EAST: Extensible Attentional Self-Learning Transformer for Medical Image Segmentation

Na Tian, Wencang Zhao

***Abstract*—Existing medical image processing models based on Transformers primarily rely on self-attention mechanisms to capture short-range and long-range visual dependencies. However, this approach has limitations in modeling the global context of full-resolution images, resulting in the loss of significant details. In order to address these issues, we propose an Extensible Attentional Self-learning Transformer (EAST) architecture for medical image segmentation. In EAST, tokenized images are input into an extensible attention module, enabling the training of multi-scale representations that effectively capture both fine-grained local interactions and coarse-grained global relationships. This allows for more comprehensive learning of semantic information. The obtained features are then passed through a self-learning module, further refining the representations of different samples to generate a more accurate feature map. To handle high-resolution images, the EAST architecture utilizes a U-shaped structure and skip connections for sequential processing of extensible attention feature maps. Experimental results on the Synapse dataset and ACDC dataset demonstrate the superior performance of our EAST architecture compared to other methods. Additionally, the EAST model is capable of capturing more detailed information, leading to precise localization of structures.**

***Index Terms*—Transformer, Medical image, Extensible attention, Self-learning.**

I. Introduction

MEDICAL image segmentation methods used in computer-aided diagnosis and image-guided surgery tasks require high accuracy and robustness [1]. The mainstream approaches for medical image segmentation primarily rely on convolutional neural networks (CNNs) [2], such as U-Net [3] and its various derivatives [4–8]. However, these methods often suffer from the limitation of convolutional localization, which hampers their ability to effectively model and understand contextual information, as shown in Fig. 1(a). To address this limitation, there is an urgent need for efficient networks in medical segmentation that can leverage the advantages of both local and global attention mechanisms. The Transformer architecture [9], known for its superior global processing capabilities, emerges as a promising alternative to CNN and has gained significant attention in recent research.

With the widespread success of Transformer architecture in natural language processing (NLP) [9, 10] and computer vision (CV) [11–13], numerous researchers have started to investigate its potential for enhancing the local modeling capabilities of CNN. Transformer architectures excel in capturing global contextual semantics, allowing them to capture both short-range and long-range visual dependencies. Indeed, this advantage often comes at the cost of requiring large-scale pre-training and involving computationally expensive quadratic operations. As a result, the processing speed of Transformer-based models may be compromised, particularly in the context of medical image analysis.

In recent studies, researchers have made attempts to integrate Transformer architectures with CNN for medical image segmentation. Chen [15] introduced Vision Transformer, combining the locality of convolution and the global strategy of Transformer to mitigate the need for large-scale training. Cao [16] explored the use of pre-training Swin Transformer for medical image segmentation, demonstrating the feasibility of replacing CNN backbones with convolution-free models. At this time, considerable researchers attempted various methods of combining CNN with Transformer [17–19] to achieve medical segmentation. However, these pure Transformer approaches have revealed weaknesses, including a tendency to overlook low-level details and high computational costs. Moreover, most Transformers focus on modeling the global context of all stages as shown in Fig. 1(b), neglecting fine-grained positioning information and the correlation between different samples, leading to coarse segmentation.

To address the global modeling limitations of Transformers, some researchers have explored the use of different attention windows as illustrated in Fig 1(c). Liu [12] introduced shifted windows, which restrict self-attention calculations to non-overlapping local windows. Dong [14] developed cross windows for parallel computing to form cross-shaped windows with Self-attention of horizontal and vertical stripes. Huang [20] proposed criss-cross attention by considering row attention and column attention alternately to capture global context. But these approaches are still limited to a few areas of attention interaction and fail to establish close relationships between samples.

With the purpose to overcome these issues, we introduce an Extensible Attentional Self-learning Transformer (EAST) for medical image segmentation. Our goal is to develop a pure Transformer architecture that is completely convolution-free and capable of capturing both short-range and long-range correlation information. The EAST model combines extensible attention to learn multi-scale attention maps and self-learning to integrate correlation information between different samples. As illustrated in Fig. 1(d), EAST overcomes the limitations of traditional transformers in medical image segmentation. It enhances the localization capabilities of convolutional methods while leveraging the benefits of a pyramid structure to learn multi-granularity features.

The core component of our EAST model is the EL block, which consists of the EA (Extensible Attention) and SL

Manuscript received November 12, 2022; revised July 12, 2023. This work was supported by the National Natural Science Foundation of China under Grant (No.61171131) and Key R&D Plan of Shandong Province under Grant (No.YD01033).

Na Tian is a Ph.D. student in the College of Automation and Electronic Engineering at Qingdao University of Science and Technology, Qingdao, 266061, China. (e-mail: tennesse863@gmail.com).

Wencang Zhao is a professor and doctoral supervisor at the College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao, 266061, China. (corresponding author, e-mail: CoinsLAB@qust.edu.cn).

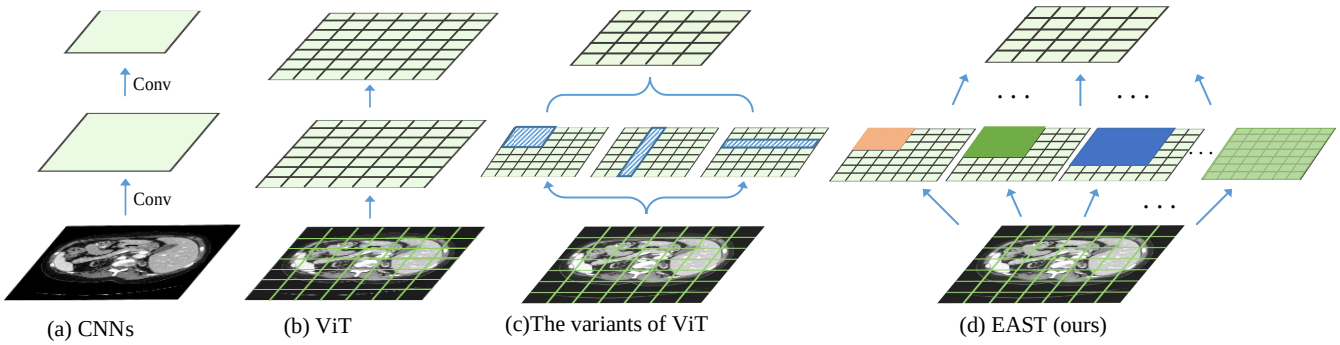

Fig. 1. The operation comparisons of different structures. From left to right: (a) CNNs: U-Net [3], Res-Unet [6], etc. (b) ViT [11], (c) The variants of ViT: Swin Transformer [12], CSwin Transformer [14], etc. (d) EAST (ours). The colored blocks in the CD represent the attention-handling process of the Transformer. By comparing the convolution operation and window attention operation of different network architectures, our EAST can realize the finer division of the window and interact with the global semantics to greatly increase the accuracy of medical image segmentation.

(Self-Learning) modules, as illustrated in Fig. 2. EA is designed to capture both local and global information through its attention-based step-wise expansion operations, as shown in Section III-A1 and Fig. 4. This allows the model to attend to different regions of the image and capture relevant features at multiple scales and resolutions. In addition, skip connections make it possible to refine low-level information extractions through a U-shaped structure, as described in Section III-B. By combining EA and skip connections, our model effectively captures both local and global information while preserving important low-level information. This can contribute to more accurate and robust segmentation results for medical images.

Furthermore, our EAST model incorporates a self-learning module, which plays a crucial role in improving the accuracy of segmentation predictions while reducing model complexity. This module leverages a self-attention mechanism to focus on different parts of the feature map and learn the relationships between them. By integrating the self-learning module into the model, we are able to refine the features and generate more representative feature maps. The details of the self-learning module are described in Section III-A2 and visualized in Fig. 5. These components are combined to form a pyramid structure within EAST, enabling the model to expand its processing from local to global image contexts. This architecture enhances the accuracy of segmentation results and improves performance on challenging medical image datasets.

Our proposed medical image segmentation model is unique in that it is the first to use a multi-scale self-learning approach without the use of convolutional neural networks. We make several significant contributions in our approach:

- We introduce the Extensible Attentional Self-learning Transformer, which enables the processing of medical feature maps at multiple scales and resolutions. This leads to more accurate and efficient feature extraction.
- The EA module is introduced to handle multi-scale attention maps, significantly improving segmentation and positioning accuracy.
- The traditional feedforward neural network is replaced with an SL module that integrates information from different samples, resulting in improved segmentation accuracy while reducing model complexity.
- We construct a U-shaped pure Transformer network specifically tailored for medical image segmentation, demonstrating excellent performance and robustness. The effectiveness of our approach is validated through experiments on the Synapse and ACDC datasets.

II. Related Work

Medical image segmentation, which involves the pixel-level separation of organs or lesions from medical images, has benefited greatly from the success of convolutional neural networks (CNNs). CNNs have played a crucial role in achieving accurate segmentation in medical images. However, researchers have been exploring the integration of transformers in medical image segmentation to address the limitations of CNNs and improve segmentation accuracy. In recent years, there have been significant efforts to introduce transformers into the medical field, aiming to maximize segmentation accuracy. In the following section, we will present and analyze the progress made by CNNs and transformers in the medical imaging domain.

A. CNN-based methods

With the development of modern science and technology, deep learning for medical segmentation has become very prevalent. CNN has been the dominant framework in the field of image vision for a long time, especially Fully Convolutional Networks (FCN). Initially, a given image has been separated into feature maps of arbitrary size by using a fully convolutional structure [2]. Inspired by FCN, U-net has undoubtedly become the optimal solution for medical image segmentation. It has made multi-scale and multi-granularity prediction possible by adding skip connections between corresponding low-level and high-level feature maps of the same spatial size. Zhao [21], Kirillov [22] and Lin [23] have designed multiple pyramid modules through a variety of different methods in order to obtain richer semantic information and segmentation results. Hu [24] has optimized the structure of CNN in a differential evolutionary algorithm way to achieve global capabilities. For a long time, everyone researcher has been devoted to exploring various optimization methods of CNN to obtain accurate results on corresponding tasks. However, the convolutional layer in CNN does not have the ability to capture long-range correlation. Even if optimization methods such as multi-scale [25, 26] are added, there are still shortcomings like

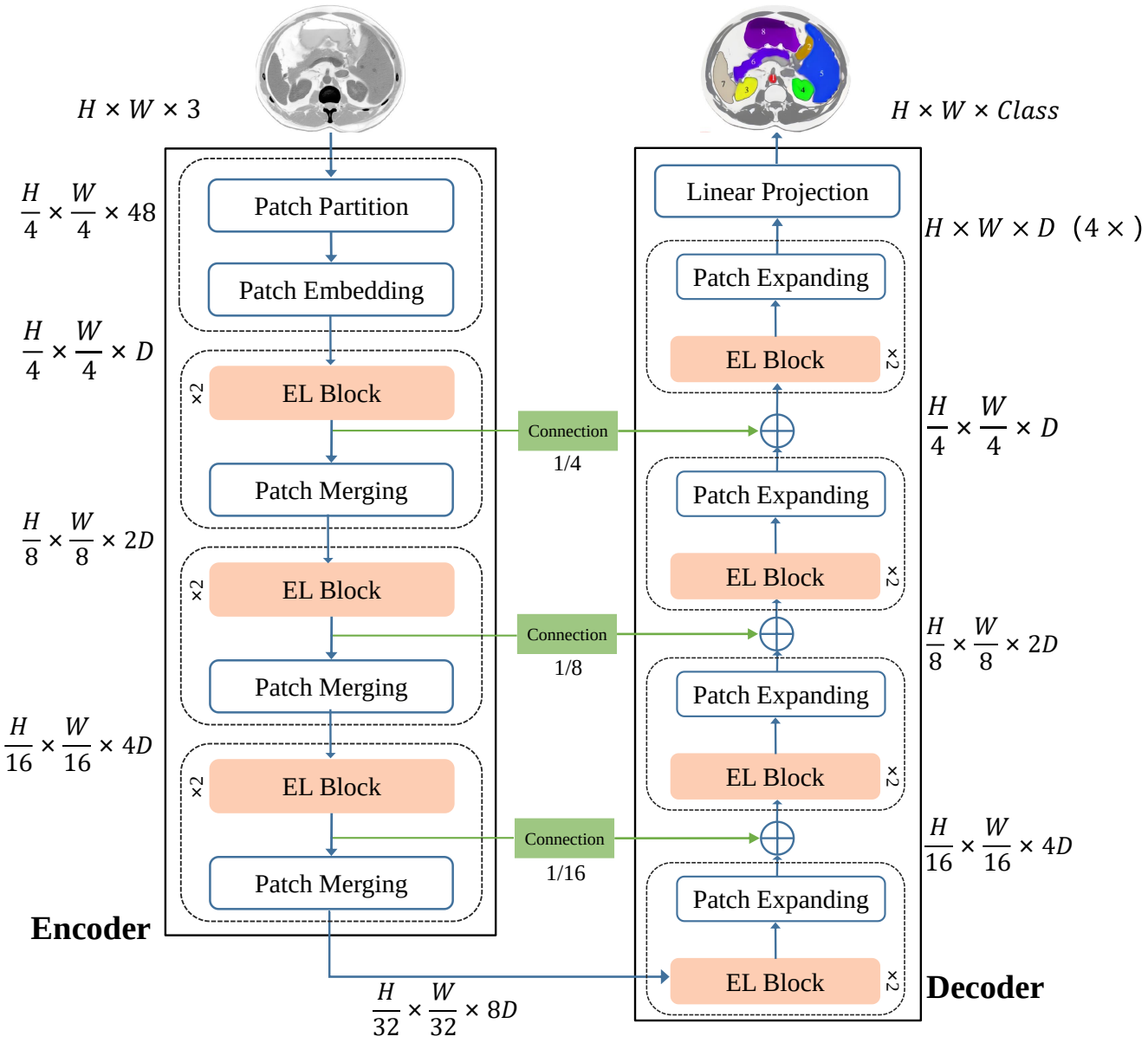

Fig. 2. The schematic of our EAST architecture, which consists of Encoder, Decoder and skip Connections. The EL block is the most crucial component of the entire framework and serves as the core part of implementing extensible attention.

local dependence, difficult training and high cost. How to obtain a more efficient and multi-scale attention network architecture has become a direction of constant exploration. The emergence of Transformer undoubtedly provides a fresh solution idea to these issues.

B. Transformers

Before the advent of Transformer, most backbones for medical segmentation were based on CNN [3, 4]. Especially BERT [10] shining in NLP, researchers have begun to explore its possibilities in CV. The emergence of ViT [11] has undoubtedly broken this gap. Although the design of ViT is a pure transformer model, it is mainly used for image classification and has certain limitations. After that, the Microsoft team and those involved in the field are designing the Transformer for superior performance, including a generic visual backbone network with partially sliding windows [12, 13] and its variants [14, 27–29]. Moreover, there are also researchers devoted to reducing the high computational costs associated with global attention while improving the accuracy of attention. For the purpose of obtaining more regional attention maps and reducing the computational cost of global self-attention, Dong [14] tried to use a cross-shaped window to improve. Yang [27] proposed a focal attention method to focus on a new mechanism of coarse and fine-grained attention. Wang [28] explored out of the pyramid visual transformer for dense prediction. Chen [30] processed image tokens using cross-attention of two independent branches to obtain a multi-scale visual transformer. Zhang [31] explored multi-scale feature maps for high-resolution encoding to improve ViT. Wang [32] employed cross-scale attention and Xu [33] utilized co-scale and convolutional attention mechanisms to enhance image transformers and more. After seeing the success of transformer and its variants on various tasks, many medical researchers also want to transfer it to the field of medical image processing. But

the accuracy of medical image processing has always been extremely demanding.

After being heartened by the positive experience of TransUnet [15], Swin-Unet [16], Msu-Net [25], and Ds-transunet [34], we are promising for explorations of Transformer in medical image applications. However, there existed several serious issues in most transformer experiments. They all have concentrated excessive attention on global features yet ignored local details. This has proven to be a significant challenge for the task of medical image segmentation. To address this problem, we introduce an extensible attention mechanism that can process feature maps at multiple granularities, allowing it to focus on both texture features and high-level semantic features of the images. By combining the proposed EAST with the classic U-Net codec and skip connections, we achieve accurate segmentation of both intra- and inter-image relationships in medical images.

III. Methods

At present, medical images are not closely related to the context so segmentation is not accurate. Most processing networks can only handle local or global situations, so we devise the Extensible Attentional Self-learning Transformer (EAST) architecture in Fig. 2 to solve these issues. This work is based on a U-shaped encoder-decoder structure with several skip connections, which can recover low-level spatial information and mix it with high-level semantic features for enhancing finer segmentation details. As is demonstrated, EAST also adopts a similar pyramid structure with [12, 28], which can help us obtain high-resolution feature maps suitable for medical image segmentation tasks. In the encoder stage, the medical image $H \times W \times 3$ is firstly divided into patches of size 4×4, while the feature dimension of each patch is $4 \times 4 \times 3 = 48$. Then these patches are projected to the hidden space of the corresponding dimension through the patch embedding layers. This spatial feature map will then be fed to EL blocks. Furthermore, the patch expanding layers[16] are devoted to implementing channel expansion and up-sampling in the decoder stage. The whole procedure is served for the fusion of contextual features and multi-scale semantic features through skip connections. Feeding the up-sampled features to the linear projection layer is the ultimate step that could output more accurate segmentation prediction results. We will explain the core block named EL, which consists of EA module and SL module. Then, we will describe other modules of the overall architecture in EAST.

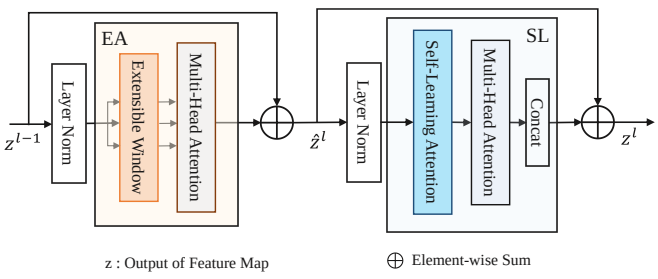

Fig. 3. The schematic diagram of the proposed EL block, which contains Layer Norm, Extensible Attention Module and Self-Learning Module with residual connections.

A. EL Block

The accuracy of segmentation targets is frequently affected by the interested region size. For a long time, Transformer has mostly focused on the global situation. It is difficult to obtain relatively refined segmentation results. Besides, ViT results in higher Flops and larger memory consumption with the fine-grained patch size focusing on images.

Based on these issues, the EL structure proposed by us includes both EA (Eq. (1)) and SL (Eq. (2)) as illustrated in Fig. 3. The self-attention is divided by using scalable windows from local to global in EA. It is a prominent way to understand the image and avoid ignoring the main features. The expanded attention windows are then fed into Multi-Head attention for interactive operations to capture the relationship between tokens. The output feature map is processed by SL after passing through LN layer[12]. This operation aims to improve communication between samples by using self-learning attention and multi-head attention. Not only does SL improve the generalization ability by introducing a self-learning matrix between samples, but also it continues to improve the ability of single-head attention by applying multi-head attention.

Thence, the output of the l-th layer in EL block can be written as follows:

$$\hat{z}^{\ell} = EA\left(LN\left(z^{\ell-1}\right)\right) + z^{\ell-1}, \tag{1}$$

$$z^{\ell} = SL\left(LN\left(\hat{z}^{\ell}\right)\right) + \hat{z}^{\ell} \tag{2}$$

where $\hat{z}^{\ell}$ represents the outputs of EA module, ℓ donates the ℓ-layer feature representation and z^{ℓ} denotes the output after SL module of the l-th block.

1) EA Module:

For the sake of completing the capture of coarse-grained and fine-grained targets to obtain attention regions of different scales, we introduce the EA module.

In the i-th stage,an input feature map F is given as $F_{i-1} \in R^{H_{i-1} \times W_{i-1} \times C_{i-1}}$. C is denoted as an arbitrary dimension projecting by a linear embedding layer. In Extensible Window, F is firstly divided into $\frac{H_{i-1}}{P_i} \times \frac{W_{i-1}}{P_i}$ patches, where P is the window partition size. Then we arrange the attention map in different scales according to the segmentation size. These tokens are extracted at multiple granularity levels. In Multi-Head Attention, each patch is flattened and projected to a C_i-dimensional embedding. Windows can be gradually expanded to the global. Window pooling is performed at each scaling level to obtain pooled tokens of different scales. Then the tokens of multiple expanded windows are concatenated together for a linear mapping to obtain the query $Q \in R^{N^2 \times d_K}$, key $K \in R^{N^2 \times d_K}$ and value $V \in R^{N^2 \times d_V}$ [9, 12]. N is the number of patches gotten from each window. d_K and d_V are $query/key$ dimension and $value$ dimension in the embedding space, respectively.

At this point, we can calculate extensible attention:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}} + B\right)V \tag{3}$$

among them, B is the learnable relative position deviation taken from a smaller-sized bias matrix $\widehat{B} \in R^{(2N-1)\times(2N+1)}$. Dividing each element of QK^T by the square root of $\sqrt{d_K}$ is

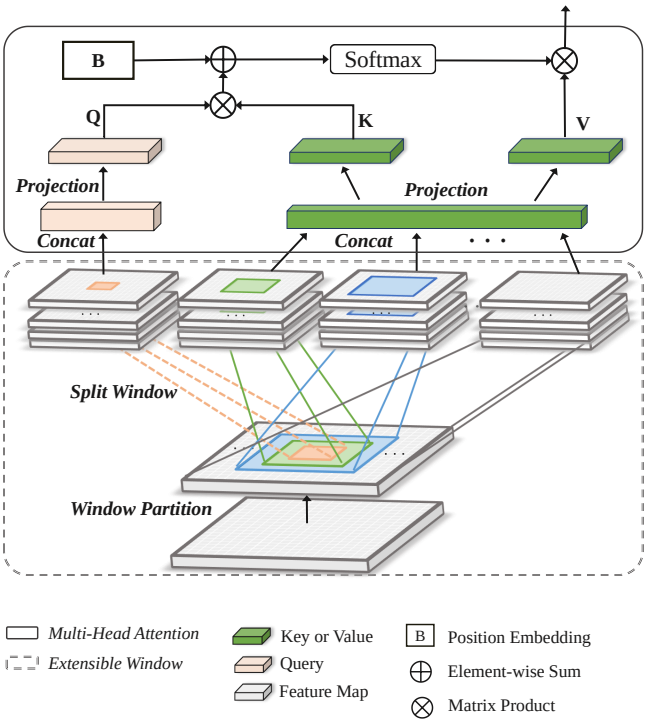

Fig. 4. An illustration of our expanded attention module. The feature maps are subdivided into different window sizes of particular granularity. We first take the most fine-grained window as the query matrix. The tokens of other relatively coarse-grained windows are connected immediately to map the key and value matrix. The obtained query, key and value matrix are combined with learnable relative positional encoding to perform softmax operation for getting attention results.

to prevent the magnitude of the values from growing wildly and to help back-propagation well.

To refine feature extraction, we introduce multi-head attention to learn attention operations, where the details are stated tersely as follows:

$$MultiHead\left(Q_i, K_i, V_i\right) = Con\left(h_0, h_1, ..., h_m\right) W^O \quad (4)$$

$$h_i = Attention\left(QW_i^Q, KW_i^K, QW_i^V\right) \quad (5)$$

where $Con\left(\cdot\right)$ is the concatenation operation as in [9]. h and m are denoted as the head and the number of the attention layer, respectively. The parameter matrices $W_i{}^Q \in R^{d\times d_K}$, $W_i{}^K \in R^{d\times d_K}$, $W_i{}^V \in R^{d\times d_V}$ and $W_i{}^O \in R^{d\times d}$ are the projections. $W_i{}^O$ as a linear transformation matrix is introduced to make the dimensions of input and output consistent.

With this design, EA could not only pay attention to fine-grained features in the local region, but also concentrate on the coarse-grained feature in the extensible region and the global scope. In addition, the heads are separated by extracting the query. It can reduce the complexity of the network while obtaining accurate segmentation.

2) SL Module:

SL module is proposed in this work to achieve more accurate and less complex segmentation tasks. Its main process is shown in Fig. 5 in detail, and then we will describe it formulaically.

Assume that the given feature map input after Layer Norm in Fig. 3 is $F_{SL} \in R^{N^{'}\times d}$, where $N^{'}$ denotes the number of pixels in images and d denotes the number of feature dimension. We follow the design rule of External Attention [35] that all samples share two different memory units $M_K \in R^{S\times d}$ and $M_V \in R^{S\times d}$, in which the main self-learning unit of EAST will be constituted.

The feature F_{SL} first obtains the query matrix $Q \in R^{N^{'}\times d}$ through the linear mapping of the self-attention mechanism. The attention between input pixels and self-learning memory cells is computed via the learnable key matrix, which is calculated as:

$$x_{i,j}^{'} = \left(QM_K{}^T\right) \quad (6)$$

where $x_{i,j}^{'}$ is the similarity between the i-th row M_K. In order to avoid the input features being too sensitive to scale and ignoring the correlation between feature maps, double normalization [36] is introduced. The operation in Equations (7) and (8) is to normalize columns normalization and rows respectively.

$$x_{i,j}^{*} = exp\left(x_{i,j}^{'}\right) / \sum_K exp\left(x_{K,j}^{'}\right) \quad (7)$$

$$x_{i,j} = x_{i,j}^{'} / \sum_K exp\left(x_{i,K}^{*}\right) \quad (8)$$

where the simplified calculation of $x_{i,j}$ is expressed as $x_{i,j} = Norm\left(QM_K^T\right)$. The obtained attention map is then calculated more accurately with the learnable value matrix M_V to improve the self-learning ability of this network as follows:

$$F_{out} = x_{i,j}M_V = \left(Norm\left(QM_K^T\right)\right) M_V \quad (9)$$

where F_{out} is the output of attentional feature map. SL continues to use multi-head attention to enhance the ability of the self-learning matrix, where each head can activate regions of interest to varying degrees. Its multi-head self-learning module can be written as:

$$F_{out} = MultiHead\left(F, M_K, M_V\right) \quad (10)$$

$$= Con\left(h_1, h_2, ..., h_H\right) W_O \quad (11)$$

where $Con\left(\cdot\right)$ is the concatenation operation. h_i denotes the i-th head and H denotes the number of heads. W_O is similar to $W_i{}^O$ in EA module to make the dimensions of input and output consistent. After this module, we could get a novel learnable attention map. The concatenated linear layer and the normalized layer are operated for connecting internal pixels to external elements.

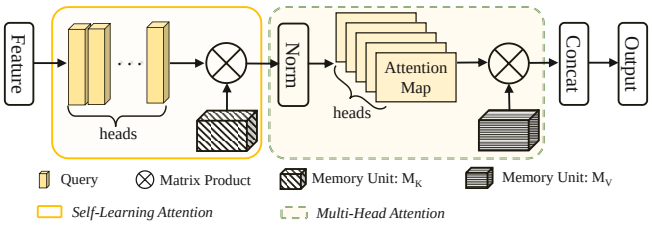

Fig. 5. The schematic process of our self-learning module. Query matrix is obtained through linear embedding, while M_K and M_V are the main learnable memory units used in this module.

B. Other Module of EAST

Our implementation of the extensible attention feature extraction model EAST relies on several modules, including EL blocks, skip connections, patch merging layer, and patch expanding layer.

1) EL Block:

The EL blocks are used to extract high-level features from the input data. These blocks use the attention of Transformer to extract features at multiple scales. The EL blocks are used in Section III-A to extract features that are sensitive to different spatial resolutions.

2) Skip Connection:

Skip connections are used to connect the output of a deep layer to a shallower layer, which allows for the integration of information from different depths. This helps to mitigate the vanishing gradient problem and allows the network to better capture both local and global information. To fuse multi-scale features obtained from the encoder with up-sampled features, skip connection is introduced. Its structure and function are mostly the same as U-Net [3]. Skip connection could reduce the loss of spatial information due to down-sampling and ensure the reusability of features. The dimension of the concatenated features has remained the same as the dimension of the up-sampled features. Moreover, we compared and discussed in detail the influence of the number of skip connections on EAST in Section IV-D3.

3) Patch Merging Layer:

The patch merging layer is used to reduce the dimensionality and merge features across different spatial locations. This is done to reduce the data volume and to allow for the efficient processing of large images. As the network gets deeper, the number of tokens increases dramatically. Therefore, we use the patch merging layer to reduce its number and achieve the purpose of generating hierarchical representations. A patch merging layer concatenates the features of each group of 2×2 neighboring patches and applies a linear layer on the $4C$-dimensional concatenated features. This operation down-samples the resolution of the features by $2\times$, and the output dimension is set to $2\times$. Then it is applied for feature transformation in EAST. This process is carried out three times in the encoder stage, and the output resolutions are $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively.

4) Patch Expanding Layer:

The patch expanding layer is responsible for restoring the dimensionality and resolution of the merged features. This is done to preserve spatial information and to allow for the extraction of features at multiple scales. It is the inverse operation of patch merging layer, which expands the resolution of the input features by $2\times$. And the size of the input features is reduced to $1/4$ of the original. This layer is employed four times in the decoder stage, and the feature outputs are $\frac{H}{16} \times \frac{W}{16}$, $\frac{H}{8} \times \frac{W}{8}$ and $\frac{H}{4} \times \frac{W}{4}$ and $H \times W$, respectively.

Overall, these modules work together to extract extensible attention features from the input data, which can be used for medical image segmentation.

IV. Experiments

A. Datasets

1) Synapse multi-organ segmentation dataset (Synapse):

We utilized the public multi-organ dataset from the MICCAI2015 Multi-Atlas Abdomen Labeling Challenge containing 30 abdominal CT scans. According to the setting of TransUnet [15], the dataset was divided into a training set with 18 samples and a test set with 12 samples.

For a fair comparison, we used the average Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95) as evaluation criteria to verify its segmentation performance for 8 abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen and stomach).

2) Automated Cardiac Diagnosis Challenge dataset (ACDC):

ACDC is also a public dataset, which is the result of cardiac MRI scans collected from different patients. The MR images of each patient were labeled with left ventricle (LV), right ventricle (RV) and myocardium (MYO).

Here, we randomly divided the dataset into 70 training samples, 10 validation samples and 20 test samples similar to TransUnet [15]. We have continued to report with DSC to validate our experiments.

B. Implementation Details

The Extensible Attentional Self-learning Transformer is executed with Pytorch and all experiments are performed on 4 NVIDIA GTX 1080Ti GPUs. We augment the data with random flips and rotations to increase the diversity of the data. The size of the input images is set to 224×224 for all methods. Our model is trained from scratch on ImageNet [40]. During training, the default batch size is 12 about 200 epochs. The model is back propagated with the Adam [41] optimizer learning rate 0.01, momentum 0.9, and weight decay 1e-4.

C. Experiment results

We experimentally compared the Synapse multi-organ segmentation dataset with the most advanced methods [3, 11, 15, 16, 37–39], as shown in Table I. First of all, it can be seen that the traditional CNN methods still have good performance. But it has been proven to be effective by adding a Transformer or using pure transformer architecture. These frameworks could achieve better results than CNN in a certain extent. Among them, R50 U-Net, R50 Att-Unet, and R50 ViT are all compared according to the setting mode of TransUnet [15]. Compared with V-Net [37], DARR [38] and ViT [11], other methods have reached more than 70% of DSC, but still have high HD. The experimental results are present that this method achieves the best segmentation effect, reaching 79.44% DSC and 19.28mm HD on Synapse. It is not difficult to see that our algorithm achieves an accuracy improvement of 0.31% and 2.27% in DSC and HD evaluation indexes. Compared with CNN models (e.g. R50 U-Net, R50 Att-Unet, U-Net, Att-Unet) or Transformer-related models (e.g. R50 ViT, TransUnet, Swin-Unet [16]), our experiments all obtained higher DSC and lower HD. And our model has achieved better segmentation results. Furthermore, it demonstrates the best segmentation performance for individual organs, namely Kidney (L), Liver, Pancreas, and Stomach, surpassing the best results by 0.63%, 0.24%, 0.74%, and 0.32%, respectively.

The progress of the experimental results in these two indicators proves that the method of gradually expanding attention proposed by us is feasible. Fig. 6 also shows some segmentation results. It can also be observed from this figure that our framework can increase the accuracy of segmentation to a certain extent. EAST architecture could learn high-level

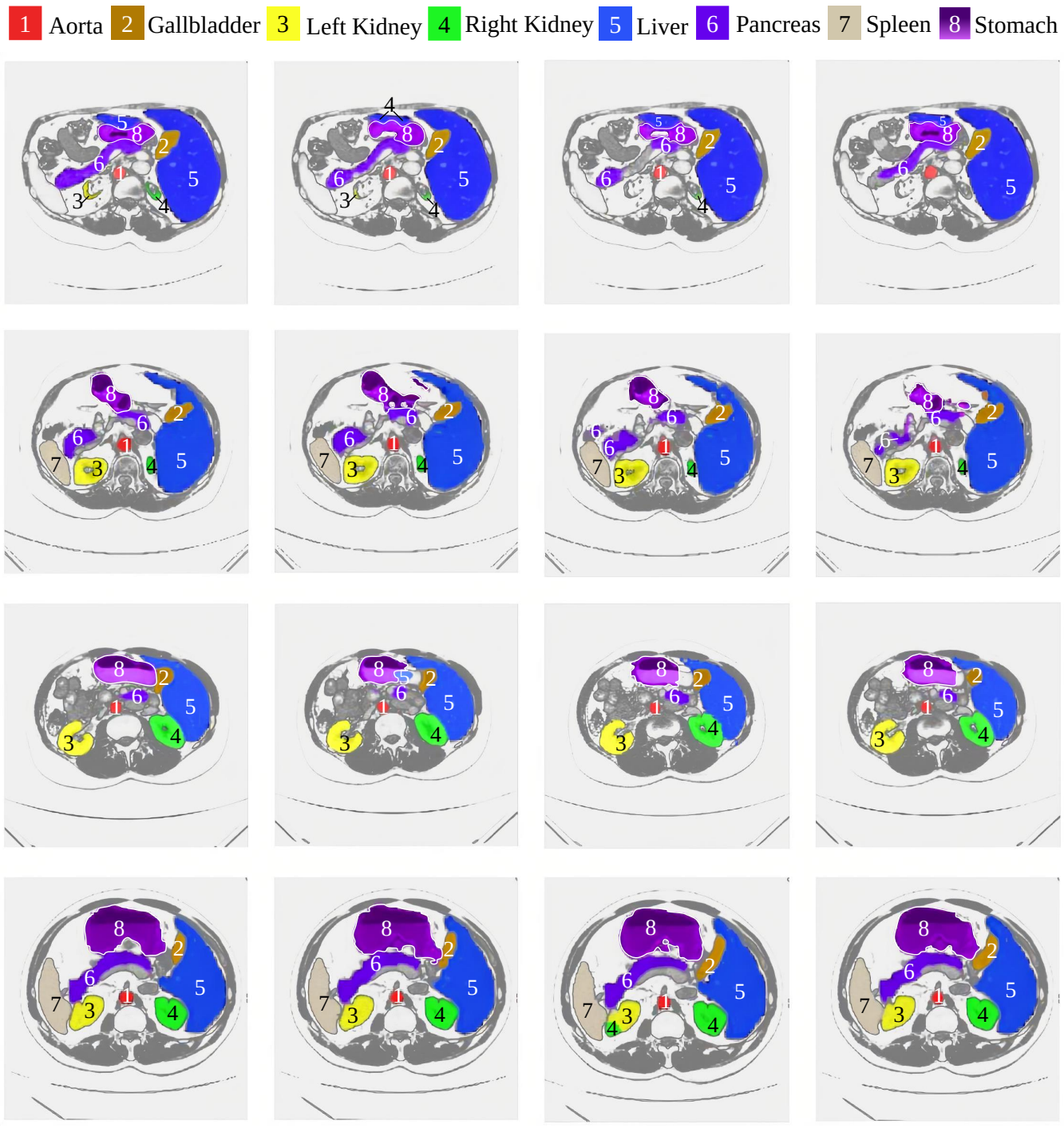

Fig. 6. The qualitative verification and comparison of different methods' segmentation results on the Synapse multi-organ CT dataset. From left to right: (a) Ground Truth, (b) Swin-UNet, (c) TransUNet, (d) EAST (ours). Our prediction results exit finer division and more accurate segmentation.

semantic features and low-level texture features at the same time, and realize accurate positioning and segmentation.

In order to evaluate the generalization ability of EAST model, we also train and test the medical image segmentation on the ACDC dataset. The results are shown in Table II, we still chose some state-of-the-art methods for comparison. The experimental results are displayed that our method has higher accuracy, which is similar to our results on Synapse. Although our method has 0.25% improvement compared with Swin-Unet [16], the success and improvement of these experiments prove that the framework has excellent generalization ability and robustness.

D. Ablation Study

We conducted ablation research on the main components of EAST to investigate the effectiveness of the proposed expanded attention and self-learning structure and explore the impact of image input scale and number of skip connections on the accuracy of model segmentation.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT SEGMENTATION EXPERIMENTAL RESULTS ON THE SYNAPSE MULTI-ORGAN SEGMENTATION DATASET. THE AVERAGE DSC %, HD IN MM AND THE AVERAGE DSC OF EACH SINGLE ORGAN ARE PRESENTED RESPECTIVELY.

Methods	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach	DSC	HD
V-Net [37]	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98	68.81	-
DARR [38]	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96	69.77	-
R50 U-Net [15]	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16	74.68	36.87
R50 Att-UNet [15]	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95	75.57	36.97
U-Net [3]	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58	76.85	39.70
Att-UNet [39]	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75	76.85	39.70
ViT [11]	44.38	39.59	67.46	62.94	89.21	43.14	75.45	69.78	61.50	39.61
R50 ViT [15]	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95	71.29	32.87
TransUNet [15]	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62	77.48	31.69
Swin-UNet [16]	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60	79.13	21.55
EAST(ours)	**87.39**	**67.57**	**83.91**	**77.88**	**94.53**	**57.46**	**89.85**	**76.92**	**79.44**	**19.28**

TABLE II
THE PERFORMANCE DEMONSTRATION OF CARDIAC SEGMENTATION IN ACDC DATASET BY USING SOME DIFFERENT METHODS. THERE ARE ALSO SHOWN THE SEGMENTATION RESULTS FOR MYO AND LV.

Methods	DSC	RV	MYO	LV
R50 U-Net	87.55	87.10	80.63	94.92
R50 Att-UNet	86.75	87.58	79.20	93.47
ViT	81.45	81.46	70.71	92.18
R50 ViT	87.57	86.07	81.88	94.75
TransUNet	89.71	88.86	84.53	95.83
Swin-UNet	90.00	88.55	85.62	95.83
EAST(ours)	**90.25**	**88.82**	**86.07**	**95.87**

1) The influence of EA / SL:

We attempt to delete EA or SL in our experimental architecture to verify the validity of the proposed module. The experimental results are listed in Table III. Experiments show that EA and SL are pretty vital for the model, and the lack of any module will lead to a decline in performance. In summary, EAST can achieve better segmentation performance, and it is indispensable to extensible attention and self-learning. The experiment illustrates the importance of inter-sample information interactions, since SL is more volatile for the results. EA, on the other hand, is less volatile for the results, which we suspect is because Transformer itself has global strengths. In addition, the addition of skip connections to lower-level information can compensate for its attention from local information in some way. A later experiment on the ablation of the number of skip connections confirmed this assumption.

TABLE III
RESULTS OF ABLATION EXPERIMENTS ON DIFFERENT MODULES OF EAST. SEPARATE TESTS ON THE INFLUENCE OF EA AND SL MODULES ON SEGMENTATION RESULTS

Methods	EA	SL	DSC	HD
EAST	√	×	79.02	20.86
EAST	×	√	78.97	21.73
EAST	√	√	79.44	19.28

2) The influence of image input scale:

Fig. 7 shows the input resolution of 224×224 and 512×512 experimental results. When we utilize 512 × 512 size as an image input, the input sequence length of Transformer will get larger. The larger size makes the experimental results more excellent and the segmentation results more accurate. However, the model accuracy is improved at the expense of computing speed and increasing computing overhead. Like TransUnet [15] and Swin-Unet [16], we still use the default resolution of 224×224 for reducing the computing overhead and improving the network computing speed.

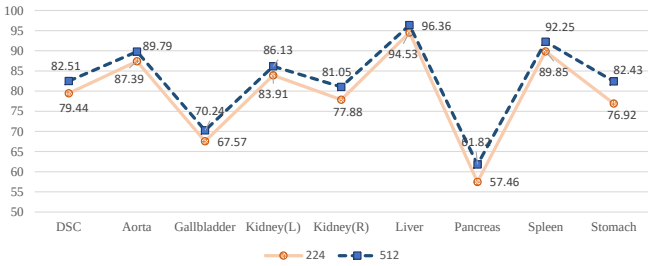

Fig. 7. Ablation study on the influence of image input scale about the average DSC (%) and the accuracy of various organs (%). It is proved that a larger input size has higher performance.

3) The influence of the number of skip connections:

It has been mentioned before that skip connection is extremely profitable for EAST. It allows the extraction of low-level spatial information to enhance the region of interest for segmentation in detail. The main interest of this ablation is to measure the effect of the number of skip connections on the segmentation performance.

The skip connections in EAST are located at resolution levels of 1/4,1/8,1/16. The average DSC and its scores on the 8 organs are compared in Fig. 8 by varying the number of skip connections to 0,1,2,3. For the "1-skip" setting, we only added skip connections in the 1/4 resolution range. For the "2-skip" setting, we added skip connections in the 1/4 and 1/8 resolution range.

The more skip connections we add, the superior the segmentation will be. It increases the segmentation of small organs even more. When skip connections are added for the first time, the segmentation performance is raised up even faster. This experiment verifies that skip connections are critical for extracting low-level detail. In fact, our validation shows that EAST (72.98%) performs much better than Swin-Unet (72.46%) without any skip connection, which demonstrates the superiority of the EA and SL modules for medical image processing. The best average DSC and HD shown in Fig. 8 could be gained by inserting skip connections into all three up-sampling steps of the EAST (i.e., in the 1/4, 1/8 and 1/16 resolution range). Therefore, we have adopted this configuration for EAST.

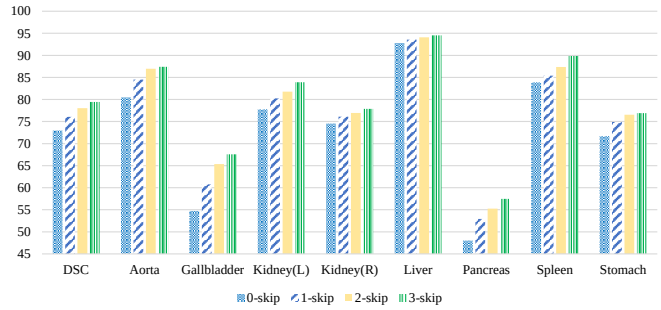

Fig. 8. Ablation study on the influence of the number of skip connections about the average DSC(%) and the accuracy of various organs (%). It shows that the best performance is achieved when the number of skip connections is 3, which is the chosen number in EAST.

V. Conclusion

Accurate image segmentation plays a crucial role in medical imaging applications, as it can greatly enhance diagnostic, therapeutic, and surgical outcomes. In order to achieve precise segmentation and improve overall effectiveness, we propose a robust and efficient visual Transformer named Extensible Attentional Self-learning Transformer (EAST) specifically designed for medical image segmentation tasks. By leveraging the Extensible Attention (EA) and Self-Learning (SL) modules, our model is capable of capturing image information accurately and comprehensively. Additionally, our model benefits from its global processing capabilities, allowing it to process image features in a sequential manner and enhance semantic understanding through the U-shaped structure. Through extensive experiments conducted on the Synapse and ACDC datasets, we have demonstrated the capable performance and generalization ability of our proposed algorithm in the auxiliary task of medical image segmentation.

VI. Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 36–46, 2021.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[4] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, 2018.

[5] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059, 2020.

[6] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 327–331, 2018.

[7] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, 2016.

[8] E. Erwin, "A hybrid clahe-gamma adjustment and densely connected u-net for retinal blood vessel segmentation using augmentation data," *Engineering Letters*, vol. 30, no. 2, pp. 485–493, 2022.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] 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," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10 012–10 022, 2021.

[13] C. Zhu, W. Ping, C. Xiao, M. Shoeybi, T. Goldstein, A. Anandkumar, and B. Catanzaro, "Long-short transformer: Efficient transformers for language and vision," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[14] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," *arXiv preprint arXiv:2107.00652*, 2021.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[16] 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," *arXiv preprint arXiv:2105.05537*, 2021.

[17] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-net transformer: Self and cross attention for medical image segmentation," *International Workshop on Machine Learning in Medical Imaging*, pp. 267–276, 2021.

[18] Y. Sha, Y. Zhang, X. Ji, and L. Hu, "Transformer-unet: Raw image processing with unet," *arXiv preprint*

arXiv:2109.08417, 2021.

[19] Y. Gao, M. Zhou, and D. N. Metaxas, "Utnet: a hybrid transformer architecture for medical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 61–71, 2021.

[20] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 603–612, 2019.

[21] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, 2017.

[22] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6399–6408, 2019.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

[24] Y. Hu, X. Zhang, J. Yang, and S. Fu, "A hybrid convolutional neural network model based on different evolution for medical image classification." *Engineering Letters*, vol. 30, no. 1, pp. 168–177, 2022.

[25] R. Su, D. Zhang, J. Liu, and C. Cheng, "Msu-net: Multi-scale u-net for 2d medical image segmentation," *Frontiers in Genetics*, vol. 12, p. 140, 2021.

[26] Z. Zhang, B. Sun, and W. Zhang, "Pyramid medical transformer for medical image segmentation," *arXiv preprint arXiv:2104.14702*, 2021.

[27] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal self-attention for local-global interactions in vision transformers," *arXiv preprint arXiv:2107.00641*, 2021.

[28] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.

[29] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[30] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 357–366, 2021.

[31] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2998–3008, 2021.

[32] W. Wang, L. Yao, L. Chen, B. Lin, D. Cai, X. He, and W. Liu, "Crossformer: A versatile vision transformer hinging on cross-scale attention," *arXiv preprint arXiv:2108.00154*, 2021.

[33] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9981–9990, 2021.

[34] A. Lin, B. Chen, J. Xu, Z. Zhang, and G. Lu, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *arXiv preprint arXiv:2106.06716*, 2021.

[35] M. Guo, Z. Liu, T. Mu, and S. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[36] M. Guo, J. Cai, Z. Liu, T. Mu, R. R. Martin, and S. Hu, "Pct: point cloud transformer." *Computational Visual Media*, vol. 7, pp. 187–199, 2021.

[37] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

[38] S. Fu, Y. Lu, Y. Wang, Y. Zhou, W. Shen, E. Fishman, and A. Yuille, "Domain adaptive relational reasoning for 3d multi-organ segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 656–666, 2020.

[39] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.

[40] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. FeiFei, "Imagenet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

Na Tian is currently pursuing a Ph.D. in Industrial Equipment and Control Engineering at the College of Automation and Electronic Engineering, Qingdao University of Science and Technology, in Qingdao, Shandong Province, China. She received a bachelor's degree in electronic information science and technology from the School of Automation, Qingdao University of Science and Technology in 2020, and will continue her academic research at Qingdao University of Science and Technology as part of a master-doctoral program. Her primary research areas include intelligent perception and machine vision, knowledge inference and intelligent recognition, and medical image processing.

Wencang Zhao received a B.Eng. degree in Automation from Qingdao University of Science and Technology, Qingdao, China, in 1995, the M.Eng. degree in Signal and Information Processing from the Shandong University, Jinan, China, in 2002, and the Ph.D. degree in Physical Ocean Science from Ocean University of China, Qingdao, China, in 2005. He was a Visiting Scholar with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA, in 2016 and 2017. Since 2005, he has been a faculty member at the College of Automation and Electronic Engineering, Qingdao University of Science and Technology, in Qingdao, China. His research interests encompass pattern recognition, image processing, and machine learning.