Journal of Computer and Communications, 2024, 12, 238-251
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

ATFF: Advanced Transformer with Multiscale Contextual Fusion for Medical Image Segmentation

Xinping Guo, Lei Wang, Zizhen Huang, Yukun Zhang, Yaolong Han

School of Computer Science and Technology, Shandong University of Technology, Zibo, China
Email: 450641388@qq.com

How to cite this paper: Guo, X.P., Wang, L., Huang, Z.Z., Zhang, Y.K. and Han, Y.L. (2024) ATFF: Advanced Transformer with Multiscale Contextual Fusion for Medical Image Segmentation. *Journal of Computer and Communications*, **12**, 238-251.
https://doi.org/10.4236/jcc.2024.123015

Received: February 21, 2024
Accepted: March 26, 2024
Published: March 29, 2024

Abstract

Deep convolutional neural network (CNN) greatly promotes the automatic segmentation of medical images. However, due to the inherent properties of convolution operations, CNN usually cannot establish long-distance interdependence, which limits the segmentation performance. Transformer has been successfully applied to various computer vision, using self-attention mechanism to simulate long-distance interaction, so as to capture global information. However, self-attention lacks spatial location and high-performance computing. In order to solve the above problems, we develop a new medical transformer, which has a multi-scale context fusion function and can be used for medical image segmentation. The proposed model combines convolution operation and attention mechanism to form a u-shaped framework, which can capture both local and global information. First, the traditional converter module is improved to an advanced converter module, which uses post-layer normalization to obtain mild activation values, and uses scaled cosine attention with a moving window to obtain accurate spatial information. Secondly, we also introduce a deep supervision strategy to guide the model to fuse multi-scale feature information. It further enables the proposed model to effectively propagate feature information across layers, Thanks to this, it can achieve better segmentation performance while being more robust and efficient. The proposed model is evaluated on multiple medical image segmentation datasets. Experimental results demonstrate that the proposed model achieves better performance on a challenging dataset (ETIS) compared to existing methods that rely only on convolutional neural networks, transformers, or a combination of both. The mDice and mIou indicators increased by 2.74% and 3.3% respectively.

Keywords

Medical Image Segmentation, Advanced Transformer, Deep Supervision,

Attention Mechanism

1. Introduction

Medical image segmentation aims to identify objects of interest from surrounding tissues and structures. It is essential for reliable diagnosis and morphological analysis of specific lesions, and can provide a reliable basis for pathological research and clinical diagnosis, so that doctors can make more accurate diagnosis. X-rays use the fact that bones decay faster than other soft tissues to obtain the state and density of bones [1]. Computed tomography (CT) is used to examine dense structures such as bones and implants [2]. Magnetic resonance imaging (MRI) can display high-resolution anatomical information of soft tissue [3]. Ultrasound can analyze the distribution of tissues because different tissues have different impedances to sound [4]. Accurate and robust medical image segmentation results play a vital role in clinical diagnosis and treatment (such as computer-aided diagnosis, preoperative evaluation and image-guided surgery), and are of great significance to the accuracy and efficiency of clinical diagnosis [5].

With the development of deep learning technology, some methods based on convolutional neural networks, such as U-Net [6], Attention U-Net [7], and nnU-Net [8], have dominated the field of medical image segmentation. Among them, in 2015, the classic U-shaped full convolutional neural network model based on codec structure was first proposed in literature [6], called UNet. The UNet increases the receptive field by constantly stacking convolution layers and downsampling. Through multi-layer convolution operations, the obtained highlevel feature map is helpful to identify the segmented target, and the jump connection method is used to add the low-level detail features of the Encoder stage to the upsampling part, which is conducive to the accurate positioning of the target. In addition, U-Net is widely used because of its simple and efficient features. Many researchers have proposed various improved models on the basis of it. For example, in 2018, reference [9] systematically evaluated the effects of different FCN variants on breast lesion segmentation for the first time, and achieved better segmentation results than traditional methods. In the literature [10], UNet + + obtains features of different depths and levels by nested and dense jump connections. In the literature [11], AttentionU-Net proposes to use the attention gate model, which can hinder the model learning of task-independent features and strengthen the learning of task-related features. In the literature [12], Unet3 + proposes an all-round jump connection, which connects the low-level features of feature maps of different scales with high-level features, so that the segmentation accuracy is improved. In [13], UNeXt uses a convolutional multi-layer perceptron (MLP)-based segmentation network to focus on learning local dependencies. Although these methods have greatly improved in accuracy and generalization ability than traditional methods, there are still some shortcomings in these methods. For example, stacked convolution and skip connec-

tion operations are used in UNet, but the resolution is reduced, resulting in the loss of many details such as edges and textures, which is detrimental to the accurate segmentation of medical images. It will lead to the problem that it is difficult to accurately locate the organ and the contour of the segmented target is vague. Moreover, due to the lack of support of context information, the accuracy and precision of semantic segmentation are not improved, and it is also greatly affected by factors such as noise interference. UNet + + has many more intermediate nodes than U-Net, but it increases the parameters of the model.

Recently, the transformer developed in the field of natural language processing has been successfully applied to various computer vision [14]. Recent studies have also shown that Transformer can achieve great success in medical image segmentation. Transformer uses the attention mechanism to establish long-distance interactions in the image to learn global context information. It is worth noting that Transformer can aggregate global information at an early stage [15], so we need to consider how to effectively transfer low-level feature information to high-level. Recent studies have shown that multi-scale connections in Transformer are more influential than multi-scale connections in CNN, which further enhances the feature similarity between low-level and high-level. [16] Therefore, making full use of multi-scale feature representation can improve the performance of the visual converter. In order to further improve performance, various U-Nets have tried to combine traditional architectures with attention mechanisms to construct attention-based U-Nets [17]. TransUNet combines U-Net and Transformer to capture local and global features for medical image segmentation, and achieves excellent segmentation performance [18]. To model local and global context, nn-Former [19] exploited the combination of interleaved convolution and self-attention operations within the encoder and decoder for volumetric medical image segmentation. Furthermore, MISSFormer [20] embedded depth-wise convolution into the transformer block for capturing local and global dependencies. In order to solve the training requirements of a large number of data sets, MedT proposed a special medical image segmentation Transformer [21], which uses axial attention in the multi-head attention block [22]. In summary, Transformer has shown great potential in medical image segmentation.

Although deep learning models, especially convolutional neural networks (CNN), have made significant progress in medical image segmentation tasks, they still have limitations in handling long-distance dependencies and complex spatial relationships. Recently, the Transformer model has attracted widespread attention due to its success in natural language processing (NLP) and has begun to be applied to computer vision tasks. By introducing an advanced Transformer architecture, ATFF may provide a new solution for capturing complex spatial relationships and long-distance dependencies in medical images. By combining Transformer's self-attention mechanism and multi-scale context fusion, ATFF may fill the gap in existing technology in enhancing the expressive ability of key features in medical images. Taking advantage of Transformer's processing ad-

vantages of long-distance dependencies and comprehensive utilization of multi-scale information, ATFF may be an important addition to the existing literature in improving the accuracy of specific and complex medical image segmentation tasks. In summary, the ATFF model is expected to provide higher accuracy, flexibility, and robustness when dealing with complex medical image segmentation tasks by combining an advanced Transformer architecture and a multi-scale context fusion strategy, thus opening up new research in the field of medical image analysis.

2. The Whole Method

Figure 1 shows the overall framework for medical image segmentation, which consists of an encoder, a decoder and a deep supervision to form a U-shaped network. Advanced Transformer Block aims to learn the long-distance context information in each feature map. In order to further enhance the spatial positioning capability, Advanced Transformer Block is combined with CNN-based Bottleneck block to construct an encoder. As shown in Figure 1, the Bottleneck Block is composed of some convolution operations with residual connections, which has been proved to be efficient. These convolution operations are designed to extract local information in each feature map. Therefore, the encoder part can effectively capture local and global features. Moreover, in ATFF, the output of the previous encoder is the input of the next encoder. The decoder consists of a convolutional layer, an upsampling layer and a ReLU activation layer. There is also a jump connection between these encoders and decoders. In addition, we also introduce a deep supervision strategy to guide the model to fuse multi-scale feature information. This will help the proposed ATFF to effectively propagate feature information across layers, thereby achieving better pixel-level segmentation accuracy.

As shown in Figure 2, the traditional transformer block is redesigned and named as the advanced transformer block. Specifically, the three-layer normalization in ReMix-FFN is simplified to two layers, called Advanced FFN. Subsequently, the exchange feedforward neural network (FFN) position and the normalized back layer normalization are used to obtain the mild activation amplitude, so as to achieve stable training. In addition, in order to obtain accurate spatial information, a scaled cosine attention with a moving window [23] is used to replace the space-reduced self-attention. The advanced transformer block is made of a multi-head self-attention module, residual connections, post-layer normalization, and Advanced MixFFN with GELU nonlinear function. We replace scaled dot product attention with scaled cosine attention for the self-attention computation. As shown in Figure 1, the window-based multi-head self-attention (W-MSA) module and the shifted window-based multi-head self-attention (SW-MSA) module are applied to two consecutive advanced transformer blocks, respectively. Based on this window division mechanism, two consecutive advanced transformer blocks can be expressed as:

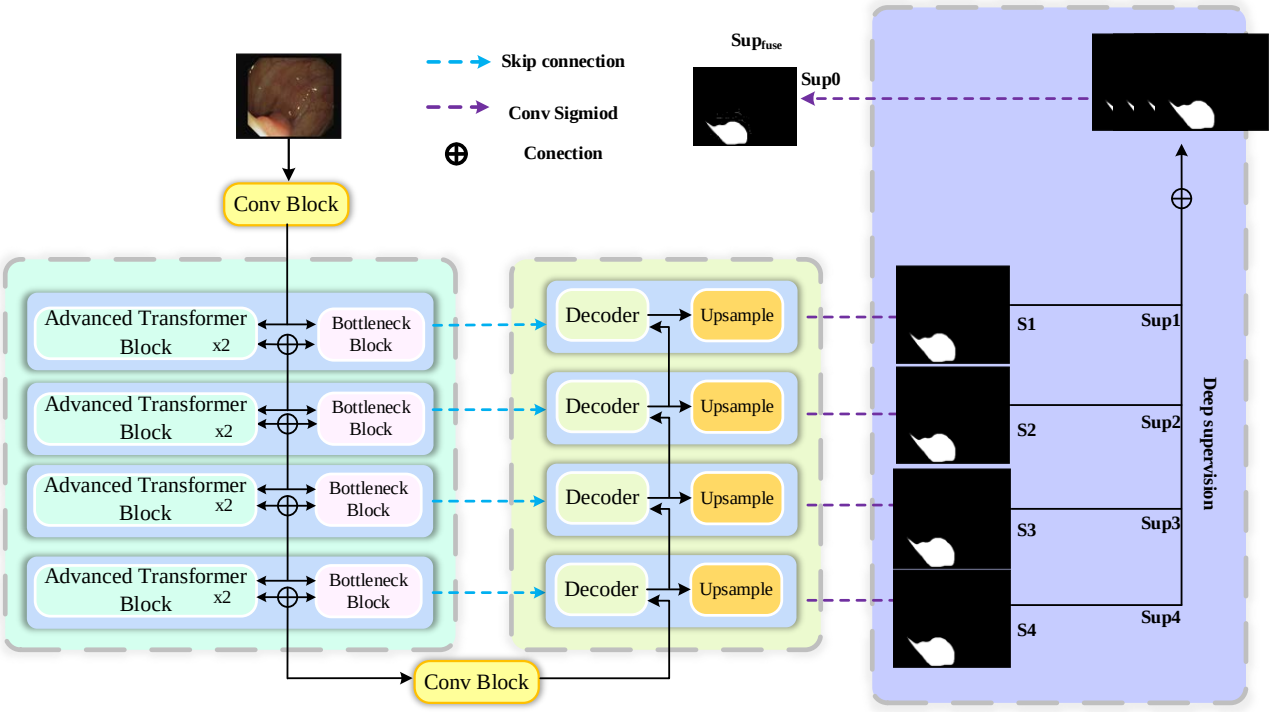

Figure 1. The Architecture of the VTANet model.

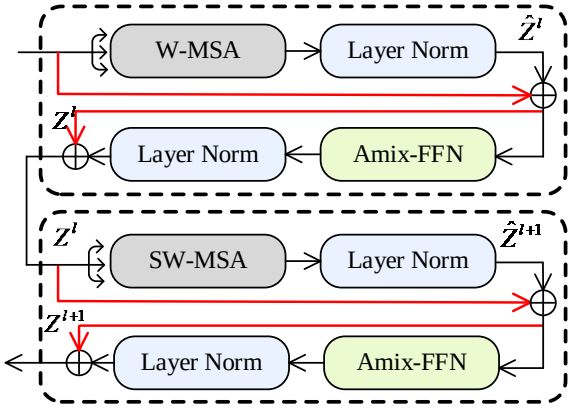

Figure 2. Two successive advanced transformer blocks. AMix-FFN stands for Advanced MixFFN.

$$\begin{aligned}
\hat{Z}^{l} &= \mathrm{LN}\left(\mathrm{W\text{-}MSA}\left(Z^{l-1}\right)\right) + Z^{l-1}, \\
Z^{l} &= \mathrm{LN}\left(\mathrm{AMix\text{-}FFN}\left(\hat{Z}^{l}\right)\right) + \hat{Z}^{l}, \\
\hat{Z}^{l+1} &= \mathrm{LN}\left(\mathrm{SW\text{-}MSA}\left(Z^{l}\right)\right) + Z^{l}, \\
Z^{l+1} &= \mathrm{LN}\left(\mathrm{AMix\text{-}FFN}\left(\hat{Z}^{l+1}\right)\right) + \hat{Z}^{l+1},
\end{aligned} \tag{1}$$

where $\hat{Z}^{l}$ and Z^{l} are the outputs of the (S)W-MSA and Advanced MixFFN modules, respectively.

In addition, a scaled cosine attention is proposed to solve the problem of numerical instability and gradient disappearance. The attention formula is as fol-

lows:

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{\cos(Q,K)}{\tau}+B\right)V \tag{2}$$

where $Q,K,V\in R^{M^2\times M^2}$ are the query, Key, and value matrices. M^2 indicates the number of patches in the window. B denotes the relative position deviation matrix, and τ is a learnable scalar. The cosine function has a lower attention weight since its value is automatically adjusted to the interval [−1, 1].

2.1. Post-Layer Normalization

Ze *et al.*'s research [24] shows that the pre-layer normalization used in traditional transformers will bring difficulties to training, because the activation value output by each residual block will be directly incorporated into the main branch. The huge difference in the amplitude of the activation values of each layer causes training difficulties, and this difficulty is more serious in the residual block containing the pre-layer normalized convolution. Therefore, we align the positions of the convolutional FFN and the layer-normalized MSA to obtain a gentle activation value amplitude, that is, post-layer normalization, thereby stabilizing the training. This seemingly simple but effective operation simplifies the normalization of the three-layer recursive layer in the original ReMix-FFN into two layers.

2.2. Advanced MixFFN

The combination of FFN and post-layer normalization allows advanced hybrid FFN to use only two layers of recursive layer normalization to learn discriminative features. In addition, deep convolution continues to be used to combine key components of the local context. As shown in **Figure 3**, the advanced MixFFN consists of two fully connected layers, a deep convolution layer, a double normalization layer and a GELU layer. A skip connection is added before the deep convolutional layer. Then, the result of the deep convolution is added to the skip connection as the input of the layer normalization. This process can be expressed as follows:

$$y_1=\text{LN}\left(\text{Conv}_{3\times3}\left(\text{FC}\left(x_{in}\right)\right)+\text{FC}\left(x_{in}\right)\right) \tag{3}$$

$$y_2=\text{LN}\left(y_1+\text{FC}\left(x_{in}\right)\right) \tag{4}$$

$$x_{out}=\text{FC}\left(\text{GELU}\left(y_2\right)\right) \tag{5}$$

where, x_{in} denotes self-attention output. $\text{Conv}_{3\times3}$ denotes a deep convolution with a kernel size of 3 × 3. LN Presentation layer normalization. GELU represents the activation function.

In this work, we further introduce a deep supervision strategy to train the model. As shown in **Figure 1**, deep supervision uses multi-scale feature representation to train the model, and the fusion result of the obtained multi-scale features is regarded as the final segmentation map. The main loss function is defined as

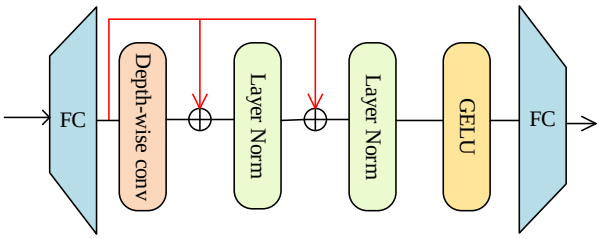

Figure 3. Advanced MixFFN structure diagram.

$$L = \sum_{m=1}^{M} W_m L_m + W_{fuse} L_{fuse} \tag{6}$$

where $L_m L_{side}^m L_{side}^m$ is the deep supervision related loss of each stage output S_m, m = 1, 2, 3, 4, *i.e.*, the deep supervision Sup 1, Sup 2, Sup 3 and Sup4. And L_{fuse} is the deep supervision related loss of final fused segmentation output S_{fuse}, *i.e.*, the deep supervision Sup 0. W_m and W_{fuse} are the weights of each loss term, respectively, which are used to control the balance between these loss terms. In general, the larger the weight of a specific loss term, the more attention will be paid in the training. Here, referring to [25], we set $W_m = 1$ and $W_{fuse} = 1$ due to that each loss term is evenly allocated and they play equal roles in the training.

Here, we use the standard bivariate cross entropy (CE) function to define the loss function, which can be expressed as

$$L_{CE} = -\frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \left\{ P(x,y) \log\left[\hat{P}(x,y)\right] + \left[1 - P(x,y)\right] \log\left[1 - \hat{P}(x,y)\right] \right\} \tag{7}$$

Here, W and H are the width and height of the image, respectively, and $P(x, y)$ is the ground truth pixel value and is the prediction pixel value at location (x, y). Deep supervision enables the proposed model to learn more low-level details in the shallow layer and high-level semantics in the multi-scale layer. This will help the proposed ATFF to effectively propagate feature information across layers. Therefore, it can eliminate the attenuation of low-level information and retain more input spatial information.

3. Experimental Results and Discussion

3.1. The Experiment Setting

Five challenging public datasets, including Kvasir-SEG [26], ClinicDB [27], ColonDB [28], Endoscene [25] and ETIS [29], are used to evaluate the proposed method. Specifically, the ClinicDB and Kvasir-SEG datasets were used to evaluate the learning ability of the model. The ClinicDB contains 612 images that were extracted from colonoscopy videos. Kvasir-SEG included 1000 polyp images. In the experiment, the same 548 and 900 images in the ClinicDB and Kvasir-SEG datasets were used as training sets, and the remaining 64 and 100 images were used as corresponding test sets.

The experiment is implemented using the pytorch framework. Considering the difference in the size of each polyp image, a multi-scale strategy is used in the training phase. In addition, the AdamW optimizer is used to update the

network parameters, which is widely used in transformer networks [22] [23]. The learning rate is set to 1e-4, and weight dacay is also adjusted to 1e-4. In addition, the size of the input image is adjusted to 352 × 352, and the minibatch size is 16 for 100 epochs. In the test section, only the image size is adjusted to 352 × 352, and there is no post-processing optimization strategy.

In the experiment, five widely used evaluation indexes are used, including Dice Coefficient (Dice), Hausdorff Distance (HD), Intersection over Union (IoU), Accuracy (ACC) and Recall (REC) are employed to evaluate the quantitative results. Here, Dice and IoU mainly focus on the internal consistency of segmentation results. ACC and REC are computed with pixel-by-pixel to assess the quantitative evaluation of segmentation performance. Dice, IoU, ACC and REC can be calculated by four parameters, *i.e.*, True-Positive (TP), False-Positive (FP), True-Negative (TN) and False-Negative (FN). They are defined as:

$$\text{Dice} = \frac{2\times \text{TP}}{2\times \text{TP} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{9}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10}$$

$$\text{REC} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

HD is used to evaluate the sensitivity of segmentation boundary. For the image segmentation, HD is calculated between boundaries of the predicted result and ground truth. Let x and y represent pixels in the predication set X and ground truth set Y, respectively. HD is defined as:

$$\text{HD}(X,Y) = \max\left(\text{maxmin}\left\|x-y\right\|_2, \text{maxmin}\left\|x-y\right\|_2\right) \tag{12}$$

where HD (X, Y) is the longest distance of a point in one set to its closest point in the other set, where, $x \in X$, $y \in Y$.

3.2. Experimental Results

In order to verify the effectiveness and robustness of the proposed network, comparative experiments were conducted with other methods in qualitative and quantitative evaluation. Six classic network models were compared. These comparison methods include CNNs-based models such as ConvUNeXt [30] EU-Net [31], and the recently leading Transformer and attention-based models such as MedT [32], TGANet [33], Swin-Unet [34].

Firstly, the image segmentation comparison experiment is carried out on the Kvasir-SEG dataset. The quantitative comparison results of the evaluation indicators are shown in **Table 1**. It is worth noting that in medical image segmentation, indicators such as Dice, HD and IoU are generally more worthy of attention. From **Table 1**, we can see the following points. The Dice value of ATFF is 89.78%, the HD value is 3.22%, the IoU value is 84.2%, the ACC value is 97.62%,

and the REC value is 91.38%. The Dice value, HD value, IoU value, ACC value, and REC value are 0.42%, 0.22%, 1.2%, 0.41%, and 0.04% higher than the latest transformer-based model Swin-UNet, respectively. From Table 2, we can also see that the Dice value, HD value, IoU value, ACC value, and REC value are respectively 0.24%, 1.07%, 2.42%, 1.09%, and 1.75% higher than the transformer-based TransUnet. Compared with other transformer-based popular models such as MedT and TransUNet, the proposed model ATFF has also been significantly improved. The main advantages are the proposed Advanced Transforner Block and deep supervision, which enables ATFF to pay more attention to important information and effectively spread the underlying feature information to the high level. Therefore, it can reduce the attenuation of the underlying information and retain more input spatial information.

In order to verify the generalization ability of the proposed model, two polyp segmentation datasets are used as tests, including ETIS and ColonDB. There are 196 images in ETIS and 380 images in ColonDB. As can be seen from Table 3 and Table 4, the Dice score on the ColonDB dataset is 2.55% ahead of the U-Net model. The IoU score on the ETIS dataset is 2.74% ahead of the U-Net model. The results show that the proposed model has strong generalization ability.

Figure 4 and Figure 5 show the visualization results of different segmentation methods in ETIS and ColonDB datasets. From left to right, the input images are MedT, ConvUNeXt, TransUNet, EU-Net, TGANet, Swin-UNet and the proposed model ATFF segmentation results. The red curve is the boundary of the true value of the lesion ground.

Table 1. The segmentation results of Kvasir-Seg dataset.

	Dice	HD	IoU	ACC	REC
MedT	89.02	4.55	83.23	97.21	90.92
ConvUNeXt	89.30	4.46	83.45	97.34	90.34
TransUNet	89.34	4.03	83.79	96.23	90.51
EU-Net	89.39	3.97	83.88	96.45	90.69
TGANet	89.47	3.95	83.91	97.46	91.02
Swin-UNet	89.56	3.44	84.40	97.21	91.34
ATFF	**89.98**	**3.22**	**85.60**	**97.62**	**91.38**

Table 2. The segmentation results of ClinicDB dataset.

	Dice	HD	IoU	ACC	REC
MedT	87.13	5.57	81.23	96.21	89.45
ConvUNeXt	87.35	5.46	81.17	96.54	89.38
TransUNet	87.34	5.34	81.14	96.63	89.43
EU-Net	87.35	5.23	81.25	96.75	89.82
TGANet	87.46	4.67	81.34	97.66	89.31
Swin-UNet	87.46	4.34	82.23	97.71	89.32
ATFF	**87.58**	**4.27**	**83.56**	**97.62**	**91.18**

Table 3. The segmentation results of ColonDB dataset.

	Dice	HD	IoU	ACC	REC
MedT	87.12	4.76	80.23	97.21	88.12
ConvUNeXt	87.32	4.62	80.17	97.34	88.23
TransUNet	88.03	4.56	80.14	97.52	88.51
EU-Net	88.19	4.31	80.25	98.23	88.62
TGANet	88.20	3.53	80.34	97.15	88.54
Swin-UNet	88.56	3.31	80.23	97.25	88.65
ATFF	**88.89**	**3.26**	**81.56**	**98.21**	**89.09**

Table 4. The segmentation results of ETIS dataset.

	Dice	HD	IoU	ACC	REC
MedT	87.13	5.57	82.23	98.07	86.03
ConvUNeXt	87.35	5.46	82.31	98.03	86.27
TransUNet	87.34	5.34	82.65	98.20	86.73
EU-Net	87.35	5.23	82.93	98.33	87.03
TGANet	87.46	4.67	83.75	98.34	87.14
Swin-UNet	87.46	4.34	83.32	98.21	87.23
ATFF	**87.58**	**4.27**	**84.74**	**98.96**	**88.08**

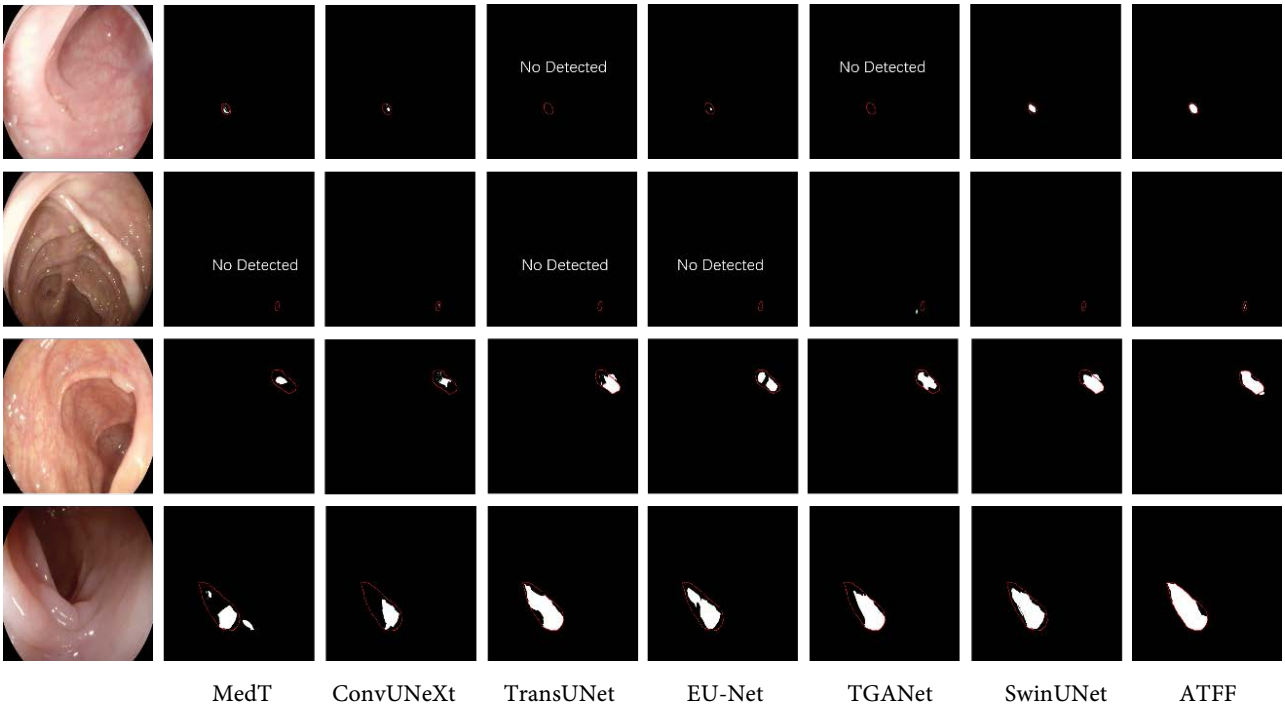

Figure 4. The visual comparison of the proposed model and the state-of-the-art methods on ETIS.

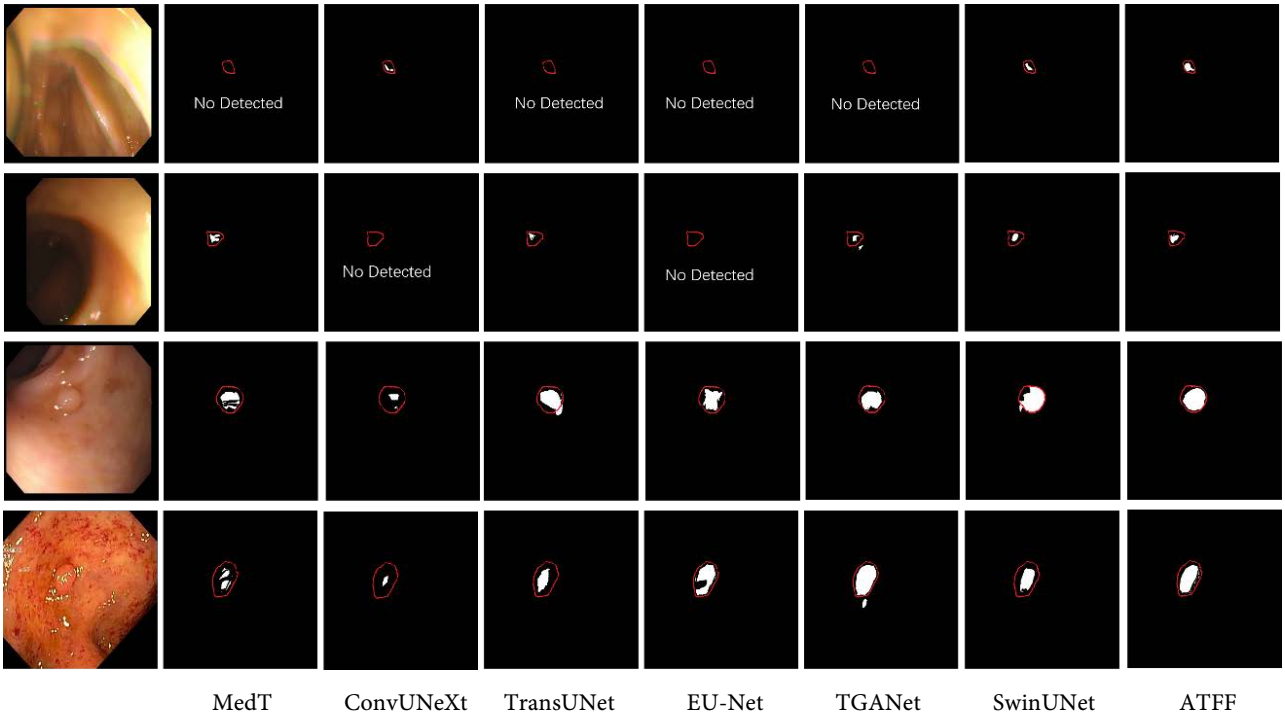

Figure 5. The visual comparison of the proposed model and the state-of-the-art methods on ColonDB.

3.3. Discussion of Experimental Results

It can be seen from **Figure 4** and **Figure 5** that compared with other segmentation results, this method pays more attention to the lesion area than MedT and AMSUNet, suppresses the unimportant feature area, and the segmentation result is more accurate than EU-Net. In the case of little difference between the color pixels of the lesion area and the color pixels of the background area, the model can pay more attention to the small edges than Swin-UNet. In general, ATFF not only effectively alleviates the disturbance of tumor size, surrounding tissues and cascades, but also obtains segmentation results closer to the real ground mask. The comprehensive evaluation results and visual effects show that this method achieves better segmentation results with less missed detection and false detection in polyp lesion segmentation.

4. Conclusion

In short, this paper constructs a new model structure ATFF and proposes a medical image segmentation method based on this model. The convolution operation is combined with the improved attention mechanism to form a U-shaped network to capture global and local feature information, with multi-scale global context fusion function. Furthermore, a deep supervision strategy is introduced to train the model. Eliminate the attenuation of low-level information and retain more input spatial information. In the future, we will draw more knowledge and inspiration from the new deep learning theory and continue to optimize the proposed model and segmentation method. For example, the influence of each

module on the performance of the model is deeply discussed and analyzed to further improve the segmentation accuracy.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Xia, S., Zhu, H., Liu, X., Gong, M., Huang, X., Xu, L., Zhang, H. and Guo, J. (2019) Vessel Segmentation of X-Ray Coronary Angiographic Image Sequence. *IEEE Transactions on Biomedical Engineering*, **67**, 1338-1348. https://doi.org/10.1109/TBME.2019.2936460

[2] Park, S. and Chung, M. (2022) Cardiac Segmentation on CT Images through Shape-Aware Contour Attentions. *Computers in Biology and Medicine*, **147**, Article ID: 105782. https://doi.org/10.1016/j.compbiomed.2022.105782

[3] Huo, Y., Liu, J., Xu, Z., Harrigan, R., Assad, A., Abramson, R. and Landman, B. (2017) Robust Multicontrast MRI Spleen Segmentation for Splenomegaly Using Multi-Atlas Segmentation. *IEEE Transactions on Biomedical Engineering*, **65**, 336-343. https://doi.org/10.1109/TBME.2017.2764752

[4] Ungi, T., Greer, H., Sunderland, K., Wu, V., Baum, Z., Schlenger, C., Oetgen, M., Cleary, K., Aylward, S. and Fichtinger, G. (2020) Automatic Spine Ultrasound Segmentation for Scoliosis Visualization and Measurement. *IEEE Transactions on Biomedical Engineering*, **67**, 3234-3241. https://doi.org/10.1109/TBME.2020.2980540

[5] Cai, Y. and Wang, Y. (2022) MA-Unet: An Improved Version of Unet Based on Multi-Scale and Attention Mechanism for Medical Image Segmentation. 3*rd International Conference on Electronics and Communication*; *Network and Computer Technology*, Volume 12167, 205-211. https://doi.org/10.1117/12.2628519

[6] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Berlin, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[7] Oktay, O., Schlemper, J., Folgoc, L.L., *et al.* (2018) Attention u-net: Learning Where to Look for the Pancreas.

[8] Isensee, F., Jaeger, P.F., Kohl, S.A.A., *et al.* (2021) nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nature Methods*, **18**, 203-211. https://doi.org/10.1038/s41592-020-01008-z

[9] Yap, M.H., Pons, G., Marti, J., *et al.* (2017) Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. *IEEE Journal of Biomedical and Health Informatics*, **22**, 1218-1226. https://doi.org/10.1109/JBHI.2017.2731873

[10] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., *et al.* (2019) Unet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging*, **39**, 1856-1867. https://doi.org/10.1109/TMI.2019.2959609

[11] Oktay, O., Schlemper, J., Folgoc, L.L., *et al.* (2018) Attention U-Net: Learning Where to Look for the Pancreas.

[12] Huang, H., Lin, L., Tong, R., *et al.* (2020) Unet3+: A Full-Scale Connected Unet for

Medical Image Segmentation. *ICASSP* 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 1055-1059. https://doi.org/10.1109/ICASSP40776.2020.9053405

[13] Valanarasu, J.M.J. and Patel, V.M. (2022) Unext: Mlp-Based Rapid Medical Image Segmentation Network. *Medical Image Computing and Computer Assisted Intervention-MICCAI* 2022: 25*th International Conference*, Singapore, 18-22 September 2022, 23-33. https://doi.org/10.1007/978-3-031-16443-9_3

[14] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. and Salakhutdinov, R. (2019) Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 2978-2988. https://doi.org/10.18653/v1/P19-1285

[15] Raghu, M., *et al.* (2021) Do Vision Transformers See like Convolutional Neural Networks? *Advances in Neural Information Processing Systems*, **34**, 12116-12128.

[16] Nguyen, T., Raghu, M. and Kornblith, S. (2020) Do Wide and Deep Networks Learn the Same Things? Uncovering How Neural Network Representations Vary with Width and Depth.

[17] Oktay, O., Schlemper, J., Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y. and Kainz, B. (2019) Attention U-Net: Learning Where to Look for the Pancreas. *Medical Image Analysis*, **53**, 197-207. https://doi.org/10.1016/j.media.2019.01.012

[18] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. and Zhou, Y. (2021) TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation.

[19] Zhou, H.Y., Guo, J., Zhang, Y., *et al.* (2021) nnformer: Interleaved Transformer for Volumetric Segmentation.

[20] Huang, X., Deng, Z., Li, D., *et al.* (2021) Missformer: An Effective Medical Image Segmentation Transformer.

[21] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. arXiv: 1706.03762.

[22] Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A. and Chen, L. (2020) Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. *European Conference on Computer Vision*, Glasgow, 23-28 August 2020, 108-126. https://doi.org/10.1007/978-3-030-58548-8_7

[23] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., *et al.* (2022) Swin Transformer v2: Scaling up Capacity and Resolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 12009-12019. https://doi.org/10.1109/CVPR52688.2022.01170

[24] Liu, Z., Hu, H., Lin, Y., *et al.* (2022) Swin Transformer v2: Scaling up Capacity and Resolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 12009-12019. https://doi.org/10.1109/CVPR52688.2022.01170

[25] Vazquez, D., Bernal, J., Sanchez, F.J., Fernandez-Esparrach, G., Lopez, A.M., Romero, A., Drozdzal, M. and Courville, A. (2017) A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *Journal of Healthcare Engineering*, **2017**, Article ID: 4037190. https://doi.org/10.1155/2017/4037190

[26] Jha, D., Smedsrud, P.H., Riegler, M.A., *et al.* (2020) Kvasir-seg: A Segmented Polyp Dataset. *MultiMedia Modeling*: 26*th International Conference*, *MMM* 2020,

Daejeon, 5-8 January 2020, 451-462. https://doi.org/10.1007/978-3-030-37734-2_37

[27] Bernal, J., Sanchez, F.J., Fernandez-Esparrach, G., Gil, D., Rodrıguez, C. and Vilarino, F. (2015) Wm-dova Maps for Accurate Polyp Highlighting in Colonoscopy: Validation vs. Saliency Maps from Physicians. *Computerized Medical Imaging and Graphics*, **43**, 99-111. https://doi.org/10.1016/j.compmedimag.2015.02.007

[28] Tajbakhsh, N., Gurudu, S.R. and Liang, J. (2015) Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE TMI*, **35**, 630-644. https://doi.org/10.1109/TMI.2015.2487997

[29] Silva, J., Histace, A., Romain, O., Dray, X. and Granado, B. (2014) Toward Embedded Detection of Polyps in WCE Images for Early Diagnosis of Colorectal Cancer. *The International Journal of Computer Assisted Radiology and Surgery*, **9**, 283-293. https://doi.org/10.1007/s11548-013-0926-3

[30] Han, Z., Jian, M. and Wang, G. (2022) ConvUNeXt: An Efficient Convolution Neural Network for Medical Image Segmentation. *Knowledge-Based Systems*, **253**, Article ID: 109512. https://doi.org/10.1016/j.knosys.2022.109512

[31] Fu, Z., Li, J. and Hua, Z. (2022) DEAU-Net: Attention Networks Based on Dual Encoder for Medical Image Segmentation. *Computers in Biology and Medicine*, **150**, Article ID: 106197. https://doi.org/10.1016/j.compbiomed.2022.106197

[32] Valanarasu, J., Oza, P., Hacihaliloglu, I. and Patel, V. (2021) Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Strasbourg, 27 September-1 October 2021, 36-46. https://doi.org/10.1007/978-3-030-87193-2_4

[33] Tomar, N., Jha, D., Bagci, U. and Ali, S. (2022) TGANet: Text-Guided Attention for Improved Polyp Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*), Singapore, 18-22 September 2022, 151-160. https://doi.org/10.1007/978-3-031-16437-8_15

[34] Cao, H., Karlinsky, L., Michaeli, T., Nishino, K., *et al.* (2023) Swin-Unet: Unet-Like Pure Trans-Former for Medical Image Segmentation. *European Conference on Computer Vision Computer Vision*, Tel Aviv, 23-27 October 2022, 205-218. https://doi.org/10.1007/978-3-031-25066-8_9