Dual-domain Feature Enhancement and Boundary-guided Multi-branch Attention Network for Polyp Segmentation in colonoscopy images

Sisi Chen, Chaoli Wang*,Zhanquan Sun†
Institute of Optical-Electrical and Computer Engineering
University of Shanghai for Science and Technology
Shanghai, China
{3041825557@qq.com, clwang@usst.edu.cn, sunzhq@usst.edu.cn}

***Abstract*—In recent years, deep learning methods have made significant progress in polyp image segmentation tasks. However, effectively modeling the complementary relationship between spatial and frequency domain features within a unified framework and accurately describing the complex and blurred polyp boundaries remains a challenging task. Most existing methods focus on spatial domain feature learning, and even when frequency domain information is introduced, the differences in spectral energy distribution across different feature layers are often neglected. Meanwhile, cross-layer feature fusion and region-specific modeling are insufficient, and the high-level semantic information is not effectively used to guide low-level spatial features, lacking differentiated modeling mechanisms for foreground, background, and boundary regions. To address these issues, we propose a Dual-domain Feature Enhancement and Boundary-guided Multi-branch Attention Network (DBMA-Net) for polyp image segmentation. The network first introduces a Multi-scale Context and Attention Fusion Module (MCAF) to enhance the multi-scale structural and texture representation of shallow features. Next, a Dual-domain Feature Enhancement Module (DFE) is designed to strengthen the expression of deep features by collaboratively modeling spatial semantics and frequency domain global texture information. In addition, a Boundary-guided Multi-branch Attention Module (BMAM) is proposed, which adaptively focuses on foreground, background, and boundary regions using high-level prediction results, thereby improving the ability to describe complex boundaries. Experimental results on five publicly available polyp segmentation datasets demonstrate that the proposed DBMA-Net outperforms existing methods across various evaluation metrics.**

***Index Terms*—polyp segmentation, medical image segmentation, dual-domain feature learning, boundary-guided attention**

I. Introduction

Colorectal cancer (CRC) is one of the most prevalent cancers world- wide, ranking second in both incidence and mortality, posing a serious threat to public health [1]. Therefore, early detection and removal of polyps are crucial for preventing CRC. Colonoscopy, as the "gold standard" for the diagnosis and treatment of CRC, allows for direct observation of intestinal mucosa and identification of potential lesions. However, this process is highly dependent on the experience of doctors and is affected by factors such as the diverse shapes of polyps, blurred boundaries, and their high similarity to surrounding mucosa [2], posing risks of false positives and missed detections. Therefore, developing a computer-aided system capable of automatically and accurately segmenting polyps in complex scenes is of great value in reducing the burden on doctors and improving the early diagnosis and treatment of CRC.

Early methods for polyp segmentation relied on manually designed low-level features such as color, texture, and shape [3], [4]. However, due to the limited expressive power of these features, they struggled to adapt to complex polyp shapes and blurred boundaries, resulting in weak generalization capabilities. With the development of deep learning, methods based on Convolutional Neural Networks (CNNs) have gradually become mainstream. U-Net [5] has become a classic framework through the effective fusion of multi-layer features, and its segmentation performance has been further improved through deeper network structures and attention mechanisms [6], [7]. Nevertheless, the local receptive field of convolution limits its ability to model global context. To address this, Transformer-based methods such as Swin-UNet [8] and Polyp-PVT [9] have been introduced, which can better capture long-range dependencies but still suffer from deficiencies in detail and boundary delineation. Recent CNN–Transformer hybrid architectures, such as TransFuse [10], have achieved better segmentation results in complex scenarios by combining the advantages of both.

The aforementioned methods have made certain progress in spatial domain feature modeling and global semantic modeling, but most of them still primarily rely on spatial domain representation, with relatively limited utilization of frequency domain information [7], [9]. Existing research has shown that frequency domain features possess unique advantages in characterizing global structure, periodic texture, and edge variations, which are helpful in distinguishing polyps from the background and mitigating low contrast and noise interference

This work was supported by the National Natural Science Foundation of China under Grant 62173232 and Grant 62003214. *†Corresponding authors: Chaoli Wang (clwang@usst.edu.cn) and Zhanquan Sun (sunzhq@usst.edu.cn).

[11], [12]. However, in current polyp segmentation methods, frequency domain information is often introduced only in a coarse-grained manner as an auxiliary form, such as through Fourier transform or spectral filtering to enhance features [12], [13]. This approach often overlooks the differences in feature spectral distribution across different semantic levels and fails to effectively distinguish between the distinct roles of low-frequency semantic structures and high-frequency boundary details. Furthermore, features in the frequency and spatial domains are often modeled in separate branches, lacking an effective collaborative interaction mechanism, which limits their discriminative ability in complex scenarios. Therefore, achieving collaborative modeling of spatial and frequency domain features within a unified framework and fully exploiting multi-level spectral information remain of great significance for enhancing the performance of colorectal polyp segmentation.

In addition to the inadequate collaborative modeling between frequency and spatial domains, the insufficient cross-layer feature fusion and regional discriminative modeling are also significant factors limiting the performance of existing polyp segmentation methods. In colonoscopic images, polyp regions often exhibit characteristics such as blurred boundaries, irregular shapes, and high similarity to surrounding mucosal tissues, leading to significant overlap between foreground, background, and boundary regions in the feature space. Although some methods enhance boundary perception by introducing boundary constraints or explicit edge supervision [14], [15], these methods often rely on fixed forms of boundary representation, making it difficult to adapt to the complex boundary variations of polyps with different scales and shapes. On the other hand, most existing segmentation networks tend to uniformly process the entire feature map during feature modeling, lacking a differentiated modeling mechanism for foreground, background, and boundary regions. This makes it difficult for the network to effectively distinguish between real polyp regions and background interference when facing low-contrast areas or uncertain boundaries, leading to issues such as boundary shifts or local discontinuities. Therefore, there is still a lack of effective solutions for fully utilizing high-level semantic guidance in the multi-layer feature interaction process to achieve adaptive attention to different regions.

To address the above issues, we propose a polyp segmentation network (DBMA-Net) that combines dual-domain feature enhancement and boundary-guided multi-branch attention. The network uses Pyramid Vision Transformer (PVT) as the encoder to model long-range dependencies and extract multi-scale semantic representations. A cascaded feature decoder (CFM) aggregates high-level features to generate initial globally consistent predictions. To tackle the challenges of scale variation and background interference in shallow features, we design a Multi-scale Context and Attention Fusion Module (MCAF) to enhance the discriminative ability of low-level features. Additionally, the Dual-domain Feature Enhancement (DFE) module adaptively divides frequency components and jointly models spatial and frequency-domain information, suppressing redundant low-frequency interference while preserving key polyp structure and boundary-related information. Finally, we propose a Boundary-guided Multi-branch Attention Module (BMAM) to fine-tune the modeling of foreground, background, and boundary regions under the guidance of high-level predictions, improving segmentation accuracy and robustness in complex boundary regions. The main contributions of this work include:

- We propose the DBMA-Net, which alleviates the issues of insufficient frequency domain modeling and inadequate cross-layer feature fusion and region discrimination modeling in existing methods.
- We have designed a Multi-scale Context and Attention Fusion (MCAF) module, which enhances the structural expression ability of shallow features through multi-receptive field context modeling and adaptive attention mechanism, effectively preserving fine-grained details and weak boundary information.
- We designed a Dual-domain Feature Enhancement (DFE) module to explicitly model the complementary relationship between spatial and frequency domain features. By adaptively partitioning frequency components based on spectral energy distribution, we achieve the enhancement of deep features.
- We designed a Boundary-guided Multi-branch Attention Module (BMAM), which utilizes high-level predictions as semantic priors to perform differentiated modeling for foreground, background, and boundary regions, and enhances boundary perception through dynamic attention fusion.

II. Methodology

A. Overall architecture

Fig. 1 illustrates the architecture of the proposed DBMA-Net, which adopts PVTv2-b2 as the backbone and consists of four main modules: Multi-scale Context and Attention Fusion Module (MCAF), Dual-domain Feature Enhancement Module (DFE), Boundary-guided Multi-branch Attention Module (BMAM) and Cascaded Fusion Module (CFM) [9].

Specifically, given an input image $I \in \mathbb{R}^{H\times W\times 3}$, it is first fed into the PVTv2-b2 [16] backbone to extract multi-level features $F_i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$, where $i \in \{1,2,3,4\}$ and $C_i \in \{64,128,320,512\}$ denote the channel dimensions of each stage. Then, the shallow feature F_1, which contains rich details related to polyps, is input into the MCAF module. By leveraging multi-scale dilated convolutions together with channel and spatial attention mechanisms, MCAF enhances low-level structural details and outputs the shallow enhanced feature F_1'. Meanwhile, the middle- and high-level features $\{F_2, F_3, F_4\}$ are respectively input into three parallel DFE modules. Through joint modeling in the spatial and frequency domains, DFE improves feature representation capability and enables the network to capture cross-scale semantic information of polyps, resulting in enhanced features F_2', F_3', and F_4'.

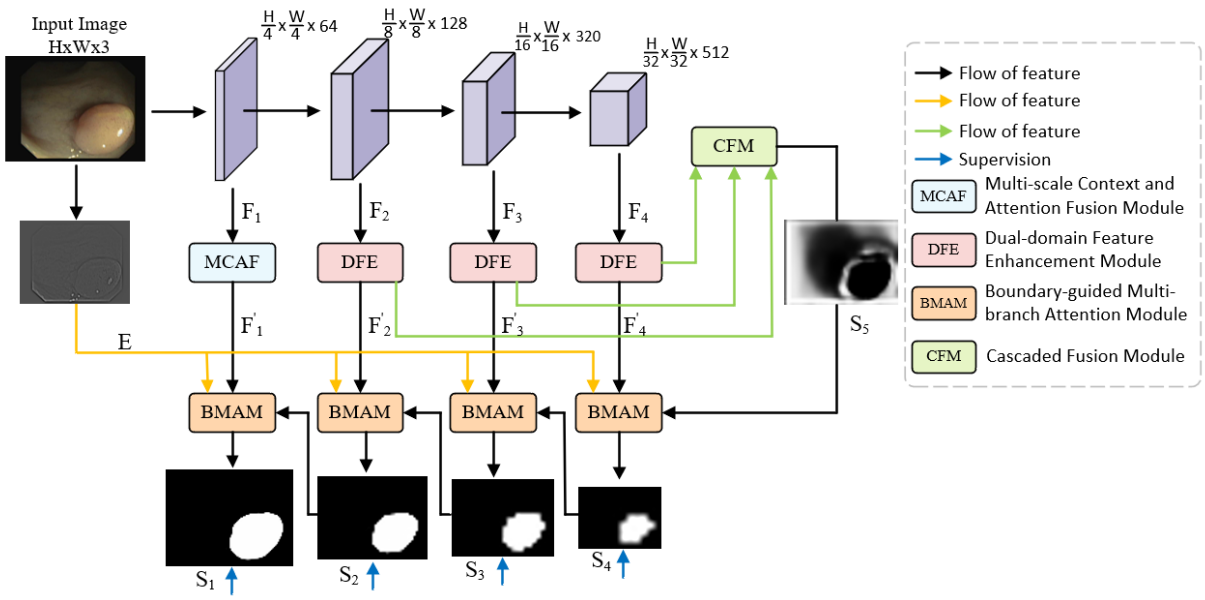

Fig. 1. Overall architecture of the proposed DBMA-Net.

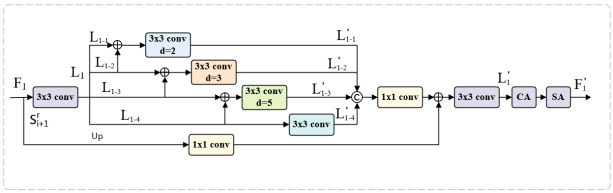

Fig. 2. Overview of Multi-scale Context and Attention Fusion Module (MCAF).

Subsequently, the enhanced features $F_i' \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 64}$, where $i \in \{2,3,4\}$, are fed into the CFM to aggregate high-level features and capture global contextual information. This process generates a coarse global localization prediction map $S_5 \in \mathbb{R}^{\frac{H}{2^3} \times \frac{W}{2^3} \times 1}$, which serves as guidance for subsequent refinement. Next, the output of the previous BMAM (with the last BMAM receiving S_5), the corresponding enhanced feature F_i' $(i \in \{1,2,3,4\})$, and the edge guidance map E extracted by a Laplacian operator are jointly fed into the BMAM. Finally, four segmentation maps with different spatial resolutions S_i $(i \in \{1,2,3,4\})$ are produced, where S_1 is regarded as the final segmentation result.

B. Multi-scale Context and Attention Fusion Module

To enhance the representation capability of shallow features for complex polyp shapes and irregular boundaries, we propose the MCAF module, as illustrated in Fig. 2. The module enlarges the receptive field via multi-branch dilated convolutions to improve contextual modeling, and integrates channel and spatial attention mechanisms to highlight discriminative features, thereby effectively improving the segmentation of small polyps and blurred boundaries.

Specifically, given the shallow feature map F_1, a 3×3 convolution is first applied to obtain a smoothed feature representation L_1, which helps capture more fine-grained spatial details and improves feature continuity. Then, L_1 is evenly split along the channel dimension into four feature groups, denoted as $\{L_{1-1}, L_{1-2}, L_{1-3}, L_{1-4}\}$. Subsequently, adjacent branch features are combined through element-wise addition and fed into 3×3 convolutions with different dilation rates to extract multi-scale contextual information. The resulting feature maps from all branches are concatenated along the channel dimension and passed through a 1×1 convolution for feature fusion, followed by a residual connection and a 3×3 convolution for further refinement. Finally, channel attention and spatial attention are employed to emphasize salient features and suppress irrelevant responses, producing the enhanced shallow feature map F_1'. The overall process can be formulated as follows.

$$L_1 = \mathrm{Conv}_{3\times3}(F_1) \tag{1}$$

$$\begin{cases} L'_{1-1} = \mathrm{Conv}^{d2}_{3\times3}(L_{1-1} + L_{1-2}) \\ L'_{1-2} = \mathrm{Conv}^{d3}_{3\times3}(L_{1-2} + L_{1-3}) \\ L'_{1-3} = \mathrm{Conv}^{d5}_{3\times3}(L_{1-3} + L_{1-4}) \\ L'_{1-4} = \mathrm{Conv}_{3\times3}(L_{1-4}) \end{cases} \tag{2}$$

$$L'_1 = \mathrm{Conv}_{3\times3}\Big(\mathrm{Conv}_{1\times1}\big(\mathrm{Cat}(L'_{1-1}, L'_{1-2}, L'_{1-3}, L'_{1-4})\big) + \mathrm{Conv}_{1\times1}(F_1)\Big) \tag{3}$$

$$F'_1 = \mathrm{SA}\big(\mathrm{CA}(L'_1)\big) \tag{4}$$

where $\mathrm{Conv}_{k\times k}$ denotes a convolution operation with kernel size k, $\mathrm{Conv}^{d}_{3\times3}$ represents a dilated convolution with dilation rate d, $\mathrm{Cat}(\cdot)$ denotes channel-wise concatenation, $\mathrm{CA}(\cdot)$

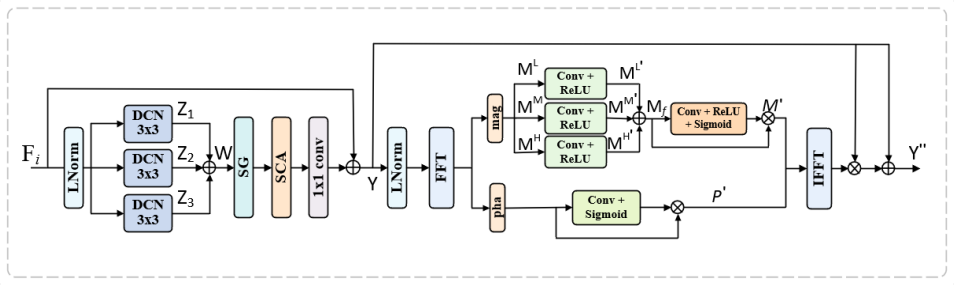

Fig. 3. Overview of Dual-domain Feature Enhancement Module (DFE).

and $\mathrm{SA}(\cdot)$ indicate channel attention and spatial attention, respectively.

C. Dual-domain Feature Enhancement Module

Existing colorectal polyp segmentation methods primarily rely on spatial domain feature modeling. However, due to the variability of polyp shapes, blurred boundaries, and low contrast, it is difficult to simultaneously account for geometric structure adaptability and global context consistency. Frequency domain features, on the other hand, can enhance the structural information of images from a global perspective, but current frequency domain methods struggle to adapt to the varying energy distributions across different images. To address these issues, we propose a Dual-domain Feature Enhancement (DFE) module, which improves the modeling accuracy of polyp regions and boundaries through adaptive spatial domain modeling and dynamic frequency domain energy perception.

Specifically, as illustrated in Fig. 3, during the spatial-domain enhancement stage, the input feature F_i $(i \in \{2,3,4\})$ is first normalized, followed by a multi-branch deformable convolution structure. Each branch independently learns deformation offsets, allowing adaptive adjustment of sampling locations to better match the irregular geometry of polyps.

$$Z_j = \mathcal{D}_j(\mathrm{LN}(F_i)), \quad j \in \{1,2,3\} \tag{5}$$

$$W = \sum_{j=1}^{3} Z_j \tag{6}$$

To further enhance feature discriminability, a gating and channel selection mechanism is introduced. Channel-wise weights are generated via global average pooling to emphasize informative channels, followed by a 1×1 convolution to restore the original channel dimension. Finally, a residual connection is employed to avoid feature distortion caused by excessive enhancement.

$$Y = \mathrm{Conv}_{1\times1}(\mathrm{SCA}(\mathrm{SG}(W))) + F_i \tag{7}$$

In the frequency-domain enhancement stage, the spatially enhanced feature Y is first normalized and transformed into the frequency domain via a 2D Fourier transform. The magnitude spectrum and phase spectrum are denoted as mag and pha, respectively. Based on the cumulative energy distribution in the frequency domain, low-, mid-, and high-frequency components are adaptively determined.

$$\mathrm{mag}, \mathrm{pha} = \mathrm{FFT}(\mathrm{LN}(Y)) \tag{8}$$

$$M_i' = \mathrm{ReLU}(\mathrm{Conv}(M_i)), \quad i \in \{L, M, H\} \tag{9}$$

$$M_f = \sum_{i\in\{L,M,H\}} M_i' \tag{10}$$

$$M' = M_f \otimes \sigma(\mathrm{ReLU}(\mathrm{Conv}(M_f))) \tag{11}$$

$$P' = \mathrm{pha} \otimes \sigma(\mathrm{Conv}(\mathrm{pha})) \tag{12}$$

$$Y' = \mathrm{IFFT}(M', P') \tag{13}$$

$$Y'' = Y' \otimes Y + Y \tag{14}$$

where $\mathrm{LN}(\cdot)$ denotes normalization, $\sigma(\cdot)$ is the Sigmoid activation function, and $\mathrm{IFFT}(\cdot)$ denotes the inverse Fourier transform that maps frequency-domain features back to the spatial domain.

D. Boundary-guided Multi-branch Attention Module

In this paper, we propose a Boundary-guided Multi-branch Attention Module (BMAM), which jointly utilizes boundary features extracted by the Laplacian operator, high-level prediction guidance maps, and semantic features. The module enhances foreground, background, and boundary information in multiple branches and performs adaptive fusion through a dynamic gating mechanism, effectively improving polyp boundary delineation and overall segmentation performance.

Specifically, as illustrated in Fig. 4, the reverse feature S_{i+1}^r is first derived from the predicted feature S_{i+1}.

$$S_{i+1}^r = 1 - \sigma(S_{i+1}) \tag{15}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function. Subsequently, the predicted feature S_{i+1}, the reverse feature S_{i+1}^r, and the boundary feature E are upsampled to the same spatial resolution as the encoder feature F_1'. These features are then element-wise multiplied with the corresponding enhanced encoder feature F_i' to generate three attention feature maps, namely Att_{i+1}, Att_{i+1}^r, and Att_{i+1}^E.

$$\begin{cases} \mathrm{Att}_{i+1} = \mathrm{Up}(S_{i+1}) \otimes F_i' \\ \mathrm{Att}_{i+1}^r = \mathrm{Up}(S_{i+1}^r) \otimes F_i' \\ \mathrm{Att}_{i+1}^E = \mathrm{Up}(E) \otimes F_i' \end{cases} \tag{16}$$

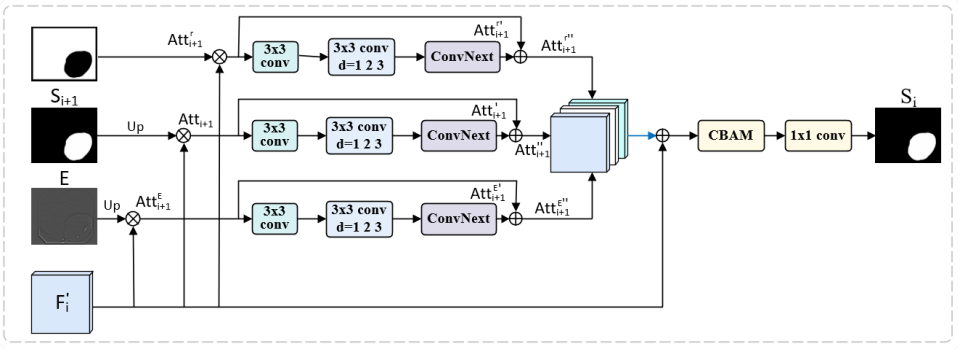

Fig. 4. Overview of Boundary-guided Multi-branch Attention Module (BMAM).

where $\mathrm{Up}(\cdot)$ denotes the upsampling operation and $\otimes$ represents element-wise multiplication. Subsequently, each attention feature map is processed by a 3×3 convolution, followed by parallel dilated convolutions with different dilation rates to capture contextual information under diverse receptive fields. A lightweight ConvNext-style module is then employed to perform nonlinear feature reorganization and stabilization.

$$\begin{cases} \mathrm{Att}'_{i+1} = \mathrm{CN}(\mathrm{DConv}_{3\times3}(\mathrm{Conv}_{3\times3}(\mathrm{Att}_{i+1}))) \\ \mathrm{Att}^{r'}_{i+1} = \mathrm{CN}\big(\mathrm{DConv}_{3\times3}\big(\mathrm{Conv}_{3\times3}(\mathrm{Att}^{r}_{i+1})\big)\big) \\ \mathrm{Att}^{E'}_{i+1} = \mathrm{CN}\big(\mathrm{DConv}_{3\times3}\big(\mathrm{Conv}_{3\times3}(\mathrm{Att}^{E}_{i+1})\big)\big) \end{cases} \tag{17}$$

where $\mathrm{DConv}_{3\times3}(\cdot)$ represents dilated 3×3 convolutions with dilation rates of 1, 2, and 3, and $\mathrm{CN}(\cdot)$ denotes a nonlinear feature reorganization module. Residual connections are then introduced by adding the original attention feature maps Att_{i+1}, Att^{r}_{i+1}, and Att^{E}_{i+1} to their corresponding enhanced features, which effectively alleviates feature distortion caused by excessive enhancement. Subsequently, the three branches are dynamically fused using a gated attention mechanism to adaptively select more discriminative branch features. After branch fusion, a 3×3 convolution is applied to further integrate the fused features, followed by a global residual connection to preserve the original high-level semantic information. Then, the CBAM is employed to recalibrate the feature responses, thereby highlighting polyp regions and their boundary structures. Finally, a 1×1 convolution is used to generate the segmentation prediction.

$$\begin{cases} \mathrm{Att}''_{i+1} = \mathrm{Att}'_{i+1} + \mathrm{Att}_{i+1} \\ \mathrm{Att}^{r''}_{i+1} = \mathrm{Att}^{r'}_{i+1} + \mathrm{Att}^{r}_{i+1} \\ \mathrm{Att}^{E''}_{i+1} = \mathrm{Att}^{E'}_{i+1} + \mathrm{Att}^{E}_{i+1} \end{cases} \tag{18}$$

$$S_i = \mathrm{Conv}_{1\times1}\Big(\mathrm{CBAM}\big(\mathrm{Conv}_{3\times3}\big(\mathrm{Att}(\mathrm{Att}''_{i+1}, \mathrm{Att}^{r''}_{i+1}, \mathrm{Att}^{E''}_{i+1})\big)\big) + F'_i\Big) \tag{19}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ denotes a 1×1 convolution, $\mathrm{Conv}_{3\times3}(\cdot)$ denotes a 3×3 convolution, $\mathrm{CBAM}(\cdot)$ represents the convolutional block attention module, and $\mathrm{Att}(\cdot)$ denotes the dynamic gated fusion mechanism.

E. Cascaded Fusion Module

In the feature decoding and fusion stage, to fully integrate high-level semantic information from different hierarchical levels, we introduce the CFM [9]. The module aggregates features in a progressive manner through successive upsampling and cascaded fusion, which effectively enhances the interaction among high-level features. Specifically, CFM is employed to fuse the enhanced features F'_2, F'_3, and F'_4, and produces an initial coarse prediction map S_5 with globally consistent semantic information, which serves as reliable guidance for subsequent refinement.

F. Loss function

In the proposed method, deep supervision is applied to the four prediction maps $\{S_1, S_2, S_3, S_4\}$ generated by the BMAM module. The overall training objective is defined as the average loss over all prediction scales, which can be formulated as

$$\mathcal{L}_{\text{total}} = \frac{1}{4}\sum_{i=1}^{4}\mathcal{L}(G, S_i) \tag{20}$$

where G denotes the ground-truth segmentation mask. The loss function $\mathcal{L}$ is a weighted combination of the IoU loss and the binary cross-entropy (BCE) loss, which is expressed as

$$\mathcal{L} = \mathcal{L}^{w}_{\text{IoU}} + \mathcal{L}^{w}_{\text{BCE}} \tag{21}$$

III. Experiments and Results

A. Datasets

The proposed method is trained and evaluated on five publicly available colorectal polyp segmentation datasets, including Kvasir-SEG [17], CVC-ClinicDB [18], CVC-300 [19], CVC-ColonDB [20], and ETIS-LaribPolypDB [21]. To ensure fairness and comparability, all experiments strictly follow the widely adopted training and testing protocol used in PraNet [7], without any manual re-splitting or additional random sampling of the datasets. The training set contains 1,450 images, including 900 images from Kvasir-SEG and 550 images from CVC-ClinicDB. All remaining images are used for performance evaluation. Specifically, the test set consists of 798 images, including 100 images from Kvasir-SEG, 62 images from CVC-ClinicDB, 60 images from CVC-300, 380

images from CVC-ColonDB, and 196 images from ETIS-LaribPolypDB.

B. Implementation Details

This study implements the proposed DBMA-Net model based on the PyTorch framework. All input images are resized to 352 × 352 pixels during both training and testing. The model uses the AdamW optimizer, with an initial learning rate and weight decay set to 1e-4. The training follows the multi-scale training strategy from PraNet [7], with scale factors set to $\{0.75, 1, 1.25\}$. To assess the effectiveness of our model, we employ six metrics: the mean dice similarity coefficient (mDice) , mean Intersection over Union (mIoU), weighted F-measure (F_β^w), S-measure (S_α), max E-measure (E_ϕ^{max}), and mean absolute error (MAE).

C. Quantitative comparison

To comprehensively evaluate DBMA-Net, we performed comparative experiments with 9 state-of-the-art segmentation methods: U-Net [5], UNet++ [6], PraNet [7], Polyp-PVT [9], CaraNet [23], C2FNet [22], CFANet [14], MEGANet [15] and SAFE-Net [24] (as shown in Tables I, II, III, IV, and V). To ensure fairness, the performance results of these methods were directly taken from their original publications, with results not provided in the original papers sourced from third-party publications. Both quantitative and qualitative analyses demonstrate the superior performance of DBMA-Net across multiple datasets, highlighting its robust learning ability and generalization capacity.

1) Learning ability: Based on the observations in Tables I and II, the proposed method outperforms other models on the Kvasir-SEG and CVC-ClinicDB datasets, achieving significant performance improvements on most evaluation metrics. Specifically, on the Kvasir-SEG dataset, the proposed method achieves mDice and mIoU scores of 92.7% and 87.8%, respectively. Compared to traditional general medical segmentation models, our method shows higher performance, with mDice improving by 10.9% and 10.6%, and mIoU improving by 13.2% and 13.5%, respectively. Compared to classical models such as PraNet and Polyp-PVT, which are specifically designed for polyp segmentation, our method also shows advantages, with mDice increasing by 2.9% and 1.0%, and mIoU increasing by 3.8% and 1.4%. When compared to the latest SAFE-Net model, mDice is improved by 1.0%, and mIoU by 1.3%; at the same time, the proposed method achieved 92.8%, 95.5%, and 98.3% in F_β^ω, S_α, and E_ξ^{max}, significantly outperforming the latest methods. On the CVC-ClinicDB dataset, mDice reaches 94.2% and mIoU reaches 89.6%, both of which are the highest values among the compared methods. Overall, the proposed model demonstrates a significant advantage in polyp segmentation learning ability.

2) Generalization ability: To validate the generalization capability of the model, we conducted evaluations on three unseen benchmark datasets: CVC-300, CVC-ColonDB, and ETIS-LaribPolypDB. The experimental results show that on the CVC-300 dataset, the mDice and mIoU reached 90.3% and 83.5%, improving by 3.2% and 3.8%, respectively, compared to the classical model PraNet. On the CVC-ColonDB dataset, despite its challenging nature, the model still achieved 81.4%

TABLE I
QUANTITATIVE RESULTS ON KVASIR-SEG (SEEN). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	mDice	mIoU	F_β^ω	S_α	E_ϕ^{max}	MAE
U-Net [5]	0.818	0.746	0.794	0.858	0.893	0.055
UNet++ [6]	0.821	0.743	0.808	0.862	0.909	0.048
PraNet [7]	0.898	0.840	0.885	0.915	0.948	0.030
Polyp-PVT [9]	0.917	0.864	0.911	0.925	0.962	0.023
CaraNet [23]	0.907	0.851	0.902	0.915	0.953	0.027
C2FNet [22]	0.900	0.842	0.894	0.912	0.949	0.030
CFANet [14]	0.915	0.861	0.903	0.924	0.942	0.023
MEGANet [15]	0.913	0.863	0.907	0.918	0.959	0.025
SAFE-Net [24]	0.917	0.865	0.908	0.931	0.963	0.025
Ours	**0.927**	**0.878**	**0.928**	**0.955**	**0.983**	**0.020**

TABLE II
QUANTITATIVE RESULTS ON CVC-CLINICDB (SEEN). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	mDice	mIoU	F_β^ω	S_α	E_ϕ^{max}	MAE
U-Net [5]	0.823	0.755	0.811	0.889	0.954	0.019
UNet++ [6]	0.794	0.729	0.785	0.873	0.931	0.022
PraNet [7]	0.899	0.849	0.896	0.936	0.979	0.009
Polyp-PVT [9]	0.937	0.889	0.936	0.949	0.989	0.006
CaraNet [23]	0.926	0.880	0.926	0.944	0.979	0.008
C2FNet [22]	0.921	0.871	0.922	0.938	0.982	0.008
CFANet [14]	0.933	0.883	0.924	0.950	0.974	0.007
MEGANet [15]	0.938	0.894	0.940	0.950	0.986	0.006
SAFE-Net [24]	0.937	0.889	0.931	0.958	0.991	0.007
Ours	**0.942**	**0.896**	**0.941**	**0.971**	**0.995**	**0.006**

TABLE III
QUANTITATIVE RESULTS ON CVC-300 (UNSEEN). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	mDice	mIoU	F_β^ω	S_α	E_ϕ^{max}	MAE
U-Net [5]	0.710	0.627	0.684	0.843	0.875	0.022
UNet++ [6]	0.707	0.624	0.624	0.624	0.898	0.018
PraNet [7]	0.871	0.797	0.843	0.925	0.972	0.010
Polyp-PVT [9]	0.900	0.833	0.884	0.935	0.981	0.007
CaraNet [23]	0.884	0.813	0.864	0.925	0.961	0.009
C2FNet [22]	0.873	0.804	0.858	0.922	0.973	0.008
CFANet [14]	0.893	0.827	0.875	0.938	0.978	0.008
MEGANet [15]	0.899	0.834	0.882	0.935	0.969	0.007
SAFE-Net [24]	0.895	0.828	0.872	0.938	0.978	0.008
Ours	**0.903**	**0.835**	**0.899**	**0.952**	**0.996**	**0.006**

TABLE IV
QUANTITATIVE RESULTS ON CVC-COLONDB (UNSEEN). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	mDice	mIoU	F_β^ω	S_α	E_ϕ^{max}	MAE
U-Net [5]	0.512	0.444	0.498	0.712	0.776	0.061
UNet++ [6]	0.483	0.410	0.467	0.691	0.760	0.064
PraNet [7]	0.712	0.640	0.699	0.820	0.872	0.043
Polyp-PVT [9]	0.808	0.727	0.795	0.865	0.919	0.031
CaraNet [23]	0.737	0.663	0.728	0.823	0.869	0.036
C2FNet [22]	0.718	0.641	0.712	0.817	0.877	0.044
CFANet [14]	0.743	0.665	0.728	0.835	0.898	0.039
MEGANet [15]	0.793	0.714	0.779	0.854	0.895	0.040
SAFE-Net [24]	0.781	0.704	0.767	0.858	0.908	0.034
Ours	**0.814**	**0.732**	**0.814**	**0.895**	**0.969**	**0.030**

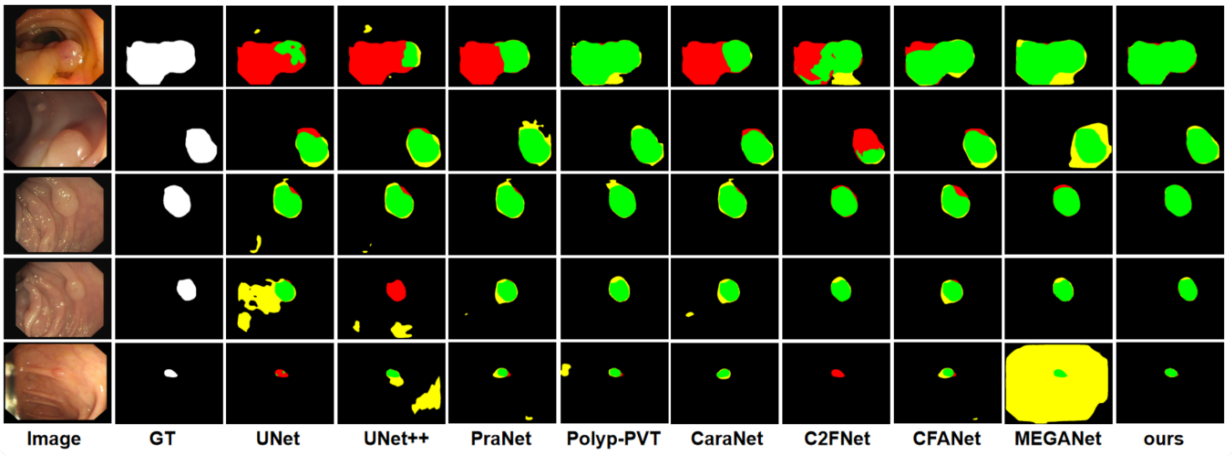

Fig. 5. Qualitative segmentation results of different methods. Green, red, and yellow represent true positives, false negatives, and false positives, respectively.

TABLE V
QUANTITATIVE RESULTS ON ETIS-LARIBPOLYPDB (UNSEEN). THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	mDice	mIoU	F_β^ω	S_α	E_ϕ^{max}	MAE
U-Net [5]	0.398	0.335	0.366	0.684	0.740	0.036
UNet++ [6]	0.401	0.344	0.390	0.683	0.776	0.035
PraNet [7]	0.628	0.567	0.600	0.794	0.841	0.031
Polyp-PVT [9]	0.787	0.706	0.750	0.871	0.910	0.013
CaraNet [23]	0.732	0.656	0.712	0.840	0.885	0.012
C2FNet [22]	0.677	0.607	0.656	0.812	0.888	0.030
CFANet [14]	0.732	0.655	0.693	0.845	0.892	0.014
MEGANet [15]	0.739	0.665	0.702	0.836	0.858	0.037
SAFE-Net [24]	0.794	0.713	0.748	0.884	0.919	0.015
Ours	**0.799**	**0.716**	**0.794**	**0.900**	**0.973**	**0.012**

mDice and 73.2% mIoU. On the ETIS-LaribPolypDB dataset, our method outperformed the comparison methods in mDice (79.9%) and mIoU (71.6%), and achieved a significant 97.3% improvement in E-measure, with MAE of 1.2%. Compared to strong models such as Polyp-PVT and SAFE-Net, our method demonstrates sustained advantages in structural consistency and prediction stability, proving its excellent cross-dataset generalization ability and robustness.

D. Qualitative comparison

Fig. 5 shows the qualitative comparison results between the proposed method and several state-of-the-art polyp segmentation models in different clinical challenge scenarios. It can be observed that, under complex endoscopic imaging conditions, the proposed method outperforms others. Specifically, in scenes with uneven illumination and strong background interference (as shown in the 1st-2nd rows of Fig. 5), existing methods generally suffer from foreground discontinuity or background misdetection, whereas the proposed method effectively suppresses highlight reflections and low-contrast noise, generating more complete structures and stronger region consistency in segmentation results. For samples with polyp-like structures or blurred boundaries (as shown in the 3rd-4th rows of Fig. 5), the proposed method more accurately recovers the true contours of the polyps and maintains good consistency with the ground truth annotations. Particularly in small polyp and low-significance target segmentation scenarios (as shown in the 5th row of Fig. 5), most methods tend to miss detections or predict smaller regions, whereas the proposed method remains stable in locating and fully segmenting small lesions. These visual results intuitively verify the effectiveness of the proposed method in frequency-domain and spatial feature modeling and regional discrimination, allowing the network to better handle key challenges commonly found in endoscopic images, such as noise interference, semantic ambiguity, and scale diversity.

E. Ablation Study

To validate the effectiveness of each module, we conducted ablation experiments on different datasets. From Table VI, we can observe that the performance of the model has declined to varying degrees after the removal of MCAF, CFM, DFE, and BMAM modules. Specifically, the removal of MCAF results in a decrease in mDice and mIoU on Kvasir-SEG and CVC-ClinicDB, highlighting the importance of multi-scale contextual information for improving segmentation accuracy and consistency. The removal of CFM leads to performance degradation across all datasets, especially on CVC-300 and ETIS-LaribPolypDB, indicating the critical role of efficient cross-layer feature fusion in regional discrimination. The removal of DFE significantly decreases the performance across all five datasets, demonstrating that the dual-domain feature enhancement mechanism effectively improves region overlap precision and boundary consistency. The removal of BMAM significantly reduces the boundary-related metrics, emphasizing the irreplaceable importance of boundary information for polyp segmentation.

IV. CONCLUSION

Polyp segmentation plays a crucial role in early screening for colorectal cancer. To address the limitations of existing

TABLE VI
ABLATION STUDY ON KVASIR-SEG, CVC-CLINICDB, CVC-300, CVC-COLONDB, AND ETIS-LARIBPOLYPDB. THE METRICS ARE MDICE AND MIOU. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Version	Kvasir-SEG		CVC-ClinicDB		CVC-300		CVC-ColonDB		ETIS-LaribPolypDB	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
w/o MCAF	0.919	0.874	0.938	0.889	0.892	0.825	0.794	0.717	0.784	0.705
w/o CFM	0.916	0.864	0.937	0.891	0.884	0.813	0.800	0.727	0.792	0.714
w/o DFE	0.918	0.864	0.929	0.880	0.872	0.796	0.805	0.724	0.772	0.692
w/o BMAM	0.901	0.856	0.927	0.879	0.895	0.803	0.801	0.716	0.779	0.695
DBMA-Net	**0.927**	**0.878**	**0.942**	**0.896**	**0.903**	**0.835**	**0.814**	**0.732**	**0.799**	**0.716**

methods in frequency domain information utilization, cross-layer feature fusion, and region discrimination modeling, we propose DBMA-Net, which combines dual-domain feature enhancement and boundary-guided multi-branch attention. This approach enhances the model's segmentation accuracy in complex endoscopic scenes. By incorporating multi-scale contextual modeling through the MCAF module, dual-domain feature collaborative enhancement through the DFE module, and boundary-aware attention modeling through the BMAM module, DBMA-Net significantly improves segmentation performance. Experimental results show that DBMA-Net outperforms existing methods on five public datasets, and ablation studies validate the effectiveness of each module. Future work will focus on improving the inference efficiency and computational cost of the model to better suit real-time clinical applications.

REFERENCES

[1] J. Zhu, M. Ge, Z. Chang, and W. Dong, "CRCNet: Global-local context and multi-modality cross attention for polyp segmentation," *Biomedical Signal Processing and Control*, vol. 83, p. 104593, 2023.

[2] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, 2021.

[3] J. Yao, M. Miller, M. Franaszek, and R. M. Summers, "Colonic polyp segmentation in CT colonography based on fuzzy clustering and deformable models," *IEEE Trans. Med. Imaging*, vol. 23, no. 11, pp. 1344–1352, 2004.

[4] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, "Texture-based polyp detection in colonoscopy," in *Proc. Bildverarbeitung für die Medizin*, Heidelberg, Germany, 2009, pp. 346–350.

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

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

[7] 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. Intervent. (MICCAI)*, 2020, pp. 263–273.

[8] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 205–218.

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

[10] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2021, pp. 14–24.

[11] W. Xu, R. Xu, C. Wang, X. Li, S. Xu, and L. Guo, "PSTNet: Enhanced polyp segmentation with multi-scale alignment and frequency domain integration," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 10, pp. 6042–6053, 2024.

[12] C. Yang and Z. Zhang, "PFD-Net: Pyramid Fourier deformable network for medical image segmentation," *Comput. Biol. Med.*, vol. 172, p. 108302, 2024.

[13] H. Wang, K.-N. Wang, J. Hua, Y. Tang, Y. Chen, G.-Q. Zhou, and S. Li, "Dynamic spectrum-driven hierarchical learning network for polyp segmentation," *Med. Image Anal.*, vol. 101, p. 103449, 2025.

[14] 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, p. 109555, 2023.

[15] 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)*, 2024, pp. 7985–7994.

[16] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "PVTv2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[17] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," *MMM 2020. Lecture Notes in Computer Science*, vol. 11962, 2019.

[18] 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 Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[19] 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," *Journal of Healthcare Engineering*, vol. 2017, no. 1, pp. 4037190, 2017.

[20] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2015.

[21] 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," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[22] G. Chen, S.-J. Liu, Y.-J. Sun, G.-P. Ji, Y.-F. Wu, and T. Zhou, "Camouflaged object detection via context-aware cross-level fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6981–6993, 2022.

[23] A. Lou, S. Guan, H. Ko, and M. H. Loew, "CaraNet: Context axial reverse attention network for segmentation of small medical objects," in *Medical Imaging 2022: Image Processing*, vol. 12032, pp. 81–92, 2022.

[24] J. Yu and L. Qi, "SAFE-Net: Shape-aware and feature enhancement network for polyp segmentation," *Biomedical Signal Processing and Control*, vol. 99, 106906, 2025.