 2.9768%, and 3.3424% higher than these three types of models, respectively. FTUNet's exceptional performance stems from its proficient handling of intricate edge semantic information and its effective segmentation of irregular shapes.

4.5.2 Comparison of Segmented Images

In our pursuit of diverse image samples to enrich model training, we have incorporated operations such as flipping, translation, and zooming. These augmentations are intended to preserve image authenticity while broadening sample diversity. The assessment of LUNG CT segmentation holds paramount significance in clinical diagnosis and treatment. Figure 5 delineates the performance of FTUNet and comparative models on this dataset, facilitating a visual juxtaposition of segmentation effects. Figure 5a depicts the original lesion image, Fig. 5b provides pixel-level annotations by experts. For an intuitive comparison, we present five image contrast domains denoted 1–5 in red boxes. Adverse effects are accentuated with yellow boxes.

Figure 5 unveils that various networks demonstrate commendable segmentation effects, satisfying Semantic Segmentation requisites. U-shaped architectures, exemplified by U-Net (Fig. 5d), facilitate feature capture and image restoration through intricate encoding and decoding structures. However, U-Net manifests a deficiency in pixel filling during upsampling (marker 3). OAU-net (Fig. 5e) and SmaAT-UNet (Fig. 5f), both U-shaped networks, excel in handling irregular edges. OAU-net employs a Sobel operator-like mechanism for precise lung edge representation, while SmaAT-UNet leverages an enhanced attention mechanism for meticulous feature control, effectively addressing the void at marker 3. SegNet (Fig. 5g) and Refine Net (Fig. 5h) also exhibit commendable performance, with SegNet demonstrating finer granularity owing to its utilization of the VGG16 encoding. However, the surplus segmentation granularity in SegNet may lead to sensitivity issues, as discerned in the yellow box highlighting fusion problems.

The DeepLab family of networks plays a key role on non-U-shaped networks. As the optimal model, DeepLab v3 + (Fig. 5m) achieves the cross-block fusion of feature maps,

Table 5 Comparison results of different models in eight categories of indicators and *P*-value (Wilcoxon Test *P*-value) in DRIVE dataset

Model	Dice (%)	Jaccard (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)	AUC (%)	Acc (%)	<i>P</i> -value
U-Net	77.3099	65.5421	79.1608	77.3523	77.3099	98.0464	86.8894	95.9125	1.04E – 35
SmaAT-UNet	72.4281	59.7944	77.7096	70.4052	72.4281	96.9914	82.6805	94.3515	3.57E – 62
OAU-net	70.9251	57.1626	76.9044	67.2116	70.9251	98.429	81.7207	95.3350	5.65E – 70
SegNet	67.2627	55.0046	83.4665	61.7164	67.2627	98.6688	79.6714	95.2539	1.87E – 70
FCN	53.1835	37.3545	67.9438	45.8914	53.1835	98.412	71.5328	93.5597	3.85E – 79
DeepLab v1	26.7896	15.8834	58.5838	18.048	26.7896	99.1331	57.5689	92.0641	3.56E – 82
DeepLab v2	12.6838	6.9404	51.0086	7.9233	12.6838	99.3404	53.6067	91.3877	1.27E – 83
DeepLab v3	48.9822	33.9962	65.1207	40.5198	48.9822	98.7557	67.8206	93.5035	1.19E – 82
DeepLab v3 +	73.8737	59.7064	75.6836	73.8229	73.8737	97.7008	84.3848	95.1453	1.52E – 54
PSPNet	10.1649	5.519	52.2313	6.8909	10.1649	98.7537	52.842	90.7459	2.17E – 83
Refine Net	32.2789	20.2978	56.8926	24.2178	32.2789	98.9573	61.0673	92.3314	3.46E – 78
FTUNet(Ours)	76.6191 (2nd)	67.7485	82.4835 (2nd)	74.9769 (2nd)	76.6191 (2nd)	99.3208 (2nd)	86.3376 (2nd)	97.0639	

Bolded font means that the data for the indicator is ranked first in all comparative models

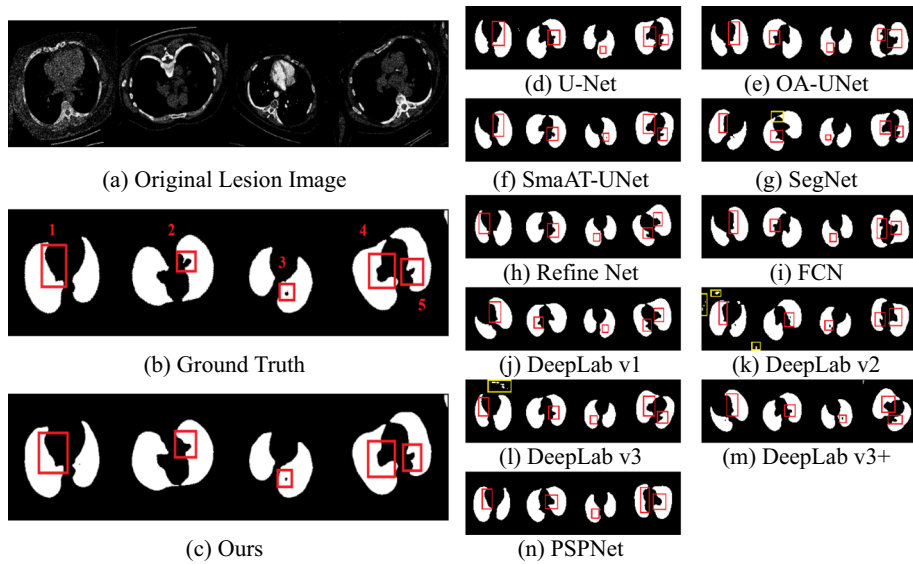


Fig. 5 Segmentation of LUNG CT images

and its segmentation effect has been perfectly displayed on issues such as edge contour and internal filling. Furthermore, as an earlier fully convolutional structure, FCN implements skip structures of different depths and integrates semantic information of different depths. As an improvement, PSPNet achieves richer feature grasping on the scale of the receptive field. However, the image shows that on the basis of the general Semantic Segmentation, the local details are still insufficient (Fig. 5n). As the network proposed in this paper, FTUNet combined with the Transformer ViT mechanism to globally realize the attention mechanism of image processing. Meanwhile, combined with the downsampling information, the generalized information mining of the Inception structure is implemented at the bottom layer with the most channels. According to Fig. 5c, our segmentation results are basically consistent with the ground truth (markers 1, 4, and 5), and can overcome the loss of upsampling information and avoid rough pixel completion (marker 3). However, there are still some flaws in marker 2.

As can be seen from the Fig. 6, the model generally has a certain ability to identify the lesion area, but the labeled part still poses a great challenge to the effect of the network. Among the contrast models, U-Net (Fig. 6d) and its variants (Fig. 6e–h) have certain advantages in edge processing, but because the colors at the junction are too close, it is easy to mistake the skin as the lesion area. For example, the judgment results of images such as U-Net, SmaAT-U-Net, SegNet and Refine Net in the fourth lesion area verify that this kind of U-shaped network is insensitive to the lesion and background, which is mainly due to the coarse-grained processing of upsampling. Specifically, Refine Net performs similarly on LUNG datasets and is very sensitive to the learning rate. Segmentation is not possible on $1e-4$, so we adopt the parameter of $1e-6$ to get the Fig. 6h. In addition, the DeepLab family still maintains excellent results in the processing effects of other networks. The image shows that the feature extraction effect is greatly affected by using the ResNet as the Encoder. This shows that ‘deep network combined with residual’ has a strong auxiliary effect on the segmentation effect. Such as the lesion segmentation image of Fig. 6l, m shows that, with the optimization of the DeepLab

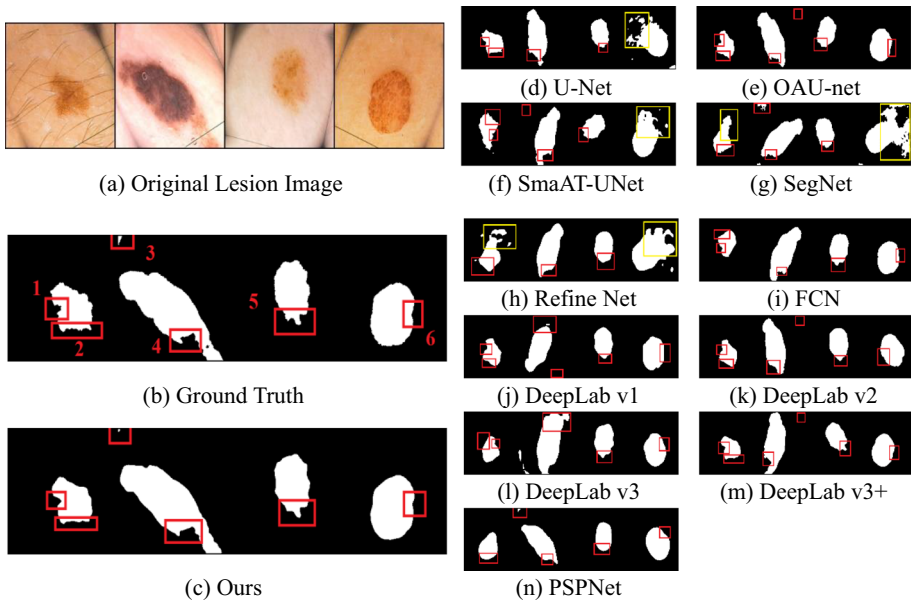


Fig. 6 Segmentation of SKIN Melanoma images

series, the capture of details will be more fine-grained. In addition, FCN and PSPNet also have a more ambiguous edge treatment.

It can be seen from the Fig. 7 that most models have good segmentation effect and basic ability to deal with fine-grained objects, but the processing effect is slightly biased. The experiment shows that the traditional lightweight U-shaped network has better detail processing ability and better discrimination on the segmentation of blood vessel trend. In contrast, a large number of improved networks perform poorly in eye blood vessel segmentation, especially in the early series of DeepLab family. Although the ASPP module is applied, the image results prove that it is difficult to achieve blood vessel segmentation by using a single hole technology. While FTUNet adopts a parallel ASPP structure, and the image proves that segmentation can be realized better with the support of rich semantics extracted from multiple dilations in parallel.

From Figs. 7, 8, 9, it is evident that FTUNet demonstrates outstanding performance in

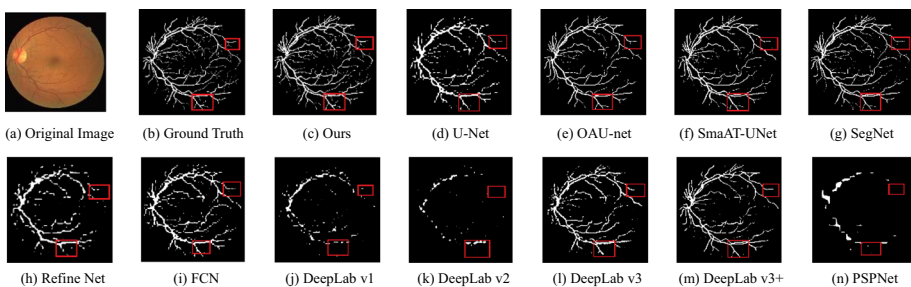
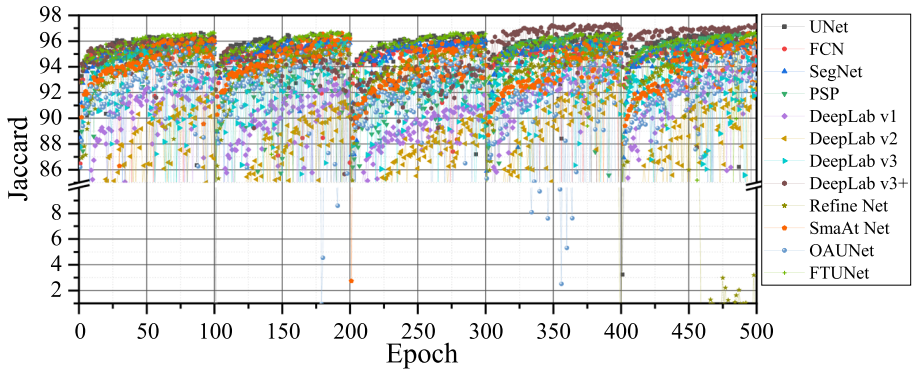
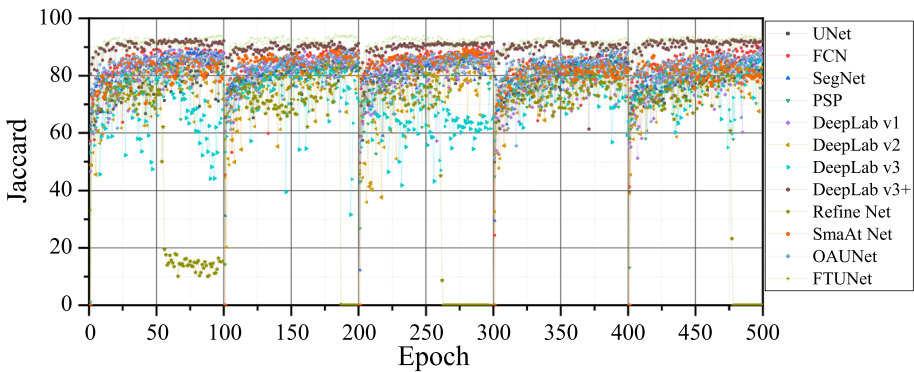


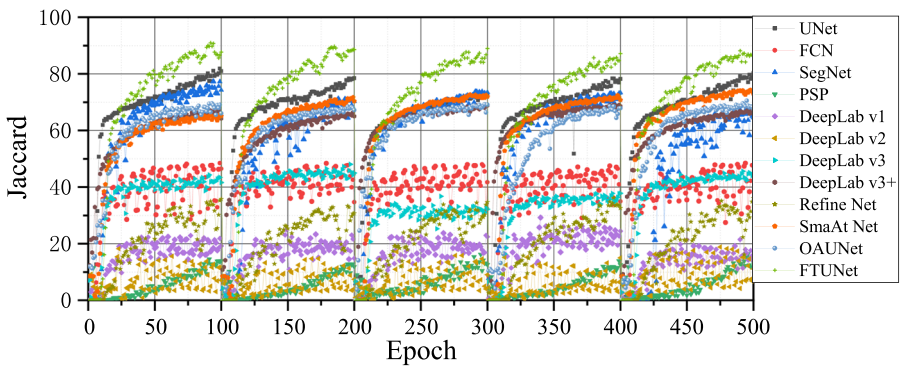
Fig. 7 Segmentation of DRIVE images



(a) The comparison results of the training process of the Jaccard indicator on LUNG.



(b) The comparison results of the training process of the Jaccard on SKIN



(c) The comparison results of the training process of the Jaccard on DRIVE

Fig. 8 Training effect of 500 epoch of different models on two datasets

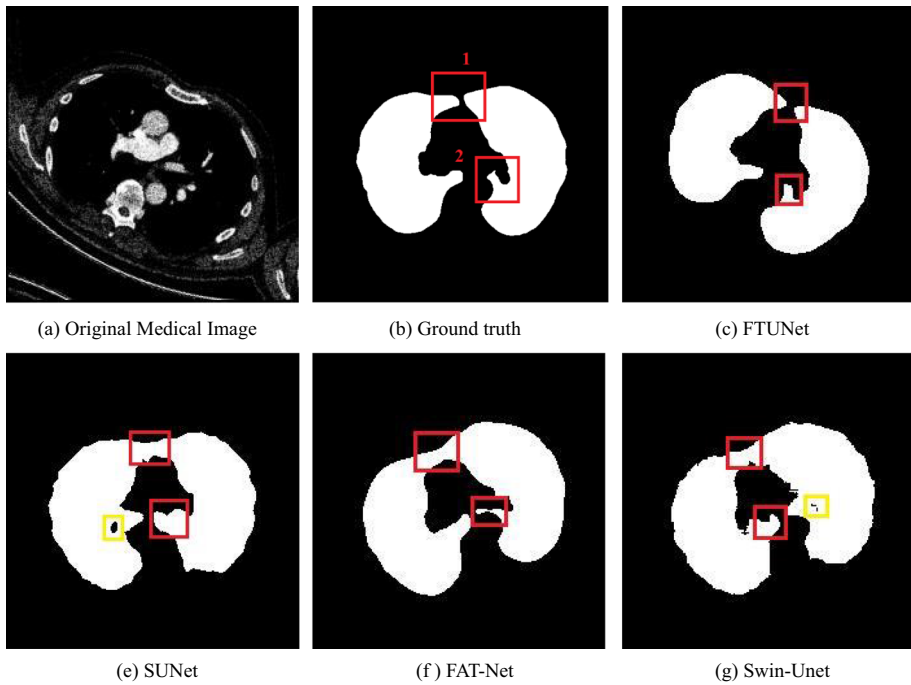


Fig. 9 Segmentation results of FTUNet and different Transformer variant models on LUNG

image segmentation, owing to its meticulously crafted network architecture that adeptly processes features encompassing large areas, ambiguous edges, and fine-grained structures in images. LTrans empowers FTUNet to handle extensive information by extracting semantic details at varying levels from different depth sequences of the Transformer layer, achieving a comprehensive understanding of the overall image structure. ASPP-Inception, with a focus on addressing fuzzy edge information, utilizes parallel microscale convolution kernels and ASPP modules to enhance semantic feature learning for target edges, thereby improving sensitivity and accuracy in fuzzy or complex edge areas. Additionally, the PA mechanism, through a reassessment of channel weights, optimizes attention to different features, enabling FTUNet to excel in handling fine-grained image structures, particularly in preserving detailed information. This multi-level design allows FTUNet to selectively process various image features, offering a comprehensive and robust solution for image segmentation tasks.

4.5.3 Comparison of Training Process

Since the Jaccard is the indicator that can best reflect the superiority of the models, Fig. 8 shows the comparison results of learning efficiency of different networks on Jaccard in 3 datasets during the learning process. In order to magnify the contrast effect, a breakpoint is set in Fig. 8a. It can be seen from the data in the figures that the model proposed by us is superior to the other comparison models, mainly reflected in two aspects: 1) Indicator values are more stable. In the analysis of different indicators, the FTUNet model may not have the highest value in a compromise of an indicator, but it is the most stable. For example, the data of DeepLap v3 + in the fourth and fifth fold are higher than that of the FTUNet model, but the

range of data fluctuation is large, and the indicator value will have a relatively serious shock, even falling to 92.82% in the third fold. 2) The indicator values are generally higher. The value of FTUNet model in different indicators is ahead of other models. Taking DRIVE data set as an example, the average of well-performing U-Net was 64.5421% (ranked second), while the average of FTUNet was 67.745%, which was 3.2064% higher than the second and 7.9541% higher than the third (SmaAT-UNet).

4.5.4 Comparison with Networks with Transformer Structure

The method proposed in this study is a Semantic Segmentation model combined with the Transformer, which is a deep network supported by global attention. Further, we hope to make a deep comparison with similar networks, that is, Semantic Segmentation networks with a Transformer as its internal structure. We believe that the exploration of the analogy network with Transformer can more authoritatively and objectively verify the advanced effects of FTUNet.

It can be seen from the Table 6 that FTUNet still has a good effect compared with different Transformer type Semantic Segmentation networks. According to the data, FTUNet achieves the optimal value in 7, 8, and 7 indicators respectively, which verifies the superior characteristics of all Transformer structural networks. In terms of performance enhancement, FTUNet not only relies on AP to redistribute channel weights and AI to mine semantic features; It has also been greatly influenced by the optimized Transformer, which can be verified by the ablation experiment in Sect. 4.6.2. In Transformer, the FTUNet correlation structure takes a hierarchical approach and we expect to mine and process different levels of attention information separately, rather than passing it to the deepest part of the U shape. Therefore, the model matches the context information of different depths in Transformer with the corresponding position of U-shaped and processes it uniformly.

Other Transformer structures have different optimizations here, but with certain limitations. For example, the FAT is a dual-Encoder model, and the results of Transformer are directly sent to Bottleneck. While SUNet and Swin-Unet are mechanisms based on shifted windows, which not sensitive to small granularity objects, such as DRIVE datasets.

Figures 9, 10, and 11 demonstrated the effects of Transformer models for Semantic Segmentation in 3 datasets.

In the LUNG dataset in Fig. 9, the Ground truth (Fig. 9b) marks two obvious differences, which respectively examine the cohesion granularity of the segmented image and the ability to capture edge features. By comparison, it can be found that in the first mark, FAT-Net, SUNet, and Swin-Unet have poor segmentation processing capabilities for fine-grained images, resulting in excessive fusion. We believe that the reasons come from many aspects, including the waste of shallow features of images and the inadequacy of deep semantic information mining. On the contrary, the ASPP-Inception module designed in our model increases the depth of the model and also increases the nonlinearity of the network, making it more sensitive to fine-grained image information. Moreover, mark 2 examines the detection of irregular edges, and each Transformer variant has different segmentation results for this local feature. Among them, SUNet and Swin-Unet using Shifted Windows mechanism are relatively rough to deal with this problem, which reflects the problem of inaccurate relationship calculation when performing global attention mechanism on each token. In addition, we also pay extra attention to the judgment of pixels in the process of image upsampling. SUNet has “holes” in filling, which reflects the imprecision of feature extraction in the early stage. In fact, we believe that semantic fusion of different depths can effectively deal with

Table 6 Comparison results of four kinds of semantic segmentation networks with transformer structure on different indicators of two kinds of datasets

Data	Model	Dice	Jaccard	Precision	Recall	F1	Spe	AUC	Acc
LUNG	FAT-Net	95.4826	92.4267	95.8641	95.5356	95.4826	99.8025	87.1437	98.3155
	SUNet	93.2681	88.4724	96.3795	91.497	93.2681	99.7269	75.3505	97.4822
	Swin-Unet	93.9205	89.4687	95.9227	93.0418	93.9205	99.7101	75.7554	97.7516
SKIN	Ours (FTUNet)	97.4635	95.1379	97.0171	98.0188	97.4635	99.7878	98.1640	98.8236
	FAT-Net	93.5732	89.1423	94.4181	93.4594	93.5732	99.5123	96.863	97.0290
	SUNet	90.6415	83.7398	91.3538	91.3296	90.6415	99.2795	95.3933	95.3309
DRIVE	Swin-Unet	89.2728	81.5952	89.9628	90.2605	89.2728	99.1774	94.3573	94.7422
	Ours (FTUNet)	95.0725	90.9817	96.4227	94.3093	95.0725	99.6840	97.0795	97.5609
	FAT-Net	51.1397	42.3132	58.6698	49.7622	51.1397	99.1641	73.5619	94.7528
	SUNet	16.4428	11.096	34.0928	12.7008	16.4428	99.6782	55.8385	92.2092
	Swin-Unet	29.0159	22.2076	46.998	25.4359	29.0159	99.5597	62.0308	93.138
	Ours (FTUNet)	76.6191	67.7485	82.4835	74.9769	76.6191	99.3208	86.3376	97.0639

Bolded font means that the data for the indicator is ranked first in all comparative models

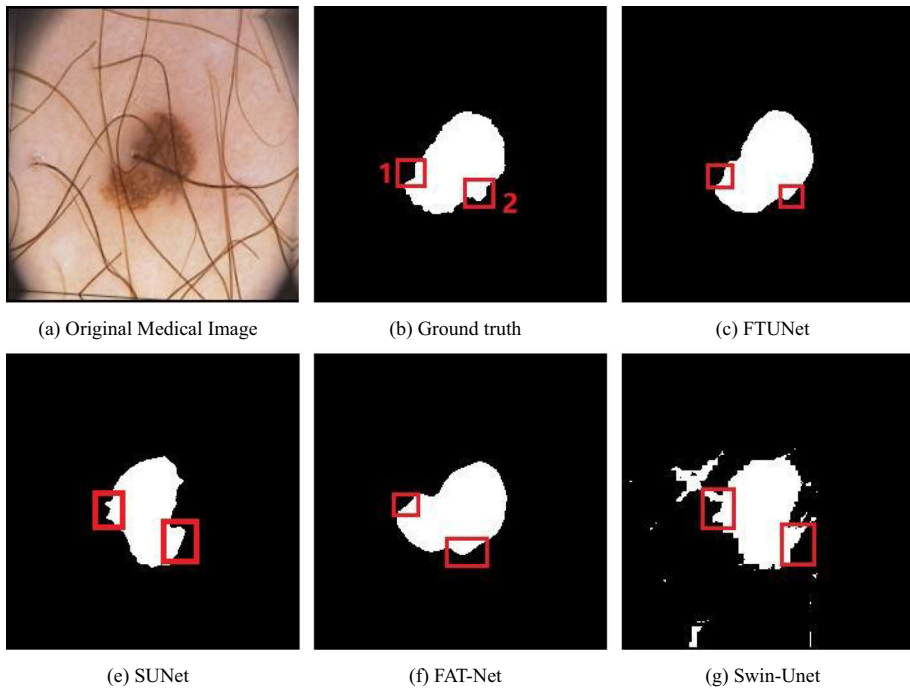


Fig. 10 Segmentation results of FTUNet and different Transformer variant models on SKIN

this problem. The LTrans adopted by FTUNet is dedicated to making full use of each layer of features and passing them to the decoding to achieve pixel classification with upsampling.

In addition, the SKIN and DRIVE datasets have similar problems. We proposed two marks in Figs. 10 and 11 to examine the segmentation granularity for irregular edges. The images demonstrated that FTUNet has better results. Other Transformer networks can generally achieve the expected results, but there are other noise defects.

4.6 Ablation Study

4.6.1 Description of Ablation Models

The Table 7 shows the idea of structural ablation. In order to verify the performance of each component on three datasets, we divided the ablation model into two parts: Main Module Ablation (MMA) and Supplementary Ablation (SA). The model involved in MMA is the main model in the ablation experiment. It combines the three components we proposed: LTrans, AI, and PA in the form of single module, double modules, and triple modules, and goes deep into each layer. In addition, considering the problem of gradient disappearance and network degradation caused by the deepening of network layers after adding new components, we added residual networks at Skip Connection and Bottleneck respectively. On this basis, we also carried out SA, which is roughly the same as the idea of MMA. However, when combining modules, we tried to add modules in different places, change the channel of module results,

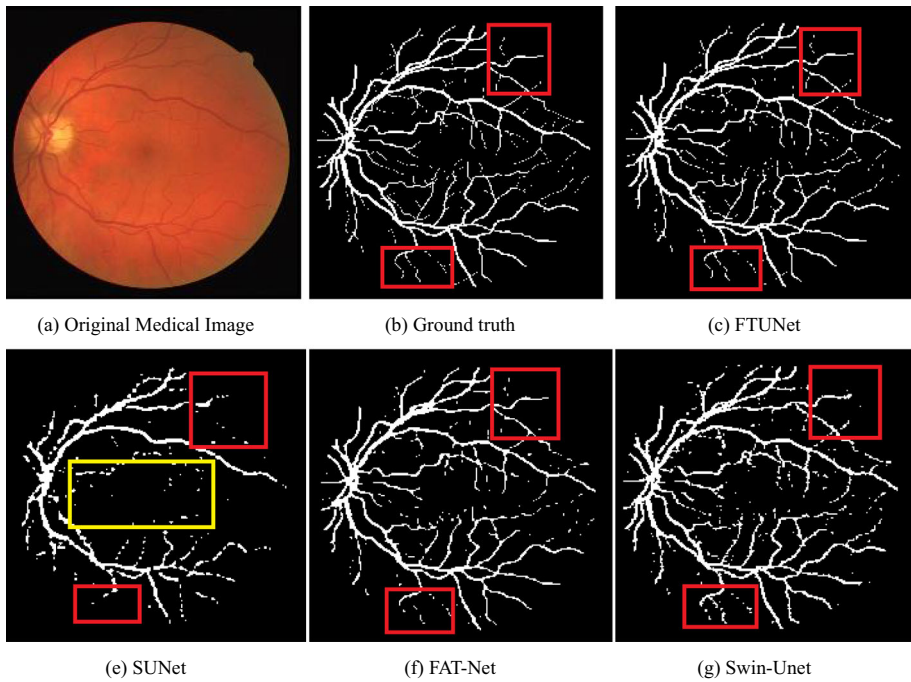


Fig. 11 Segmentation results of FTUNet and different Transformer variant models on DRIVE

and increase the number of modules, such as adding AI modules to the Encoder part and adding two AI components.

4.6.2 Ablation Analysis

Tables 8 and 9 respectively show the ablation analysis of FTUNet on different datasets and demonstrate the optimal characteristics of the current model structure. The ablation is divided into two parts, MMA (blue) and SA (orange), each with a gradient color representing the number of components inside. From the macro point of view, FTUNet has achieved the best performance in 5 and 4 indicators, respectively. Specifically, the following conclusions can be fully verified by the comparative experimental data.

- (1) The effectiveness of three modules. The data demonstrated that the three types of modules embedded respectively on the Baseline can improve the performance of each indicator. The experimental data in the above tables verified the strong supporting effect of LTrans, AI, and PA modules on semantic inference. It is worth discussing that different modules have different degrees of improvement under different data sets, so it cannot be simply confirmed that which module has the best performance improvement effect. Taking the Dice index as an example, in the LUNG data set, AI has the best improvement effect, which is 10.0092% higher than Baseline, while the PA module has the best effect on SKIN, which is 2.9723% higher than Baseline. Moreover, taking Jaccard as an example, LTrans and PA showed the best performance in the two datasets, with an increase of 9.5193% and 5.0373%, respectively.

Table 7 Design of ablation models

Type	Model formulation	Model perturbations	Simplified image	Component
MMA	Baseline	Transformer + unet		–
	Baseline + LTrans	Traditional transformer in Baseline becomes LTrans		With a single module
	Baseline + AI	The double convolution module at Bottleneck in Baseline becomes AI module		
	Baseline + PA	Add PA attention mechanism module at Skip Connection in Baseline		
	Baseline + LTrans + AI + Res(SC)	Add LTrans module, AI module to Baseline and add residual network at Skip Connection		With double modules
	Baseline + AI + Res(PA)	Baseline adds PA module and AI module, and adds residual network at both ends of PA module		
	Baseline + LTrans + AI + PA	Add LTrans module, AI module, PA module to Baseline		With triple modules
	Baseline + LTrans + Res(AI) + PA	Add LTrans module, PA module, AI module to Baseline and add residual network at both ends of AI module		
	Baseline + LTrans + AI + Res(PA), FTUNet	Add LTrans module, AI module, PA module to Baseline and add residual network at both ends of PA module		

Table 7 (continued)

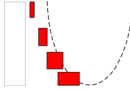
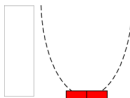
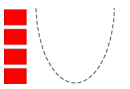
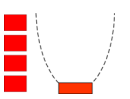
Type	Model formulation	Model perturbations	Simplified image	Component
SA	Baseline + Encoder(AI)	Add the ASPP-Inception module in the Encoder part of the Baseline		With a single module
	Baseline + Double(AI)	Double convolution at Baseline's Bottleneck is replaced by two AI modules		
	Baseline + UnifiedCh(LTrans)	Channel of the images obtained by the LTrans module is upsampled to 192 and then spliced into the Baseline		
	Baseline + LTrans + AI	Add LTrans module and AI module to Baseline		With double modules

Table 8 Ablation validation on LUNG datasets

Networks	Dice	Jaccard	Precision	Recall	F1	Spe.	AUC	Acc.
Baseline	86.9501	85.0955	89.6984	86.2294	86.9501	99.8977	92.6250	96.4910
Baseline + LTrans	96.9593	94.6148	96.7457	97.5755	96.9593	99.7797	97.9399	98.7017
Baseline + AI	96.9312	94.5181	96.2708	97.7421	96.9312	99.7600	97.7302	98.6347
Baseline + PA	96.6687	94.4872	96.1116	97.4309	96.6687	99.7839	97.9850	98.6734
Baseline+Ltrans+AI+Res(SC)	97.3863	95.1033	97.1174	97.8971	97.3863	99.7991	98.1615	98.8167
Baseline + AI + PA	97.0077	94.7930	97.1900	97.4155	97.0077	99.8055	98.0173	98.7504
Baseline + Ltrans + AI + PA	97.1876	94.6609	96.5165	98.0384	97.1876	99.7511	97.7461	98.6879
Baseline+Ltrans+Res(AI)+PA	96.4076	93.6294	95.8311	97.6812	96.4076	99.675	97.4747	98.3734
Baseline + Ltrans + AI + Res(PA) (Ours: FTUNet)	97.4635	95.1379	97.0171	98.0188	97.4635	99.7878	98.1640	98.8236
Baseline + Encoder(AI)	97.2688	94.9575	97.6180	97.0471	97.2688	99.8413	97.9804	98.7716
Baseline + Double(AI)	96.7688	94.3799	96.1904	97.4903	96.7688	99.7744	97.5507	98.6208
Baseline + UnifiedCh(Ltrans)	25.1263	20.9807	56.2271	21.0091	25.1263	99.9977	59.6134	81.4467
Baseline + Ltrans + AI	97.2594	94.7453	96.5556	98.0821	97.2594	99.7581	97.8076	98.7110

Bolded font means that the data for the indicator is ranked first in all comparative models

Table 9 Ablation validation on SKIN datasets

Networks	Dice	Jaccard	Precision	Recall	F1	Spe.	AUC	Acc.
Baseline	91.6900	85.6057	92.8760	91.9076	91.6900	99.2810	96.6658	95.6869
Baseline + LTrans	94.1648	89.9137	95.0968	94.0093	94.1648	99.5941	97.0763	97.2892
Baseline + AI	94.3094	90.2276	95.7754	93.2678	94.3094	99.6883	96.9292	97.3750
Baseline + PA	94.6623	90.6430	95.8203	93.9831	94.6623	99.6073	97.0721	97.4063
Baseline+LTrans+AI+Res(SC)	95.0349	90.8754	96.1567	94.4617	95.0349	99.6408	97.3563	97.5028
Baseline + AI + PA	94.7354	90.5152	96.4343	93.8281	94.7354	99.6931	96.8760	97.4094
Baseline + LTrans + AI + PA	94.1702	89.9399	95.0261	94.2190	94.1702	99.5237	96.9829	97.2601
Baseline+LTrans+Res(AI)+PA	93.5893	89.0949	94.0854	94.0306	93.5893	99.4594	96.9148	96.9058
Baseline + LTrans + AI + Res(PA) (Ours: FTUNet)	95.0725	90.9817	96.4227	94.3093	95.0725	99.6840	97.0795	97.5609
Baseline + Encoder(AI)	94.6244	90.1382	96.4713	93.4051	94.6244	99.6987	97.1646	97.2362
Baseline + Double(AI)	94.2375	90.0105	96.2196	92.8826	94.2375	99.6963	96.7241	97.2591
Baseline + UnifiedCh(LTrans)	71.9030	62.1797	96.4189	62.7264	71.9030	99.9420	81.6529	89.4978
Baseline + LTrans + AI	94.0640	89.8538	94.8540	94.0188	94.0640	99.5761	97.0339	97.2763

Bolded font means that the data for the indicator is ranked first in all comparative models

- (2) Effectiveness of module composition. Furthermore, we discuss the combination effect of these three modules, and find that in most cases, the combination embedding of any two modules is better than the individual embedding of a single one. Both datasets validated the above conclusions. For example, the ablation model “Baseline + LTrans + AI + Res (SC)” outperformed the model embedding alone by 8 and 7 metrics, respectively, in both datasets. This shows that the combination of the two modules allows the network to further enhance the features over a wider range of image receptive field. Meanwhile, with the support of the Transformer, the attention mechanism has been raised to a higher level. Further, we combined three types of modules to achieve three combinations, and demonstrated more outstanding results. These combinations respectively consider the *common embedding of the three modules*, the *addition of residuals at AI under common embedding*, and the *addition of residuals at PA under common embedding*. It can be seen from the results that FTUNet has the best effect among the above methods. Generally, with the increase of network depth, the improvement trend of learning effect will be gradually slow, and even the phenomenon of learning degradation will occur. In order to avoid this problem, we introduce the residual mechanism, so that the model tries to realize learning reinforcement through the strategy of “short-circuit”. The experimental data show that the residual added at PA is better than that at AI, because the attention mechanism may cause “too much amplification or reduction” in the process of realizing weight polarization. In short, the model finally adopts the result of “Baseline + LTrans + AI + Res(PA)” to obtain a better effect.
- (3) It is difficult to achieve optimal results for different locations and different numbers of modules. To further complement the validation, we made some other attempts, including adjusting the position of the lightweight AI, adding its serial number, and so on. However, experiments show that AI is difficult to maximize its value in the Encoding part, while the traditional double convolution effect is better than “deeper” or “wider” networks. Moreover, the continuous addition of AI modules at Bottleneck does not enhance

network learning ability. This is because the features of the network have reached saturation at this time, and too many $1 * 1$ convolution kernels will destroy the integrity and correlation of the original features, so the serial setting does not seem to be the optimal option either.

5 Conclusion

This study proposes a U-shaped Semantic Segmentation network FTUNet with double Encoder branches, which achieves accurate semantic inference of medical images by enhancing features. This network combines the ViT, the proprietary model of Transformer architecture in CV domain, and establishes a hierarchical context information transfer relationship with different depths from U-shaped Network. In addition, an inception-like ASPP structure is proposed to capture important local features in order to further explore the semantic information of LTrans and U-shaped encoding. Finally, a lightweight attention mechanism is proposed to strengthen and suppress features and redistribute the weights of different channels. FTUNet and 14 kinds of comparison networks have realized the comprehensive comparison of process data, image data, and overall data on 3 kinds of data sets, which verifies the superiority of the network. Meanwhile, 13 ablation networks constructed also proved the best structure of the current model.

In the experiment, we also found that FTUNet has certain defects. For example, compared with other lightweight networks such as U-Net, its parameter quantity has a significant increase, which means that the training time of the model needs to be longer.

A large number of experiments in this study demonstrated the superiority of the combination of convolution and Attention. The next step of work will continue to focus on the optimization of the Transformer, paying attention to the internal structure of the Transformer Encoder components, and connections to achieve multiple types of attempts. In addition, the processing of image patches will also try to implement different granularity strategies to achieve better segmentation.

Author contributions YW wrote the main manuscript; XY revised paper; YX, RF wrote all Figs; SZ and YY realised the experimental part. All authors reviewed the manuscript.

Data Availability The images used in this study were obtained from publicly available medical image datasets at: SKIN: <https://www.kaggle.com/code/hashbanger/skin-lesion-segmentation-using-segnet/notebook>. LUNG: <https://www.kaggle.com/datasets/kmader/finding-lungs-in-ct-data>. DRIVE: <https://drive.grand-challenge.org/>. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest No potential conflict of interest was reported by the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Voulodimos A, Doulamis N, Doulamis A, et al (2018) Deep learning for computer vision: a brief review. *Comput Intell Neurosci* 1–13
2. Garcia-Garcia A, Orts-Escolano S, Oprea S, et al (2017) A review on deep learning techniques applied to semantic segmentation. *arXiv preprint [arXiv:1704.06857](https://arxiv.org/abs/1704.06857)*
3. Mo Y, Wu Y, Yang X et al (2022) Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* 493:626–646
4. Hao S, Zhou Y, Guo Y (2020) A brief survey on semantic segmentation with deep learning. *Neurocomputing* 406:302–321
5. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507
6. Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis. *Med Image Anal* 42:60–88
7. Jiang F, Grigorev A, Rho S et al (2018) Medical image semantic segmentation based on deep learning. *Neural Comput Appl* 29(5):1257–1265
8. Asgari Taghanaki S, Abhishek K, Cohen JP et al (2021) Deep semantic segmentation of natural and medical images: a review. *Artif Intell Rev* 54(1):137–178
9. Shamsad F, Khan S, Zamir SW, et al (2022) Transformers in medical imaging: a survey. *arXiv preprint [arXiv:2201.09873](https://arxiv.org/abs/2201.09873)*
10. Haralick RM, Shapiro LG (1992) *Computer and robot vision*. Addison-wesley, Reading
11. Monteiro M, Newcombe VFJ, Mathieu F et al (2020) Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study. *Lancet Digital Health* 2(6):e314–e322
12. Yu J, Rui Y, Tao D (2014) Click prediction for web image reranking using multimodal sparse coding. *IEEE Trans Image Process* 23(5):2019–2032
13. Tang P, Liang Q, Yan X et al (2019) Efficient skin lesion segmentation using separable-Unet with stochastic weight averaging. *Comput Methods Programs Biomed* 178:289–301
14. Hasan MK, Dahal L, Samarakoon PN et al (2020) DSNet: automatic dermoscopic skin lesion segmentation. *Comput Biol Med* 120:103738
15. Huang Z, Miao J, Song H et al (2022) A novel tongue segmentation method based on improved U-Net. *Neurocomputing* 500:73–89
16. Kaganami H G, Beiji Z (2009) Region-based segmentation versus edge detection. In: 2009 fifth international conference on intelligent information hiding and multimedia signal processing. *IEEE*, pp 1217–1221
17. Zhang M, Zhou Y, Zhao J et al (2020) A survey of semi-and weakly supervised semantic segmentation of images. *Artif Intell Rev* 53(6):4259–4288
18. Zhang J, Yang J, Yu J et al (2022) Semisupervised image classification by mutual learning of multiple self-supervised models. *Int J Intell Syst* 37(5):3117–3141
19. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE Press, NJ, pp 3431–3440
20. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans on Pattern Anal Mach Intell* 39(12):2481–2495
21. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: *International conference on medical image computing and computer-assisted intervention*, Springer, Cham, pp 234–241
22. Yu J, Tan M, Zhang H et al (2019) Hierarchical deep click feature prediction for fine-grained image recognition. *IEEE Trans Pattern Anal Mach Intell* 44(2):563–578
23. Han K, Wang Y, Chen H et al (2022) A survey on vision transformer. *IEEE Tran Pattern Anal Mach Intell* 45:87–110
24. Zhou D, Kang B, Jin X, et al (2021) Deepvit: towards deeper vision transformer. *arXiv preprint [arXiv:2103.11886](https://arxiv.org/abs/2103.11886)*
25. Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. *Adv Neural Inf Process Syst* 30
26. Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)*
27. Carion N, Massa F, Synnaeve G, et al (2020) End-to-end object detection with transformers. In: *European conference on computer vision*, Springer, Cham, pp 213–229
28. Zhou L, Zhou Y, Corso JJ, et al (2018) End-to-end dense video captioning with masked transformer. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 8739–8748

29. Wu H, Chen S, Chen G et al (2022) FAT-Net: feature adaptive transformers for automated skin lesion segmentation. *Med Image Anal* 76:102327
30. Touvron H, Cord M, Douze M, et al (2021) Training data-efficient image transformers & distillation through attention. In: *International conference on machine learning*. PMLR, pp 10347–10357
31. Cao H, Wang Y, Chen J, et al (2021) Swin-UNET: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*
32. Du G, Cao X, Liang J et al (2020) Medical image segmentation based on u-net: a review. *J Imaging Sci Technol* 64:1–12
33. Zhang J, Cao Y, Wu Q (2021) Vector of locally and adaptively aggregated descriptors for image feature representation. *Pattern Recogn* 116:107952
34. Zhao H, Shi J, Qi X, et al (2017) Pyramid scene parsing network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 2881–2890
35. Chen LC, Papandreou G, Kokkinos I, et al (2014) Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*
36. Chen LC, Papandreou G, Kokkinos I et al (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans Pattern Anal Mach Intell* 40(4):834–848
37. Chen LC, Papandreou G, Schroff F, et al (2017) Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*
38. Chen L C, Zhu Y, Papandreou G, et al (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European conference on computer vision (ECCV)*, pp 801–818
39. Azad R, Asadi-Aghbolaghi M, Fathy M, et al (2020) Attention deeplabv3+: multi-level context attention mechanism for skin lesion segmentation. In: *European conference on computer vision*, Springer, Cham, pp 251–266
40. Lin G, Milan A, Shen C, et al (2017) Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1925–1934
41. Xia X, Kulis B (2017) W-net: a deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*
42. Qi K, Yang H, Li C, et al (2019) X-net: brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In: *International conference on medical image computing and computer-assisted intervention*, Springer, Cham, pp 247–255
43. Gu Z, Cheng J, Fu H et al (2019) Ce-net: context encoder network for 2d medical image segmentation. *IEEE Trans Med Imaging* 38(10):2281–2292
44. Song H, Wang Y, Zeng S et al (2023) OAU-net: outlined attention U-net for biomedical image segmentation. *Biomed Signal Process Control* 79:104038
45. Trebing K, Stańczyk T, Mehrkanoo S (2021) SmaAt-UNet: precipitation nowcasting using a small attention-UNet architecture. *Pattern Recogn Lett* 145:178–186
46. Lou A, Guan S, Loew M (2021) DC-UNet: rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation. In: *Medical imaging 2021: image processing*. SPIE, vol 11596, pp 758–768
47. Huang L, Tan J, Liu J, et al (2020) Hand-transformer: non-autoregressive structured modeling for 3d hand pose estimation. In: *European conference on computer vision*, Springer, Cham, pp 17–33
48. Huang L, Tan J, Meng J, et al (2020) Hot-net: non-autoregressive transformer for 3d hand-object pose estimation. In: *Proceedings of the 28th ACM international conference on multimedia*, pp 3136–3145
49. Lin K, Wang L, Liu Z (2021) End-to-end human pose and mesh reconstruction with transformers. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 1954–1963
50. Dai Z, Cai B, Lin Y, et al (2021) Up-detr: unsupervised pre-training for object detection with transformers. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 1601–1610
51. Zhu X, Su W, Lu L, et al (2020) Deformable detr: deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*
52. Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. PMLR, pp 8748–8763
53. Devlin J, Chang MW, Lee K, et al (2018) Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*
54. He K, Chen X, Xie S, et al (2022) Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 16000–16009
55. Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp 10012–10022

56. Li Z, Chen G, Zhang T (2020) A CNN-transformer hybrid approach for crop classification using multi-temporal multisensor images. *IEEE J Selected Topics Appl Earth Obs Remote Sens* 13:847–858
57. Li Q, Chen Y, Zeng Y (2022) Transformer with transfer CNN for remote-sensing-image object detection. *Remote Sens* 14(4):984
58. Liu Y, Sun G, Qiu Y, et al (2021) Transformer in convolutional neural networks. *arXiv preprint arXiv:2106.03180*
59. Azad R, Heidari M, Shariatnia M, et al (2022) TransDeepLab: convolution-free transformer-based DeepLab v3+ for medical image segmentation. *arXiv preprint arXiv:2208.00713*
60. Kim D, Xie J, Wang H, et al (2022) TubeFormer-DeepLab: video mask transformer. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 13914–13924
61. Sanderson E, Matuszewski BJ (2022) FCN-transformer feature fusion for polyp segmentation. In: *Annual conference on medical image understanding and analysis*, Springer, Cham, pp 892–907
62. He X, Tan EL, Bi H et al (2022) Fully transformer network for skin lesion analysis. *Med Image Anal* 77:102357
63. Xie Y, Zhang J, Shen C, et al (2021) Cotr: efficiently bridging cnn and transformer for 3d medical image segmentation. In: *International conference on medical image computing and computer-assisted intervention*, Springer, Cham, pp 171–180
64. Wang H, Zhu Y, Adam H, et al (2021) Max-deeplab: end-to-end panoptic segmentation with mask transformers. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 5463–5474
65. Yu Q, Wang H, Kim D, et al (2022) CMT-DeepLab: clustering mask transformers for panoptic segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 2560–2570
66. Hatamizadeh A, Tang Y, Nath V, et al (2022) Unetr: transformers for 3d medical image segmentation. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp 574–584
67. Fan CM, Liu TJ, Liu KH (2022) SUNet: swin transformer unet for image denoising. *arXiv preprint arXiv:2202.14009*
68. Wang H, Xie S, Lin L, et al (2022) Mixed transformer u-net for medical image segmentation. In: *ICASSP 2022–2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp 2390–2394
69. Gao Y, Zhou M, Metaxas DN (2021) UTNet: a hybrid transformer architecture for medical image segmentation. In: *International conference on medical image computing and computer-assisted intervention*, Springer, Cham, pp 61–71
70. Valanarasu JMJ, Oza P, Hacihaliloglu I, et al (2021) Medical transformer: gated axial-attention for medical image segmentation. In: *International conference on medical image computing and computer-assisted intervention*, Springer, Cham, pp 36–46
71. Chen J, Lu Y, Yu Q, et al (2021) Transunet: transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*
72. Xu G, Wu X, Zhang X, et al (2021) Levit-unet: make faster encoders with transformer for medical image segmentation. *arXiv preprint arXiv:2107.08623*
73. Petit O, Thome N, Rambour C, et al (2021) U-net transformer: self and cross attention for medical image segmentation. In: *International workshop on machine learning in medical imaging*, Springer, Cham, pp 267–276
74. Wang Y, Yu X, Yang Y et al (2023) A multi-branched semantic segmentation network based on twisted information sharing pattern for medical images. *Comput Methods Programs Biomed* 243:107914
75. Wang Y, Yu X, Guo X et al (2023) A dual-decoding branch U-shaped semantic segmentation network combining transformer attention with decoder: DBUNet. *J Visual Commun Image Represent* 95:103856
76. He K, Zhang X, Ren S et al (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 37(9):1904–1916
77. Lee HJ, Kim HE, Nam H (2019) Srm: a style-based recalibration module for convolutional neural networks. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp 1854–1862