Journal of Computer and Communications, 2026, 14(6), 55-70
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Novel Polyp Segmentation Method Based on the Vision Transformer and Attention Mechanism

Xinping Guo, Yongqi Nie, Xiuzhu Jia, Mengying Lou, Zhiyuan Li, Xiaoyu Han, Lu Yu

School of Medical Imaging, Qilu Medical University, Zibo, China
Email: 450641388@qq.com

How to cite this paper: Guo, X.P., Nie, Y.Q., Jia, X.Z., Lou, M.Y., Li, Z.Y., Han, X.Y. and Yu, L. (2026) A Novel Polyp Segmentation Method Based on the Vision Transformer and Attention Mechanism. *Journal of Computer and Communications*, **14**, 55-70.
https://doi.org/10.4236/jcc.2026.146005

Received: May 19, 2026
Accepted: June 20, 2026
Published: June 23, 2026

Abstract

The accurate segmentation of the polyp is very important for the diagnosis and treatment plans of colorectal cancer. Although the UNet model and models with the U-shaped structure have achieved great success in polyp image segmentation, they are still limited by the colors, sizes, and shapes of polyps, as well as the low contrast, various noise, and blurred edges of the colonoscopy, which can easily result in a large amount of redundant information, weak complementarity between different levels of features, and inaccurate polyp localization. To deal with the special characteristics of the polyp images and improve the segmentation performance, a new segmentation model named VTANet, which is based on the pyramid vision transformer and BAM (Bottleneck Attention Module), is developed. The proposed model consists of four modules: the pyramid vision transformer (PVT) encoder, the Feature Aggregation Module (FAM), the Adaptive Attention Fusion Module (AAFM), and the Aggregation Similarity Module (ASM). The PVT learns a more robust representation model; the FAM enhances the complementarity between features by cascading the encoder features and acquiring richer context and fine-grained features. The AAFM makes polyp localization more accurate by introducing the BAM attention module to obtain richer details of the polyps. To verify effectiveness and accuracy, experiments on five popularly used datasets are carefully designed and implemented. The proposed VTANet achieves competitive and generally superior performance across five public datasets. Although it does not obtain the best score on every metric, especially on several MAE or E-measure results, it consistently improves the main overlap-based metrics such as mDice and mIoU on most datasets. This indicates that VTANet provides a favorable balance between region-level accuracy, boundary preservation, and generalization ability.

Keywords

Polyp Segmentation, UNet Model, The Attention Mechanism, The Pyramid Vision Transformer

1. Introduction

As shown in Figure 1, colorectal cancer is the third most prevalent and second most lethal cancer in the world. According to statistical reports, colorectal cancer also shows an increasing trend in prevalence and mortality rate in China, accounting for almost 23.7% of the 4.57 million new cancer cases each year [1]. Polyps, which grow abnormally in the colon and rectum over time, are the main cause of colorectal cancer. When cells in the colon or rectum grow out of control, it is easy for cancer to develop, which can even lead to death. Thus, the ability to quickly and accurately detect the location of polyps and provide treatments, such as colonoscopy and resection operation at an early stage, is very important for the health of patients.

The precise localization and extraction of the polyps are crucial steps to make the diagnosis and treatment plans. Medical image segmentation provides a strong and helpful tool for doctors to carefully observe the lesion and accurately implement the operation [2]. However, as shown in Figure 2, the segmentation of the polyp is a challenging task. Firstly, even if they are of the same type, the size, color, and texture are different. Secondly, due to the reflection of the intestinal mucus and polyps under colonoscopy, the contrast between polyps and the surrounding mucosa is not strong enough, and the boundary is not very clear. The two above reasons may cause missed detection and inaccurate segmentation of polyps. Thus, an accurate segmentation method for potential polyps in the early stage is of great significance for preventing colorectal cancer.

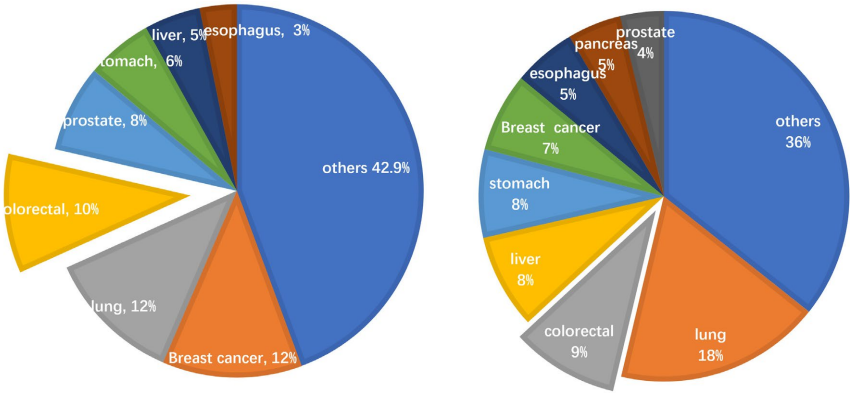

Figure 1. The prevalence and mortality of colorectal cancer.

In traditional medical image segmentation, the early conventional colon polyp image segmentation algorithm mainly analyzes the characteristics of colon polyps. Mamonov *et al.* designed a binary classifier to mark each frame of an image as

containing or not containing polyps according to the geometric analysis and texture content on each edge of a colonoscopy video [3]. Bernal *et al.* obtained the polyp boundary information using a Window Median Depth of Valleys Accumulation (WM-DOVA) energy map, and detected intestinal polyps through polyp texture to complete the detection of the polyp area [4]. Sasmal *et al.* proposed a detection method based on the shape of polyps, mainly using the principal component analysis (PCA) method and the region-based active contour model to complete the segmentation [5]. For these segmentation algorithms, they heavily relied on the manual extraction of features, and the polyps and their surrounding tissues are distinguished by the training classifiers. The expressive ability of the manually extracted features is quite limited. Although the traditional algorithms are relatively simple in implementation, they cannot consider the effective features of the polyp area at the same time and cannot combine these features. Therefore, the segmentation results of them are not satisfying.

Recently, it has been reported that, compared with traditional segmentation methods, deep learning-based segmentation methods perform better. The main principle of the deep learning-based colon polyp image segmentation algorithms is to design a convolutional neural network model, use colon polyp images and labels to train the model, and then use the trained model for segmentation. The typical models include CNN, GAN, and the UNet model [6]. For example, Ronneberger *et al.* proposed a fully symmetric UNet network with an encoder-decoder structure [7]. The UNet network uses skip connections between the encoder and the decoder for feature fusion, which performs well in cell segmentation tasks. Inspired by the successful application of the UNet network in biomedical image segmentation, more and more related works on the UNet model and its variant structures are used to segment polyps. For example, Zhang *et al.* proposed a U-shaped network ResUNet with a deep residual system [8]. The residual connection is introduced into each convolution module of the UNet to obtain deeper image features, thereby improving the accuracy of segmentation results. Zhou *et al.* proposed the UNet++ model by reducing the depth of the unknown network; it redesigns the jump connection and designs a scheme to prune the network to improve the performance of UNet [9]. Fan *et al.* proposed a parallel reverse attention network, PraNet, for accurate segmentation of polyps [10]. Jha *et al.* proposed a double UNet network. By cascading two variants of the UNet structure to form a dual-network structure, the entire network has more robust feature extraction capabilities and a larger receptive field [11].

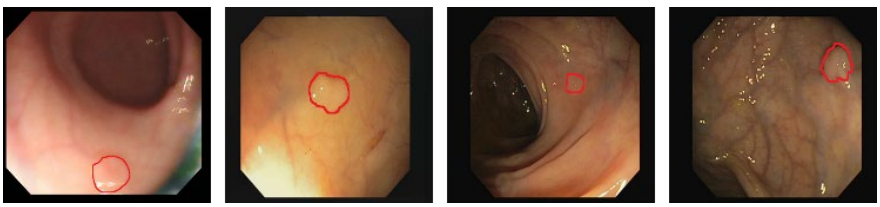

Figure 2. Examples of the colorectal polyp in endoscopic images. In each image, the red line indicates the polyp contour. The polyp varies in morphology, size, and brightness.

Though good results have been obtained by the above-mentioned deep learning-based segmentation algorithms, there is still much room for improvement to accurately segment the polyps due to the special characteristics of the polyps: (1) the colors of polyps and surrounding tissues are extremely similar; (2) a diversity of sizes, shapes, and textures of the polyps; (3) some polyps may be hidden in the folds of the colon.

In order to address the special characteristics of polyp segmentation, a new segmentation model named VTANet, which is based on the pyramid vision transformer and BAM attention, is proposed in this paper. Experiments on five public polyp image datasets demonstrate that the proposed model greatly improves polyp image segmentation performance. Compared with existing transformer-based or attention-based polyp segmentation methods, VTANet is not a simple replacement of the CNN backbone with a transformer encoder. Instead, it integrates PVTv2, BAM, FAM, AAFM, and ASM into a unified segmentation framework. PVTv2 is used to extract multi-scale global contextual representations with a relatively low computational cost. FAM aggregates high-level encoder features to enhance semantic consistency and lesion localization. AAFM introduces BAM-based channel-spatial attention to strengthen low-level boundary, texture, and color representations. ASM further models the similarity relationship between low-level detail features and high-level semantic cues, allowing detailed appearance information to be injected into global semantic representations. Therefore, the novelty of VTANet lies in the cooperative design of transformer-based hierarchical encoding, attention-guided low-level feature refinement, high-level feature aggregation, and similarity-based feature interaction for accurate polyp segmentation.

2. The Whole Method

2.1. The Architecture of the VTANet

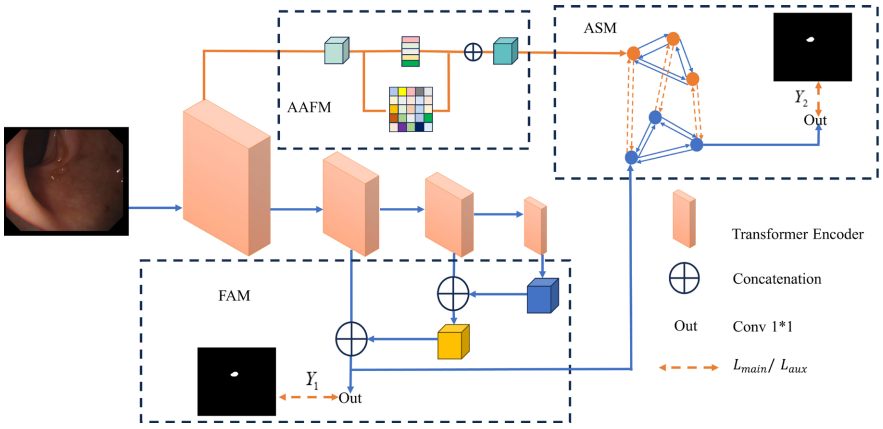

Figure 3. The architecture of the VTANet model.

As shown in **Figure 3**, the proposed VTANet model consists of four key modules: the pyramid vision transformer (PVT) encoder, the Feature Aggregation Module

(FAM), the Attention Fusion Module (AAFM), and the Aggregation Similarity Module (ASM). PVT is used to obtain the long-range dependency features from the encoder. FAM aggregates the high-level features to obtain the semantic and location information of polyps. AAFM removes noise and enhances the low-level polyp representation information, including texture, color, and edge. ASM combines the low-level and high-level features provided by AAFM and FAM, effectively transmitting information to the whole region.

2.2. The Transformer Encoder

Due to uncontrollable factors in the collection of polyp images, they tend to contain significant noise, such as motion blur, rotation, and reflection. Some recent work has reported that the vision transformer shows more robust performance and better robustness than CNNs [12]-[14]. Inspired by these ideas, the vision transformer is used as the backbone network to extract more robust and powerful polyp information. Different from the fixed columnar structure or shift window, the PVT is a pyramid architecture whose representation is calculated with spatial-reduction attention operations. Thus, it can reduce resource consumption. Specifically, the encoder part uses an improved version of PVT, namely PVTV2, with more powerful feature extraction capabilities. In order to make PVTV2 suitable for the segmentation task of polyps, the last classification layer is deleted, and four multi-scale feature maps (x1, x2, x3, x4) are generated at different stages. In these feature maps, x1 provides polyp information in the underlying features; x2, x3, x4 provide the advanced features of polyps, such as semantic and location information.

2.3. Feature Aggregation Module

The primary purpose of the FAM module is to combine high-level features from the encoder into a better feature representation to improve performance. It captures different information through different features to enhance the robustness of the model and reduce overfitting. Specifically, we define $F(\cdot)$ as a convolutional unit composed of a 3 × 3 convolutional layer with padding set to 1. Batch normalization and ReLU [15] are used. We define $G(\cdot)$ as a convolutional unit composed of a 1 × 1 convolutional layer with padding set to 1 and ReLU. Firstly, the highest-level feature map X_4 is up-sampled, and the up-sampled results are passed through the convolution unit $F_1(\cdot)$ to obtain X_{4-t}. Then, the obtained result X_{4-t} is spliced with the feature mapping X_3 in the encoder to generate the fusion feature X_{3-a}. The result obtained by X_{3-a} through the convolution unit $G_1(\cdot)$ is up-sampled, and the up-sampled result is passed through the convolution unit $F_2(\cdot)$ to obtain X_{3-t}. Then, the obtained result X_{3-t} is spliced with the feature map X_2 in the encoder to generate the fusion feature X_{2-a}. Finally, the obtained feature fusion X_{2-a} obtains the final output feature famfeature of the module through the convolution unit G2(·). The process is described by the following equations.

$$a = G_1\left(X_{3-a}\right)\left\{Concat\left\{X_3, up\left\{F_1\left\{X_4\right\}\right\}\right\}\right\}. \tag{1}$$

$$feature = G_2\left(X_{2-a}\right)\left\{Concat\left\{X_2, up\left\{F_2\left(a\right)\right\}\right\}\right\}. \tag{2}$$

2.4. The Adaptive Attention Fusion Module

The low-level features usually contain rich details, such as the texture, color, and edge of polyps. However, polyps are often very similar in appearance to the background. Therefore, a powerful extractor is needed to identify the details of polyps. As shown in **Figure 4**, an adaptive attention fusion module is introduced to capture the details of polyps from different dimensions of the low-level feature map X_1. Precisely, the adaptive attention fusion module consists of the channel attention operation Attc(·) and the spatial attention operation Atts(·) [16]. Firstly, the feature map X_1 generated by the encoder is encoded into a one-dimensional feature vector through global average pooling so that each channel has a global receptive field; then, the fully connected layer is used to reduce the dimension of the one-dimensional feature vector, and the ReLU activation function is used for nonlinear processing. Then the fully connected layer is used to increase the dimension. Finally, the corresponding weight $M_c\left(X_1\right)$ is obtained by batch normalization. In summary, the calculation formula for the channel attention operation is described by Equation (3):

$$M_c\left(X_1\right) = BN\left\{W_1\left\{W_0 AvgPool\left(X_1\right)\right\}\right\}. \tag{3}$$

The spatial attention operation process in parallel with the channel attention is as follows: Firstly, the X_1 feature map is reduced by a 11 convolution, and then the feature information is extracted by two dilated convolutions with a convolution kernel size of 33. The dilated convolution has a larger receptive field. Finally, the feature map is mapped to $1 \times W \times H$ by a 1 * 1 convolution, and the spatial attention map $M_s\left(X_1\right)$ is obtained. The calculation formula is described by Equation (4):

$$M_s\left(X_1\right) = BN\left\{f_3^{11}\left\{f_2^{33}\left\{f_1^{33}\left\{f_0^{11}\left\{f\right\}\right\}\right\}\right\}\right\}. \tag{4}$$

When the channel attention and spatial attention are fused, the $M_c\left(X_1\right)$ and $M_s\left(X_1\right)$ are extended to the same latitude through the broadcast mechanism. Then, the weights are added to obtain the attention vector $M\left(X_1\right)$. Finally, the input feature graph X_1 is multiplied by $M\left(X_1\right)$ element-wise and then added to X_1 through the residual structure. The formula is as follows:

$$X_1' = X_1 + X_1 \times \sigma\left(M_c\left(X_1\right) + M_s\left(U\right)\right). \tag{5}$$

2.5. The Aggregation Similarity Module

The non-local operation is introduced into the graph convolution domain to implement the aggregate similarity module, which explores the relationship between the low-level local features from AAFM and the high-level cues from FAM. Therefore, ASM can use global attention to inject detailed appearance features into high-

level semantic features. Given a feature map Y_1 containing high-level semantic information and Y_2 with rich appearance details, they are merged through self-attention. Firstly, the summation of the two linear mapping functions $W_\theta(\cdot)$ and $W_\Phi(\cdot)$ is applied on Y_1. The dimension of Y_1 is reduced to obtain the feature mapping $Q \in R^{\frac{H}{8}\times\frac{W}{8}\times 16}$ and $K \in R^{\frac{H}{8}\times\frac{W}{8}\times 16}$. Then, the convolution operation with a kernel size of 1 * 1 is used as the linear mapping process [17]. The process can be expressed as:

$$Q = W_\theta(Y_1), K = W_\Phi(Y_1). \tag{6}$$

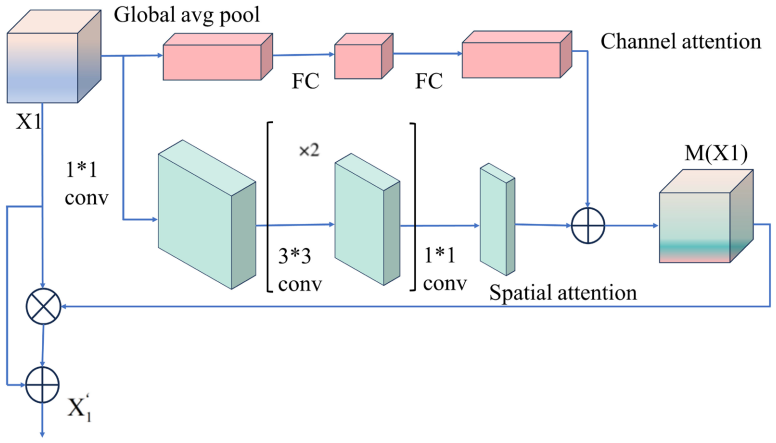

Figure 4. The network structure of BAM.

For Y_2, we use the convolution unit W_g to reduce the channel dimension to 32 and interpolate it to the same size as Y_1. Then, the softmax function is applied to the channel dimension, and the second channel is selected as the attention mapping to obtain $T_2' \in R^{\frac{H}{8}\times\frac{W}{8}\times 1}$. Next, calculate the Hadamard product between K and T_2'. This operation assigns different weights to the pixels, thereby increasing the weight of the edge pixels. After that, the adaptive pooling operation is used to reduce the displacement of the feature, and center clipping is applied to obtain the feature map $V \in R^{4\times 4\times 16}$. The process can be expressed as follows:

$$V = AP\left(K \times F\left(W_g\left(Y_2\right)\right)\right). \tag{7}$$

where AP represents pooling and clipping operations.

3. Experimental Results and Discussion

3.1. The Experimental Setting

To evaluate the proposed method, five public polyp datasets, *i.e.*, the Kvasir-SEG [18], ClinicDB [19], ColonDB [20], Endoscene [21], and ETIS [22], are used. Specifically, the ClinicDB and Kvasir-SEG datasets are used to assess the learning ability of the model. The ClinicDB contains 612 images that are extracted from colonoscopy videos. Kvasir-SEG includes 1000 polyp images. In the experiment, the same 548 and

900 images in the ClinicDB and Kvasir-SEG datasets are used as the training sets, and the remaining 64 and 100 images are used as the corresponding testing sets.

During training, the official training images from Kvasir-SEG and ClinicDB were used. A validation subset was separated from the training set to monitor model convergence and select the best checkpoint. The checkpoint with the highest validation mDice was used for final testing. All experiments were conducted with a fixed random seed to reduce the influence of random initialization and data shuffling. The PVTv2 encoder was initialized with ImageNet-pretrained weights, while the newly added FAM, AAFM, ASM, and prediction layers were randomly initialized. The multi-scale training strategy refers to random resizing of the input images within a predefined scale range before cropping or resizing them to 352 × 352, so that the model can learn scale-robust representations for polyps of different sizes.

All the experiments are implemented using the PyTorch framework. Considering the difference in the size of each polyp image, a multi-scale strategy is used in training. In addition, the AdamW optimizer is used to update the network parameters, which is widely used in transformer networks [23] [24]. The learning rate is set to 1e−4, and the weight decay 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.

The training process uses two loss functions to optimize the output model, which can be expressed by the following formula:

$$L = L_{main} + L_{aux}. \tag{8}$$

where L_{main} and L_{aux} are the primary and auxiliary loss functions, respectively.

The main loss function calculates the loss between the final segmentation result and the ground truth. The formula can be written as:

$$L_{main} = L_{IoU}^{w}\left(Y_2, G\right) + L_{BCE}^{w}\left(Y_2, G\right). \tag{9}$$

The auxiliary loss function calculates the loss between the intermediate result from FAM and the ground truth. The formula can be written as:

$$L_{aux} = L_{IoU}^{w}\left(Y_1, G\right) + L_{BCE}^{w}\left(Y_1, G\right). \tag{10}$$

where L_{IoU}^{w} and L_{BCE}^{w} are the weighted intersection over union (IoU) loss and weighted binary cross-entropy (BCE) loss.

The prediction graph is limited in terms of global structure (object level) and local detail (pixel level) perspectives, which is different from the standard BCE loss function (treating all pixels equally).

Six popularly used evaluation indices, including the Dice, IoU, mean absolute error (MAE), weighted F-measure (F_{β}^{w}), S-measure (Sα) [25], and E-measure (E_{ε}) [26], are adopted to evaluate the performances. The Dice and IoU are region-level similarity measures that mainly focus on the internal consistency of segmented objects. We use the average values of Dice and IoU, denoted as mDice and mIoU, respectively. MAE measures the difference between the model predic-

tion results and the actual labels. The weighted F-measure (F_β^w) comprehensively considers the recall and precision, MAE measures the average pixel-wise absolute error between the normalized saliency prediction map and the binary ground-truth mask, and S-measure evaluates the structural similarity between the real-valued saliency map and the binary ground-truth. It considers object-aware and region-aware structure similarities. E-measure considers the global means of the image and local pixel matching simultaneously [27].

3.2. Experimental Results

In order to verify the effectiveness and robustness of the proposed model, 7 famous network models are compared, namely: UNet [7], UNet++ [9], MSEG [28], ACSNet [29], PraNet [10], EU-Net [30], and SANet [31].

As can be seen from **Table 1**, the mDice, mIoU, Sα, m, MAE scores of the proposed model on the ETIS dataset are higher than those of UNet by 2.89%, 2.74%, 2.7%, 1.09%, 1.64%, and 0.3%, respectively. In addition, it can be seen from **Tables 2-5** that the six evaluation metrics also achieve good results on the other four datasets. The combined results show that the model has better learning ability.

All comparison methods were retrained using the same data preprocessing, input resolution, training schedule, optimizer, and evaluation metrics.

Table 1. The segmentation results of endoscene dataset.

	mDice	*mIoU*	F_β^w	S_α	mE_ε	*MAE*
UNet	0.710	0.627	0.684	0.843	0.847	0.022
UNet++	0.707	0.624	0.687	0.839	0.834	0.018
MSEG	0.874	0.804	0.852	0.924	0.948	0.009
ASCNet	0.863	0.787	0.825	0.923	0.939	0.013
PraNet	0.871	0.797	0.843	0.925	0.950	0.010
SANet	0.837	0.765	0.805	0.904	0.919	0.015
EU-Net	0.888	0.815	0.859	0.928	0.962	0.008
VTANet	**0.904**	**0.826**	**0.872**	**0.941**	**0.978**	**0.009**

Table 2. The segmentation results of kvasir-seg dataset.

	mDice	*mIoU*	F_β^w	S_α	mE_ε	*MAE*
UNet	0.818	0.746	0.794	0.858	0.881	0.055
UNet++	0.821	0.743	0.808	0.862	0.886	0.048
MSEG	0.897	0.839	0.885	0.912	0.942	0.028
ASCNet	0.898	0.838	0.882	0.920	0.941	0.032
PraNet	0.898	0.840	0.885	0.915	0.944	0.030
SANet	0.904	0.847	0.892	0.915	0.949	0.027
EU-Net	0.908	0.854	0.893	0.917	0.951	0.028
VTANet	**0.921**	**0.865**	**0.912**	**0.923**	**0.956**	**0.023**

Table 3. The segmentation results of the clinicdb dataset.

	$mDice$	$mIoU$	F_β^w	S_α	mE_ε	MAE
UNet	0.823	0.755	0.811	0.889	0.913	0.019
UNet++	0.794	0.729	0.785	0.873	0.891	0.022
MSEG	0.909	0.864	0.907	0.938	0.961	0.007
ASCNet	0.882	0.826	0.873	0.927	0.947	0.011
PraNet	0.899	0.849	0.896	0.936	0.979	0.009
SANet	0.912	0.856	0.907	0.929	0.968	0.012
EU-Net	0.902	0.846	0.891	0.936	0.959	0.011
VTANet	**0.916**	**0.867**	**0.916**	**0.943**	**0.972**	**0.010**

Table 4. The segmentation results of ColonDB dataset.

	$mDice$	$mIoU$	F_β^w	S_α	mE_ε	MAE
UNet	0.512	0.432	0.498	0.713	0.696	0.061
UNet++	0.483	0.410	0.467	0.691	0.680	0.064
MSEG	0.735	0.666	0.724	0.834	0.859	0.038
ASCNet	0.716	0.649	0.697	0.829	0.839	0.039
PraNet	0.712	0.640	0.699	0.820	0.847	0.043
SANet	0.753	0.670	0.726	0.837	0.869	0.043
EU-Net	0.756	0.681	0.730	0.831	0.863	0.045
VTANet	**0.767**	**0.694**	**0.743**	**0.856**	**0.876**	**0.041**

Table 5. The segmentation results of etis dataset.

	$mDice$	$mIoU$	F_β^w	S_α	mE_ε	MAE
UNet	0.398	0.335	0.366	0.684	0.643	0.036
UNet++	0.401	0.344	0.390	0.683	0.629	0.035
MSEG	0.700	0.630	0.671	0.828	0.854	0.015
ASCNet	0.578	0.509	0.530	0.754	0.737	0.059
PraNet	0.628	0.567	0.600	0.794	0.808	0.031
SANet	0.687	0.609	0.636	0.793	0.807	0.067
EU-Net	0.750	0.654	0.685	0.849	0.881	0.015
VTANet	**0.763**	**0.669**	**0.693**	**0.855**	**0.884**	**0.038**

Table 6. The ablation results of the etis dataset.

	$mDice$	$mIoU$	F_β^w	S_α	mE_ε	MAE
PVT	0.712	0.623	0.609	0.821	0.807	0.046
PVT + AAFM	0.734	0.624	0.687	0.839	0.834	0.038
PVT + FAM	0.754	0.604	0.652	0.824	0.848	0.049
PVT + ASM	0.715	0.657	0.625	0.823	0.839	0.043
VTANet	**0.763**	**0.669**	**0.693**	**0.855**	**0.884**	**0.038**

Figure 5 and Figure 6 show the visualization results of different segmentation methods on the two datasets, ClinicDB and ColonDB. Figure 7 shows the visualization results of different segmentation methods on the other datasets, Kvasir-seg and ENDOSCENE. From left to right, the segmentation results are obtained by UNet, UNet++, MSEG, ASCNet, PraNet, SANet, EU-Net, and the proposed model, respectively. The red curve is the boundary of the actual value of the lesion ground [32] [33]. It can be seen from Figure 4 and Figure 5 that compared with other segmentation results, the proposed method pays more attention to the lesion area than UNet and UNet++, suppresses the unimportant feature area, and the segmentation result is more accurate than UNet. With 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 trim edges than PraNet. In general, VTANet 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 and visual effects show that the proposed method achieves better segmentation results with less missed and false detection in polyp lesion segmentation.

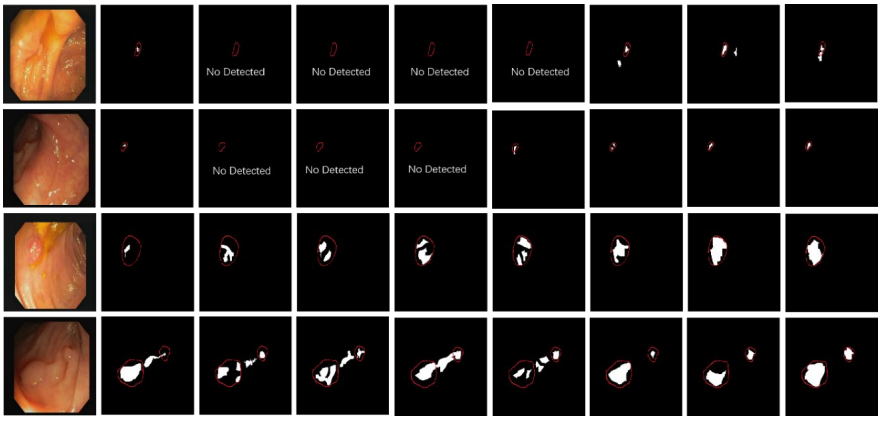

Figure 5. The visual comparison of the proposed model and the state-of-the-art methods on ClinicDB.

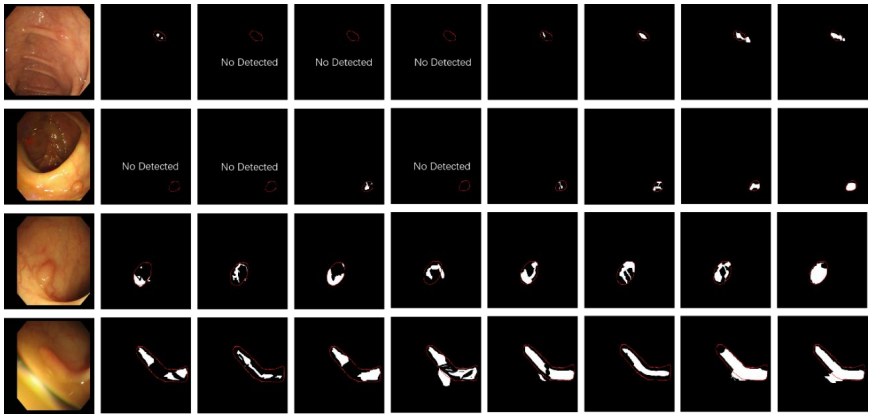

Figure 6. The visual comparison of the proposed model and the state-of-the-art methods on ColonDB.

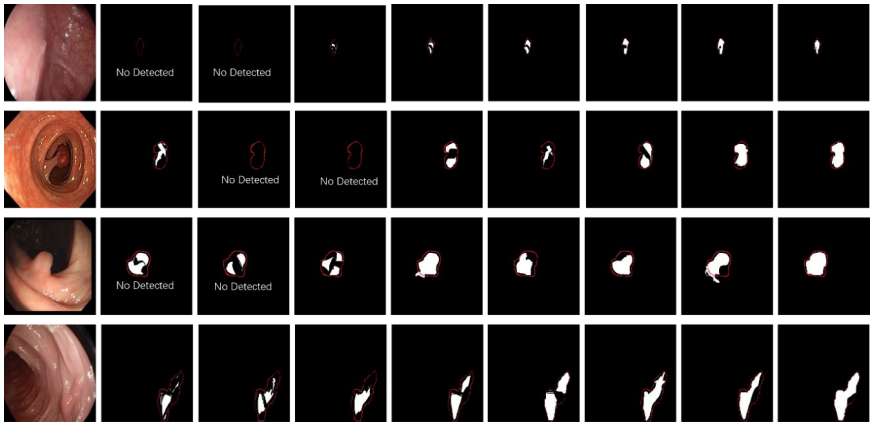

Figure 7. The visual comparison of the proposed model and the state-of-the-art methods on Kvasir-seg, Endoscene, and ETIS.

To verify the generalization ability of the proposed model, three polyp segmentation datasets, including ETIS, ColonDB, and EndoScene, are used for testing. There are 196 images in ETIS, 380 in ColonDB, and 60 in Endoscene, respectively. It can be seen from **Table 1**, **Tables 4-5** that the mDice score on the ColonDB dataset is 2.55% higher than that of the UNet model. The mIoU score on the ETIS dataset is 2.74% higher than that of the U-Net model. The score on the Endoscene dataset is 2.55% higher than that of the UNet model [34]-[37]. The results show that the proposed model has strong generalization ability.

Finally, the effectiveness of each component in the overall model is described in detail, and the settings for training, testing, and hyperparameters are consistent with the previous ones. We use PVTv2 as a baseline and verify the performance of the proposed model by removing modules from the proposed model. The experimental results of different modules in **Table 6** show that these modules have played a role in improving network performance. It can be seen from **Table 6** that after the introduction of the FAM module, the mDice score is 0.42% higher than that of the original basic PVTv2 network. The introduction of the AAFM module also improves the performance of the original PVT network [38] [39]. The ablation results show that each module contributes to VTANet from different perspectives, but the improvement is not uniform across all metrics. Compared with the PVT baseline, AAFM improves mDice and weighted F-measure, indicating that BAM-based channel-spatial attention is helpful for enhancing low-level texture and boundary-related features. However, its limited improvement in mIoU suggests that low-level attention alone is insufficient to fully improve region-level consistency. FAM increases mDice by aggregating high-level semantic features, which helps locate polyp regions more accurately, but the decrease in mIoU and MAE indicates that feature aggregation alone may also introduce coarse responses or redundant semantic information. ASM brings a more obvious improvement in mIoU, suggesting that similarity-based interaction between low-level details and high-level semantics helps refine the global structure of the predicted region. The full VTANet achieves the best overall balance among mDice, mIoU, S-measure,

and E-measure, demonstrating that these modules are complementary rather than independently optimal.

4. Conclusions

Although VTANet achieves competitive results on five public polyp datasets, several limitations remain. First, the model may still fail when polyps have extremely low contrast, very small size, severe motion blur, specular reflection, or ambiguous boundaries close to the surrounding mucosa. Second, the training and testing datasets are all public benchmark datasets, and the generalization ability of VTANet on multi-center clinical data, different endoscopy devices, and real-time video sequences still needs further validation. Third, the combination of PVTv2, FAM, AAFM, and ASM improves segmentation accuracy but also increases model complexity compared with simpler CNN-based models. Therefore, there is still a practical trade-off between segmentation accuracy and computational efficiency. Future work will focus on lightweight deployment, external clinical validation, and real-time video polyp segmentation.

Experimental results on five public datasets show that VTANet achieves competitive performance compared with several established segmentation models. The improvement is especially evident in mDice and mIoU, while some metrics on specific datasets remain slightly lower than those of the best competing methods. These results suggest that the proposed feature aggregation and attention fusion strategy is effective, but further optimization is still needed for all-metric superiority.

Acknowledgements

This study was supported by project ZR2021MF017 supported by the Shandong Provincial Natural Science Foundation; project ZR2020MF147 supported by the Shandong Provincial Natural Science Foundation; and the National Natural Science Foundation of China (62273155).

Conflicts of Interest

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

References

[1] Shi, J.-H., Zhang, Q., Tang, Y.-H. and Zhang, Z.-Q. (2022) Polyp-Mixer: An Efficient Context-Aware MLP-Based Paradigm for Polyp Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, **33**, 30-42. https://doi.org/10.1109/tcsvt.2022.3197643

[2] Yu, T. and Wu, Q. (2023) HarDNet-CPS: Colorectal Polyp Segmentation Based on Harmonic Densely United Network. *Biomedical Signal Processing and Control*, **85**, Article 104953. https://doi.org/10.1016/j.bspc.2023.104953

[3] Mamonov, A.V., Figueiredo, I.N., Figueiredo, P.N. and Richard Tsai, Y. (2014) Automated Polyp Detection in Colon Capsule Endoscopy. *IEEE Transactions on Medical Imaging*, **33**, 1488-1502. https://doi.org/10.1109/tmi.2014.2314959

[4] Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C. and Vilariño, 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

[5] Sasmal, P., Iwahori, Y., Bhuyan, M.K. and Kasugai, K. (2018) Active Contour Segmentation of Polyps in Capsule Endoscopic Images. 2018 *International Conference on Signals and Systems* (*ICSigSys*), Bali, 1-3 May 2018, 201-204. https://doi.org/10.1109/icsigsys.2018.8372666

[6] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/cvpr.2015.7298965

[7] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W. and Frangi, A., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[8] Zhang, Z., Liu, Q. and Wang, Y. (2018) Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, **15**, 749-753. https://doi.org/10.1109/lgrs.2018.2802944

[9] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N. and Liang, J. (2020) 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

[10] Fan, D., Ji, G., Zhou, T., Chen, G., Fu, H., Shen, J., *et al*. (2020) PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In: Martel, A.L., *et al*., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 263-273. https://doi.org/10.1007/978-3-030-59725-2_26

[11] Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P. and Johansen, H.D. (2020) DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. 2020 *IEEE* 33*rd International Symposium on Computer-Based Medical Systems* (CBMS), Rochester, 28-30 July 2020, 558-564. https://doi.org/10.1109/cbms49503.2020.00111

[12] Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., *et al*. (2021) Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 568-578. https://doi.org/10.1109/iccv48922.2021.00061

[13] Gan, C., Li, Y., Li, H., Sun, C. and Gong, B. (2017) VQS: Linking Segmentations to Questions and Answers for Supervised Attention in VQA and Question-Focused Semantic Segmentation. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 1811-1820. https://doi.org/10.1109/iccv.2017.201

[14] Ioffe, S. and Szegedy, C. (2015) Batch normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the* 32*nd International Conference on Machine Learning*, Lille, 6-11 July 2015, 448-456.

[15] Glorot, X., Bordes, A. and Bengio, Y. (2011) Deep Sparse Rectifier Neural Networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, **15**, 315-323.

[16] Park, J., Woo, S., Lee, J.-Y. and Kweon, I.S. (2018) Bam: Bottleneck Attention Module. arXiv:1807.06514.

[17] Su, Y., Cheng, J., Zhong, C., Zhang, Y., Ye, J., He, J., *et al*. (2023) FeDNet: Feature Decoupled Network for Polyp Segmentation from Endoscopy Images. *Biomedical Signal Processing and Control*, **83**, Article 104699. https://doi.org/10.1016/j.bspc.2023.104699

[18] Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., *et al*. (2019) Kvasir-SEG: A Segmented Polyp Dataset. In: Ro, Y., *et al*., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 451-462. https://doi.org/10.1007/978-3-030-37734-2_37

[19] Sharma, P., Gautam, A., Maji, P., Pachori, R.B. and Balabantaray, B.K. (2023) Li-segpnet: Encoder-Decoder Mode Lightweight Segmentation Network for Colorectal Polyps Analysis. *IEEE Transactions on Biomedical Engineering*, **70**, 1330-1339. https://doi.org/10.1109/tbme.2022.3216269

[20] Tajbakhsh, N., Gurudu, S.R. and Liang, J. (2016) Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE Transactions on Medical Imaging*, **35**, 630-644. https://doi.org/10.1109/tmi.2015.2487997

[21] Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., *et al*. (2017) A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *Journal of Healthcare Engineering*, **2017**, 1-9. https://doi.org/10.1155/2017/4037190

[22] Silva, J., Histace, A., Romain, O., Dray, X. and Granado, B. (2013) Toward Embedded Detection of Polyps in WCE Images for Early Diagnosis of Colorectal Cancer. *International Journal of Computer Assisted Radiology and Surgery*, **9**, 283-293. https://doi.org/10.1007/s11548-013-0926-3

[23] Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., *et al*. (2022) PVT V2: Improved Baselines with Pyramid Vision Transformer. *Computational Visual Media*, **8**, 415-424. https://doi.org/10.1007/s41095-022-0274-8

[24] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 10012-10022. https://doi.org/10.1109/iccv48922.2021.00986

[25] Milletari, F., Navab, N. and Ahmadi, S. (2016) V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 *Fourth International Conference on 3D Vision* (*3DV*), Stanford, 25-28 October 2016, 565-571. https://doi.org/10.1109/3dv.2016.79

[26] Fan, D., Cheng, M., Liu, Y., Li, T. and Borji, A. (2017) Structure-Measure: A New Way to Evaluate Foreground Maps. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 4548-4557. https://doi.org/10.1109/iccv.2017.487

[27] Wang, K., Liu, L., Fu, X., Liu, L. and Peng, W. (2023) RA-DENet: Reverse Attention and Distractions Elimination Network for Polyp Segmentation. *Computers in Biology and Medicine*, **155**, Article 106704. https://doi.org/10.1016/j.compbiomed.2023.106704

[28] Huang, C.-H., Wu, H.-Y. and Lin, Y.-L. (2021) HarDNet-MSEG: A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS. arXiv:2101.07172.

[29] Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H. and Yang, R. (2022) Salient Object Detection in the Deep Learning Era: An In-Depth Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 3239-3259. https://doi.org/10.1109/tpami.2021.3051099

[30] Patel, K., Bur, A.M. and Wang, G. (2021) Enhanced U-Net: A Feature Enhancement Network for Polyp Segmentation. 2021 18*th Conference on Robots and Vision* (*CRV*), Burnaby, 26-28 May 2021, 181-188. https://doi.org/10.1109/crv52889.2021.00032

[31] Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K. and Cui, S. (2021) Shallow Attention Network for Polyp Segmentation. In: de Bruijne, M., *et al*., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 699-708. https://doi.org/10.1007/978-3-030-87193-2_66

[32] Li, S., Feng, Y., Xu, H., Miao, Y., Lin, Z., Liu, H., *et al*. (2023) Caenet: Contrast Adaptively Enhanced Network for Medical Image Segmentation Based on a Differentiable Pooling Function. *Computers in Biology and Medicine*, **167**, Article 107578. https://doi.org/10.1016/j.compbiomed.2023.107578

[33] Liu, L., Li, Y., Wu, Y., Ren, L. and Wang, G. (2023) LGI Net: Enhancing Local-Global Information Interaction for Medical Image Segmentation. *Computers in Biology and Medicine*, **167**, Article 107627. https://doi.org/10.1016/j.compbiomed.2023.107627

[34] Xu, S., Xiao, D., Yuan, B., Liu, Y., Wang, X., Li, N., *et al*. (2023) FAFuse: A Four-Axis Fusion Framework of CNN and Transformer for Medical Image Segmentation. *Computers in Biology and Medicine*, **166**, Article 107567. https://doi.org/10.1016/j.compbiomed.2023.107567

[35] Li, Z., Zhang, N., Gong, H., Qiu, R. and Zhang, W. (2023) MFA-Net: Multiple Feature Association Network for Medical Image Segmentation. *Computers in Biology and Medicine*, **158**, Article 106834. https://doi.org/10.1016/j.compbiomed.2023.106834

[36] Zou, Y., Ge, Y., Zhao, L. and Li, W. (2023) MR-Trans: Multiresolution Transformer for Medical Image Segmentation. *Computers in Biology and Medicine*, **165**, Article 107456. https://doi.org/10.1016/j.compbiomed.2023.107456

[37] Zhang, J., Qin, Q., Ye, Q. and Ruan, T. (2023) ST-UNet: Swin Transformer Boosted U-Net with Cross-Layer Feature Enhancement for Medical Image Segmentation. *Computers in Biology and Medicine*, **153**, Article 106516. https://doi.org/10.1016/j.compbiomed.2022.106516

[38] Zhang, Z., Sun, G., Zheng, K., Yang, J., Zhu, X. and Li, Y. (2023) Tc-Net: A Joint Learning Framework Based on CNN and Vision Transformer for Multi-Lesion Medical Images Segmentation. *Computers in Biology and Medicine*, **161**, Article 106967. https://doi.org/10.1016/j.compbiomed.2023.106967

[39] Li, P., Zhou, R., He, J., Zhao, S. and Tian, Y. (2023) A Global-Frequency-Domain Network for Medical Image Segmentation. *Computers in Biology and Medicine*, **164**, Article 107290. https://doi.org/10.1016/j.compbiomed.2023.107290