IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 July 2024, accepted 11 August 2024, date of publication 19 August 2024, date of current version 27 August 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3445428*

RESEARCH ARTICLE

# PDLFBR-Net: Partial Decoder Localization and Foreground-Background Refinement Network for Polyp Segmentation

**YANBIN PENG, (Member, IEEE), MINGKUN FENG, ZHINIAN ZHAI, AND ZHIJUN ZHENG**
School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China

Corresponding author: Yanbin Peng (pyb2010@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972357, and in part by the Basic Public Welfare Research Program of Zhejiang Province under Grant LGF22F020017.

**ABSTRACT** Polyp segmentation is vital for early detection and treatment of colorectal cancer, significantly improving patient prognosis. This paper proposes an efficient and precise polyp segmentation model called the Partial Decoder Localization and Foreground-Background Refinement Network (PDLFBR-Net), which simulates the human object recognition process. Specifically, PDLFBR-Net comprises three key modules: the Cross-level Attention-enhanced Fusion Module (CAFM), the Position Recognition Module (PRM), and the Foreground-Background Refinement Module (FBRM). The CAFM enhances feature representation by fusing information from adjacent levels, providing more discriminative features. The PRM module simulates the human recognition process by using a partial decoder to locate potential polyp tissues from a global perspective. Subsequently, the FBRM is used to perform specific recognition, gradually refining the initial prediction results through foreground and background focusing to achieve precise recognition. Extensive experiments demonstrate that the proposed PDLFBR-Net model significantly outperforms existing state-of-the-art models on five challenging datasets. On the Kvasir-SEG benchmark dataset, the mean Dice and mean IoU values reached 93.7% and 89.5%, respectively, which represents an improvement of 0.4% and 0.6% compared to the best-performing state-of-the-art (SOTA) method.

**INDEX TERMS** Attention, convolutional neural network, polyp segmentation, partial decoder.

## I. INTRODUCTION

Colorectal cancer is one of the malignant tumors with high incidence and mortality rates worldwide. Early detection and treatment are crucial for improving patient survival rates. As an important precursor to colorectal cancer, the detection and segmentation of polyps are critical steps in clinical practice [1]. Automated polyp segmentation technology assists doctors in quickly and accurately identifying polyps during endoscopic examinations, significantly enhancing the efficiency of diagnosis and treatment. However, as shown in Figure 1, there are a series of complex challenges in the polyp segmentation task, which impose higher requirements on the design of segmentation algorithms.

Firstly, polyps exhibit significant diversity in terms of size, shape, and color. Some polyps may appear spherical, flat, or irregularly shaped, and their surface textures may vary depending on the type. For example, adenomatous polyps usually have a smooth surface, while serrated polyps have a more complex surface structure. This diversity demands that segmentation algorithms have strong generalization capabilities to adapt to various polyp morphologies. Secondly, the boundaries between polyps and surrounding tissues are often unclear, making them visually difficult to distinguish. In endoscopic images, polyp edges may appear blurry or discontinuous due to lighting, viewing angles, or tissue penetration, complicating accurate boundary detection. This blurriness makes traditional edge detection methods inadequate for precisely locating polyp boundaries in complex endoscopic environments. Additionally, endoscopic images

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

have complex backgrounds, including structures like intestinal folds, blood vessels, and luminal contents, which can closely resemble polyps in color and texture. Variations in lighting conditions and noise interference, such as liquid or reflections on the lens, further complicate the background, making it harder to differentiate polyps from the background. Motion blur during endoscope operation and the dynamic intestinal environment also leads to continuously changing background structures, increasing the complexity of the segmentation task. Therefore, developing algorithms that can effectively handle these complexities is crucial for improving the accuracy of polyp detection [2], [3].

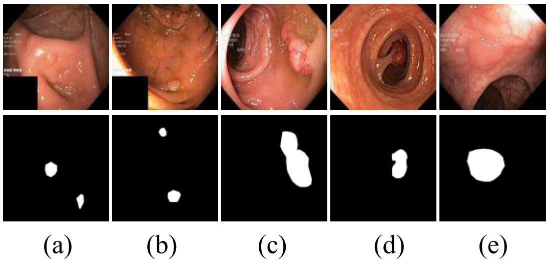


**FIGURE 1. Examples of polyp images: (a) and (b) small polyps. (c) and (d) complex background (folds). (e) polyps with unclear boundaries with surrounding tissue. The first row is the image, and the second row is the Ground Truth.**

Traditional polyp segmentation methods often rely on manually designed features, making it challenging to handle various complex scenarios and diverse polyp morphologies. For example, early methods based on texture and shape features [4], [5], while capable of segmenting polyps to some extent, struggle with the diverse nature of polyps and complex backgrounds due to the limitations of hand-crafted features. In recent years, convolutional neural networks (CNNs) have achieved great success in many fields [6], [7], [8], [9], [10], [11], significantly enhancing the performance of various tasks. Specifically, in the field of image segmentation, the application of CNNs has markedly improved the performance of polyp segmentation. Classic architectures such as U-Net [12] and DeepLabV3+ [13] effectively extract multi-scale features through fully convolutional network structures, achieving good segmentation results. However, these methods primarily focus on global feature representation, neglecting the capture of local details, particularly in handling polyp boundary details and complex backgrounds, resulting in noticeable shortcomings. Attention mechanism-based methods provide new insights for polyp segmentation. For instance, Jha et al. [14] proposed ResUNet++, which enhances feature representation by incorporating residual units and attention mechanisms, but its segmentation performance in complex backgrounds remains limited. Zhai et al. [15] introduced attention fusion modules, attention enhancement modules, and multi-view attention modules, effectively enriching and optimizing feature representation, but boundary details and handling blurry regions still require improvement. Additionally, Mahmud et al. [16] designed PolypSegNet, which enhances detail capture through multi-scale feature fusion but still has room for improvement in handling complex backgrounds and detail preservation.

In this paper, we propose an innovative Partial Decoder Localization and Foreground-Background Refinement Network (PDLFBR-Net), aiming to enhance polyp segmentation accuracy by simulating the global localization and local refinement capabilities in human object recognition processes. PDLFBR-Net consists of three key modules: Cross-level Attention-enhanced Fusion Module (CAFM), Position Recognition Module (PRM), and Foreground-Background Refinement Module (FBRM). The CAFM module enhances feature expression by integrating features from different levels, enabling the model to capture more discriminative cross-level information. The PRM module uses partial decoders to globally locate potential polyp regions, generating initial prediction maps. The FBRM module further focuses on foreground and background, refining initial predictions through contextual reasoning to achieve precise segmentation of polyp boundaries and details. In summary, the main contributions of this paper are as follows:

1) This paper introduces the concept of contextual reasoning into the polyp segmentation task and develops a new foreground and background refinement strategy, which refines and distinguishes uncertain areas, thereby improving the segmentation effect of polyp tissues.

2) This paper proposes a novel polyp segmentation method named PDLFBR-Net. This method first performs cross-level feature fusion, then uses partial decoders to locate polyp tissues from a global perspective, and finally progressively refines uncertain areas to obtain optimized polyp segmentation results.

3) The PDLFBR-Net method outperforms existing state-of-the-art models on five benchmark datasets. Experimental results demonstrate the superiority and robustness of our method in complex scenarios.

## II. RELATED WORKS

In the field of medical image analysis, polyp segmentation has always been an important and challenging research topic [17], [18], [19]. Early detection and accurate segmentation of colorectal polyps are crucial for the prevention of colorectal cancer. In recent years, with the rapid development of deep learning, many deep learning-based methods have been introduced into the task of polyp segmentation and have achieved significant results.

### A. TRADITIONAL POLYP SEGMENTATION METHODS

Early methods for polyp segmentation were primarily based on image processing and machine learning techniques. These methods often relied on manually designed features and rules, including edge detection, threshold segmentation, and morphological operations. For example, Xia et al. [20] first detected the preliminary region of interest (pROI) using a technique similar to the Hough transform for preprocessing to eliminate the background. The image was then segmented through two steps: a relaxation process to identify homogeneous regions and a tightening process to merge unnecessary regions based on color differences in the CIE color space. Wang [21] et al. proposed a new computer-aided detection (CAD) technique for detecting colorectal polyps. This method utilized morphological and texture information of the colon wall, quickly identifying suspicious regions through local and global geometric information. Edge detection techniques and a hypothetical elliptical polyp model were used to quantitatively analyze the growth regions of suspicious polyps, and false positives were eliminated by combining extracted morphological and texture features. Jerebko et al. [22] proposed a new method for colorectal polyp detection based on symmetric curvature patterns. This method distinguished polyps from other structures by extracting symmetric curvature features from candidate regions and combined other features to improve the overall sensitivity of the detection system. However, due to the complex and variable shapes and textures of polyps, these methods showed limited performance in practical applications.

### B. POLYP SEGMENTATION METHODS BASED ON CONVOLUTIONAL NEURAL NETWORKS

The introduction of convolutional neural networks (CNNs) has significantly propelled advancements in various fields [23], [24]. In the field of medical image analysis, the application of CNNs has also led to substantial progress in polyp segmentation techniques [25]. U-Net [12] is one of the earliest and most classic fully convolutional networks. Its encoder-decoder structure effectively combines local and global information, making it widely used in medical image segmentation tasks. The basic architecture of U-Net includes a symmetrical encoder and decoder structure, connecting feature maps of corresponding layers in the encoder and decoder through skip connections. This design allows U-Net to retain more detailed information during decoding, enabling it to capture high-resolution local details while understanding global context information, thus performing excellently in various segmentation tasks.

Building on the successful application of U-Net, many improved models have been proposed to further enhance segmentation performance. For example, UNet++ [26] redesigned the skip connections of the original U-Net by introducing nested skip connection paths, further reducing the semantic gap between the encoder and decoder, thereby improving the model's expressive capabilities. The multi-resolution feature fusion strategy of UNet++ allows it to better capture fine-grained features when dealing with complex medical images, significantly improving segmentation accuracy and robustness.

Additionally, ResUNet [27] is another improved model based on U-Net. It combines the advantages of residual networks (ResNet) by introducing residual blocks in the encoder and decoder, addressing the issue of gradient vanishing during the training of deep networks. This design of ResUNet not only improves the training efficiency of the network but also enhances its ability to extract complex features, making it particularly effective in handling medical images with complex backgrounds and subtle structural differences.

These improved models based on U-Net, by introducing techniques such as skip connections, multi-scale feature fusion, and residual blocks, have greatly enhanced the ability to capture fine-grained features, significantly improving the performance of polyp segmentation. However, despite these improvements performing well in specific scenarios, researchers are still exploring other methods to further enhance the accuracy and robustness of polyp segmentation. These methods include the design of new network architectures, multi-modal fusion, the application of Generative Adversarial Networks (GAN), and the introduction of attention mechanisms. For example, Poorneshwaran et al. [28] innovatively applied conditional Generative Adversarial Networks (cGAN) to the polyp segmentation task, leveraging the powerful capabilities of GANs to generate more accurate segmentation results. Banik et al. [29] proposed a multi-modal fusion network called Polyp-Net for the automatic segmentation of polyps in colonoscopy images. Polyp-Net combines Dual-Tree Wavelet Pooling Convolutional Neural Network (DT-WpCNN) and Local Gradient Weighted Embedding Level Set Method (LGWe-LSM), improving segmentation accuracy while reducing false positives. This method achieved excellent performance on the CVC-Clinic dataset. Yue et al. [30] proposed an Adaptive Context Exploration Network (ACENet) for polyp segmentation in colonoscopy images. This network adopts an encoder-decoder architecture, combining the Attentional Atrous Spatial Pyramid Pooling Module (AASPP) module and Adaptive Context Extraction (ACE) module to effectively capture and fuse multi-scale features, improving the localization and detection of polyp regions. Wei et al. [31] proposed a new model called Shallow Attention Network (SANet) for polyp segmentation in colonoscopy images. SANet reduces the impact of image color on model training through color exchange operations, enhances the segmentation quality of small polyps using shallow attention modules, and introduces a probability correction strategy during inference to balance the imbalance of foreground and background pixels. Yin et al. [32] proposed a Duplex Contextual Relation Network (DCR-Net) for polyp segmentation in colonoscopy images. DCRNet captures contextual relationships within and across images, enhancing feature representation and thereby improving segmentation accuracy. Zhou et al. [33] proposed a Cross-level Feature Aggregation Network (CFA-Net) for polyp segmentation. This network includes a boundary prediction network to generate boundary-aware features and incorporates these features into the segmentation network through a layer-wise strategy. CFA-Net designs a two-stream segmentation network that effectively handles scale variations and boundary blurriness issues through Cross-level Feature Fusion (CFF) and Boundary Aggregated Module (BAM).

Overall, these diverse polyp segmentation methods, through various technical means, further advance the field and demonstrate the great potential and broad application prospects of deep learning in medical image processing.

### C. TRANSFORMER-BASED METHODS FOR POLYP SEGMENTATION

Recent studies have introduced Transformer architectures, combining them with CNNs to leverage their powerful global feature extraction capabilities, further improving segmentation performance. Dong et al. [34] proposed a novel polyp segmentation framework based on the Pyramid Vision Transformer (PVT), called Polyp-PVT. Unlike traditional CNN-based methods, Polyp-PVT employs a Transformer encoder to extract stronger and more robust features. Additionally, the framework introduces three standard modules: the Cascade Fusion Module (CFM), the Camouflage Identification Module (CIM), and the Similarity Aggregation Module (SAM), which capture high-level and low-level features of polyps and effectively fuse them. Zhang et al. [35] proposed a Hybrid Semantic Network (HSNet) for automatic polyp segmentation. HSNet combines the advantages of Transformers and Convolutional Neural Networks (CNNs) and captures long-range dependencies and local details through the Cross Semantic Attention Module (CSA), Hybrid Semantic Complement Module (HSC), and Multi-Scale Prediction Module (MSP). Liu et al. [36] proposed a novel Cross Attention and Feature Exploration Network (CAFE-Net) for polyp segmentation. CAFE-Net addresses issues such as small polyp target loss, fine-grained detail recovery, and limited multi-scale feature aggregation capability through the Feature Supplement and Exploration Module (FSEM), Cross Attention Decoding Module (CADM), and Multi-Scale Feature Aggregation Module (MFA). Yue et al. [37] proposed a Boundary Uncertainty Aware Network (BUNet) for automated polyp segmentation. BUNet employs the Pyramid Vision Transformer (PVTv2) as the encoder, combined with the Boundary Exploration Module (BEM) and the Boundary Uncertainty Aware Module (BUM) to handle the diversity in polyp size and shape, as well as boundary ambiguity.

Although these methods have achieved good results in polyp segmentation, there is still room for improvement. In this study, we developed a polyp segmentation model based on the Transformer architecture, utilizing partial decoders and attention mechanisms, surpassing existing methods on multiple public datasets.

## III. THE PROPOSED METHOD

This paper designs a Partial Decoder Localization and Foreground-Background Refinement Network (PDLFBR-Net), which includes three key modules: the Cross-level Attention-enhanced Fusion Module (CAFM), the Position Recognition Module (PRM), and the Foreground-Background Refinement Module (FBRM). The CAFM enhances feature discrimination by fusing information from adjacent levels. The PRM uses a partial decoder to locate potential polyp tissues from a global perspective. The FBRM gradually refines the initial prediction results through foreground and background focusing, thereby achieving precise polyp segmentation.

### A. OVERVIEW

The proposed structure of the PDLFBR-Net network is shown in Figure 2. Given an RGB image with a resolution of [h, w], it is first input into the PVTv2 [38] backbone network to extract multi-level features, denoted as $\mathbf{F}_i$, where $i \in \{1, 2, 3, 4\}$. The resolution of the feature maps is [h/$2^{i+1}$, w/$2^{i+1}$], and the number of channels of the feature maps is $\{64, 128, 320, 512\}$, respectively. These features are then input into the CAFM module for feature enhancement. Subsequently, a partial decoder is applied to the high-level features $\{\mathbf{F}_4, \mathbf{F}_3, \mathbf{F}_2\}$ to locate potential polyp tissues. Finally, the Foreground-Background Refinement Module is used to gradually refine the boundaries and details of the polyp tissues, achieving accurate polyp segmentation.

### B. CROSS-LEVEL ATTENTION-ENHANCED FUSION MODULE

We propose a module named Cross-level Attention-enhanced Fusion Module. CAFM enhances the model's ability to capture cross-level contextual semantic information by fusing feature maps from adjacent levels. Through this fusion method, CAFM not only improves the diversity and accuracy of feature representation but also enhances the model's ability to handle various complex polyp shapes. Specifically, the CAFM module uses upper-level features to generate attention maps, which are then expanded and used to enhance the current level's features. As shown in Figure 3, we consider both channel attention mechanism and spatial attention mechanism [8]. In the channel attention mechanism, channel attention is applied to the upper-level features to calculate the weight of each channel. Through this step, the model can identify which channels contain more useful information and enhance the features of these channels. In the spatial attention mechanism, spatial attention is applied to the upper-level features to calculate the weight of each pixel's position. Through this step, the model can focus more on key positions in the image and ignore irrelevant background areas. In the channel attention mechanism, we use global max pooling and global average pooling. In the spatial attention mechanism, we use global max pooling along the channel and global average pooling along the channel.

Taking $\mathbf{F}_a$ and $\mathbf{F}_b$ as examples, where $\mathbf{F}_a$ represents the current level features and $\mathbf{F}_b$ represents the upper-level features, the channel attention map $\mathbf{AM}_c$ is expressed as:

$$\mathbf{AM}_C = \text{Softmax}(\mathbf{AM}_{CA} \oplus \mathbf{AM}_{CM}) \tag{1}$$

$$\mathbf{AM}_{CA} = \text{Conv1}(\text{Conv1}(\text{A}(\mathbf{F}_b))) \tag{2}$$

$$\mathbf{AM}_{CM} = \text{Conv1}(\text{Conv1}(\text{M}(\mathbf{F}_b))) \tag{3}$$

where A() denotes the global average pooling operation, M() denotes the global max pooling operation, Softmax() denotes

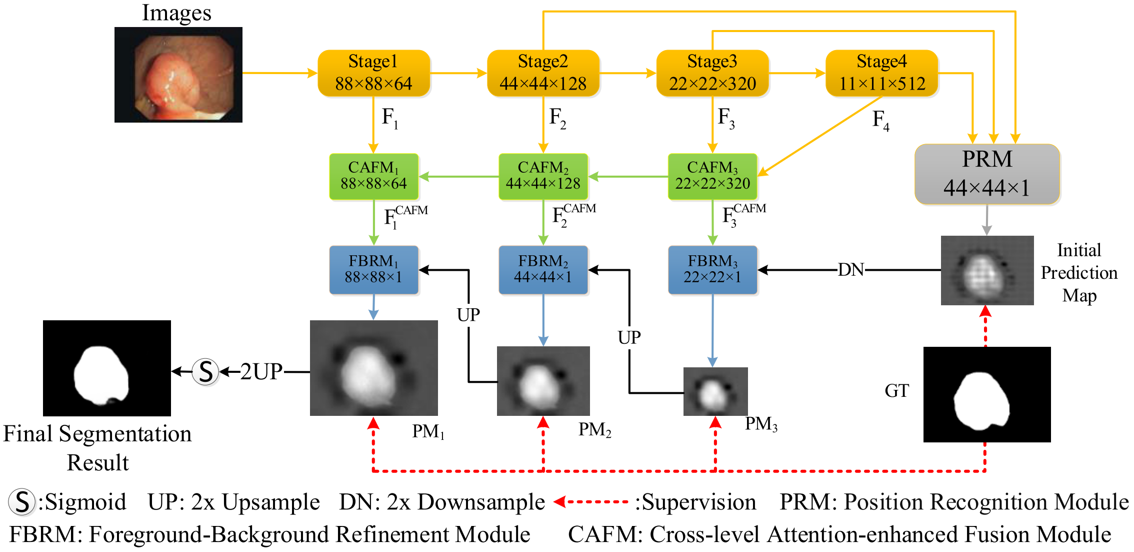


**FIGURE 2.** Overview of the proposed PDLFBR-Net.

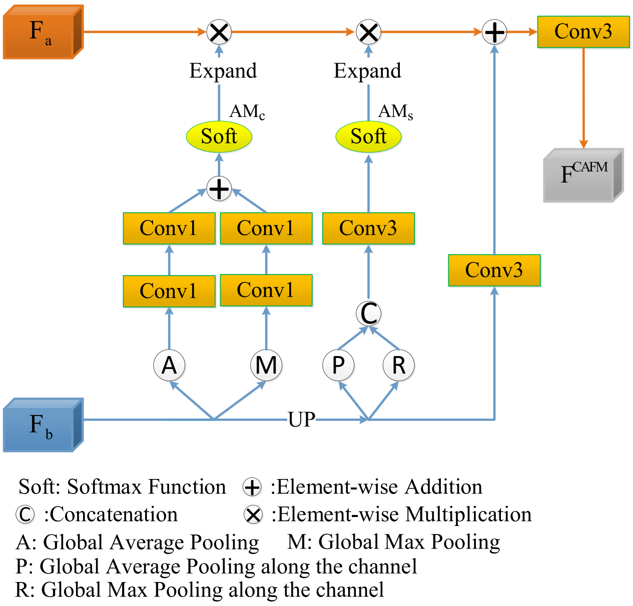

Fa
Expand
Expand
AMc
AMs
Soft
Soft
Conv1
Conv1
Conv1
Conv1
Conv3
Conv3
Conv3
C
A
M
P
R
UP
Fb
F CAFM
Soft: Softmax Function + :Element-wise Addition
C :Concatenation × :Element-wise Multiplication
A: Global Average Pooling M: Global Max Pooling
P: Global Average Pooling along the channel
R: Global Max Pooling along the channel

**FIGURE 3.** Cross-level attention-enhanced fusion module.

the softmax function, and Conv1() denotes the convolution operation with a kernel size of 1 × 1. ⊕ denotes element-wise addition.

The spatial attention map $\mathbf{AM}_s$ is expressed as:

$$\mathbf{AM}_S = \text{Softmax}(\text{Conv3}(\text{Concat}(\mathbf{AM}_{SA}, \mathbf{AM}_{SM}))) \quad (4)$$

$$\mathbf{AM}_{SA} = \text{P}(\text{UP}(\mathbf{F}_b)) \quad (5)$$

$$\mathbf{AM}_{SM} = \text{R}(\text{UP}(\mathbf{F}_b)) \quad (6)$$

where P() denotes the global average pooling operation along the channel, R() denotes the global max pooling operation along the channel, Conat() denotes the concatenation operation, UP() denotes the 2x upsampling operation, and Conv3() denotes the convolution operation with a kernel size of 3 × 3.

Next, expand the channel attention map $\mathbf{AM}_c$ and the spatial attention map $\mathbf{AM}_s$ to the same size as $\mathbf{F}_a$, and then perform element-wise multiplication with $\mathbf{F}_a$ respectively to obtain the attention-enhanced feature map $\mathbf{F}^{AE}$. Then, perform a residual connection with $\mathbf{F}_b$ and conduct a convolution operation to obtain the final output $\mathbf{F}^{CAFM}$. The formal representation is as follows:

$$\mathbf{F}^{CAFM} = \text{Conv3}(\mathbf{F}^{AE} \oplus \text{Conv3}(\text{UP}(\mathbf{F}_b))) \quad (7)$$

$$\mathbf{F}^{AE} = \mathbf{F}_a \otimes \text{Expand}(\mathbf{AM}_C) \otimes \text{Expand}(\mathbf{AM}_S) \quad (8)$$

Here, Expand() represents the expansion operation, which expands the size of the input feature map to be the same as $\mathbf{F}_a$. ⊗ represents the element-wise multiplication operation. As shown in Figure 2, the obtained enhanced features are denoted as $\mathbf{F}_i^{CAFM}$, where $i \in \{1,2,3\}$.

### C. POSITION RECOGNITION MODULE

Given the four levels of features extracted by the backbone network, the purpose of the Position Recognition Module (PRM) is to locate the polyp tissue and generate the initial prediction map. This paper adopts a partial decoder for localization. As described in the literature [39], the low-level features, due to their high resolution, typically require more computational resources but contribute less to performance improvement. In contrast, high-level features are rich in semantic information and can accurately determine the location of the polyps. Therefore, as shown in Figure 4, the PRM module integrates the high-level features $\{\mathbf{F}_4, \mathbf{F}_3, \mathbf{F}_2\}$ to calculate the initial polyp segmentation map. Specifically, to improve computational efficiency, we first use a CR convolutional block to adjust the channel numbers of the three feature maps to 32, resulting in the adjusted feature maps $\{\mathbf{RF}_4, \mathbf{RF}_3, \mathbf{RF}_2\}$. Next, we use upsampling and

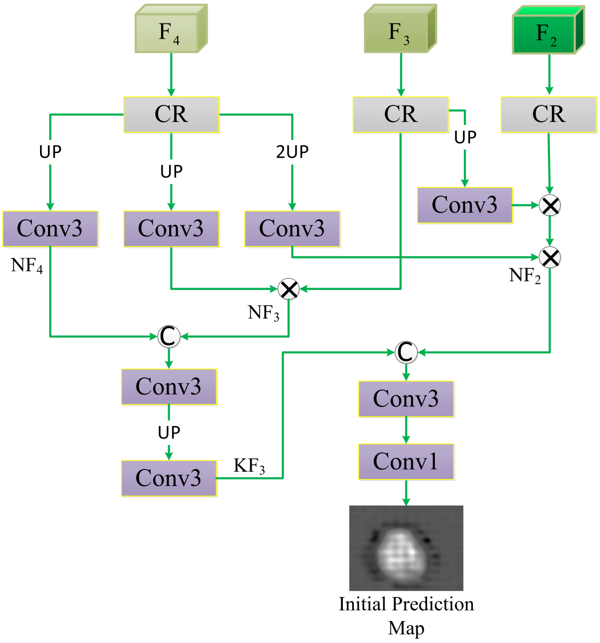


**FIGURE 4.** **Position recognition module.**

element-wise multiplication operations to reduce the discrepancies among the multi-level features, obtaining the feature maps $\{\mathbf{NF}_4, \mathbf{NF}_3, \mathbf{NF}_2\}$, which are formalized as follows:

$$\mathbf{NF}_4 = \text{Conv3}(\text{UP}(\mathbf{RF}_4)) \tag{9}$$
$$\mathbf{NF}_3 = \text{Conv3}(\text{UP}(\mathbf{RF}_4)) \otimes \mathbf{RF}_3 \tag{10}$$
$$\mathbf{NF}_2 = \text{Conv3}(\text{UP}(\text{UP}(\mathbf{RF}_4))) \otimes \text{Conv3}(\text{UP}(\mathbf{RF}_3)) \otimes \mathbf{RF}_2 \tag{11}$$

Next, through upsampling and concatenation operations, the initial prediction map $\mathbf{P}$ is generated. The formal representation is as follows:

$$\mathbf{KF}_3 = \text{Conv3}(\text{UP}(\text{Conv3}(\text{Concat}(\mathbf{NF}_4, \mathbf{NF}_3)))) \tag{12}$$
$$\mathbf{P} = \text{Conv1}(\text{Conv3}(\text{Concat}(\mathbf{KF}_3, \mathbf{NF}_2))) \tag{13}$$

### D. FOREGROUND-BACKGROUND REFINEMENT MODULE

The primary function of the FBRM module is to refine the prediction results of cross-level fused features. Given the different roles of the foreground and background in saliency prediction, we use foreground focus maps and background focus maps to refine the cross-level fused features, and then combine these two refined results for saliency prediction. As shown in Figure 5, the foreground focus map is the output obtained by applying a sigmoid function to the upper-level prediction map, denoted as $\mathbf{FM}_i$. The background focus map is obtained by subtracting the foreground focus map from an all-ones matrix and is denoted as $\mathbf{BM}_i$. The formal representation is as follows:

$$\mathbf{FM}_i = \text{UP}(\text{Sigmoid}(\mathbf{PM}_{i+1})) \tag{14}$$
$$\mathbf{BM}_i = \mathbf{E} - \mathbf{FM}_i \tag{15}$$

where $\mathbf{E}$ is an all-ones matrix, Sigmoid() is the Sigmoid function. $\mathbf{PM}_{i+1}$ is the prediction map of the $i+1$ level.

Next, we expand the foreground focus map and background focus map along the channel dimension to match the size of the $i$-th cross-level fused feature $\mathbf{F}_i^{\text{CAFM}}$. Then, we perform element-wise multiplication between these focus maps and the cross-level fused feature $\mathbf{F}_i^{\text{CAFM}}$ to obtain the foreground focus feature $\mathbf{FF}_i$ and background focus feature $\mathbf{BF}_i$, respectively. The formal representation is as follows:

$$\mathbf{FF}_i = \mathbf{F}_i^{\text{CAFM}} \otimes \text{Expand}(\mathbf{FM}_i) \tag{16}$$
$$\mathbf{BF}_i = \mathbf{F}_i^{\text{CAFM}} \otimes \text{Expand}(\mathbf{BM}_i) \tag{17}$$

Humans can achieve precise object recognition through careful observation. This observational process typically involves comparing the texture and semantic features of the object region with suspected regions, ultimately making a judgment. This paper simulates this process by constructing a context reasoning module, which performs context reasoning on both foreground focus features and background focus features. The foreground focus features primarily concentrate on the foreground region, using foreground cues to focus more on uncertain areas within the foreground, thus making more accurate judgments. Correspondingly, the background focus features mainly concentrate on the background region, utilizing background cues to identify polyp pixels in the background.

As shown in Figure 2, in our PDLFBR-Net model, the decoding process involves progressively predicting and upsampling from higher to lower levels. During this process, the resolution of the prediction map increases, but the boundaries of polyp become increasingly blurred. The foreground focus and background focus methods can effectively address these blurred boundary regions, refining them to obtain a clear polyp segmentation map.

As shown in Figure 6, the Context Reasoning Module (CRM) consists of three reasoning branches. Each reasoning branch includes a $1 \times 1$ convolution for channel reduction, an $\text{h}_k \times 1$ convolution and a $1 \times \text{h}_k$ convolution for local feature perception, and a $3 \times 3$ dilated convolution with a dilation rate of $\text{d}_k$ for context exploration. Here, $\text{h}_k$ is set to 3, 5, and 7, and $\text{d}_k$ is set to 2, 4, and 8 for $k \in \{1, 2, 3\}$, respectively. Next, the input feature map $\mathbf{F}_{\text{input}}$ first undergoes a $1 \times 1$ convolution for channel reduction, producing a residual branch. The residual branch is then concatenated with the outputs of the three reasoning branches along the channel dimension, followed by a $3 \times 3$ convolution for feature fusion, ultimately generating the output $\mathbf{F}_{\text{output}}$ of the CRM. This design allows the CRM module to achieve a larger receptive field, enabling it to perceive rich contextual information. The formal representation is as follows:

$$\mathbf{F}_{\text{output}} = \text{Conv3}(\text{Concat}(\text{Conv1}(\mathbf{F}_{\text{input}}), \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)) \tag{18}$$
$$\mathbf{b}_k = \text{DConv3}_{\text{d}_k}(\text{Conv}_{\text{h}_k \times 1}(\text{Conv}_{1 \times \text{h}_k}(\text{Conv1}(\mathbf{F}_{\text{input}})))) \quad k = \{1, 2, 3\} \tag{19}$$

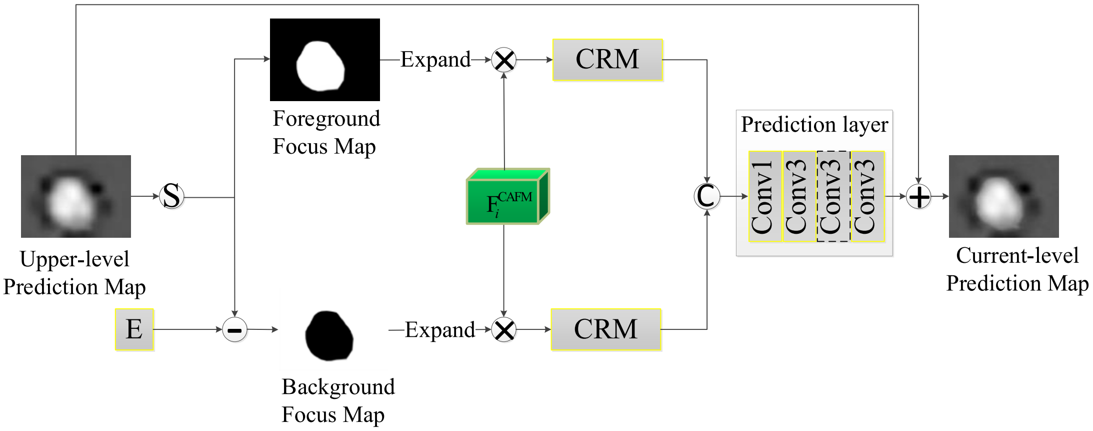


**FIGURE 5.** Foreground-background refinement module.

where $\mathbf{b}_k$ represents the inference result of the $k$-th reasoning branch, $\mathrm{Convh}_{h_k \times 1}()$ denotes the convolution operation with a kernel size of $h_k \times 1$, and $\mathrm{DConv3}_{d_k}()$ represents the dilated convolution with a kernel size of $3 \times 3$ and a dilation rate of $d_k$.

Finally, the foreground focus features and background focus features are processed separately by the context reasoning module. The results of these two inferences are then concatenated, and the concatenated result is processed by the prediction layer. The processed result is then combined with the upper-level prediction map $\mathbf{PM}_{i+1}$ through element-wise addition, ultimately yielding the current level prediction map $\mathbf{PM}_i$. The initial prediction map corresponds to the fourth level prediction map $\mathbf{PM}_4$.

As shown in Figure 5, the prediction layer consists of four convolutional layers. The first convolutional layer (Conv1) reduces the number of channels by half, while the last convolutional layer (Conv3) adjusts the number of channels to 1, resulting in the prediction layer output. It is noteworthy that, to save computational cost, the prediction layer in the $\mathrm{FBRM}_1$ module includes only three convolutional layers; thus, the third convolutional layer is depicted with a dashed line in the figure.

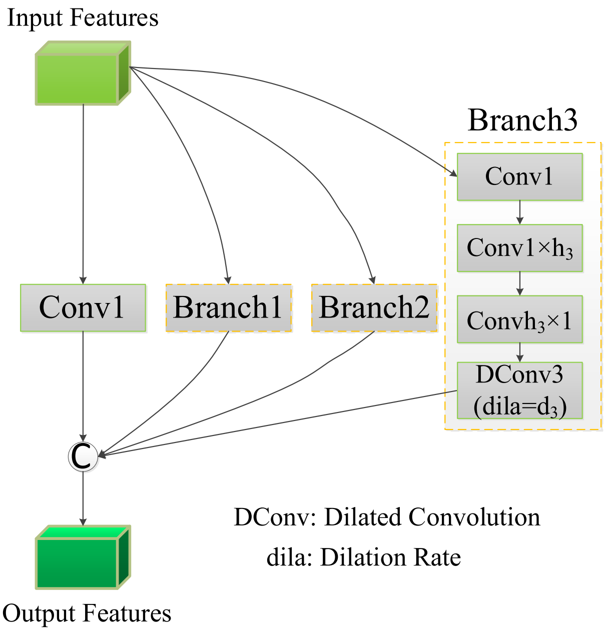


**FIGURE 6.** Context reasoning module.

### E. LOSS FUNCTION

Our loss function is defined as follows:

$$\mathrm{L} = \mathrm{L_P} + \mathrm{L_{PM}} \tag{20}$$

Here, $\mathrm{L_P}$ is the initial prediction map loss function, and $\mathrm{L_{PM}}$ is the prediction map loss function. $\mathrm{L_P}$ is used to calculate the loss between the initial prediction map $\mathbf{P}$ generated by the PRM module and the ground truth map $\mathbf{GT}$. $\mathrm{L_{PM}}$, on the other hand, is used to calculate the loss between the three prediction maps $\{\mathbf{PM}_1, \mathbf{PM}_2, \mathbf{PM}_3\}$ generated by the FBRM module and the ground truth map $\mathbf{GT}$. Their definitions are as follows:

$$\mathrm{L_P} = \mathrm{L_{wIoU}}(\mathbf{P}, \mathbf{GT}) + \mathrm{L_{wBCE}}(\mathbf{P}, \mathbf{GT}) \tag{21}$$

$$\mathrm{L_{PM}} = \sum_{i=1}^{3} (\mathrm{L_{wIoU}}(\mathbf{PM}_i, \mathbf{GT}) + \mathrm{L_{wBCE}}(\mathbf{PM}_i, \mathbf{GT})) \tag{22}$$

Here, $\mathrm{L_{wIoU}}$ represents weighted intersection over union loss function. $\mathrm{L_{wBCE}}$ represents weighted binary cross-entropy loss function [40].

## IV. EXPERIMENTS

In this section, we provide a detailed description of the datasets, implementation details, and evaluation metrics. We compare our model with existing state-of-the-art (SOTA) models in terms of learning ability, generalization capability, and qualitative results. Additionally, we present the comparison results and validate the effectiveness of each module through ablation experiments.

### A. DATASETS

In our experiments, we used five polyp datasets, including Kvasir-SEG [41], EndoScene [42], CVC-ColonDB [43], CVC-ClinicDB [44], and ETIS-Larib [45].

Kvasir-SEG is an open-access dataset of gastrointestinal polyp images and their corresponding segmentation masks. These data are manually annotated by doctors and verified by experienced gastroenterology experts. EndoScene is a cross-domain dataset covering various lesions of the digestive tract, making it an important resource for lesion detection and segmentation research. CVC-ColonDB is an endoscopic image dataset focused on polyps found during colonoscopy. This dataset includes 380 different images, covering various types of polyp appearances, and is used to evaluate the performance of polyp detection and segmentation algorithms. CVC-ClinicDB is an endoscopic image dataset provided by the Central Vision Center (CVC) in Barcelona, Spain, containing a large number of polyp images. This dataset is widely used for training and evaluating polyp segmentation algorithms, featuring good representativeness and diversity. ETIS-Larib is a high-resolution endoscopic image dataset specifically designed for polyp detection and segmentation. This dataset features high-quality images and precise annotations, making it suitable for complex polyp segmentation tasks. The detailed information of these datasets is summarized in Table 1.

**TABLE 1. The five datasets used in this work.**

| Datasets | Year | Number | Resolution |
|---|---|---|---|
| Kvasir-SEG | 2020 | 1000 | 332 × 487~<br>1920 × 1072 |
| EndoScene | 2017 | 60 | 500 × 574 |
| CVC-ColonDB | 2012 | 380 | 500 × 574 |
| CVC-ClinicDB | 2015 | 612 | 384 × 288 |
| ETIS-Larib | 2014 | 196 | 1225 × 966 |

### B. IMPLEMENTATION DETAILS

We adopted PVTv2 [38] as the backbone network for PDLFBR-Net and implemented our model in PyTorch. All models were trained and tested on a system with a single NVIDIA RTX 4090 GPU with 24 GB of memory. All images were resized to 352 × 352, and each batch contained 16 images. Random rotation, random scaling, horizontal flipping, and vertical flipping were used as data augmentation strategies. We optimized the parameters using the Adam optimizer with an initial learning rate of 1e-4. In the Kvasir-SEG and CVC-ClinicDB datasets, we randomly selected 80% of the images for training, 10% for validation, and 10% for testing. All images from the remaining three datasets were used exclusively for testing.

### C. EVALUATION METRICS

This paper utilizes seven commonly used evaluation metrics, including the mean Dice coefficient (meanDice) [46], the mean intersection over union (meanIoU), the mean absolute error (MAE), the weighted F-measure (wFm) [47], the S-measure (Sm) [48], the mean E-measure (meanEm), and the maximum E-measure (maxEm) [49]. Among these, a smaller MAE value indicates better performance, while larger values for the other metrics suggest better segmentation quality.

### D. COMPARISON WITH STATE-OF-THE-ART METHODS

To fully verify the effectiveness of our proposed PDLFBR-Net model, we compared it with twelve existing polyp segmentation methods, including Unet [12], UNet++ [26], PraNet [50], PolypPVT [34], SANet [31], UACANet [51], CCLDNet [52], CFANet [33],M2SNet [53], CAFE-Net [36], MLFF-Net [54] and MEGANet [55].

#### 1) QUANTITATIVE COMPARISON

To validate the performance of the model, we conducted quantitative comparative experiments on five datasets. The Kvasir-SEG and CVC-ClinicDB datasets are ones the model has been exposed to during the training phase (known datasets). Parts of these datasets have been used to train the model, so the model's performance on these datasets can reflect its learning capability. On the other hand, the EndoScene, CVC-ColonDB, and ETIS-Larib datasets are entirely new (unknown datasets) and were not seen by the model during training. By testing the model's performance on these datasets, we can evaluate its generalization ability, i.e., its capability to handle unseen data. By testing on both types of datasets, we can comprehensively assess the model's performance in both known and unknown environments.

Tables 2 and 3 respectively present the quantitative comparison results of the proposed PDLFBR-Net model and twelve state-of-the-art models on the Kvasir-SEG and CVC-ClinicDB datasets. For ease of comparison, the best two results are highlighted in red and blue. As shown in the tables, our PDLFBR-Net model outperforms other models across all evaluation metrics. For instance, compared to the classic U-Net, the meanDice and meanIoU values on the Kvasir-SEG dataset are approximately 0.119 and 0.149 higher, respectively; on the CVC-ClinicDB dataset, the meanDice and meanIoU values are approximately 0.126 and 0.148 higher, respectively. Similar advantages are observed for other metrics.

Moreover, compared to the latest state-of-the-art model CAFE-Net, our PDLFBR-Net model excels across all evaluation metrics. For example, on the Kvasir-SEG dataset, the meanDice and meanIoU values are approximately 0.004 and 0.006 higher, respectively; on the CVC-ClinicDB dataset, the meanDice and meanIoU values are approximately 0.006 and 0.004 higher, respectively. Similar advantages are observed for other metrics as well. The new method employs the powerful PVT as the backbone network to extract features and utilizes well-designed modules to locate polyp positions and refine polyp boundaries, resulting in superior performance.

To validate the generalization ability of the PDLFBR-Net model, we conducted quantitative comparison experiments

**TABLE 2.** Quantitative comparison results on the Kvasir-SEG dataset. ↑ indicates that higher values are better, while ↓ indicates that lower values are better. The best two results are highlighted in red and blue, respectively.

| Methods | meanDice ↑ | meanIoU ↑ | wFm ↑ | Sm ↑ | meanEm ↑ | maxEm ↑ | MAE ↓ |
|---|---|---|---|---|---|---|---|
| Unet(2015) | 0.818 | 0.746 | 0.794 | 0.858 | 0.881 | 0.893 | 0.055 |
| UNet++(2018) | 0.821 | 0.744 | 0.808 | 0.862 | 0.886 | 0.909 | 0.048 |
| PraNet(2020) | 0.898 | 0.841 | 0.885 | 0.915 | 0.944 | 0.948 | 0.030 |
| PolypPVT(2021) | 0.917 | 0.864 | 0.911 | 0.925 | 0.956 | 0.962 | 0.023 |
| SANet(2021) | 0.904 | 0.847 | 0.892 | 0.915 | 0.949 | 0.953 | 0.028 |
| UACANet(2021) | 0.905 | 0.852 | 0.897 | 0.914 | 0.949 | 0.951 | 0.026 |
| CCLDNet(2022) | 0.909 | 0.859 | 0.902 | 0.927 | 0.952 | 0.960 | 0.022 |
| CFANet(2023) | 0.915 | 0.862 | 0.903 | 0.924 | 0.956 | 0.962 | 0.023 |
| M2SNet(2023) | 0.906 | 0.850 | 0.899 | 0.923 | 0.948 | 0.956 | 0.026 |
| MEGANet(2024) | 0.911 | 0.859 | 0.904 | 0.916 | - | 0.954 | 0.026 |
| MLFF-Net(2024) | 0.919 | 0.866 | 0.908 | 0.925 | 0.958 | 0.962 | 0.023 |
| CAFE-Net(2024) | 0.933 | 0.889 | 0.927 | 0.939 | 0.967 | 0.971 | 0.019 |
| Ours | 0.937 | 0.895 | 0.934 | 0.943 | 0.969 | 0.982 | 0.018 |

**TABLE 3.** Quantitative comparison results on the CVC-ClinicDB dataset.

| Methods | meanDice ↑ | meanIoU ↑ | wFm ↑ | Sm ↑ | meanEm ↑ | maxEm ↑ | MAE ↓ |
|---|---|---|---|---|---|---|---|
| Unet(2015) | 0.823 | 0.755 | 0.811 | 0.890 | 0.913 | 0.953 | 0.019 |
| UNet++(2018) | 0.794 | 0.729 | 0.785 | 0.873 | 0.891 | 0.931 | 0.022 |
| PraNet(2020) | 0.899 | 0.849 | 0.896 | 0.937 | 0.963 | 0.979 | 0.009 |
| PolypPVT(2021) | 0.937 | 0.889 | 0.936 | 0.950 | 0.985 | 0.989 | 0.006 |
| SANet(2021) | 0.916 | 0.859 | 0.909 | 0.940 | 0.972 | 0.976 | 0.012 |
| UACANet(2021) | 0.916 | 0.870 | 0.917 | 0.940 | 0.965 | 0.968 | 0.008 |
| CCLDNet(2022) | 0.932 | 0.886 | 0.932 | 0.953 | 0.979 | 0.983 | 0.007 |
| CFANet(2023) | 0.933 | 0.883 | 0.924 | 0.951 | 0.981 | 0.989 | 0.007 |
| M2SNet(2023) | 0.916 | 0.867 | 0.915 | 0.946 | 0.971 | 0.979 | 0.009 |
| MEGANet(2024) | 0.930 | 0.885 | 0.931 | 0.950 | - | 0.980 | 0.008 |
| MLFF-Net(2024) | 0.943 | 0.897 | 0.944 | 0.952 | 0.988 | 0.992 | 0.006 |
| CAFE-Net(2024) | 0.943 | 0.899 | 0.941 | 0.957 | 0.986 | 0.991 | 0.006 |
| Ours | 0.949 | 0.903 | 0.951 | 0.960 | 0.987 | 0.994 | 0.006 |

with state-of-the-art models on the EndoScene, CVC-ColonDB, and ETIS-Larib datasets. The results are shown in Tables 4, 5, and 6, respectively. As can be seen from the tables, our method achieved the best performance on both the CVC-ColonDB and ETIS-Larib datasets. On the CVC-ColonDB dataset, the meanDice and meanIoU values of our method are 0.007 and 0.006 higher, respectively, than those of the second-ranked MLFF-Net model. On the EndoScene dataset, although our method did not achieve the best results in wFm, meanEm and maxEm, it still ranked second.

### 2) QUALITATIVE COMPARISON

The visual comparison of segmentation performance between PDLFBR-Net and other SOTA models is shown in Figure 7. PDLFBR-Net performs superiorly in various complex scenarios. For example, in small polyp images (rows 1 and 2), PDLFBR-Net can sensitively capture small polyp targets while effectively eliminating background noise, resulting in higher accuracy. This is due to the model's multi-scale feature extraction capability, which accurately captures fine-grained clues for target recognition. In complex background scenarios (rows 3 and 4), PDLFBR-Net can more precisely distinguish between foreground and background information. This is because our FBAM module infers from both foreground and background within low-level features, refining uncertain areas, and thereby achieving accurate polyp tissue segmentation. Additionally, in scenarios where the foreground and background are similar (row 5), uneven lighting (rows 6 and 7), and multiple polyps (row 8), PDLFBR-Net consistently performs outstandingly. These performance

**TABLE 4.** Quantitative comparison results on the EndoScene dataset.

| Methods | meanDice ↑ | meanIoU ↑ | wFm ↑ | Sm ↑ | meanEm ↑ | maxEm ↑ | MAE ↓ |
|---|---|---|---|---|---|---|---|
| Unet(2015) | 0.710 | 0.627 | 0.684 | 0.843 | 0.848 | 0.875 | 0.022 |
| UNet++(2018) | 0.707 | 0.624 | 0.687 | 0.839 | 0.834 | 0.898 | 0.018 |
| PraNet(2020) | 0.871 | 0.797 | 0.843 | 0.925 | 0.950 | 0.972 | 0.010 |
| PolypPVT(2021) | 0.900 | 0.833 | 0.884 | 0.935 | 0.973 | 0.981 | 0.007 |
| SANet(2021) | 0.888 | 0.815 | 0.859 | 0.928 | 0.962 | 0.972 | 0.008 |
| UACANet(2021) | 0.903 | 0.837 | 0.886 | 0.934 | 0.974 | 0.976 | 0.006 |
| CCLDNet(2022) | 0.876 | 0.806 | 0.848 | 0.929 | 0.949 | 0.978 | 0.008 |
| CFANet(2023) | 0.893 | 0.827 | 0.875 | 0.938 | 0.962 | 0.978 | 0.008 |
| M2SNet(2023) | 0.897 | 0.832 | 0.880 | 0.940 | 0.969 | 0.985 | 0.007 |
| MEGANet(2024) | 0.887 | 0.818 | 0.863 | 0.924 | - | 0.959 | 0.009 |
| MLFF-Net(2024) | 0.904 | 0.841 | 0.891 | 0.939 | 0.969 | 0.977 | 0.007 |
| CAFE-Net(2024) | 0.901 | 0.834 | 0.882 | 0.939 | 0.971 | 0.981 | 0.006 |
| Ours | 0.906 | 0.841 | 0.887 | 0.941 | 0.973 | 0.983 | 0.006 |

**TABLE 5.** Quantitative comparison results on the CVC-ColonDB dataset.

| Methods | meanDice ↑ | meanIoU ↑ | wFm ↑ | Sm ↑ | meanEm ↑ | maxEm ↑ | MAE ↓ |
|---|---|---|---|---|---|---|---|
| Unet(2015) | 0.504 | 0.436 | 0.491 | 0.710 | 0.691 | 0.781 | 0.059 |
| UNet++(2018) | 0.481 | 0.408 | 0.467 | 0.693 | 0.680 | 0.764 | 0.061 |
| PraNet(2020) | 0.712 | 0.640 | 0.699 | 0.820 | 0.847 | 0.872 | 0.043 |
| PolypPVT(2021) | 0.808 | 0.727 | 0.795 | 0.865 | 0.913 | 0.919 | 0.031 |
| SANet(2021) | 0.752 | 0.669 | 0.725 | 0.837 | 0.867 | 0.875 | 0.043 |
| UACANet(2021) | 0.783 | 0.704 | 0.772 | 0.848 | 0.894 | 0.897 | 0.034 |
| CCLDNet(2022) | 0.781 | 0.706 | 0.765 | 0.859 | 0.888 | 0.893 | 0.031 |
| CFANet(2023) | 0.743 | 0.665 | 0.728 | 0.835 | 0.869 | 0.898 | 0.039 |
| M2SNet(2023) | 0.753 | 0.676 | 0.735 | 0.843 | 0.871 | 0.879 | 0.038 |
| MEGANet(2024) | 0.781 | 0.706 | 0.766 | 0.845 | - | 0.899 | 0.038 |
| MLFF-Net(2024) | 0.820 | 0.742 | 0.805 | 0.870 | 0.914 | 0.917 | 0.035 |
| CAFE-Net(2024) | 0.820 | 0.740 | 0.803 | 0.874 | 0.914 | 0.918 | 0.026 |
| Ours | 0.827 | 0.748 | 0.810 | 0.876 | 0.919 | 0.923 | 0.025 |

results fully demonstrate the superiority of PDLFBR-Net in various complex scenarios.

### *E. ABLATION STUDY*

In this section, we tested the various modules of our model on five datasets to gain a deeper understanding of our model's performance. The three candidate modules are the Cross-level Attention-enhanced Fusion Module (CAFM), the Position Recognition Module (PRM), and the Foreground-Background Refinement Module (FBRM). In Tables 7 and 8, the results for different configurations are presented as follows: the #1 row shows the experimental results after removing the CAFM module, the #2 row shows the results after removing the PRM module, the #3 row shows the results after removing the FBRM module, and the #4 row shows the experimental results of the complete PDLFBR-Net model.

From the data in the tables, it is evident that the results in the #4 row are significantly better than those in the #3 row, with the meanDice and meanIoU metrics on the CVC-ClinicDB dataset improving by 0.036 and 0.049 respectively, and the structured measure Sm increasing by 0.033. These improvements indicate that the inclusion of the Foreground-Background Refinement Module helps the model to more accurately distinguish polyp tissues, resulting in clearer segmentation maps. Additionally, the results in the #4 row are also better than those in the #1 and #2 rows, further proving the importance of the CAFM and PRM modules.

**TABLE 6.** Quantitative comparison results on the ETIS-Larib dataset.

| Methods | meanDice ↑ | meanIoU ↑ | wFm ↑ | Sm ↑ | meanEm ↑ | maxEm ↑ | MAE ↓ |
|---|---|---|---|---|---|---|---|
| Unet(2015) | 0.398 | 0.335 | 0.366 | 0.684 | 0.643 | 0.740 | 0.036 |
| UNet++(2018) | 0.401 | 0.343 | 0.390 | 0.683 | 0.629 | 0.776 | 0.035 |
| PraNet(2020) | 0.628 | 0.567 | 0.600 | 0.794 | 0.808 | 0.841 | 0.031 |
| PolypPVT(2021) | 0.787 | 0.706 | 0.750 | 0.871 | 0.906 | 0.910 | 0.013 |
| SANet(2021) | 0.750 | 0.654 | 0.685 | 0.849 | 0.881 | 0.897 | 0.015 |
| UACANet(2021) | 0.694 | 0.616 | 0.650 | 0.816 | 0.848 | 0.851 | 0.023 |
| CCLDNet(2022) | 0.778 | 0.704 | 0.743 | 0.872 | 0.888 | 0.902 | 0.016 |
| CFANet(2023) | 0.733 | 0.655 | 0.693 | 0.846 | 0.881 | 0.892 | 0.014 |
| M2SNet(2023) | 0.742 | 0.666 | 0.709 | 0.853 | 0.875 | 0.886 | 0.017 |
| MEGANet(2024) | 0.789 | 0.709 | 0.753 | 0.866 | - | 0.915 | 0.015 |
| MLFF-Net(2024) | 0.784 | 0.707 | 0.750 | 0.866 | 0.891 | 0.903 | 0.025 |
| CAFE-Net(2024) | 0.822 | 0.738 | 0.775 | 0.898 | 0.917 | 0.940 | 0.014 |
| Ours | 0.824 | 0.745 | 0.781 | 0.899 | 0.921 | 0.952 | 0.013 |

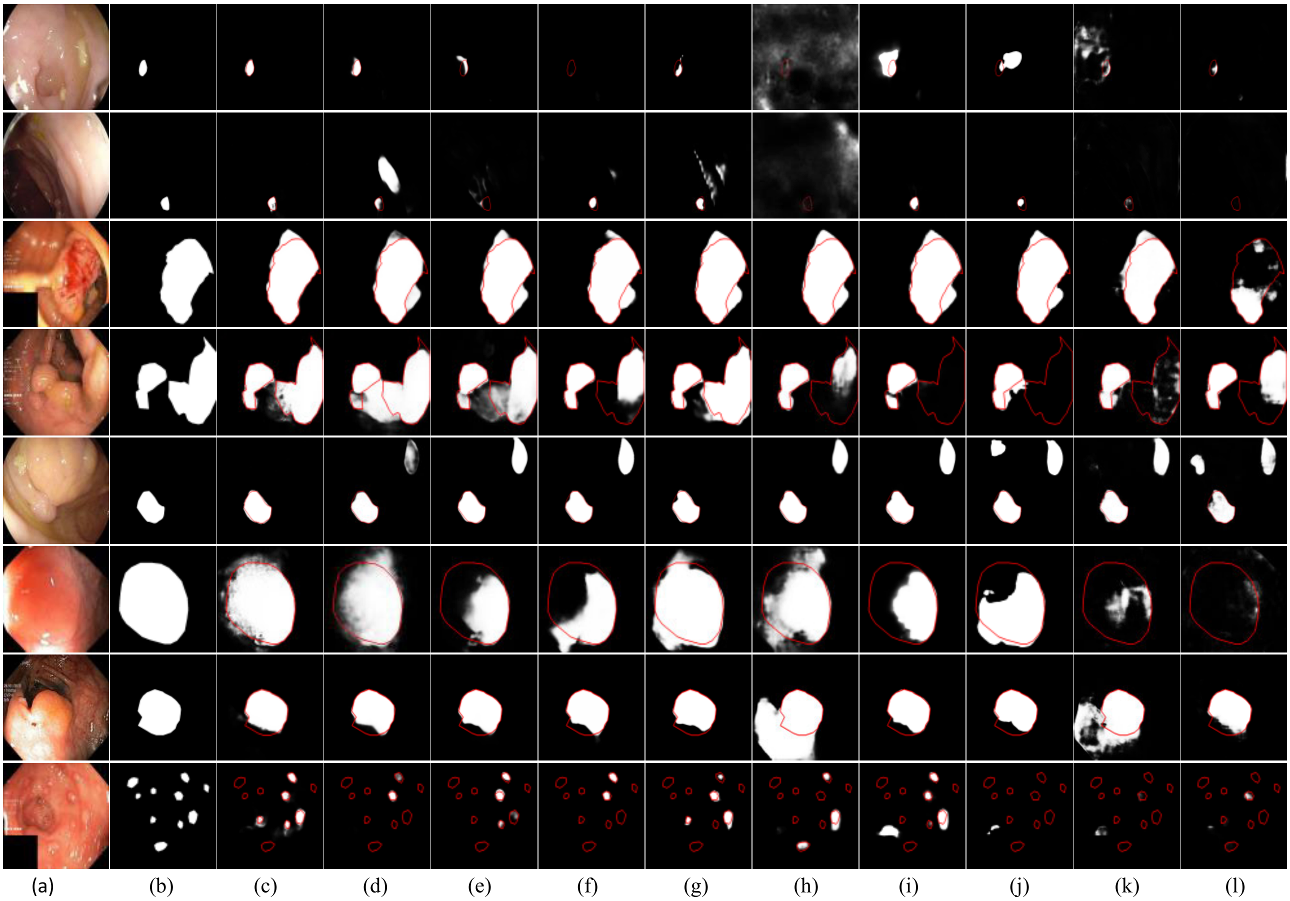


**FIGURE 7.** Visual comparison between PDLFBR-Net and the latest polyp segmentation models: (a) Images, (b) GT, (c) Ours, (d) CCLDNet, (e) CFANet, (f) M2SNet, (g) PolypPVT, (h) PraNet, (i) SANet, (j) UACANet, (k) UNet, (l) UNet++. Polyp regions are marked in white, non-polyp regions in black, and the ground truth polyp contour is annotated with a red line.

In the final row, we can see that the PDLFBR-Net model, which combines all three modules, achieves the best results.

The Grad-CAM heatmaps in Figure 8 clearly demonstrate the good performance of the model. Through these heatmaps,

**TABLE 7.** Ablation study results of different modules on CVC-ClinicDB and EndoScene Datasets. The best performance in each column is highlighted in bold.

| Candidates | | | | CVC-ClinicDB | | | EndoScene | | |
|---|---|---|---|---|---|---|---|---|---|
| # | CAFM | PRM | FBRM | meanDice ↑ | meanIoU ↑ | Sm ↑ | meanDice ↑ | meanIoU ↑ | Sm ↑ |
| 1 | - | √ | √ | 0.922 | 0.876 | 0.933 | 0.883 | 0.810 | 0.904 |
| 2 | √ | - | √ | 0.935 | 0.892 | 0.941 | 0.902 | 0.837 | 0.911 |
| 3 | √ | √ | - | 0.913 | 0.854 | 0.927 | 0.857 | 0.782 | 0.907 |
| 4 | √ | √ | √ | **0.949** | **0.903** | **0.960** | **0.906** | **0.841** | **0.941** |

**TABLE 8.** Ablation study results of different modules on Kvasir-SEG, CVC-ColonDB, and ETIS-Larib datasets.

| Candidates | | | | Kvasir-SEG | | | CVC-ColonDB | | | ETIS-Larib | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # | CAFM | PRM | FBRM | meanDice ↑ | meanIoU ↑ | Sm ↑ | meanDice ↑ | meanIoU ↑ | Sm ↑ | meanDice ↑ | meanIoU ↑ | Sm ↑ |
| 1 | - | √ | √ | 0.916 | 0.873 | 0.923 | 0.796 | 0.719 | 0.856 | 0.796 | 0.715 | 0.872 |
| 2 | √ | - | √ | 0.928 | 0.886 | 0.931 | 0.815 | 0.732 | 0.863 | 0.814 | 0.730 | 0.886 |
| 3 | √ | √ | - | 0.910 | 0.861 | 0.914 | 0.778 | 0.703 | 0.845 | 0.772 | 0.693 | 0.860 |
| 4 | √ | √ | √ | **0.937** | **0.895** | **0.943** | **0.827** | **0.748** | **0.876** | **0.824** | **0.745** | **0.899** |

**TABLE 9.** Ablation study results of different supervision methods. The best performance in each column is highlighted in bold.

| supervision method | Kvasir-SEG | | CVC-ClinicDB | | EndoScene | | CVC-ColonDB | | ETIS-Larib | |
|---|---|---|---|---|---|---|---|---|---|---|
| | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ |
| only LP | 0.901 | 0.852 | 0.902 | 0.831 | 0.851 | 0.771 | 0.773 | 0.683 | 0.789 | 0.704 |
| only LPM | 0.933 | 0.889 | 0.943 | 0.895 | 0.895 | 0.826 | 0.818 | 0.742 | **0.826** | 0.742 |
| ALL | **0.937** | **0.895** | **0.949** | **0.903** | **0.906** | **0.841** | **0.827** | **0.748** | 0.824 | **0.745** |

**TABLE 10.** Ablation study comparison results of different loss functions. The best performance in each column is highlighted in bold.

| Loss Function | Kvasir-SEG | | CVC-ClinicDB | | EndoScene | | CVC-ColonDB | | ETIS-Larib | |
|---|---|---|---|---|---|---|---|---|---|---|
| | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ |
| only BCE | 0.916 | 0.869 | 0.914 | 0.864 | 0.894 | 0.812 | 0.806 | 0.726 | 0.790 | 0.712 |
| only IoU | 0.932 | 0.888 | 0.924 | 0.881 | 0.891 | 0.818 | 0.813 | 0.737 | 0.803 | 0.728 |
| BCE+IoU | **0.937** | **0.895** | **0.949** | **0.903** | **0.906** | **0.841** | **0.827** | **0.748** | **0.824** | **0.745** |

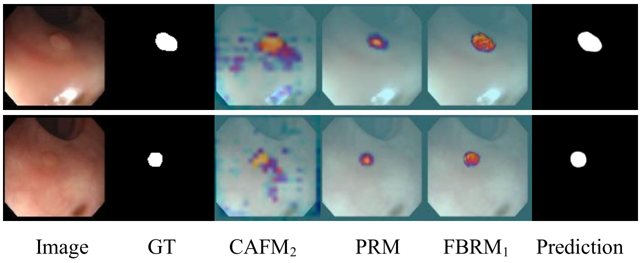


**FIGURE 8.** Grad-CAM heatmaps of $\text{CAFM}_2$, PRM, and $\text{FBRM}_1$ modules in the model.

we can see that the model accurately focuses on the critical areas of the images, indicating its excellent performance in identifying and capturing important features. This visualization not only validates the model's effectiveness but also further proves its potential in practical applications.

To assess the impact of different supervision methods on network performance, we conducted an ablation study. Table 9 presents the comparative results of this study. In the table, "only LP" indicates the results when only the initial prediction map is supervised, "only LPM" refers to the results when only the three prediction maps (PM1, PM2, PM3) generated by the FBRM module are supervised, and "ALL" represents the results when all the prediction maps are supervised simultaneously. As shown in Table 9, supervising only the initial prediction map or only the prediction maps leads to a decline in performance, while supervising both simultaneously yields the best results.

To validate the effectiveness of the proposed hybrid loss function, we conducted an ablation study [56]. Table 10 presents the comparative results of this ablation study. In the table, "only BCE" represents the results using only the weighted Binary Cross-Entropy (BCE) loss function, "only IoU" represents the results using only the weighted Intersection over Union (IoU) loss function, and "BCE+IoU" represents the results using our proposed hybrid loss function. As shown in Table 10, the overall performance of the IoU loss function is slightly better than that of the BCE loss function, while the hybrid loss function achieved the best results.

To further validate the necessity of the CRM module, we conducted an ablation study. Table 11 presents the comparative results of this study. "CRM1" represents the performance when only a single 3 × 3 convolution is used within the CRM module, while "CRM2" denotes the network performance when using dilated convolutions with

**TABLE 11.** Ablation study results of different CRM modules. The best performance in each column is highlighted in bold.

| CRM Module | Kvasir-SEG | | CVC-ClinicDB | | EndoScene | | CVC-ColonDB | | ETIS-Larib | |
|---|---|---|---|---|---|---|---|---|---|---|
| | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ | meanDice ↑ | meanIoU ↑ |
| CRM1 | 0.932 | 0.886 | 0.939 | 0.892 | 0.898 | 0.830 | 0.820 | 0.739 | 0.810 | 0.729 |
| CRM2 | 0.936 | 0.891 | **0.950** | 0.900 | 0.902 | 0.837 | 0.823 | 0.745 | 0.817 | 0.737 |
| CRM | **0.937** | **0.895** | 0.949 | **0.903** | **0.906** | **0.841** | **0.827** | **0.748** | **0.824** | **0.745** |

the same dilation rate (dilation rate of 2) and different kernel sizes (3 × 3, 5 × 5, and 7 × 7) in three branches. ‘‘CRM’’ represents the CRM module proposed in this paper. As shown in Table 11, CRM2 achieves comparable results to CRM. This is because, despite having a fixed dilation rate, CRM2’s increased kernel sizes allow for a larger receptive field, enabling the capture of contextual information over a broader range, thus resulting in good performance. In contrast, CRM1, which uses only a single 3 × 3 convolution, shows a significant performance drop compared to CRM.

## V. CONCLUSION

This paper presents a novel and efficient method for polyp segmentation, named the Partial Decoder Localization and Foreground-Background Refinement Network (PDLFBR-Net). By designing the Cross-level Attention-enhanced Fusion Module (CAFM), the Position Recognition Module (PRM), and the Foreground-Background Refinement Module (FBRM), PDLFBR-Net effectively simulates the human object recognition process, achieving high-precision polyp segmentation through global localization and refined processing. In the CAFM module, we introduce a mechanism that integrates features from adjacent levels, enhancing the model’s ability to capture cross-level semantic information, thereby improving its capacity to identify diverse polyp characteristics. The PRM module utilizes partial decoders to extract potential polyp regions from high-level features, effectively locating the target areas. The FBRM module further refines the foreground and background, enhancing segmentation boundary accuracy through contextual reasoning, significantly improving the precision of the final prediction. Extensive experiments on five challenging datasets demonstrate that our PDLFBR-Net model significantly outperforms existing state-of-the-art methods in terms of segmentation accuracy and generalization performance. This result validates the superiority and robustness of our method in handling complex polyp morphologies and backgrounds.

In future research, we will focus on optimizing the computational efficiency of PDLFBR-Net while maintaining high segmentation accuracy. Additionally, we will explore techniques such as self-supervised learning and domain adaptation to enhance the robustness of this method under different imaging conditions. Specifically, we plan to adopt the following strategies to optimize the computational efficiency of PDLFBR-Net:

1) Model Pruning and Quantization: By employing model pruning and quantization techniques, we aim to remove redundant parameters and connections in the network, reducing the model’s complexity and computational load. Through these techniques, we hope to significantly decrease the computational resource requirements without significantly affecting segmentation accuracy.

2) Lightweight Network Design: We will replace certain modules of PDLFBR-Net with more lightweight network structures. These lightweight networks can maintain high accuracy while offering lower computational costs and faster inference speeds.

3) Multi-Resolution Processing: We will implement a multi-resolution processing strategy, where images are initially processed at a low resolution to quickly obtain preliminary segmentation results. Subsequently, fine processing will be performed on local high-resolution areas. This approach can reduce the overall computational burden while ensuring segmentation accuracy.

Additionally, to enhance the robustness of PDLFBR-Net under different imaging conditions, we plan to conduct the following research:

1) Self-Supervised Learning: We will explore self-supervised learning techniques, leveraging large amounts of unlabeled medical images for pre-training. This can help us learn more feature representations and improve the model’s performance when labeled data is limited.

2) Domain Adaptation: We will investigate domain adaptation techniques by introducing domain-adversarial training or domain transfer learning. These techniques will enable the model to adapt to data distribution changes under different imaging devices and conditions, thereby enhancing the model’s robustness.

3) Data Augmentation Techniques: We will employ various data augmentation techniques, such as rotation, translation, scaling, and contrast adjustment, to generate diverse training data. This will enhance the model’s generalization ability under different imaging conditions.

## REFERENCES

[1] P. Wang, X. Xiao, J. R. Glissen Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang, X. Yang, L. Li, J. He, X. Yi, J. Liu, and X. Liu, “Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy,” *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 741–748, Oct. 2018.

[2] L. F. Sánchez-Peralta, L. Bote-Curiel, A. Picón, F. M. Sánchez-Margallo, and J. B. Pagador, “Deep learning to find colorectal polyps in colonoscopy: A systematic literature review,” *Artif. Intell. Med.*, vol. 108, Aug. 2020, Art. no. 101923.

[3] J. Mei, T. Zhou, K. Huang, Y. Zhang, Y. Zhou, Y. Wu, and H. Fu, “A survey on deep learning for polyp segmentation: Techniques, challenges and future trends,” 2023, *arXiv:2311.18373*.

[4] S. Y. Park, D. Sargent, I. Spofford, K. G. Vosburgh, and Y. A-Rahim, "A colon video analysis framework for polyp detection," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1408–1418, May 2012.

[5] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE Trans. Med. Imag.*, vol. 33, no. 7, pp. 1488–1502, Jul. 2014.

[6] Y. Peng, Z. Zhai, and M. Feng, "SLMSF-Net: A semantic localization and multi-scale fusion network for RGB-D salient object detection," *Sensors*, vol. 24, no. 4, p. 1117, Feb. 2024.

[7] R. Guan, Z. Li, W. Tu, J. Wang, Y. Liu, X. Li, C. Tang, and R. Feng, "Contrastive multiview subspace clustering of hyperspectral images based on graph convolutional networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5510514.

[8] S. Woo, J. Park, J.-Y. Lee, and I.-S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[9] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1949–1961, 2021.

[10] Y. Liu, Y. Yang, Y. Jiang, and Z. Xie, "Multi-view orientational attention network combining point-based affinity for polyp segmentation," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123663.

[11] T. Zhou, Y. Zhou, C. Gong, J. Yang, and Y. Zhang, "Feature aggregation and propagation network for camouflaged object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 7036–7047, 2022.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[14] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 225–2255.

[15] C. Zhai, L. Yang, Y. Liu, and H. Yu, "DBMA-Net: A dual-branch multi-attention network for polyp segmentation," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–16, 2024.

[16] T. Mahmud, B. Paul, and S. A. Fattah, "PolypSegNet: A modified encoder–decoder architecture for automated polyp segmentation from colonoscopy images," *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104119.

[17] M. Gupta and A. Mishra, "A systematic review of deep learning based image segmentation to detect polyp," *Artif. Intell. Rev.*, vol. 57, no. 1, pp. 1–10, Jan. 2024.

[18] K. Hu, L. Zhao, S. Feng, S. Zhang, Q. Zhou, X. Gao, and Y. Guo, "Colorectal polyp region extraction using saliency detection network with neutrosophic enhancement," *Comput. Biol. Med.*, vol. 147, Aug. 2022, Art. no. 105760.

[19] V. Sharma, A. Kumar, D. Jha, M.-K. Bhuyan, P.-K. Das, and U. Bagci, "ControlPolypNet: Towards controlled colon polyp synthesis for improved polyp segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 2325–2334.

[20] S. Xia, S.-M. Krishnan, M.-P. Tjoa, and P.-M.-Y. Goh, "A novel methodology for extracting colon's lumen from colonoscopic images," *J. Syst. Cybern. Inform.*, vol. 1, no. 2, pp. 7–12, 2003.

[21] Z. Wang, L. Li, J. Anderson, D.-P. Harrington, and Z. Liang, "Computer-aided detection and diagnosis of colon polyps with morphological and texture features," in *Medical Imaging 2004: Image Processing*. Bellingham, WA, USA: SPIE, 2004, pp. 972–979.

[22] A. Jerebko, S. Lakare, P. Cathier, S. Periaswamy, and L. Bogoni, "Symmetric curvature patterns for colonic polyp detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2006, pp. 169–176.

[23] S. Zhang, X. Zhang, S. Wan, W. Ren, L. Zhao, and L. Shen, "Generative adversarial and self-supervised dehazing network," *IEEE Trans. Ind. Informat.*, vol. 20, no. 3, pp. 4187–4197, Mar. 2024.

[24] S. Zhang, W. Ren, X. Tan, Z.-J. Wang, Y. Liu, J. Zhang, X. Zhang, and X. Cao, "Semantic-aware dehazing network with adaptive feature fusion," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 454–467, Jan. 2023.

[25] S.-H. Kassani, P.-H. Kassani, M.-J. Wesolowski, K.-A. Schneider, and R. Deters, "Automatic polyp segmentation using convolutional neural networks," in *Proc. Can. Conf. Artif. Intell.*, 2020, pp. 290–301.

[26] Z. Zhou, M.-M.-R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, 2018, pp. 3–11.

[27] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[28] J. M. Poorneshwaran, S. Santhosh Kumar, K. Ram, J. Joseph, and M. Sivaprakasam, "Polyp segmentation using generative adversarial network," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 7201–7204.

[29] D. Banik, K. Roy, D. Bhattacharjee, M. Nasipuri, and O. Krejcar, "Polyp-Net: A multimodel fusion network for polyp segmentation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[30] G. Yue, S. Li, T. Zhou, M. Wang, J. Du, Q. Jiang, W. Gao, T. Wang, and J. Lv, "Adaptive context exploration network for polyp segmentation in colonoscopy images," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 2, pp. 487–499, Apr. 2023.

[31] J. Wei, Y. Hu, R. Zhang, Z. Li, S.-K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 699–708.

[32] Z. Yin, K. Liang, Z. Ma, and J. Guo, "Duplex contextual relation network for polyp segmentation," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[33] T. Zhou, Y. Zhou, K. He, C. Gong, J. Yang, H. Fu, and D. Shen, "Cross-level feature aggregation network for polyp segmentation," *Pattern Recognit.*, vol. 140, Aug. 2023, Art. no. 109555.

[34] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," 2021, *arXiv:2108.06932*.

[35] W. Zhang, C. Fu, Y. Zheng, F. Zhang, Y. Zhao, and C.-W. Sham, "HSNet: A hybrid semantic network for polyp segmentation," *Comput. Biol. Med.*, vol. 150, Nov. 2022, Art. no. 106173.

[36] G. Liu, S. Yao, D. Liu, B. Chang, Z. Chen, J. Wang, and J. Wei, "CAFE-Net: Cross-attention and feature exploration network for polyp segmentation," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121754.

[37] G. Yue, G. Zhuo, W. Yan, T. Zhou, C. Tang, P. Yang, and T. Wang, "Boundary uncertainty aware network for automated polyp segmentation," *Neural Netw.*, vol. 170, pp. 390–404, Feb. 2024.

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

[39] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3902–3911.

[40] J. Wei, S. Wang, and Q. Huang, "F$^3$net: Fusion, feedback and focus for salient object detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 12321–12328, Apr. 2020.

[41] D. Jha, P.-H. Smedsrud, M.-A. Riegler, P. Halvorsen, T.-D. Lange, D. Johansen, and H.-D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Model., Jan.*, 2020, pp. 451–462.

[42] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthcare Eng.*, vol. 1, pp. 1–9, Jun. 2017.

[43] J. Bernal, J. Sánchez, and F. Vilariño, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, Sep. 2012.

[44] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. Saliency maps from physicians," *Computerized Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.

[45] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 9, no. 2, pp. 283–293, Mar. 2014.

[46] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[47] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

[48] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.

[49] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," 2018, *arXiv:1805.10421*.

[50] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2020, pp. 263–273.

[51] T. Kim, H. Lee, and D. Kim, "UACANet: Uncertainty augmented context attention for polyp segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2167–2175.

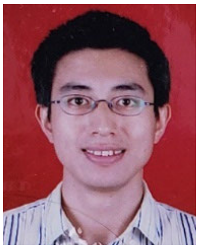
[52] H. Yang, Q. Chen, K. Fu, L. Zhu, L. Jin, B. Qiu, Q. Ren, H. Du, and Y. Lu, "Boosting medical image segmentation via conditional-synergistic convolution and lesion decoupling," *Computerized Med. Imag. Graph.*, vol. 101, Oct. 2022, Art. no. 102110.

[53] X. Zhao, H. Jia, Y. Pang, L. Lv, F. Tian, L. Zhang, W. Sun, and H. Lu, "M$^2$SNet: Multi-scale in multi-scale subtraction network for medical image segmentation," 2023, *arXiv:2303.10894*.

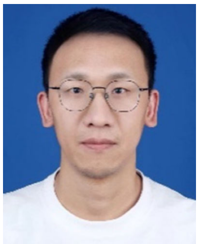
[54] J. Liu, Q. Chen, Y. Zhang, Z. Wang, X. Deng, and J. Wang, "Multi-level feature fusion network combining attention mechanisms for polyp segmentation," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102195.

[55] N.-T. Bui, D.-H. Hoang, Q.-T. Nguyen, M.-T. Tran, and N. Le, "MEGANet: Multi-scale edge-guided attention network for weak boundary polyp segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 7985–7994.

[56] S. Rezvani, M. Fateh, and H. Khosravi, "ABANet: Attention boundary-aware network for image segmentation," *Expert Syst.*, vol. 41, no. 9, pp. 1–26, Sep. 2024.



**YANBIN PENG** (Member, IEEE) received the Ph.D. degree from Zhejiang University, China, in 2008. He is currently an Associate Professor with Zhejiang University of Science and Technology. His current research interests include computer vision, image processing, deep learning, object detection, and their applications.



**MINGKUN FENG** received the Ph.D. degree in information and communication engineering from Nanjing University of Posts and Telecommunications, China, in 2016. He is currently an Associate Professor with Zhejiang University of Science and Technology, China. His research interests include pattern recognition, machine learning, and artificial intelligence.

**ZHINIAN ZHAI** received the Ph.D. degree from South China University of Technology. He has experience working as a C++ Developer at Guangdong Beidian Communication Equipment Company Ltd. He is currently a Lecturer with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology. His research interests include information security, machine learning, and deep learning.

**ZHIJUN ZHENG** received the Ph.D. degree from Xi'an Jiaotong University, China. He is currently an Associate Professor with Zhejiang University of Science and Technology. His research interests include machine learning, multimedia analysis retrieval, image retrieval, and statistical learning.