ITC 2/54

Information Technology and Control
Vol. 54 / No. 2/ 2025
pp. 489-503
DOI 10.5755/j01.itc.54.2.41107

Bi-Encoder Polyp Net: A Novel Architecture for Enhanced Polyp Segmentation in Endoscopic Images

Received 2025/04/08 | Accepted after revision 2025/05/19

HOW TO CITE: Duan, Q., Gu, C. (2025). Bi-Encoder Polyp Net: A Novel Architecture for Enhanced Polyp Segmentation in Endoscopic Images. *Information Technology and Control,* 54(2), 489-503. https://doi.org/10.5755/j01.itc.54.2.41107

Bi-Encoder Polyp Net: A Novel Architecture for Enhanced Polyp Segmentation in Endoscopic Images

Qiqiang Duan, Cong Gu

School of Mathematics and Information Science, Zhongyuan University of Technology,
Zhengzhou, 451191, Henan, China

Corresponding author: gucong@zut.edu.cn

Automatic polyp segmentation in endoscopic images holds critical clinical value for early colorectal cancer diagnosis. While existing segmentation models have achieved notable progress, two key challenges persist in algorithmic performance improvement. First, dynamic adjustments of colonoscope tip orientation during examinations induce viewpoint variations, which amplify polyp appearance diversity and hinder robust feature learning. Second, the inherent similarity between polyps and surrounding tissues leads to blurred boundaries. Although convolutional neural networks (CNNs) have demonstrated significant advancements, their limitations in modeling global dependencies and reliance on aggressive downsampling operations often cause redundant network structures and local detail loss. To address these bottlenecks, we propose Bi-Encoder Polyp Net – a novel parallel architecture integrating Pyramid Vision Transformer and ResNet. This dual-branch design effectively captures global contextual dependencies while preserving low-level spatial details. A feature alignment module bridges the semantic gap between dual-branch feature maps, and an iterative semantic embedding unit further injects high-level semantic information into aligned low-level features. Extensive experiments across five public polyp segmentation benchmarks validate the network's effectiveness, demonstrating superior capability in processing real-world colonoscopy images.

KEYWORDS: Transformer, CNN, Polyp Segmentation, Image Segmentation

1. Introduction

Colorectal cancer (CRC) is a common gastrointestinal malignant tumor, and in terms of digestive system cancers, its incidence ranks only behind liver, gastric, and esophageal cancers. In the United States, colorectal cancer is one of the leading causes of cancer-related deaths. According to statistics from

2019, there were approximately 145,600 new cases of colorectal cancer, with over 50,000 people losing their lives as a result [23]. However, if detected and treated at an early stage, the five-year survival rate for colorectal cancer can reach about 90% [15]. This highlights the critical importance of early diagnosis and intervention for improving patient outcomes.

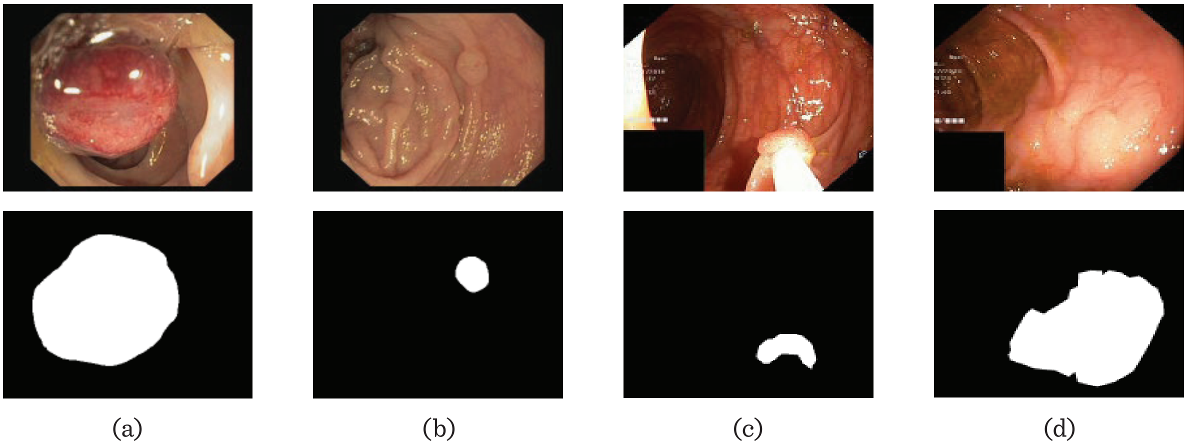
Colorectal polyps are protruding lesions that form on the surface of the intestinal mucosa and have the potential to evolve into colorectal cancer. Timely detection and removal of these polyps are crucial for preventing such a transformation. Currently, colonoscopy is widely recognized as the most effective tool for screening and preventing colorectal cancer. Studies indicate that during routine colonoscopies, approximately 25% of polyps may be missed [17], which undoubtedly increases the risk of individuals developing colorectal cancer. Furthermore, due to the often unclear boundaries of polyps in videos, doctors may face difficulties in locating and excising these abnormal tissues, especially when dealing with complex cases. Hence, developing an algorithm that enables precise and effective automatic segmentation of polyps during colonoscopies is essential for increasing the accuracy of diagnoses and the efficacy of treatments within clinical environments.

Automatic polyp segmentation remains an extremely challenging task to this day, primarily due to the following reasons: (1) The variety of polyp types, with their sizes and shapes varying significantly (Figure 1a-b), and (2) the complexity of the anatomical environment in colonoscopy videos, which makes it particularly difficult to extract features that can effectively distinguish polyps from the background (Figure 1(c)-(d)).

Considerable efforts have been dedicated to tackling these issues. Conventional techniques usually focus on extracting different features for polyp detection, such as geometric attributes [41], [21], intensity patterns [12], and volumetric properties [40], [9]. To be specific, Yoshida [41] and Näppi [21] applied geometric features, namely shape and curvature, in identifying polyps. Conversely, Jerebko et al. [12] proposed leveraging integral transforms and edge detection operators to accurately define the contours of polyps. In summary, manual feature-based methods do not achieve clinical-grade accuracy in segmentation.

The development of CNNs has led to considerable advancements in numerous areas within computer vision over the last few years [16], [18], [25], [26], [10], [39], [11]. The key factor behind this success is the convolution process, which collects local features hierarchically to create powerful representations of images. Although CNNs excel at extracting local features, they find it challenging to capture global representations, such as the long-range dependencies between visual elements, which are often essential for advanced computer vision tasks. To tackle this issue, one intuitive strategy is to widen the receptive field, although this could require more intensive and possibly destructive pooling methods.

Figure 1
Polyp Diversity and Background Complexity.


(a) (b) (c) (d)

The Transformer architecture [31] has recently been integrated into visual tasks [7], [35], [28], [43], [4], [5], [1], [45], [14]. The Vision Transformer(ViT) approach [7] involves splitting images into patches and adding position embeddings. This method forms sequences of tokens and employs successive Transformer blocks, utilizing them for the generation of vectors with parameters, enhancing effective visual representations. The self-attention mechanism enables these models to manage complex spatial transformations and long-range feature dependencies, facilitating the creation of global representations. Additionally, the MLP architecture contributes to this capability by further processing these features. However, it has been noted that visual Transformers often struggle to capture local feature details, which hinders their ability to effectively differentiate between background and foreground elements.

Effective differentiation between foreground and background is crucial for the success of polyp segmentation tasks. Understanding the characteristics and challenges of both and adopting appropriate strategies can improve the segmentation accuracy and reliability of models. However, integrating local features with global representations accurately continues to be a challenge that has yet to be fully addressed.

In the current study, to combine the advantages of CNN and Transformer, a dual-network architecture called BiEncoder-Polyp-Net (BiPolypNet) is proposed. This network aims to integrate the local feature extraction capabilities of ResNet [10] with the global representation learning abilities provided by PVT (Pyramid Vision Transformer) [33], thereby enhancing the model's understanding and expression of polyp images. Through this design, BiPolypNet not only captures the fine boundaries and local details of polyps but also effectively models their global contextual information, thus improving the accuracy and robustness of the polyp segmentation task. Considering the differences between ResNet and PVT features, a feature alignment module is designed to align the extracted ResNet and PVT features, ensuring consistency during the fusion process and ultimately producing outputs that embody the characteristics of both networks. Additionally, to address discrepancies between feature maps at different levels, an iterative semantic embedding unit is designed. This unit combines spatial detail information from low-level feature maps with semantic information from high-level feature maps, further enhancing the model's performance. This design not only improves the effectiveness of feature fusion but also provides more precise feature representations for the polyp segmentation task.

The contributions of this paper include the following:

- **Dual Encoder:** combines the local feature extraction capabilities of ResNet with the global representation learning ability provided by PVT, thereby enhancing the model's understanding and expression of polyp images.
- **Feature Alignment Unit:** fuses the local feature information from ResNet with the global feature representations from PVT using an interactive approach, thereby aligning the features extracted by ResNet and PVT, ensuring consistency during the fusion process and ultimately producing outputs that embody the characteristics of both networks.
- **Iterative Semantic Embedding Unit:** combines spatial detail information from low-level feature maps with semantic information from high-level feature maps to further enhance the model's performance.

2. Related Work

Initially, medical image segmentation relied primarily on previously established machine learning algorithms [29]. However, with the introduction of deep convolutional neural networks (CNNs) [20], this field has moved beyond the limitations of traditional methods, achieving significant improvements. Deep CNNs have not only greatly enhanced the accuracy and processing efficiency of image segmentation but also reduced reliance on manually designed features. Effective segmentation outcomes have increasingly depended on the use of CNNs and their variants, which have improved both accuracy and efficiency. Due to its clear-cut design and remarkable effectiveness, the U-shaped architecture [22] has become the preferred choice for medical image segmentation tasks. It boasts a balanced encoder-decoder structure enhanced with skip connections, which have demonstrated considerable efficacy. Techniques such as dilated convolutions [42] and context-aware modeling approaches [44] further enhance performance. Encoder-decoder frameworks with a U-shaped architecture, such as U-Net and its variants, have attracted significant attention for their ability to improve medical image analysis. Following this successful model, several similar architectures like ResUNet [38], DenseUNet [3], KiuNet [30], and UNet++ [46] have been developed. These variants introduce specific enhancements and adjustments targeted at diverse medical imaging applications. For example, UNet++ features dense inter-module skips connections that yield better segmentation outcomes. Hence, these architectural designs have proven highly effective across various medical domains. Although CNN-based methods have led to significant progress in medical image segmentation, the accuracy and efficiency remain limited by the local nature of convolution operations and complicated data access patterns.

Current research focuses on overcoming these limitations to further improve the performance of medical image segmentation models. The Transformer model, which utilizes self-attention mechanisms and multi-layer perceptrons (MLPs), has achieved outstanding success in various fields, particularly in natural language processing. It has set new standards for performance in applications like machine translation [31]. Moreover, the advent of the ViT [7] has expanded the application of Transformers to a wide range of visual tasks. ViT achieves an excellent balance between speed and accuracy in image recognition, though it necessitates pre-training on large datasets. To address this limitation, researchers proposed DeiT [28], which enhances ViT's performance on ImageNet through advanced training strategies. The Swin Transformer [19] stands out as another important visual Transformer, offering a hierarchical structure that serves as an efficient and powerful foundation for visual processing. With remarkable success in image classification, object detection, and semantic segmentation, the Swin Transformer represents a major advancement. However, Transformer-based techniques may still face challenges in certain scenarios, particularly due to variations in viewpoint and the subtle appearance of polyps, which can hinder robust feature extraction.

3. Method

As shown in Figure 2, the BiPolypNet framework is made up of two parallel branches, each intended to handle information processing in a distinct manner. By leveraging its architecture of deep convolutional layers, the ResNet branch effectively captures the local features of the input image, such as edges, textures, and other low-level visual elements essential for polyp detection. On the other hand, the Pyramid Vision Transformer (PVT) branch empowers the model to grasp long-range dependencies across any two points in the image without being constrained by local neighborhoods. This capability allows it to excel in processing global information and efficiently recognizing large-scale spatial patterns.

The features derived from the two branches possess identical resolutions but differ in their channel numbers. These corresponding hierarchical levels are then input into a feature alignment module, where the input feature maps experience alignment and fusion. This process leverages spatial and channel attention mechanisms and a Hadamard product to purposefully integrate information. Subsequently, the multi-level aligned and fused feature maps are fed into an iterative semantic embedding unit and combined with both shallow and deep aligned and fused feature maps to ensure more precise capture of the polyp's edge details.

3.1 PVT Branch

A hierarchical Transformer architecture with a pyramid structure forms the PVT branch, enabling the generation of multi-scale feature maps critical for dense prediction tasks like object detection and semantic segmentation. The PVT consists of four stages designed to generate multi-scale feature maps. All stages share a similar architecture, incorporating overlapping patch embedding layers and Transformer encoder layers. In the first stage, given an input image of size $H \times W \times 3$, the image is first divided into patches. These patches are fed into a linear projection to obtain embedded patches. Subsequently, the embedded patches are input into the Transformer Encoder layer. Within the Transformer Encoder, the patch vectors first undergo a self-attention mechanism to capture global dependencies between patches. This is followed by a feedforward neural network for nonlinear transformations and feature extraction. Through this architecture, the model can flexibly adjust the scale of feature maps at each encoder stage to accommodate diverse tasks and data.

Specifically, in the initial stage, the input image $H \times W \times 3$ is partitioned into $\frac{HW}{4^2}$ overlapping patches of size $4 \times 4 \times 3$. These patches undergo linear projection to form $\frac{HW}{4^2} \times C_1$ embedded vectors. The

Figure 2

Overview of the proposed BiPolypNet.

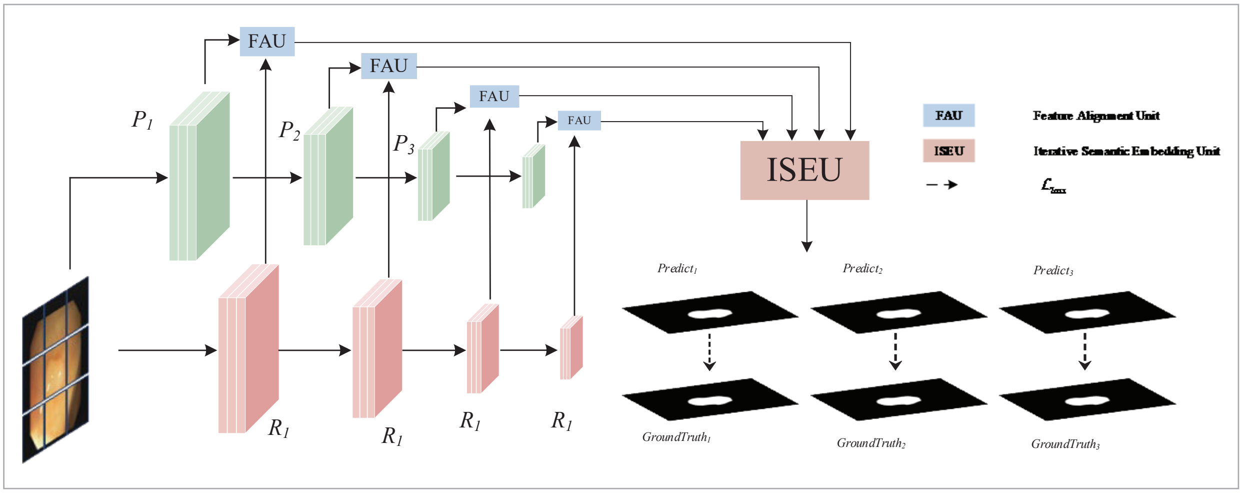
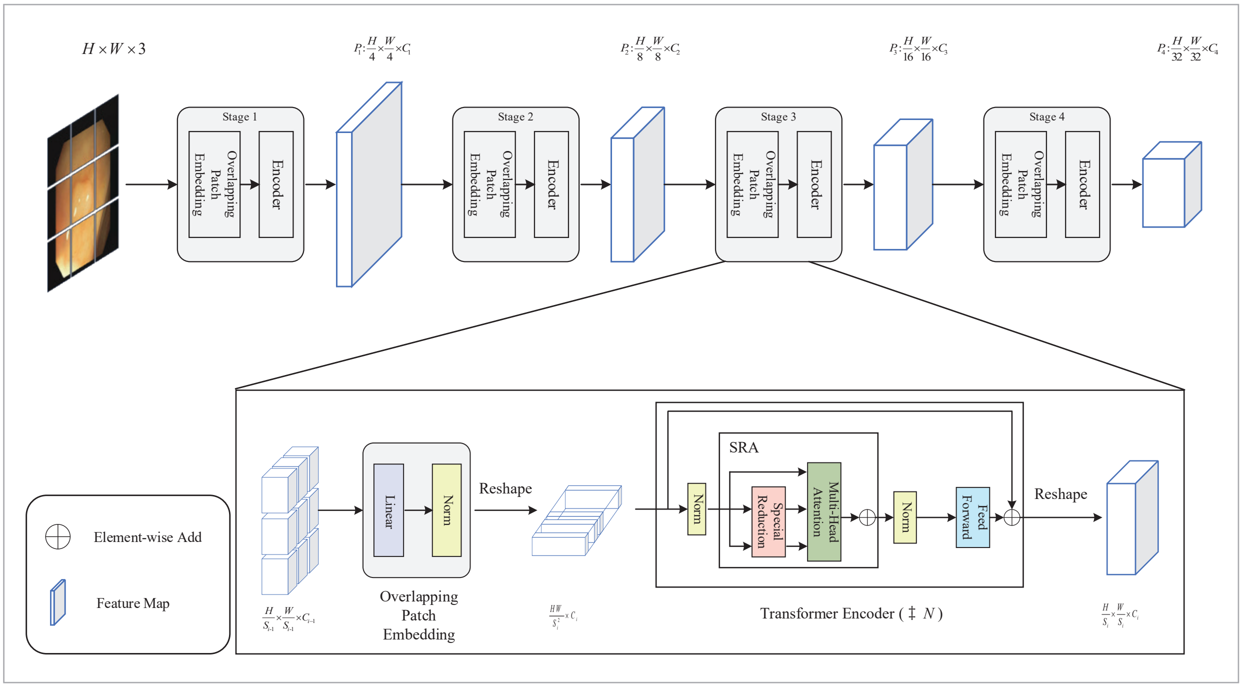


Figure 3

PVT Backbone.


$H \times W \times 3$
$P_1: \frac{H}{4} \times \frac{W}{4} \times C_1$
$P_2: \frac{H}{8} \times \frac{W}{8} \times C_2$
$P_3: \frac{H}{16} \times \frac{W}{16} \times C_3$
$P_4: \frac{H}{32} \times \frac{W}{32} \times C_4$
Stage 1
Overlapping Patch Embedding
Encoder
Stage 2
Overlapping Patch Embedding
Encoder
Stage 3
Overlapping Patch Embedding
Encoder
Stage 4
Overlapping Patch Embedding
Encoder
Element-wise Add
Feature Map
$\frac{H}{S_{i-1}} \times \frac{W}{S_{i-1}} \times C_{i-1}$
Linear
Norm
Overlapping Patch Embedding
Reshape
$\frac{HW}{S_i^2} \times C_i$
SRA
Norm
Special Reduction
Multi-Head Attention
Norm
Feed Forward
Transformer Encoder (‡ N)
Reshape
$\frac{H}{S_i} \times \frac{W}{S_i} \times C_i$

embedded vectors are processed through a Transformer Encoder layer. The output is reshaped into a feature map $P_1 \in \frac{H}{4} \times \frac{W}{4} \times C_1$. Subsequent stages recursively apply this pipeline, generating multi-scale feature maps $P_2 \in \frac{H}{16} \times \frac{W}{16} \times C_2$, $P_3 \in \frac{H}{16} \times \frac{W}{16} \times C_3$ and $P_4 \in \frac{H}{32} \times \frac{W}{32} \times C_4$, which C_1, C_2, C_3 and C_4 are 64, 128, 320, and 512, respectively. This hierarchical feature pyramid enables the PVT to adapt to diverse downstream tasks requiring multi-scale context.

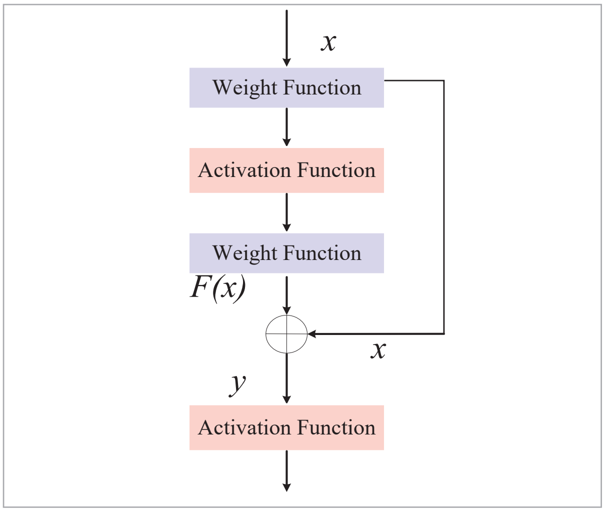
3.2 ResNet Branch

ResNet builds upon the architecture of VGG19 but introduces residual block via shortcut paths to address the challenges of training deep networks, such as vanishing/exploding gradients. This breakthrough was first demonstrated in the context of large-scale ImageNet classification tasks, where it significantly improved model performance. As shown in Figure 4, residual blocks are designed to learn residual functions by incorporating input features into deeper layers through additive interactions. This mechanism enhances gradient propagation and enables more efficient optimization, thereby boosting overall model performance.

A typical residual block is composed of convolutional layers, batch normalization, and non-linear activation functions. The residual block's design enables the input to bypass intermediate layers and be summed with the output, allowing the network to focus on learning residual patterns. This skip connection mechanism ensures gradients flow directly to earlier layers during backpropagation, addressing the vanishing gradient issue and improving training stability in deep architectures. The operation of a residual block is defined by the equation presented:

Figure 4

Rediual Block.



$$y = F(x) + x. \tag{1}$$

In the process of feature extraction, ResNet gradually builds multi-level feature representations through layer-by-layer convolution and batch normalization operations. Each layer not only captures local features but also retains information from previous layers through skip connections. This design allows the feature maps to be more refined spatially, while effectively merging multi-scale information between different layers. The features are gradually downsampled to $\frac{H}{32}\times\frac{W}{32}$, Based ResNet models typically have four feature maps: $R_1 \in \frac{H}{4}\times\frac{W}{4}\times C_1$, $R_2 \in \frac{H}{8}\times\frac{W}{8}\times C_2$, $R_3 \in \frac{H}{16}\times\frac{W}{16}\times C_3$ and $R_4 \in \frac{H}{32}\times\frac{W}{32}\times C_4$, here C_1, C_2, C_3 and C_4 are 256, 512, 1024, and 2048, whose outputs are fused with the results from PVT. Additionally, the CNN branch is very flexible and can utilize any off-the-shelf convolutional network.

Figure 5

Feature Alignment Unit.

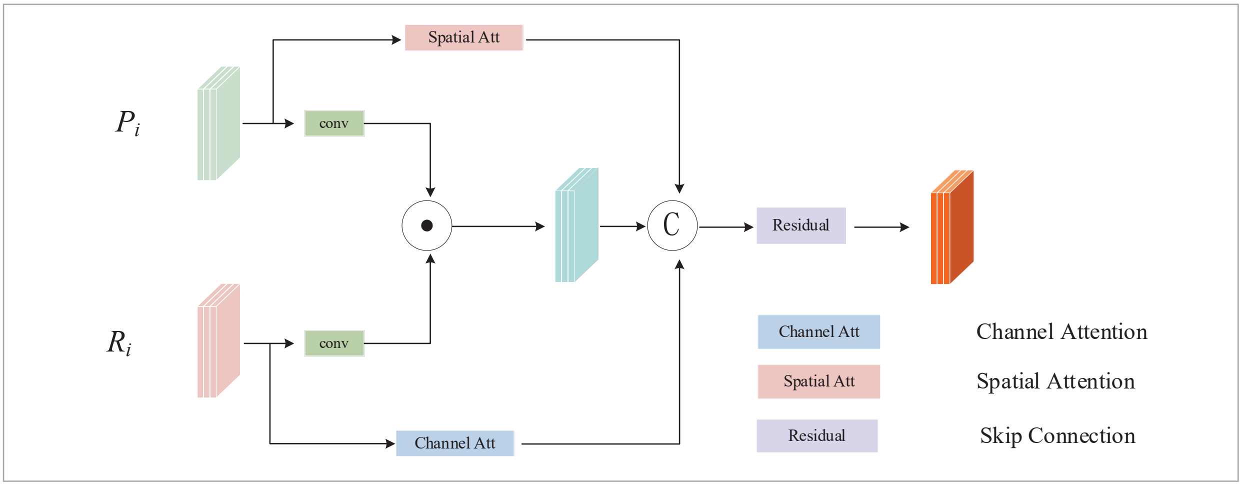


Figure 6

Polyp background contrast.

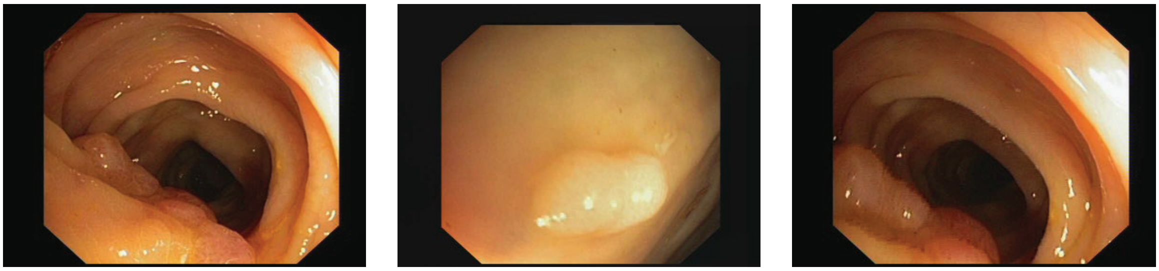


Figure 7

Efficient Global Context.

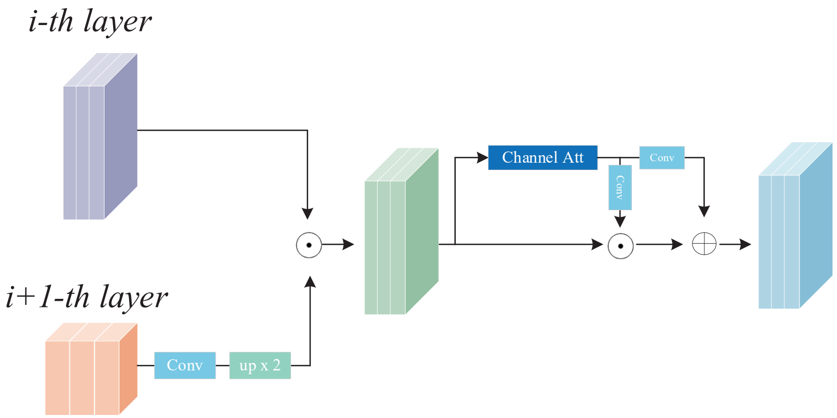
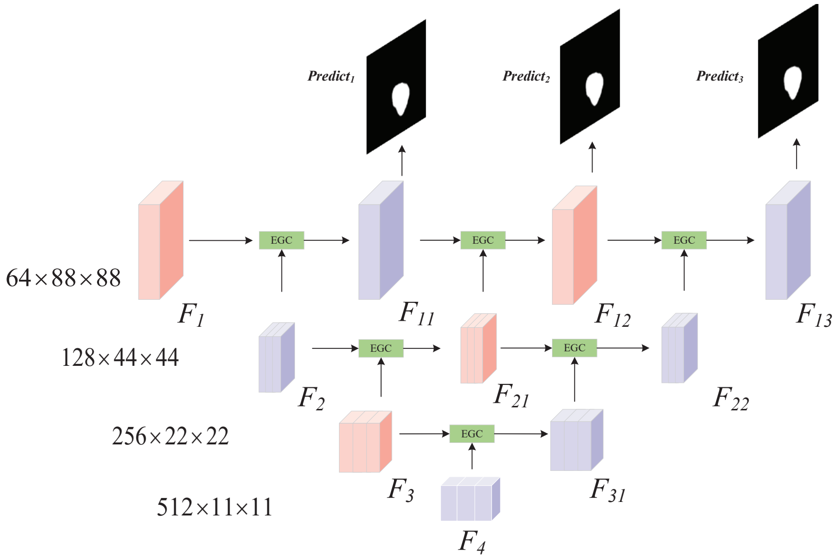


Figure 8

Iterative Semantic Embedding Unit.



3.3 Feature Alignment Unit

For efficiently integrating the encoded feature maps generated by PVT and ResNet, a feature alignment unit (see Figure 5) has been designed, which includes spatial and channel attention mechanisms as well as a multi-backbone fusion mechanism. To fuse features of two different styles, the feature alignment module first aligns the channel dimensions. Considering that the ResNet and PVT branches are designed to capture local and global features respectively, the feature alignment module is embedded across various hierarchical blocks to progressively minimize the semantic discrepancies between them through interaction. This merging process notably improves the global awareness of local features and enhances the local detail of global representations. Following alignment, the resultant features possess attributes from both networks.

Specifically, to fuse the original feature maps extracted by Pi and Ri, convolution is first performed to align the channel dimensions, followed by the Hadamard product. The feature maps that have undergone spatial attention [34], channel attention [11], and the Hadamard product are then concatenated. The aligned feature map Fi is acquired after passing through a residual block. The fused feature representation Fi, for *i*=1,2,3,4, is acquired through these operations.

$$H = Conv1(P_i) \odot Conv1(R_i) \quad (2)$$

$$C = ChannelAtt(P_i) \quad (3)$$

$$S = SpatialAtt(R_i) \quad (4)$$

$$F_i = Residual(Concat[H, C, S]), \quad (5)$$

where *Conv*1 denotes 1×1 convolutional layer and batch normalization layer, ensuring channel alignment of the dual-branch feature maps. Although both operations are labeled as *Conv*1 for notational convenience, it is important to clarify that they are implemented with distinct convolutional parameters. $\odot$ represents the Hadamard product, which leverages element-wise multiplication to efficiently select and integrate critical information from the aligned feature maps while suppressing redundant components. $ChannelAtt(\cdot)$ denotes channel attention mechanism applied to ResNet feature maps. It dynamically allocates inter-channel weights to refine the spatial representation capability of each channel, thereby emphasizing critical semantic information through channel-wise recalibration. $SpatialAtt(\cdot)$ represents spatial attention mechanism operating on PVT feature maps. By modeling long-range dependencies, it strengthens global contextual correlations and fully leverages PVT′s inherent advantage in capturing distant spatial relationships.

3.4 Iterative Semantic Embedding Unit

The BiPolypNet extracts high-level features that are rich in semantic information. However, after encoding through multiple encoders, these high-level features may lose a significant amount of spatial detail, which is crucial for improving segmentation accuracy. In the task of polyp segmentation, this issue is particularly pronounced due to the typically low contrast between the polyps and the background, as shown in Figure 6.

The iterative semantic embedding unit progressively embeds higher-level semantic information into low-level features through multiple iterations. The design of the iterative semantic embedding module contains two key components: the semantic embedding unit and the iteration process. To embed richer semantics into low-level features, this is achieved by stacking multiple semantic embedding units.

The semantic embedding functionality in this work is achieved through the Efficient Global Context (EGC) [37], whose structure is shown in Figure 7. Feature maps from adjacent layers are fed into the semantic embedding unit as input for the next round of iteration. Through multiple rounds of iteration, the model can gradually embed richer and more abstract semantic information into the low-level feature maps. The design of the semantic embedding module helps the model better understand the input data and provides more powerful and accurate semantic representations across various tasks.

As shown in Figure 8, the semantic embedding unit embeds adjacent high-level semantic information into low-level feature maps in an iterative manner. More specifically, in order to enrich low-level feature maps with comprehensive semantic information, the semantic embedding module uses the low-level feature map $X_i \in R^{C\times H\times W}$ from the layer indexed by *i* and the higher-level feature map $X_{i+1} \in R^{2C\times\frac{H}{2}\times\frac{W}{2}}$ from the layer indexed by $i+1$ as inputs. By passing through the EGC semantic embedding unit, these feature maps will continue to the subsequent iteration phase. The iteration can be performed repeatedly based on task demands, continuously infusing richer contextual meanings into the feature representations. The design of the entire iterative semantic embedding unit aims to enhance the semantic representation capability of low-level features through successive layers of feature fusion and semantic enhancement. In this way, the model reinforces the edges and borders of the targeted areas while simultaneously clarifying many vague image structures and backgrounds.

To improve the model's performance and generalization ability on complex tasks, a supervisory signal mechanism is introduced at different levels of the Semantic Embedding Module. After embedding semantic information into the low-level features, the feature maps F_{11}, F_{12} and F_{13} respectively generate prediction maps $Predict_1$, $Predict_2$, and $Predict_3$. The final prediction map is generated based on the aforementioned prediction maps, as shown in equation below:

$$P_{main} = AVG\left(Predict_1 + Predict_2 + Predict_3\right) \tag{6}$$

3.5 loss function

The loss function is given by Equation (7):

$$\mathcal{L} = \mathcal{L}_{\text{predict}_1} + \mathcal{L}_{\text{predict}_2} + \mathcal{L}_{\text{predict}_3}\ , \tag{7}$$

which calculates the difference between the predicted segmentation result Predict and the ground truth GroundTruth, as shown in Equation (8):

$$\begin{aligned}\mathcal{L}_{\text{predict}} &= L_{IoU}^{w}(Predict, GroundTruth) + \\ &+ L_{BCE}^{w}(Predict, GroundTruth)\end{aligned}\ , \tag{8}$$

which $\mathcal{L}_{IoU}^{w}$, $\mathcal{L}_{BCE}^{w}$ represent the weighted IoU loss and the weighted BCE loss, respectively. In polyp segmentation tasks, there is often an imbalance between the number of samples for normal tissue and polyp tissue. By combining these two loss functions, weighted IoU loss and weighted BCE loss, different weights are assigned to each sample based on the importance of the classes, allowing the model to focus more on the target regions. Lower weights are assigned to normal tissue samples, while higher weights are allocated to pixels in critical areas, thereby enhancing the model's attention to the target regions. This weight allocation strategy helps improve the performance and accuracy of the segmentation model in polyp segmentation tasks.

4.Experimental Results and Analysis

4.1 Dataset

To validate and evaluate the network's effectiveness and generalization capacity, this work uses five frequently adopted public datasets: Kvasir-SEG [13], CVC-ClinicDB [2], CVC-ColonDB [27], ETIS [24], and EndoScene [32]. The Kvasir-SEG dataset features 1,000 polyp images, and the ClinicDB dataset includes 612 images. Randomly, 900 and 550 images were selected from Kvasir-SEG and CVC-ClinicDB for the training set, respectively, while the remaining 100 and 62 images were used for the test set. To evaluate the model's generalization capabilities, tests were conducted on datasets from multiple centers, including CVC-ColonDB (which contains 380 images), ETIS (which has 196 images), and EndoScene (comprising 60 images). Given that these datasets stem from different medical centers, the model had not been exposed to them during training. By testing these unseen datasets, the model's generalization and robustness were effectively evaluated, showcasing its performance across different contexts.

4.2 Experimental Setup and Evaluation Metrics

The experiments carried out in this research were based on the Ubuntu 20.04 platform and employed PyTorch 2.0.0. To boost computing performance, a setup including an RTX 4090 GPU with 24GB of video memory was arranged. In terms of optimization, the AdamW optimizer was chosen, with both the learning rate and the weight decay factor set at 1×10–4. As for the learning rate scheduling, a CosineAnnealingLR scheduler was implemented to gradually lower the learning rate down to as low as 1×10^{-6}. Throughout the experimentation, a batch size of 12 was used, and the training lasted through 50 epochs. Considering the diversity of image sizes within the dataset in practical applications, all input images are uniformly resized to 352×352 pixels. However, to avoid potential information loss caused by single-scale processing, a multi-scale training strategy is implemented during network training. Specifically, the input images are scaled using scale factors of [0.75, 1, 1.25]. This approach helps alleviate the impact of variations in polyp sizes across different datasets, enhances the model's ability to recognize polyps at various scales, and improves the overall performance and robustness of the model.

Six widely used evaluation metrics were employed: $mDice$, $mIoU$, F_β^w, S_α, mE_ξ and MAE. Among them, higher values of $mDice$, $mIoU$, F_β^w, S_α and mE_ξ indicate better model performance, while MAE is the opposite. $mDice$ and $mIoU$ are common similarity metrics for evaluating segmentation models. They assess the consistency between predicted segmentation results and ground truth segmentation results at the region level, primarily focusing on the internal consistency of the segmented objects. F_β^w takes into account both recall and precision, eliminating the equal consideration of each pixel in traditional metrics. S_α emphasizes the structural similarity of the target foreground at the region and object levels. mE_ξ is used to evaluate segmentation results at the pixel and image levels. MAE is a metric for per-pixel comparison, representing the average absolute error between predicted values and ground truth values. In the evaluation of segmentation models, MAE is used to measure the degree of difference between predicted segmentation results and ground truth segmentation results.

4.3 Qualitative Analysis of Model Performance

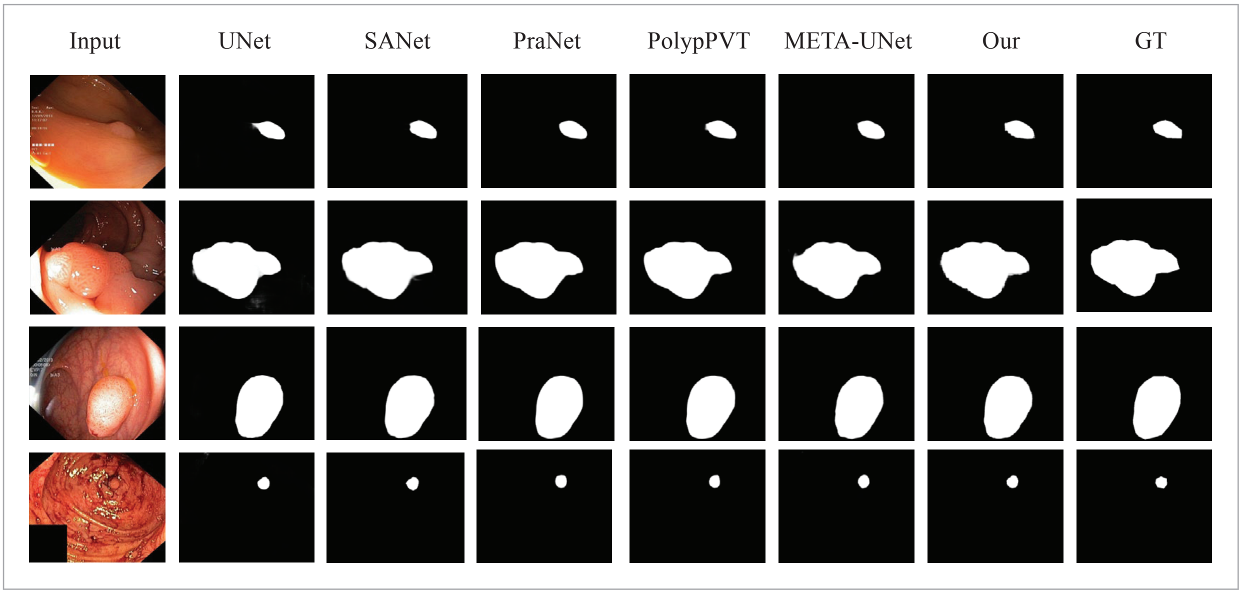
Taking the segmentation results of the algorithms on the Kvasir-SEG dataset as an example, the figure is organized as follows (see Figure 9): the first column displays the input images; columns two to six showcase the segmentation outcomes of other algorithms; the seventh column presents the segmentation results of the proposed network in this paper; and the eighth column, labeled GT (Ground Truth), represents the manually annotated ground-truth maps. This arrangement facilitates a direct visual comparison between the proposed method and competing algorithms against the reference standard.

4.4 Quantitative Analysis of Model Performance

The experiments validated BiPolypNet using the Kvasir-SEG and ClinicDB datasets, with the results shown in Tables 1 and 2. To evaluate the segmentation performance of BiPolypNet, widely adopted methods such as U-Net [22], PraNet [8], Polyp-PVT [6], and META-Unet [36] were also included in the experiments. To ensure fairness, these architectures were implemented based

Figure 9

Segmentation on the Kvasir-SEG Dataset.



on the published code and trained under the same conditions. The weights from the model that achieved the best performance on the validation set were recorded for evaluation, with the quantitative results displayed in the tables below. The best performance is highlighted in bold, while the second-best performance is underlined, allowing readers to quickly discern the relative strengths of each algorithm.

From the data in Tables 1-2, it can be observed that on the Kvasir-SEG dataset, four metrics ($mDice$, $mIoU$, F_{β}^{w}, S_{α}) surpassed the best-performing models in comparison, while evaluation metrics mE_{ξ} and MAE achieved second-best performance.

On the ClinicDB dataset, the BiPolypNet outperformed the second-best algorithms by 1.8%, 0.2%, 2.3%, 0.9%, and 1.1% in the metrics of $mDice$, $mIoU$, F_{β}^{w}, S_{α}, and mE_{ξ}, respectively. This demonstrates that BiPolypNet exhibits exceptional learning capabilities and significant advantages in addressing polyp image segmentation tasks.

Table 1

Comparison of the models on Kvasir-SEG dataset.

Model	$mDice$	$mIoU$	F_{β}^{w}	S_{α}	mE_{ξ}	MAE
U-Net	0.820	0.735	0.762	0.860	0.890	0.057
PraNet	0.900	0.843	0.888	0.911	0.947	0.026
SANet	0.896	0.837	0.880	0.916	0.945	0.029
Polyp-PVT	0.920	0.873	0.912	0.927	**0.969**	**0.019**
META-Unet	0.878	0.818	0.866	0.900	0.924	0.037
BiPolypNet	**0.922**	**0.874**	**0.913**	**0.934**	0.961	0.022

Table 2

Comparison of the models on ClinicDB dataset.

Model	$mDice$	$mIoU$	F_{β}^{w}	S_{α}	mE_{ξ}	MAE
U-Net	0.832	0.757	0.807	0.890	0.913	0.027
PraNet	0.899	0.849	0.898	0.926	0.957	0.015
SANet	0.902	0.849	0.894	0.936	0.961	0.011
Polyp-PVT	0.910	0.860	0.904	0.939	0.965	**0.008**
META-Unet	0.906	0.854	0.900	0.934	0.964	0.010
BiPolypNet	**0.928**	**0.880**	**0.927**	**0.948**	**0.977**	0.013

4.5 Comparison and Analysis of Generalization Performance

Since this paper uses Kvasir-SEG and ClinicDB as the training sets, generalization performance tests are conducted on the CVC-ColonDB, ETIS, and EndoScene datasets. Through the analysis of Tables 3-5, it can be observed that compared to existing methods, the proposed BiPolypNet exhibits weaker generalization performance on multi-center datasets from heterogeneous domains. According to our analysis, this issue is primarily attributed to data domain shift. The Kvasir-SEG and ClinicDB datasets used for training differ significantly from the CVC-ColonDB, ETIS, and EndoScene datasets used for testing in terms of image resolution, lighting conditions, annotation methods, and lesion appearance, leading to performance degradation when the model encounters data from new domains.

Furthermore, the architectural design of BiPolypNet tends to learn feature representations specific to the training data, lacking robust modeling of cross-domain variations, which may also exacerbate the problem. To alleviate this phenomenon, future work could consider incorporating domain adaptation strategies or enhancing data diversity.

Table 3

Comparison of the models on CVC-ColonDB dataset.

Model	*mDice*	*mIoU*	F_{β}^{w}	S_{α}	mE_{ξ}	*MAE*
U-Net	0.629	0.535	0.581	0.766	0.797	0.057
PraNet	0.755	0.674	0.736	0.837	0.876	0.037
SANet	0.774	0.692	0.754	0.857	0.884	0.034
Polyp-PVT	0.806	0.727	0.786	0.865	0.906	0.031
META-Unet	0.725	0.641	0.708	0.822	0.854	0.039
BiPolypNet	0.781	0.701	0.759	0.856	0.893	0.036

Table 4

Comparison of the models on ETIS dataset.

Model	*mDice*	*mIoU*	F_{β}^{w}	S_{α}	mE_{ξ}	*MAE*
U-Net	0.390	0.320	0.319	0.657	0.655	0.059
PraNet	0.732	0.653	0.695	0.840	0.884	0.014
SANet	0.715	0.637	0.673	0.833	0.861	0.028
Polyp-PVT	0.798	0.716	0.754	0.881	0.906	0.016
META-Unet	0.726	0.645	0.688	0.841	0.864	0.020
BiPolypNet	0.781	0.701	0.736	0.876	0.893	0.017

Table 5

Comparison of the models on EndoScene dataset.

Model	*mDice*	*mIoU*	F_{β}^{w}	S_{α}	mE_{ξ}	*MAE*
U-Net	0.757	0.672	0.708	0.855	0.897	0.017
PraNet	0.874	0.803	0.847	0.922	0.948	0.008
SANet	0.889	0.819	0.866	0.934	0.963	0.007
Polyp-PVT	0.887	0.818	0.863	0.931	0.956	0.008
META-Unet	0.887	0.813	0.860	0.930	0.962	0.008
BiPolypNet	0.880	0.806	0.850	0.929	0.955	0.009

Table 6

Comparison results of Ablation Study.

Model	**Kvasir-SEG**		**CVC-ColonDB**	
	mDice	*mIoU*	*mDice*	*mIoU*
BiPolypNet	0.926	0.875	0.818	0.734
BiPolypNet(w/o Bi)	0.916	0.865	0.787	0.705
BiPolypNet(w/oFAU)	0.904	0.848	0.806	0.727
BiPolypNet(w/oISEU)	0.917	0.863	0.795	0.709

4.6 Analysis of Ablation Experiment Results

To further validate the effectiveness of the method proposed in this paper, ablation experiments were conducted using the Kvasir-SEG and CVC-ColonDB datasets.

As shown in Table 6, in these experiments, the training, testing, and hyperparameter settings were the same as those mentioned above, with BiPolypNet serving as the baseline. For these experiments, the configurations for training and testing were identical to those described earlier, and the hyperparameters were set in a similar manner. BiPolypNet served as the baseline model. To evaluate the impact of components on model performance, they were gradually removed or replaced. To analyze the effectiveness of the dual encoder, feature alignment unit, and iterative semantic embedding unit, ResNet, FAU, and ISEU were respectively removed from BiPolypNet, denoted as BiPolypNet (w/o Bi), BiPolypNet (w/o FAU), and BiPolypNet (w/o ISEU). The experimental results are shown in Table 6. The results indicate that the removal of the ResNet branch, FAU, and ISEU from BiPolypNet led to varying degrees of performance decline in the model.

5.Conclusion

This paper proposes an innovative framework for polyp segmentation named BiPolypNet. The network integrates ResNet and PVT backbone as encoders to extract multi-level feature maps. These feature maps are processed by feature alignment units to generate aligned feature maps, which are then fed into our designed polyp decoding head to obtain precise segmentation predictions. The BiPolypNet architecture fully leverages CNN's inductive bias for modeling spatial correlations and the powerful capability of Transformers in capturing global relationships. Compared with existing methods widely used for colorectal polyp segmentation, BiPolypNet demonstrates superior performance across multiple evaluation metrics. This achievement not only validates the effectiveness of our approach but also highlights the potential of combining CNN and Transformer architectures in medical image segmentation tasks.

Future studies could further optimize the structure of BiPolypNet, explore more efficient feature alignment mechanisms, and investigate its applications in other medical image analysis tasks. Through continuous innovation and improvement, we believe that methods based on the combination of CNN and Transformer will play a significant role in advancing medical image processing technologies.

Acknowledgement

This work was supported by the Henan Province Graduate Education Reform and Quality Improvement Project (Grand No. YJS2025GZZ14).

References

1. Beal, J., Kim, E., Tzeng, E., Park, D. H., Zhai, A., Kislyuk, D. Toward Transformer-Based Object Detection. arXiv Preprint arXiv:2012.09958, 2020.
2. Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rordiguez, C., Vilarino, F. WM-DOVA Maps for Accurate Polyp Highlighting in Colonoscopy: Validation vs. Saliency Maps from Physicians. Computerized Medical Imaging and Graphics, 2015, 43, 99-111. https://doi.org/10.1016/j.compmedimag.2015.02.007
3. Cai, S., Tian, Y., Lui, H., Zeng, H., Wu, Y., Chen, G. Dense-UNet: A Novel Multiphoton In Vivo Cellular Image Segmentation Model Based on a Convolutional Neural Network. Quantitative Imaging in Medicine and Surgery, 2020, 10(6), 1275. https://doi.org/10.21037/qims-19-1090
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoryuko, S. End-to-End Object Detection with Transformers. European Conference on Computer Vision, 2020, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13
5. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W. Pre-Trained Image Processing Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 12299-12310. https://doi.org/10.1109/CVPR46437.2021.01212
6. Dong, B., Wang, W., Fan, D.-P., Li, J., Fu, H., Shao, L. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. arXiv Preprint arXiv:2108.06932, 2021.
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Deghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. arXiv Preprint arXiv:2010.11929, 2020.

8. Fan, D.-P., Ji, G.-P., Zhou, T., et al. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020, 263-273. https://doi.org/10.1007/978-3-030-59725-2_26
9. Gross, S., Kennel, M., Stehle, T., Wulff, J., Tischendorf, J., Trautwein, C., Aach, T. Polyp Segmentation in NBI Colonoscopy. Bildverarbeitung für die Medizin 2009: Algorithmen-Systeme-Anwendungen, 2009, 252-256. https://doi.org/10.1007/978-3-540-93860-6_51
10. He, K., Zhang, X., Ren, S., Sun, J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90
11. Hu, J., Shen, L., Sun, G. Squeeze-and-Excitation Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745
12. Jerebko, A., Franaszek, M., Summers, R. Radon Transform-Based Polyp Segmentation Method for CT Colonography Computer-Aided Diagnosis. Radiology, 2002, 257-258. https://doi.org/10.1117/12.480696
13. Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., De Lang, T., Johansen, D., Johansen, H. D. Kvasir-Seg: A Segmented Polyp Dataset. MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II, 2020, 26, 451-462. https://doi.org/10.1007/978-3-030-37734-2_37
14. Jiang, Y., Chang, S., Wang, Z. TransGAN: Two Transformers Can Make One Strong GAN. arXiv Preprint arXiv:2102.07074, 2021, 1(3).
15. Kolligs, F. T. Diagnostics and Epidemiology of Colorectal Cancer. Visceral Medicine, 2016, 32(3), 158-164. https://doi.org/10.1159/000446488
16. Krizhevsky, A., Sutskever, I., Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 2012, 25.
17. Leufkens, A., Van Oijen, M., Vleggaar, F., Siersema, P. D. Factors Influencing the Miss Rate of Polyps in a Back-to-Back Colonoscopy Study. Endoscopy, 2012, 44(05), 470-475. https://doi.org/10.1055/s-0031-1291666
18. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986
20. Long, J., Shelhamer, E., Darrell, T. Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965
21. Näppi, J., Yoshida, H. Automated Detection of Polyps with CT Colonography: Evaluation of Volumetric Features for Reduction of False-Positive Findings. Academic Radiology, 2002, 9(4), 386-397. https://doi.org/10.1016/S1076-6332(03)80184-8
22. Ronneberger, O., Fischer, P., Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III, 2015, 18, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28
23. Siegel, R. L., Fedewa, S. A., Anderson, W. F., Miller, K. D., Ma, J., Rosenberg, P. S., Jemal, A. Colorectal Cancer Incidence Patterns in the United States, 1974-2013. JNCI: Journal of the National Cancer Institute, 2017, 109(8), djw322. https://doi.org/10.1093/jnci/djw322
24. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B. Toward Embedded Detection of Polyps in WCE Images for Early Diagnosis of Colorectal Cancer. International Journal of Computer Assisted Radiology and Surgery, 2014, 9, 283-293. https://doi.org/10.1007/s11548-013-0926-3
25. Simonyan, K., Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv Preprint arXiv:1409.1556, 2014.
26. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. Going Deeper with Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594
27. Tajbakhsh, N., Gurudu, S. R., Liang, J. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. IEEE Transactions on Medical Imaging, 2015, 35(2), 630-644. https://doi.org/10.1109/TMI.2015.2487997
28. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayorelles, A., Jegou, H. Training Data-Efficient Image Transformers & Distillation Through Attention. International Conference on Machine Learning, 2021, 10347-10357.
29. Tsai, A., Yezzi, A., Wells, W., Tampany, C., Tucker, D., Fan, A., Grimson, W. E., Willsky, A. A Shape-Based Approach

to the Segmentation of Medical Imagery Using Level Sets. IEEE Transactions on Medical Imaging, 2003, 22(2), 137-154. https://doi.org/10.1109/TMI.2002.808355

30. Valanarasu, J. M. J., Sindagi, V. A., Hacihaliloglu, I., Patel, V. M. KIU-Net: Toward Accurate Segmentation of Biomedical Images Using Over-Complete Representations. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020, 363-373. https://doi.org/10.1007/978-3-030-59719-1_36
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, 30.
32. Vázquez, D., Bernal, J., Sánchez, F. J., Fernandez-Esparrach, G., Lopez, A. M., Romero, A., Drozdzal, M., Courville, A. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. Journal of Healthcare Engineering, 2017, 2017(1), 4037190. https://doi.org/10.1155/2017/4037190
33. Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L. PVT v2: Improved Baselines with Pyramid Vision Transformer. Computational Visual Media, 2022, 8(3), 415-424. https://doi.org/10.1007/s41095-022-0274-8
34. Woo, S., Park, J., Lee, J.-Y., Kweon, I. S. CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV), 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1
35. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P. Visual Transformers: Token-Based Image Representation and Processing for Computer Vision. arXiv Preprint arXiv:2006.03677, 2020.
36. Wu, H., Zhao, Z., Wang, Z. META-UNet: Multi-Scale Efficient Transformer Attention UNet for Fast and High-Accuracy Polyp Segmentation. IEEE Transactions on Automation Science and Engineering, 2023. https://doi.org/10.1109/TASE.2023.3292373
37. Wu, H., Zhao, Z., Zhong, J., Wang, W., Wen, Z., Qin, J. PolypSeg+: A Lightweight Context-Aware Network for Real-Time Polyp Segmentation. IEEE Transactions on Cybernetics, 2022, 53(4), 2610-2621. https://doi.org/10.1109/TCYB.2022.3162873
38. Xiao, X., Lian, S., Luo, Z., Li, S. Weighted Res-UNet for High-Quality Retina Vessel Segmentation. 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, 327-331. https://doi.org/10.1109/ITME.2018.00080
39. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K. Aggregated Residual Transformations for Deep Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 1492-1500. https://doi.org/10.1109/CVPR.2017.634
40. Yao, J., Miller, M., Franaszek, M., Summers, R. M. Colonic Polyp Segmentation in CT Colonography Based on Fuzzy Clustering and Deformable Models. IEEE Transactions on Medical Imaging, 2004, 23(11), 1344-1352. https://doi.org/10.1109/TMI.2004.826941
41. Yoshida, H., Masutani, Y., Maceneaney, P., Rubin, D. T., Dachman, A. H. Computerized Detection of Colonic Polyps at CT Colonography on the Basis of Volumetric Features: Pilot Study. Radiology, 2002, 222(2), 327-336. https://doi.org/10.1148/radiol.2222010506
42. Yu, F., Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv Preprint arXiv:1511.07122, 2015.
43. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z. H., Tay, F. E. H., Feng, J., Yan, S. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 558-567. https://doi.org/10.1109/ICCV48922.2021.00060
44. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. Pyramid Scene Parsing Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2881-2890. https://doi.org/10.1109/CVPR.2017.660
45. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., Zhang, L. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 6881-6890. https://doi.org/10.1109/CVPR46437.2021.00681
46. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, 2018, 3-11. https://doi.org/10.1007/978-3-030-00889-5_1

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).