Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 2, 2024

CSA-Net: A Transformer-based Polyp Segmentation Network

Chunkai Qi[1], Jian Di[1, 2, *]

[1] School of Control and Computer Engineering, North China Electric Power University, Baoding, China
[2] Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, Baoding, China
[*]Corresponding author Email: dijian6880@163.com

Abstract: It is well known that colorectal polyps are a precursor to colorectal cancer. Accurate segmentation of polyp images from colonoscopy can assist clinicians in localizing polyp regions and reduce the occurrence of misdiagnosis accurately. Many existing methods achieve good results in the polyp segmentation task, but their extraction of global and local features is often insufficient. In this paper, we propose a transformer-based polyp segmentation network (CSA-Net) that utilizes two types of attention modules- spatial attention and channel attention-to further adaptively fuse local features with their global features. The proposed network is validated on five polyp datasets. Experimental results show that our model outperforms previously proposed models.

Keywords: Deep learning, Polyp segmentation, transformer.

1. Introduction

Colorectal cancer (CRC), the third most common tumor among men and women in the U.S., has led to the use of endoscopic image-based polyp segmentation technology, which helps physicians detect early pre-cancerous lesions, known as polyps, in the intestines. This technology can accurately identify individuals with risky diseases, and by detecting polyps and abnormal cellular tissue growth, it not only improves the cure rate and survival rate of patients but also reduces the suffering of patients and the cost required for treatment and simplifies the treatment process. In addition, accurate segmentation of polyps in colorectal endoscopic images can provide physicians with more quantitative information, such as size, shape, and location, which can help them more accurately assess the condition and tailor a treatment plan that is more appropriate to the patient's situation. Organization of the Text.

With the rise of deep learning technology, many computer vision tasks, including medical image segmentation, are gradually stepping into the deep learning era. However, compared with natural image segmentation, medical image segmentation faces unique challenges, such as high noise, large scale span, uneven illumination, low contrast, etc. These characteristics make accurate segmentation of medical images more difficult, which is especially obvious for the task of endoscopic polyp segmentation.

Currently, U-Net [1] shows good applicability in the semantic segmentation of medical images, and numerous deep learning methods based on U-Net have been widely applied to medical image segmentation tasks. These network architectures all adopt a similar encoder-decoder model, where the encoder consists of multiple convolutional and pooling layers to gradually expand the perceptual range of the network and extract high-level semantic information. The decoder, on the other hand, is responsible for generating the segmentation results. Although these networks have made significant progress in terms of performance, their architectures still have some potential shortcomings that warrant further investigation. One of the obvious problems is that the remote dependency between learned pixels still needs to be improved to some extent [2]. To address this limitation, some studies [3] introduced an attention mechanism into the model for augmenting the feature map for better pixel-level classification of medical images. PraNet [4], a classic method in the field of polyp segmentation, introduced salient target detection into polyp segmentation, demonstrating excellent learning results and strong generalization capabilities. Although these attention-based methods can improve performance, they still suffer from the shortcomings of remote dependencies.

Recent advances in visual transformers [5] have overcome the limitations mentioned above in capturing remote dependencies, especially in medical image segmentation [2] [6]. However, the self-attentive mechanism used in the Transformer model suffers from a limited ability to learn local (contextual) relationships [7]. In view of the above, this paper proposes a Transformer-based channel and spatial attention network (CSA-Net), which uses PVTv2 as the backbone network and designs a feature extraction module that parallelizes the spatial attention module and the channel attention module, thus effectively capturing global and local contextual relationships between pixels. Moreover, the network has been shown to accurately segment polyp locations on multiple datasets and the model has good generalization. The innovations of this paper are as follows:(1) A new segmentation method with a self-attentive mechanism is proposed; (2) CSA-Net can efficiently extract and fuse feature information from both global and local contexts; (3) t The results on five polyp datasets demonstrate state-of-the-art performance.

2. Transformer-based Polyp Segmentation Network

Our proposed network consists of three stages: encoder module, decoder module, and feature extraction module (FEM), as shown in Fig. We use a pyramid transformer as the feature encoder, and Pvtv2 uses convolutional operations instead of the patch embedding module of the traditional

transformer as a way to capture spatial information more efficiently. The features from the encoder are fed into the FEM, which consists of two parallel attention modules, the channel attention module (CAM) and the spatial attention module (SAM).

The network architecture is shown in Fig. 1:

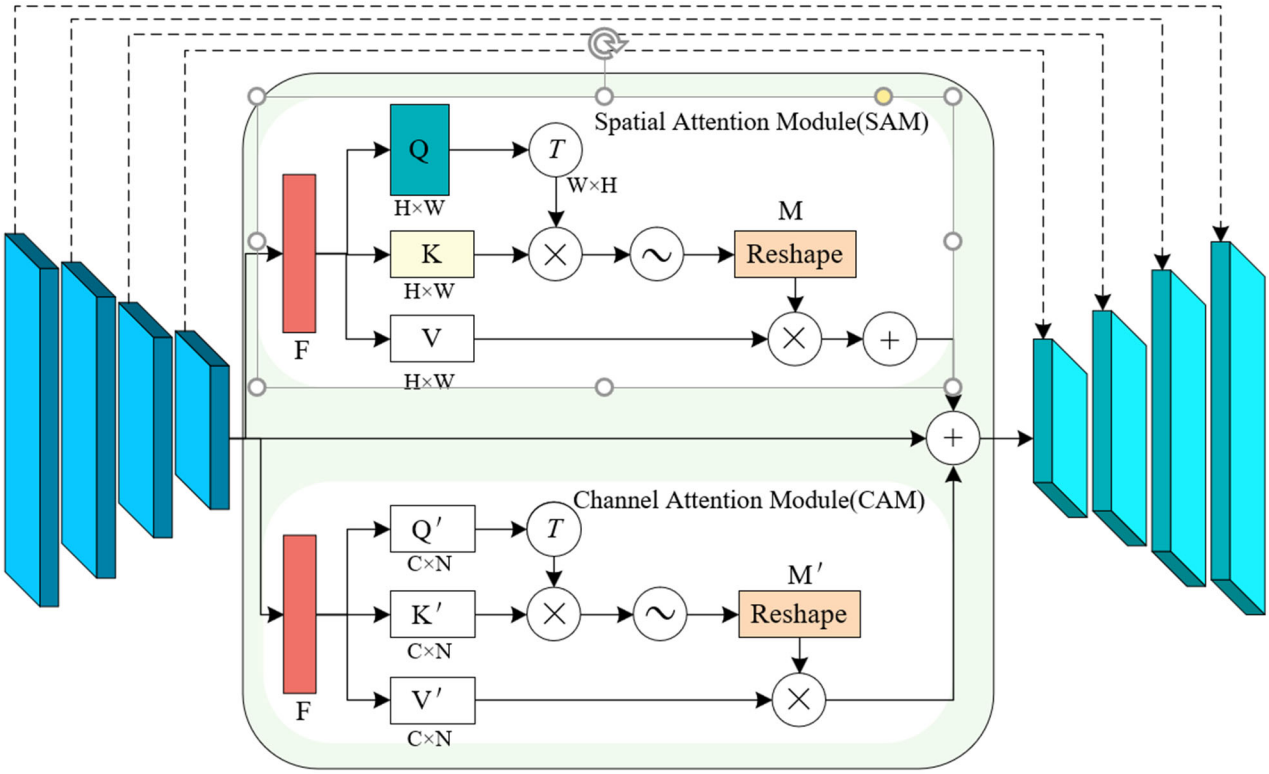

Figure 1. Diagram of the CSA-Net structure. The network consists of three stages: a feature encoder module, a channel and spatial attention module, and a feature decoder module

2.1. Spatial Attention Module.

Many recent studies have shown that local feature representations for traditional fully convolutional networks (FCNs) may lead to object classification errors [8][9]. In order to establish rich contextual dependencies on local feature representations, a spatial attention matrix needs to be constructed to model the spatial relationship between any two-pixel points for the spatial relationship between pixel features [10][11]. Therefore, we introduce SAM into the model to encode broader contextual information into local features to improve its ability to extract global contextual information.

First, we use batch normalization and ReLU layer or channel transformation processing to input features $F \in \mathbb{R}^{C\times H\times W}$. Here, C denotes the number of input channels, while H and W represent the height and width of F, respectively. Next, we use a 1×3 and a 3×1 kernel convolutional layer to generate two new feature maps, $Q \in \mathbb{R}^{C\times H\times W}$ and $K \in \mathbb{R}^{C\times H\times W}$, respectively, which capture the edge information of the semantic feature maps in the horizontal and vertical directions. Afterwards, these two new feature graphs are reshaped into the form of $\mathbb{R}^{C\times N}$ RC×N, where $N = H \times W$ represents the number of features. The transpositions of Q and K are further fused by matrix multiplication, and intra-class spatial associations can be obtained by applying a softmax layer:

$$S_{(x,y)} = \frac{\exp\left(K_{(x)} \cdot Q^{T}{}_{(y)}\right)}{\sum_{x=1}^{N} \exp\left(K_{(x)} \cdot Q^{T}{}_{(y)}\right)} \tag{1}$$

where $S_{(x,y)}$ denotes the effect of the x^{th} position on the y^{th} position.

Meanwhile, the feature map F is input to a 1×1 convolutional layer to produce a reduced-dimensional feature map $V \in \mathbb{R}^{C\times H\times W}$, and then S is reshaped to $\mathbb{R}^{C\times H\times W}$. Matrix multiplication is performed between V and S to obtain a pixel-level spatial affinity $\mathrm{M} \in \mathbb{R}^{C\times H\times W}$. Finally, we perform a pixel-level summation of F and M.

SAM can effectively acquire global and local contextual feature information and selectively aggregate contexts based on spatial attention feature maps. For the fuzzy edges of polyps and surrounding background regions, it can get more accurate segmentation results.

2.2. Channel Attention Module.

Each channel of a high-level feature contains class-specific responses [10], therefore, in this paper, we will further explore the interdependence of channel mappings and improve feature representation by emphasizing the interdependence of feature maps.

The input feature map $F \in \mathbb{R}^{C\times H\times W}$ is passed through a 1×1 convolutional layer to obtain three channel attention maps $Q' \in \mathbb{R}^{C\times H\times W}$, $K' \in \mathbb{R}^{C\times H\times W}$ and $V' \in \mathbb{R}^{C\times H\times W}$.

Similar to SAM, after we reshape into $\mathbb{R}^{C\times N}$, we perform a multiplication operation between F and its transpose and obtain the channel affinity map $M' \in \mathbb{R}^{C\times C}$ by applying a softmax layer.

$$C_{(x,y)} = \frac{\exp\left(F_{(x)} \cdot F_{(y)}\right)}{\sum_{x=1}^{C} \exp\left(F_{(x)} \cdot F_{(y)}\right)} \quad (2)$$

where $C_{(x,y)}$ represents the similarity between the x^{th} channel and the y^{th} channel. The final output can be obtained by performing a matrix multiplication between the C and V' transpositions, which will be reshaped to $\mathbb{R}^{C\times H\times W}$. Such an operation highlights the category-dependent feature mappings and helps to improve feature discrimination.

To recover the dimensionality of high-level semantic features, this paper proposes a novel feature decoder module. We design a series of upsampling layers in the decoder stage to progressively upsample the features of the current layer and align the dimensions with those of the next feature layer. Each upsampling module contains the upsampling layer, 3×3 convolution, batch normalization, and ReLU activation function. Between the encoder and decoder, we introduce jump connections, similar to U-Net [1]. Finally, we obtain the final segmentation map by applying a 1×1 convolutional layer and a Sigmoid layer to the output of the feature decoder module.

3. Experiments

This chapter first describes the datasets used for the evaluation, implementation details and evaluation criteria. Then, a series of ablation experiments are performed on five datasets, including Kvasir, CVC-ClinicDB, CVC-ColonDB, ETIS-LaribDB, and EndoScene.Finally, the effectiveness of each module in the proposed framework is demonstrated and analyzed, and the results are provided for visualization.

3.1. Datasets.

The segmentation dataset used in this paper is mainly colon polyps, and the polyp dataset contains five polyp data, which are ETIS-LaribDB [12], CVC-ClinicDB [13], CVC-ColonDB [14], EndoScene [15] and Kvasir [16].

The ETIS-LaribDB dataset is an early constructed dataset mainly used for early diagnosis of colorectal cancer and contains 196 images with an image size of 1225×966.

CVC-ClinicDB, also known as CVC-612, contains 612 images taken from 31 colonoscopy video sequences with an input size of 384×288.

CVC-ColonDB contains 380 images, mainly from 15 short colonoscopy video sequences, with an input image size of 574×500.

EndoScene is a combination of CVC-612 and CVC-300, which is divided into training and test sets as used by Fang et al. [66], but since CVC-612 has already been used to train the model, here we only use the CVC-300 test set.

The Kvasir dataset contains 1000 images with an image size of 626×546.

In this paper, this 5 dataset is used in the same way as Fan et al. [4], where 90% of the images in the training set are drawn from the Kvasir and CVC-ClinicDB datasets, and the remaining 10% are used for testing. The trained model is also tested on three never-before-seen datasets, i.e., CVC-ColonDB, ETIS-LaribDB, and CVC-300 test set, so as to validate the generalization performance of the network. For the image input for the network, we uniformly use 352×352.

3.2. Implementations details.

The methodology used in this paper is based on the PyTorch 1.13.1 implementation and utilizes NVIDIA RTX 3090 GPUs for training. We used a small batch Adam optimizer with momentum set to 0.9 to achieve optimization of the entire end-to-end network. Before inputting the network, the images were resized to a size of 352 × 352. the batch size was set to 16. all models were trained for 120 cycles with an initial learning rate of 0.0001. we used the pre-trained weights from ImageNet for the backbone networks. To increase the diversity of samples, two data enhancement strategies were used, including randomly flipping the input images vertically and horizontally and randomly rotating the images by 90 degrees. In the testing phase, this paper does not use any data enhancement or post-processing. The model uses a multi-stage loss function to train the network, which is calculated as follows:

$$loss = \lambda \times loss_{x1} + \beta \times loss_{x2} + \mu \times loss_{x3} + \omega \times loss_{x4}, \quad (3)$$

where x1, x2, x3, and x4 are the feature maps of the four predictor heads; and the losses of the four predictor heads are calculated as $loss_{x1}$, $loss_{x2}$, $loss_{x3}$, and $loss_{x4}$, respectively. λ, β, μ, and ω are the weights of the losses of each predictor head; λ, β, μ, and ω are set to 1 in the experiment.

3.3. Results.

In this study, common medical image segmentation models such as U-Net [1], U-Net++ [17], PraNet [4], TransUnet [18], SANet [19], SFA [20], and EU-Net [21] were used for quantitative comparison. Among them, U-Net and U-Net++ are the common models for medical segmentation, and their Yu methods are all polyp segmentation methods. The segmentation results of the polyp segmentation method proposed in this study with other methods on five publicly available datasets are presented in Table 1. Through Table 1, it can be observed that the polyp segmentation network proposed in this study shows excellent learning ability on Kvasir and CVC-ClinicDB datasets and basically achieves the best results in segmentation results. Only on the CVC-ClinicDB dataset, the method slightly underperforms SANet by 0.6 in terms of the Dice index and outperforms the second-ranked EU-Net by 1.3 in terms of the Dice index on the Kvasir dataset.

Similarly, we can compare the other three datasets and observe that our proposed polyp segmentation network also has a very good generalizability. Good results are obtained on three unseen datasets, ColonDB, ETIS and CVC-300, and optimal results are obtained on all three datasets. For example, on the dice metric of the ColonDB dataset, the method in this paper outperforms the

Table 1. Quantitative results for the polyp dataset,with the best results in bold and the second best results in italics.

Methods	EndoScene		CVC-ClinicDB		Kvasir		ColonDB		ETIS-LaribDB	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net[1]	71.0	62.7	82.3	75.0	81.8	74.6	51.2	44.4	39.8	33.5
U-Net++[17]	70.7	62.4	79.4	72.9	82.1	74.3	48.3	41.0	40.1	34.4
PraNet[4]	87.1	79.7	89.9	84.9	89.8	84.0	70.9	64.0	62.8	56.7
TransUnet[18]	87.5	80.0	89.1	83.0	89.1	82.9	76.9	68.6	73.7	65.8
SANet[19]	88.8	81.5	91.6	85.9	90.4	84.7	75.3	67.0	75.0	65.4
SFA[20]	46.7	32.9	70.0	60.7	72.3	61.1	46.9	34.7	29.7	21.7
EU-Net[21]	-	-	90.2	85.9	90.8	85.4	75.6	68.1	68.7	60.9
Ours	88.8	81.7	90.8	86.1	92.1	86.9	79.6	71.8	77.5	69.4

second-place TransUnet by 2.7. On the dice metric of ETIS-LaribDB, it outperforms the second-place SANet by 2.5. Taken together, the polyp segmentation method proposed in this paper outperforms the other comparative methods.

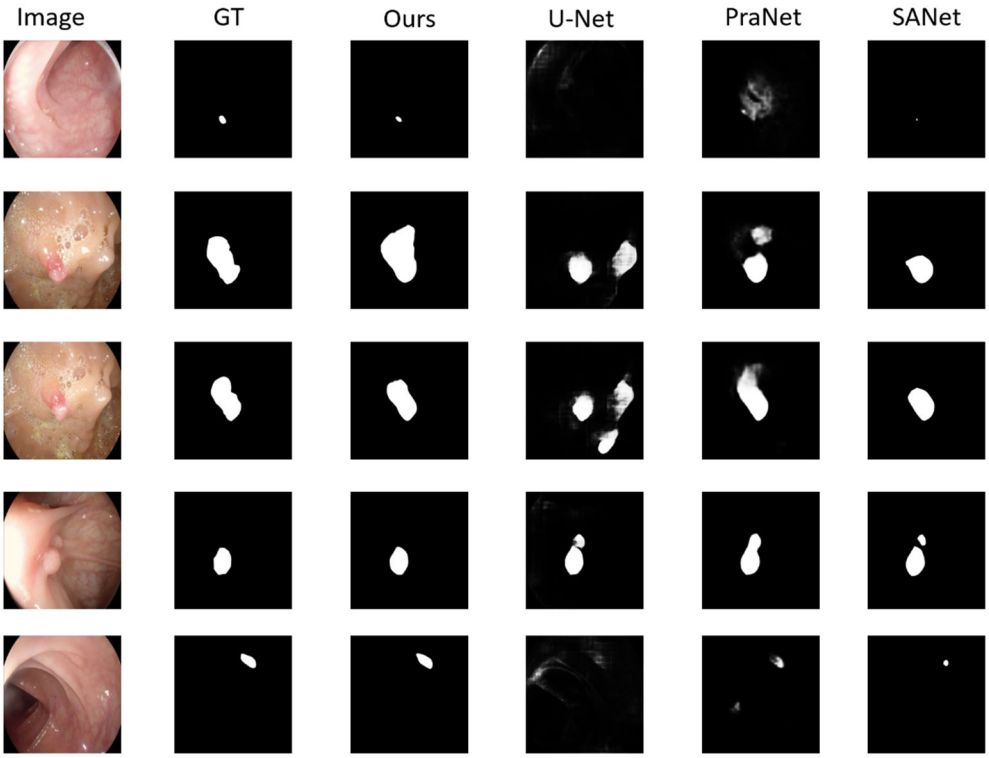

Figure 2. Qualitative results of the polyp segmentation dataset. From left to right: original image, Ground Truth, segmentation results generated by Ours (MBC-Net), U-Net, PraNet, SANet respectiv-ely.

In addition to the quantitative experiments previously described, this paper also carries out a large number of qualitative experiments. The methods proposed in this paper are compared experimentally with U-Net [1], U-Net++ [17], PraNet [4], and SANet [20], and the experimental results are analyzed. For methods for which experimental results are provided, we directly use their results. In contrast, for methods for which experimental results are not provided, we will follow the corresponding code to conduct experiments to derive segmentation results. In Fig. 2, we show a graph comparing the qualitative experimental results of the methods in this paper with those of other methods. The first and second columns of the figure show the test images and their truth labels selected in the experiments, and the third column shows the polyp segmentation method proposed in this paper. In contrast, the fourth to sixth columns show the experimental results of the other compared methods. The results of the qualitative comparison experiments show that the polyp segmentation network based on transformer encoder feature enhancement presents good segmentation results.

4. Summary

The convolution operation used in convolutional neural networks can only focus on limited local information and is restricted by the global modelling capability; at the same time, the pooling operation applied in the network also causes loss of information to a certain extent, which negatively affects the segmentation of small polyps and polyp boundaries. To address this problem, this paper proposes an improved polyp segmentation network based on Transformer. The proposed feature extraction module effectively extracts both global and local contextual feature information, overcoming the local limitations in the convolutional neural network mechanism. CSA-Net obtains the best segmentation results on several publicly available datasets, outperforming the other models with which it is compared.

References

[1] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical

image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015: 234-241.

[2] Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 205-218.

[3] Chen S, Tan X, Wang B, et al. Reverse attention for salient object detection[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 234-250.

[4] Fan D P, Ji G P, Zhou T, et al. Pranet: Parallel reverse attention network for polyp segmentation[C]//International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing, 2020: 263-273.

[5] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[6] Wang J, Huang Q, Tang F, et al. Stepwise feature fusion: Local guides global[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2022: 110-120.

[7] Islam M A, Jia S, Bruce N D B. How much position information do convolutional neural networks encode?[J]. arXiv preprint arXiv:2001.08248, 2020.

[8] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

[9] Peng C, Zhang X, Yu G, et al. Large kernel matters--improve semantic segmentation by global convolutional network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4353-4361.

[10] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146-3154.

[11] Mou L, Zhao Y, Chen L, et al. CS-Net: Channel and spatial attention network for curvilinear structure segmentation[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Springer International Publishing, 2019: 721-730.

[12] Silva J, Histace A, Romain O, et al. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer[J]. International journal of computer assisted radiology and surgery, 2014, 9: 283-293.

[13] Bernal J, Sánchez F J, Fernández-Esparrach G, et al. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians[J]. Computerized medical imaging and graphics, 2015, 43: 99-111.

[14] Tajbakhsh N, Gurudu S R, Liang J. Automated polyp detection in colonoscopy videos using shape and context information[J]. IEEE transactions on medical imaging, 2015, 35(2): 630-644.

[15] Vázquez D, Bernal J, Sánchez F J, et al. A benchmark for endoluminal scene segmentation of colonoscopy images[J]. Journal of healthcare engineering, 2017, 2017.

[16] Jha D, Smedsrud P H, Riegler M A, et al. Kvasir-seg: A segmented polyp dataset[C]//MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. Springer International Publishing, 2020: 451-462.

[17] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization[C]//International Conference on Learning Representations. 2018.

[18] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[19] Wei J, Hu Y, Zhang R, et al. Shallow attention network for polyp segmentation[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer International Publishing, 2021: 699-708.

[20] Fang Y, Chen C, Yuan Y, et al. Selective feature aggregation network with area-boundary constraints for polyp segmentation[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Springer International Publishing, 2019: 302-310.

[21] Patel K, Bur A M, Wang G. Enhanced u-net: A feature enhancement network for polyp segmentation[C]//2021 18th Conference on Robots and Vision (CRV). IEEE, 2021: 181-188.