

Article

Research on Medical Image Segmentation Based on Frequency-Domain Enhancement and Edge Awareness

Jiamin Li ^{1,2}, Yazhi Liu ^{1,2} and Wei Li ^{1,2,*} 

¹ College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063210, China; lijiamin@stu.ncst.edu.cn (J.L.); liuyazhi@ncst.edu.cn (Y.L.)

² The Hebei Key Laboratory of Industrial Intelligent Perception, Tangshan 063210, China

* Correspondence: lw@ncst.edu.cn

Abstract

Medical images commonly exhibit low contrast, weak boundaries, and complex textures. In addition, significant semantic differences exist between deep-level semantic features and shallow-level detail features, posing challenges for multi-scale feature fusion in terms of detail preservation and structural consistency. To address these issues, a frequency-enhanced and bidirectional feature-guided segmentation network (FBNet) is proposed. The network comprises two core components. The frequency-based enhancement (FBE) module employs the Fast Fourier Transform and applies adaptive modulation to the amplitude spectrum through a content-aware gating mechanism, enhancing detail expression and inter-structural contrast. The Bidirectional Guided Feature Fusion module (BGF) enables bidirectional interaction between shallow and deep features. Additionally, the Structure and Edge Awareness (SEA) module is constructed using directional and variance attention mechanisms to achieve collaborative optimization of structural modeling and edge perception. Experiments on four medical image segmentation datasets show that, compared to the second-best method, FBNet achieves improvements of 2.12, 1.57, 1.37, and 1.56 percentage points on the mIoU metric and 1.54, 1.11, 0.84, and 1.03 percentage points on the mDice metric.

Keywords: frequency enhancement; Fast Fourier Transform; structural and edge aware; bidirectional feature-guided

1. Introduction

Medical image segmentation is an essential technology for aiding disease diagnosis, guiding surgical procedures, and evaluating treatment outcomes. However, due to imaging conditions and tissue characteristics, medical images often exhibit low inter-tissue contrast, blurred target boundaries, and artifact interference, making it challenging to precisely identify key anatomical structures. Furthermore, significant variations in size, shape, and spatial distribution among different tissues pose challenges for constructing robust and precise segmentation models [1,2].

Convolutional Neural Networks (CNNs) [3] and Transformer [4] architectures have achieved notable advancements in feature extraction and semantic understanding in recent years, becoming the mainstream technological paradigms in medical image segmentation. CNNs possess inherent advantages in extracting local, detailed features such as edges and textures due to their localized receptive fields. Transformers utilize self-attention



Academic Editor: Maryam Ravan

Received: 8 March 2026

Revised: 3 April 2026

Accepted: 10 April 2026

Published: 12 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

mechanisms to model global contextual information within images. However, both approaches have certain limitations [5,6]. Due to the limited receptive field inherent in convolutional kernels, CNNs have difficulty in capturing the global correlation between features. Transformers, meanwhile, struggle with capturing fine-grained local details, particularly in complex structural regions where edge information is prone to loss. In response, researchers have proposed various hybrid architectures and attention mechanisms. For instance, models like HiFormer [7], TransFuse [8], and CSWin-UNet [9] integrate CNN and Transformer architectures to fuse local details with global semantics, achieving complementary advantages. Among these, TransFuse employs a parallel dual-branch structure for multi-level fusion of features extracted by CNN and Transformer, representing one of the most representative architectures in this direction. On the other hand, models such as GA2Net [10], TransUNETR [11], and LKCA-Net [12] incorporate diverse attention mechanisms, including spatial attention, channel attention, and deformable large-kernel attention. These mechanisms allow the models to adaptively focus on key regions and improve their ability to recognize complex shapes.

Although existing methods have improved overall target segmentation performance, issues such as inaccurate edge segmentation and loss of fine details persist when handling complex textures and weak boundaries. Crucially, most current hybrid models are primarily confined to feature modeling within the spatial domain, rendering them inherently susceptible to the effects of low contrast and noise interference commonly encountered in medical imaging. In addition, to further strengthen the capacity to perceive and localize target structures, multi-scale feature fusion strategies have been widely adopted. For example, methods such as SECA-Net [13], HiFormer [7], and UCTransNet [14] enhance structural perception by aggregating features from different hierarchical levels. SECA-Net combines attention gating layers with parallel dilated convolutions to achieve multi-scale feature fusion. UCTransNet introduces a channel-wise cross-attention mechanism to model inter-level dependencies. However, because shallow features contain richer detailed information while deep features focus on high-level semantic information, there exists a significant semantic gap between them. As a result, existing approaches still face challenges in cross-level feature fusion, making it difficult to simultaneously maintain structural consistency and detail integrity in complex scenarios.

To address the above challenges, a frequency-enhanced and bidirectional feature-guided segmentation network (FBNet) is proposed for medical image segmentation, which jointly improves model performance from both the frequency and spatial domains. Unlike traditional hybrid architectures that rely solely on spatial-domain feature modeling, FBNet introduces a frequency-domain enhancement mechanism. The core of this design lies in the direct modulation of the magnitude spectrum via content-aware gating, which physically enhances structural contrast and texture intensity, thereby overcoming the intrinsic limitations of spatial-domain operations in handling low contrast and ambiguous boundaries. Specifically, a frequency-based enhancement module (FBE) is constructed to strengthen feature detail representation and contrast by leveraging the Fast Fourier Transform and a content-aware gating mechanism. In addition, a Bidirectional Guided Feature Fusion module (BGF) is constructed to establish bidirectional interactions between shallow and deep features, enabling complementary integration of spatial details and semantic information. Furthermore, a Structure and Edge Awareness (SEA) module is introduced, which employs directional and variance-based attention mechanisms to explicitly strengthen sensitivity to structural variations and boundary regions. The contributions are summarized as follows:

1. We propose a frequency-enhanced and bidirectional feature-guided segmentation network (FBNet), which jointly addresses the challenges of low contrast, weak boundaries, and multi-scale feature integration.

2. We design a frequency-based enhancement (FBE) module. It leverages the Fast Fourier Transform and content-aware gating to adaptively enhance feature details and inter-structural contrast in the frequency domain.
3. A bidirectional feature fusion module (BGF) is developed to facilitate mutual guidance between shallow and deep features. Within this module, a Structure and Edge Awareness (SEA) mechanism is integrated to jointly optimize structural modeling and edge detection.
4. Extensive experiments in four medical image segmentation datasets (Kvasir-SEG, CVC-ClinicDB, Glas, and ISIC2018) demonstrate that FBNet achieves superior performance over existing state-of-the-art methods, with improvements of up to 2.12% in mIoU and 1.54% in mDice compared to the second-best approach.

2. Related Work

In the field of medical image segmentation, CNN- and Transformer-based techniques have achieved considerable advancements [15–18]. Convolutional architectures represented by U-Net [19], together with their subsequent derivatives like UNet++ [20], UNet3+ [21], and nnUNet [22], have been extensively employed in diverse medical tasks. However, the inherent shortcomings of CNNs in modeling distant spatial dependencies were later mitigated by the advent of Transformer models, which drove further progress in this domain. For example, models such as UNETR [23] and Swin-UNet [24] enhanced the modeling of global contextual information through self-attention mechanisms. Furthermore, Transformer layers were incorporated into the CNN backbone in Trans-UNet [25] to enable the joint modeling of both local and global information. However, in complex scenarios where the boundaries between target tissues and surrounding regions are ambiguous, these methods struggle to accurately distinguish foreground from background, resulting in limitations in structural consistency and segmentation precision.

To strengthen the capacity to recognize target regions, many studies have introduced various attention mechanisms or incorporated multi-scale feature fusion. Attention-UNet [26] introduced attention gating mechanisms in skip connections to focus on key target regions. GA2Net [10] enhanced the salient features of target regions through a mask-guided attention module while effectively suppressing irrelevant information. TransUNETR [11] combined gated convolution with local self-attention mechanisms to strengthen the modeling of pathological structural details. Moreover, HiFormer [7] employed cross-attention to fuse shallow local information with deep semantic information. UCTransNet+ [14] utilized a channel-wise cross-attention mechanism to fuse multi-scale features and capture channel dependencies across different scales. M²SNet [27] aggregated features from different hierarchical levels through multi-scale subtraction operations, emphasizing discriminative information. Although these methods improve feature saliency and target recognition to some extent, they still suffer from inaccurate edge delineation and insufficient detail preservation when handling weak boundaries and low-contrast regions.

On the other hand, some studies have enhanced boundary segmentation performance by introducing edge priors or frequency-domain information. MEGANet [28] and FoBS [29] constructed edge-guided attention by employing the Laplacian operator and deformable Laplacian convolution, respectively, with the aim of strengthening responses in boundary regions. Concurrently, frequency-domain information has been gradually introduced into medical segmentation tasks. AdaptUNet [30] enhanced image edge and texture features using a two-dimensional discrete wavelet transform. MADGNet [31] proposed a multi-frequency channel attention mechanism based on the 2D DCT, which enhanced edge details by extracting diverse features from different frequency bands. FRUNet [32] captured frequency-domain information across different channels through Fourier channel

attention to improve the perception of edge details. Although the aforementioned methods demonstrate the potential of edge operators and frequency-domain information for boundary modeling, the lack of decoupled analysis and targeted processing of frequency components makes it difficult to enhance edge contrast while preserving structural integrity. Consequently, challenges remain in achieving robust segmentation performance in low-contrast regions.

To provide a more intuitive comparison of existing studies, Table 1 summarizes the strengths and weaknesses of the various methods discussed above. Based on this analysis, we propose a frequency-enhanced and bidirectional feature-guided segmentation network (FBNet), which aims to achieve the synergistic optimization of structural modeling and edge perception. The proposed FBNet employs frequency-based enhancement modules (FBE) to perform content-aware adaptive modulation of the amplitude spectrum, thereby improving image contrast and structural details at the frequency-domain level. Meanwhile, a Bidirectional Guided Feature Fusion module (BGF) is devised to enable deeper structural modeling and enhanced edge awareness.

Table 1. Strengths and limitations of different segmentation architectures.

Category	Representative Models	Advantages	Limitations
CNN-based methods	U-Net [19], UNet++ [20], nnU-Net [22]	Efficient local feature extraction; widely adopted.	Limited receptive field; inability to model long-range global dependencies.
Transformer-based methods	UNETR [23], Swin-UNet [24], Trans-UNet [25]	Global context modeling via self-attention.	Insufficient local detail capture; challenges in regions with ambiguous boundaries.
Attention & multi-scale fusion	Attention-UNet [26], GA2Net [10], HiFormer [7], UCTransNet+ [14]	Enhanced target saliency; multi-scale feature integration.	Insufficient edge delineation and detail preservation in weak-boundary, low-contrast regions.
Edge-enhanced methods	MEGANet [28], MADGNet [31], FRUNet [32]	Strengthened boundary responses; improved perception of texture details.	Lack of decoupled analysis and targeted processing of frequency components; difficulty balancing edge enhancement and structural integrity in low-contrast regions.

3. Methods

3.1. Overall Architecture

The overall architecture of FBNet is shown in Figure 1. It follows the classic encoder–decoder paradigm and mainly consists of the hierarchical feature encoder, the frequency-based enhancement module (FBE), and the cascaded decoder. To effectively capture global contextual information, we adopt the pre-trained PVT-v2-B2 [33] as the backbone during the feature encoding stage. Given an input image $I \in \mathbb{R}^{B \times 3 \times H \times W}$, the encoder uses its hierarchical Transformer structure to progressively generate feature maps $\{t_1, t_2, t_3, t_4\}$ at four different scales, with spatial resolutions of $\left[\frac{H}{4} \times \frac{W}{4}, \frac{H}{8} \times \frac{W}{8}, \frac{H}{16} \times \frac{W}{16}, \frac{H}{32} \times \frac{W}{32}\right]$ and corresponding channel numbers of $[C, 2C, 4C, 8C]$, where the default channel dimension is 64. To further enhance structural contrast and texture clarity, a frequency-based enhancement module (FBE) is introduced after each level of the encoder output. This module performs adaptive weighting and modulation of the amplitude spectrum based on frequency-domain decomposition and a content-aware gating mechanism, significantly enhancing structural and texture features of the target

regions in the frequency domain. Features enhanced by the FBE module not only retain spatial-domain semantic information but also possess clearer structural and detailed characteristics, providing richer representations for subsequent decoding.

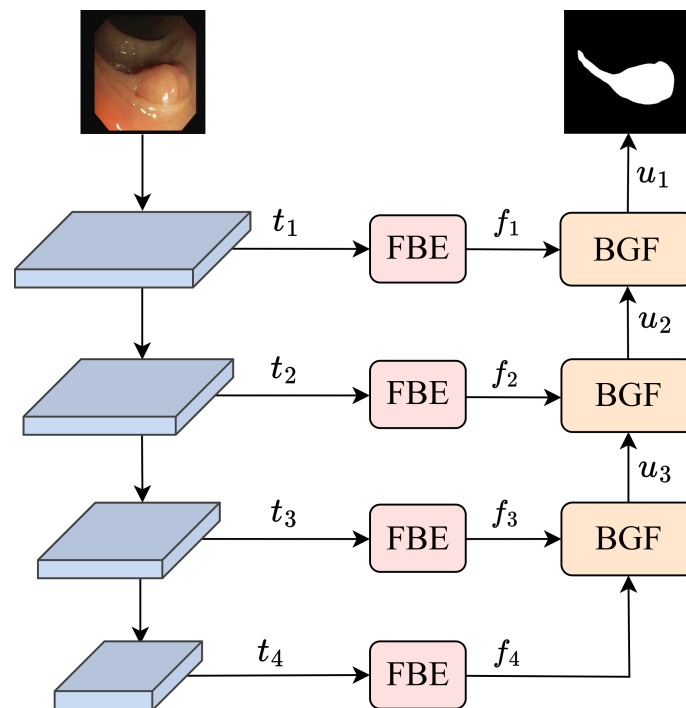


Figure 1. Overall architecture of FBNet.

The decoder is configured to facilitate effective integration of features across multiple scales and fine-grained reconstruction, consisting of three decoding stages, each composed of the Bidirectional Guided Feature Fusion module (BGF). To address the semantic gap between shallow and deep features, the BGF module establishes bidirectional interactions between shallow detail and deep semantic features. Simultaneously, the Structure and Edge Awareness module (SEA) explicitly models directional characteristics of target structures and strengthens responses in boundary regions. Finally, by progressively recovering the spatial resolution of the feature maps, the decoder produces more accurate segmentation results. Algorithm 1 summarizes the complete forward pass of the proposed FBNet, detailing each phase from input to output in a step-by-step manner.

Algorithm 1 Procedure of the FBNet.

```

1: Input: Input image  $I \in \mathbb{R}^{B \times 3 \times H \times W}$ ; Encoder: PVT-v2-B2  $E$ 
2: Output: Final segmentation mask  $S \in \mathbb{R}^{B \times 1 \times H \times W}$ 
3:  $\{t_1, t_2, t_3, t_4\} \leftarrow E(I)$  // Phase 1: Multi-scale Feature Extraction
4: for  $i = 1$  to 4 do
5:    $M_i, P_i \leftarrow \text{FFT}(t_i)$ 
6:    $M'_i \leftarrow \text{Gating}(M_i)$ 
7:    $e_i \leftarrow \text{IFFT}(M'_i, P_i)$ 
8: end for // Phase 2: Frequency-domain Feature Enhancement
9:  $d_4 \leftarrow e_4$ 
10: for  $j = 3$  down to 1 do
11:    $d_{\text{up}} \leftarrow \text{UpSample}(d_{j+1})$ 
12:    $f_j \leftarrow \text{BGF}(e_j, d_{\text{up}})$ 
13:    $d_j \leftarrow \text{SEA}(f_j)$ 
14: end for // Phase 3: Cascaded Decoding and Bidirectional Fusion
15:  $S \leftarrow \sigma(\text{Conv}(\text{UpSample}(d_1)))$  // Phase 4: Final Reconstruction and Output
16: return  $S$ 

```

3.2. FBE Module

The frequency-based enhancement module (FBE) is one of the core components of FBNet. It is designed to explore frequency-domain information of image features and enhance the capability to recognize critical structures and fine-grained boundaries in medical images. As shown in Figure 2, the FBE module maps feature representations to the frequency domain via the Fourier Transform. Leveraging the decoupled characteristics of the amplitude and phase spectra, it adaptively modulates the amplitude spectrum through a content-aware gating mechanism to enhance feature contrast and edge clarity.

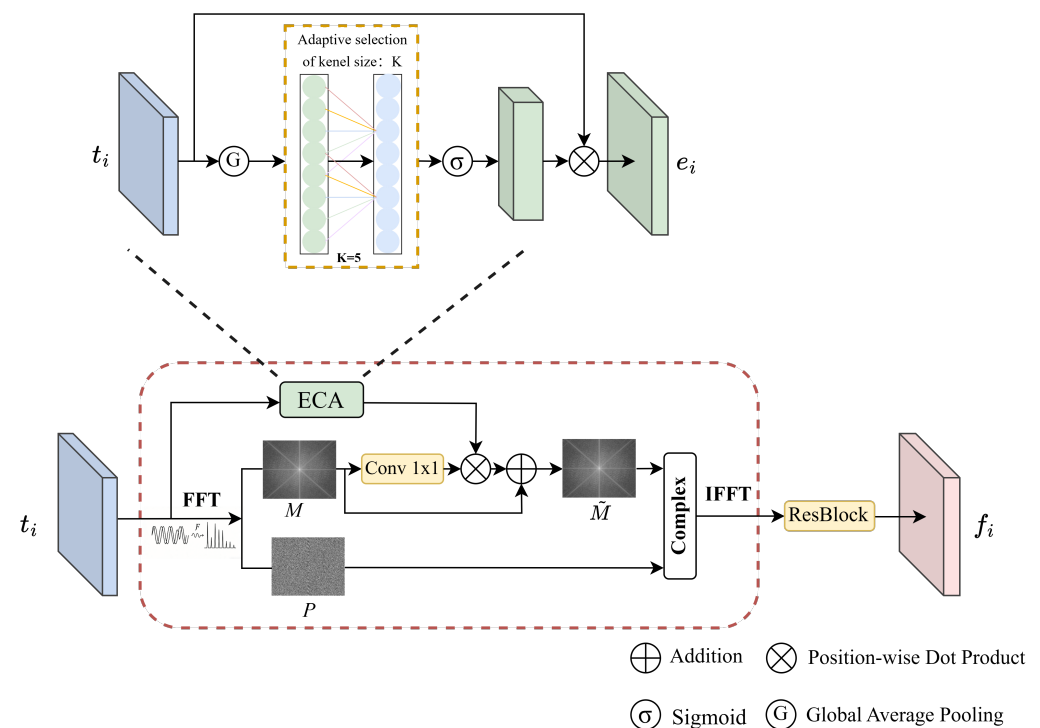


Figure 2. Architecture of the FBE Module.

Specifically, the FBE module first applies FFT to the spatial features $t_i \in \mathbb{R}^{B \times C \times H \times W}$ from the encoder at stage i , mapping them from spatial representation to frequency representation. According to Euler's formula, the frequency-domain features can be decomposed into an amplitude spectrum M and a phase spectrum P .

$$F(t_i) = \text{FFT}(t_i) = M \cdot e^{jP} \quad (1)$$

where, the amplitude spectrum M represents the energy distribution of the image across different frequency components, reflecting the image's contrast and texture strength. The phase spectrum P encodes the geometric structure and positional information of the image. Modulating the amplitude spectrum, while preserving the phase information, can enhance the saliency of lesion structures without introducing geometric distortion. Therefore, to achieve adaptive enhancement of frequency-domain features, a content-aware gating mechanism is introduced in the frequency domain to modulate the amplitude spectrum, as depicted in Figure 2. Specifically, an efficient channel attention (ECA) module [34] is used to generate channel-wise weights $G_{ECA} \in [0, 1]^{H \times W}$, and global contextual information $e_i \in \mathbb{R}^{B \times C \times H \times W}$ is leveraged to adaptively weight the responses at different frequencies. Lesion regions in medical images typically manifest as abnormal energy distributions within specific frequency ranges; thus, channel-wise modulation of frequency

responses can effectively highlight structural features associated with lesions. This process can be formalized as

$$G_{ECA} = \sigma(\text{Conv}_{5 \times 5}(\text{GAP}(t_i))) \quad (2)$$

$$e_i = t_i \otimes G_{ECA} \quad (3)$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function, and GAP represents the global average pooling operation. Since the amplitude spectrum encodes the image's contrast and texture strength, weighting and modulating the amplitude spectrum using global features $e_i \in R^{B \times C \times H \times W}$ enhances the energy response of key frequency components, thereby strengthening critical structural features. This enables the model to focus on regions with complex textures and low contrast. This process can be formalized as

$$\tilde{M} = \text{Conv}_{1 \times 1}(M) \otimes e_i + M \quad (4)$$

where, $\otimes$ denotes element-wise multiplication, M represents the original amplitude spectrum, and $\tilde{M}$ denotes the modulated and enhanced amplitude spectrum.

Finally, the enhanced amplitude spectrum $\tilde{M}$ is recombined with the original phase spectrum P and transformed back to the spatial domain via the Inverse Fourier Transform (IFFT). To mitigate potential semantic shifts caused by the frequency-domain reconstruction and to maintain stable gradient propagation, a residual block is further applied to the reconstructed features for integration. This process can be formalized as

$$f_i = \text{Res}\left(\text{Re}\left(\text{IFFT}\left(\tilde{M} \cdot e^{jP}\right)\right)\right) \quad (5)$$

where $\text{Re}(\cdot)$ denotes taking the real part. Res represents the residual block, which includes two convolutional layers and a skip connection. $f_i \in R^{B \times C \times H \times W}$ is the feature after frequency-domain enhancement. This design not only enhances structural details and feature contrast in the frequency domain but also preserves semantic consistency in the spatial domain, providing more discriminative feature representations for subsequent cross-level fusion and fine-grained segmentation.

3.3. BGF Module

The Bidirectional Guided Feature Fusion module (BGF) is devised to achieve effective fusion of multi-level features between the encoder and decoder, while incorporating directional and variance-based attention mechanisms to jointly strengthen the structural modeling and edge-awareness capabilities.

As depicted in Figure 3, this module achieves complementary fusion between structural details and semantic context through bidirectional interactions between shallow feature f_i and deep feature f_{i+1} . Specifically, deep features are first processed through average pooling, convolution, and an activation function to extract global semantic information and compress spatial dimensions, thereby enhancing their semantic guidance capability. Shallow features are processed via convolution and an activation function to retain rich local structural information. Next, shallow and deep features interact bidirectionally through element-wise multiplication, allowing shallow features to gain semantic guidance while deep features are supplemented with structural details, achieving complementary enhancement across feature levels. The operation can be expressed as:

$$h_i = \sigma(\text{Avgpool}(\text{Conv}_{1 \times 1}(\text{Up}(f_{i+1})))) \otimes f_i \quad (6)$$

$$h_{i+1} = \sigma(\text{Conv}_{3 \times 3}(f_i)) \otimes \text{Conv}_{1 \times 1}(\text{Up}(f_{i+1})) \quad (7)$$

where, Up denotes the upsampling operation. $Avgpool$ represents average pooling. h_i and h_{i+1} represent the features after bidirectional interaction.

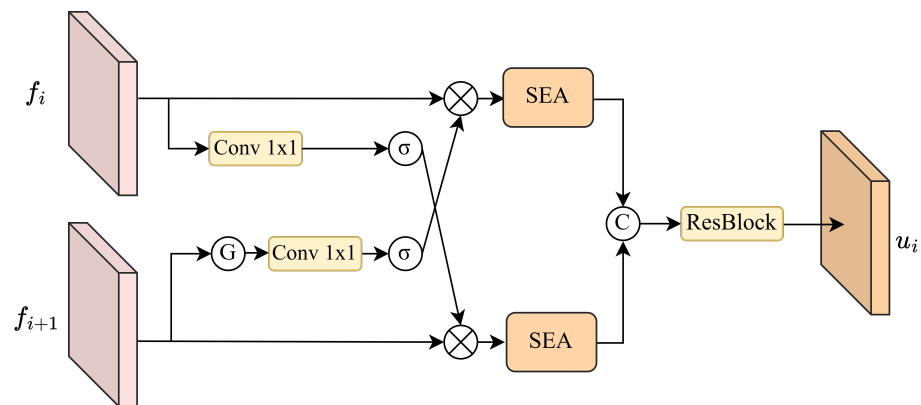


Figure 3. Architecture of the BGF Module.

To address the anisotropy of target structures and blurred boundaries in medical images, the SEA module is designed, as shown in Figure 4, to enhance features from the perspectives of directional and variance-based uncertainty. Directional attention explicitly models spatial dependencies along different directions, improving the ability to perceive texture and structural distributions across multiple orientations. Specifically, feature dimensions are rearranged, treating the spatial dimensions (H or W) as “pseudo-channel” dimensions to model spatial dependencies along the row and column directions separately.

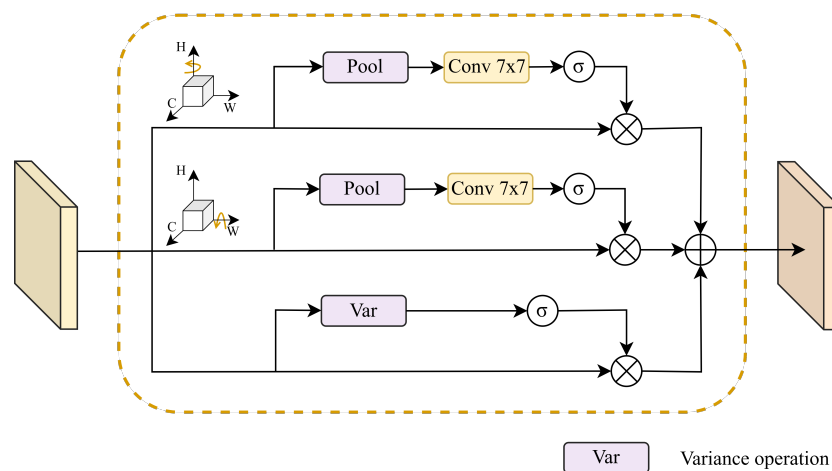


Figure 4. Architecture of the SEA Module.

The input feature tensor $h \in R^{B \times C \times H \times W}$ is converted into row-oriented tensor $f_H \in R^{B \times H \times C \times W}$ and column-oriented tensor $f_W \in R^{B \times W \times H \times C}$, which are then aggregated using pooling operations and gating function $\sigma(\cdot)$ to capture directional information $f'_{H;W}$. This process can be formalized as

$$f_{H;W} = \text{Permute}(h, [(0, 2, 1, 3); (0, 3, 2, 1)]) \quad (8)$$

$$f'_{H;W} = \sigma(\text{Conv}_{1 \times 1}(\text{Pool}(f_{H;W}))) \otimes f_{H;W} \quad (9)$$

Finally, the features are mapped back to their original dimensions through an inverse permutation, yielding the final directionally enhanced features $\tilde{f}_{H;W}$:

$$\tilde{f}_{H;W} = \text{Permute}(f'_{H;W}, [(0, 2, 1, 3); (0, 3, 2, 1)]), \tilde{f}_{H;W} \in R^{B \times C \times H \times W} \quad (10)$$

High-variance regions typically correspond to structural boundaries, lesions, or areas with abrupt texture changes. Based on this property, an attention mechanism based on variance is employed within the SEA module to identify and enhance high-uncertainty regions. Specifically, for the input features h , the variance at each spatial position (H, W) is computed along the channel dimension C to obtain variance maps $V \in \mathbb{R}^{B \times 1 \times H \times W}$. This process can be formalized as

$$V_{B,1,H,W} = \frac{1}{C} \sum_{C=1}^C (h_{B,C,H,W} - \mu_{B,H,W})^2 \quad (11)$$

where $\mu_{B,H,W}$ denotes the channel-wise mean at the spatial position. The Sigmoid function is subsequently used to generate a position-sensitive uncertainty weight map A_V , which is multiplied by the input features h element-by-element to explicitly enhance the response of high-variance regions.

$$A_V = \sigma(V) \quad (12)$$

$$F_V = h \odot A_V \quad (13)$$

where, F_V represents the enhanced features. By reinforcing high-variance regions, this mechanism significantly enhances sensitivity to structural changes and boundary areas.

Finally, the SEA module fuses the directionally and variance-enhanced features to produce the output feature F_{SEA} , which is further integrated through a residual block to maintain semantic consistency and enhance feature robustness.

$$F_{SEA} = \frac{1}{3} (\tilde{F}_H + \tilde{F}_W + F_V) \quad (14)$$

3.4. Loss Function

To comprehensively evaluate the similarity between model predictions and ground truth labels, an objective function combining Binary Cross Entropy (BCE) loss [35] and Dice loss is adopted for network training. The rationale for this combination lies in its intrinsic complementarity for medical image segmentation tasks. Specifically, the Dice loss quantifies the overlap between regions, helping to enhance the overall structural consistency of the segmentation results, which aligns with the structural modeling objective of the BGF module. Meanwhile, BCE loss calculates the discrepancy between predicted masks and ground-truth labels at the individual pixel level. By focusing on pixel-level accuracy, it provides precise supervision for the fine-grained textures and edge information enhanced by the FBE module. The integration of these two losses enables the model to achieve an optimal balance between regional consistency and pixel-level precision.

$$L_{Dice} = 1 - \frac{2 \sum_{i=1}^N (O_i \cdot G_i)}{\sum_{i=1}^N O_i^2 + \sum_{i=1}^N G_i^2} \quad (15)$$

$$L_{bce} = -\frac{1}{N} \sum_{i=1}^N [G_i \log(O_i) + (1 - G_i) \log(1 - O_i)] \quad (16)$$

where, N denotes the total number of samples. G_i and O_i represent the ground-truth value and predicted value of the i -th pixel, respectively.

The overall loss function can be expressed as:

$$L_{total} = L_{Dice} + L_{bce} \quad (17)$$

where, L_{Dice} indicates the Dice loss, L_{bce} denotes the BCE loss, and L_{total} represents the overall loss.

4. Experiments

4.1. Experimental Settings

All experiments were conducted on a server equipped with an Intel Xeon Gold 6226R CPU, 256 GB RAM, and an NVIDIA RTX 3090 GPU with 24 GB of memory. The models were implemented using the PyTorch 1.13 deep learning framework with CUDA 11.7. With a batch size of 8, our model was trained for 200 epochs. The Adam optimizer was employed with a weight decay of 1×10^{-4} and learning rate of 1×10^{-4} . During data preprocessing, since the original Kvasir-SEG, CVC-ClinicDB, and ISIC2018 datasets contain images of varying resolutions, all images from these datasets were uniformly resized to 384×384 , while those from the Glas dataset were adjusted to 512×512 . To enhance data diversity and improve model robustness, training images were augmented using techniques such as random rotation, horizontal and vertical flipping, and color jittering.

4.2. Datasets

Kvasir-SEG dataset [36]: This open-access dataset contains gastrointestinal polyp images along with their corresponding segmentation masks, containing 1000 images. The original images have varying resolutions, ranging from 332×487 to 1920×1072 . Following previous research settings [37], 880 images are used for training, while 120 images are reserved for testing.

CVC-ClinicDB dataset: Extracted from colonoscopy videos, this dataset includes 612 images with their corresponding masks [38]. The original image resolution is primarily 384×288 . Following previous research settings for dataset splitting [39], 428 images are used for training, while 184 are reserved for testing.

Glas dataset [40]: This dataset consists of 165 microscopic images stained with hematoxylin and eosin, with resolutions ranging from 581×442 to 775×522 . The optical magnification factor, while important for assessing microspecimen quality, is not specified in the original dataset. This limitation is common among public histopathology datasets. Following standard partitioning methods [41], 85 images are assigned for the training set, and 80 images are used for testing.

ISIC2018 dataset: This dataset is derived from the 2018 International Skin Imaging Collaboration (ISIC) Public Dataset for Skin Lesion Segmentation, which comprises 2594 images and corresponding ground-truth labels. The original images have varying resolutions, ranging from 450×600 to 2016×3024 . Following the setup of previous studies [42], these images were randomly split into 2074 training images and 520 test images.

4.3. Evaluation Metrics

Four commonly used metrics are adopted to comprehensively evaluate the segmentation performance of our proposed method from different perspectives, including the mean Intersection over Union (mIoU), mean Dice coefficient (mDice), Accuracy (ACC), and mean absolute error (MAE). Among them, mDice quantifies the overlap between predicted segmentation and ground truth, reflecting the similarity between them. mIoU calculates the intersection over union between predicted segmentation and ground truth, evaluating the coverage accuracy of the target area. ACC reflects the model's overall prediction capability. Boundary segmentation precision is measured using the MAE metric, which

calculates the average absolute difference between predicted values and ground-truth labels. The formulation is as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^N |O_i - G_i| \quad (18)$$

where O_i represents the prediction for the i -th pixel, and G_i indicates the associated ground-truth value.

4.4. Analysis of Experimental Results

We conducted qualitative and quantitative comparisons between our FBNet and eight state-of-the-art methods, including Swin-UNet [24], HiFormer [7], MEGANet [28], MADGNet [31], GA2-Net [10], TransUNETR [11], ConDSeg [43], and LKCA-Net [12].

Quantitative Evaluation

Table 2 reports the segmentation results on the Kvasir dataset, where the best results are highlighted in bold. Across all four metrics, the proposed FBNet demonstrated superior performance, achieving 92.82% mDice, 87.92% mIoU, and 98.18% ACC, while also obtaining the lowest MAE of 0.0184. In comparison to the second-best model, FBNet improved mDice by 1.54 percentage points and mIoU by 2.12 percentage points. Furthermore, in comparison with the frequency-domain methods MEGANet and MADGNet, FBNet shows increases of 2.56 and 2.01 percentage points in mDice, as well as 3.61 and 2.85 percentage points in mIoU. This performance gain primarily stems from the model's enhanced capability to represent complex lesion regions. By employing the frequency-domain enhancement module for content-aware adaptive modulation of the amplitude spectrum, FBNet can more robustly identify and localize lesion bodies from a frequency perspective. Meanwhile, the MAE significantly decreases from 0.0258 (GA2-Net) to 0.0184, which intuitively reflects the high consistency between the predicted masks and ground truth in terms of boundary morphology. This demonstrates that the BGF module, through its orientation and variance attention mechanisms, can sensitively capture gradient changes at lesion boundaries, significantly strengthening the perception of weak boundaries.

Table 2. Quantitative results on the Kvasir dataset.

Methods	Years	mDice	mIoU	ACC	MAE↓
Swin-UNet [24]	2022	89.70	83.66	97.20	0.0280
HiFormer [7]	2023	89.26	82.54	96.87	0.0356
MEGANet [28]	2024	90.26	84.31	97.31	0.0270
MADGNet [31]	2024	90.81	85.07	97.35	0.0309
GA2-Net [10]	2024	91.28	85.80	97.42	0.0258
TransUNETR [11]	2026	89.92	83.67	96.99	0.0301
ConDSeg [43]	2025	90.13	84.60	96.89	0.0311
LKCA-Net [12]	2026	90.13	84.17	97.14	0.0314
Ours	-	92.82	87.92	98.18	0.0184

Note: ↓ indicates lower values are better. Best results are in bold.

Tables 3–5 further validate the generalization capabilities of FBNet across different scenarios. Table 3 presents the segmentation results on the CVC-ClinicDB dataset. FBNet also performed exceptionally well, achieving 94.97% mDice and 90.69% mIoU—improvements of 1.11 percentage points and 1.57 percentage points, respectively, over the second-best model, GA2-Net. Furthermore, it achieved a minimum MAE of 0.0081, demonstrating excellent pixel-level accuracy in the segmentation of polyp lesions. FBNet continues to perform robustly when applied to the Glas and ISIC2018 datasets, which feature more complex tissue structures and greater inter-individual variability. Unlike endoscopic polyp datasets,

the Glas dataset contains glandular structures with highly irregular topologies, characterized by blurred boundaries and dense inter-glandular spaces, which makes boundary segmentation extremely challenging. Furthermore, the limited number of training samples in the Glas dataset constrains the feature representation learning capabilities of deep models. Despite these inherent complexities and data scarcity, FBNet still outperforms other state-of-the-art models, demonstrating its robustness and high accuracy when handling small sample sizes and complex morphologies. On the Glas dataset, FBNet achieved 94.03% mDice and 89.14% mIoU, outperforming the second-best model, MEGANet, by 0.84 percentage points and 1.37 percentage points, respectively, while achieving a minimum MAE of 0.0593. Compared to MADGNet, FBNet achieves improvements of 1.02% and 1.75% in mDice and mIoU, respectively. Although the aforementioned comparison methods also incorporate frequency-domain operations, their enhancement processes are primarily confined to local regions or specific frequency bands. In contrast, the amplitude spectrum adaptive modulation strategy employed by FBNet demonstrates superior segmentation performance in addressing the challenges posed by the complex structures of the Glas dataset. On the ISIC2018 dataset, FBNet achieved 91.95% mDice and 86.17% mIoU, representing improvements of 1.03 and 1.56 percentage points, respectively, over the second-best model. These results further validate FBNet's robustness in handling the challenges of weak boundaries and low contrast in medical images by enhancing its ability to model structural and edge information.

Table 3. Quantitative results on the CVC-ClinicDB dataset.

Methods	mDice	mIoU	ACC	MAE↓
Swin-UNet [24]	92.49	86.79	98.77	0.0123
HiFormer [7]	91.15	85.18	98.27	0.0188
MEGANet [28]	91.70	86.06	98.27	0.0173
MADGNet [31]	93.27	88.08	98.81	0.0377
GA2-Net [10]	93.86	89.12	98.85	0.0115
TransUNETR [11]	91.63	86.25	98.45	0.0155
ConDSeg [43]	92.10	86.94	98.64	0.0136
LKCA-Net [12]	93.60	89.12	98.77	0.0132
Ours	94.97	90.69	99.20	0.0081

Note: ↓ indicates lower values are better. Best results are in bold.

Table 4. Quantitative results on the Glas dataset.

Methods	mDice	mIoU	ACC	MAE↓
Swin-UNet [24]	90.66	83.60	90.71	0.0929
HiFormer [7]	91.53	85.11	91.66	0.0994
MEGANet [28]	93.19	87.77	93.44	0.0661
MADGNet [31]	93.01	87.39	93.16	0.1219
GA2-Net [10]	93.11	87.66	93.40	0.0660
TransUNETR [11]	91.77	85.35	91.86	0.0814
ConDSeg [43]	92.53	86.57	92.78	0.0722
LKCA-Net [12]	93.00	87.58	93.33	0.0798
Ours	94.03	89.14	94.17	0.0593

Note: ↓ indicates lower values are better. Best results are in bold.

Additionally, five independent experiments were conducted on four medical datasets to evaluate the performance benefits of FBNet. Statistical comparisons were subsequently performed using one-way analysis of variance (ANOVA) followed by post hoc Dunnett's test. In the figure, the symbol **** indicates that the difference is statistically significant ($p < 0.0001$) Compared to the latest segmentation model and suboptimal models across var-

ious datasets, our approach demonstrates statistically significant differences ($p < 0.0001$) on all datasets, as detailed in Figure 5. These results further validate the statistical significance and reliability of our approach. To provide a more comprehensive evaluation, we have included two supplementary tables in Appendix A that present the complete statistical comparison results for all models, further validating the robust superiority of FBNet.

Table 5. Quantitative results on the ISIC2018 dataset.

Methods	mDice	mIoU	ACC	MAE↓
Swin-UNet [24]	89.61	82.88	95.91	0.0487
HiFormer [7]	90.75	84.41	96.49	0.0408
MEGANet [28]	90.92	84.61	96.69	0.0334
MADGNet [31]	89.33	82.22	93.73	0.0707
GA2-Net [10]	90.50	84.13	96.54	0.0392
TransUNETR [11]	90.87	84.56	96.61	0.0339
ConDSeg [43]	90.21	83.81	96.29	0.0460
LKCA-Net [12]	90.89	84.47	96.70	0.0380
Ours	91.95	86.17	97.23	0.0282

Note: ↓ indicates lower values are better. Best results are in bold.

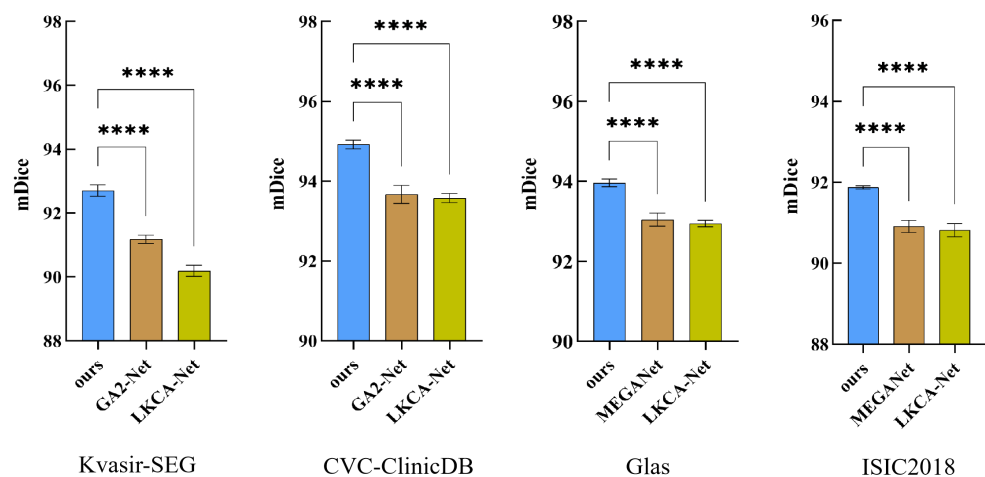


Figure 5. Statistical analysis results for four datasets. **** denotes $p < 0.0001$.

Qualitative Evaluation.

Figures 6 and 7 present qualitative comparisons on the Kvasir and CVC-ClinicDB datasets, respectively. As illustrated in the figures, other methods such as Swin-UNet and GA2-Net exhibit boundary leakage or false segmentation of large background areas in low-contrast regions, primarily due to their over-reliance on spatial-domain pixel features. Although MEGANet and MADGNet incorporate frequency-domain operations to enhance edge response, they still suffer from boundary discontinuities and over-segmentation in low-contrast areas. This indicates that their local or band-limited frequency-domain enhancement is insufficient for handling complex boundary scenarios. In contrast, FBNet leverages the FBE module to boost the amplitude spectrum response, thereby strengthening the signal at low-contrast edges. This enables the model to effectively maintain the integrity of target contours in complex scenes. Furthermore, regarding the minute lesions in Figures 6 and 7, other models show significant omissions or false identifications of the background, whereas FBNet accurately captures and localizes these tiny regions. These results indicate that the proposed SEA module effectively improves the model's sensitivity to regions with high local variation and enhances its ability to perceive fine structural details.

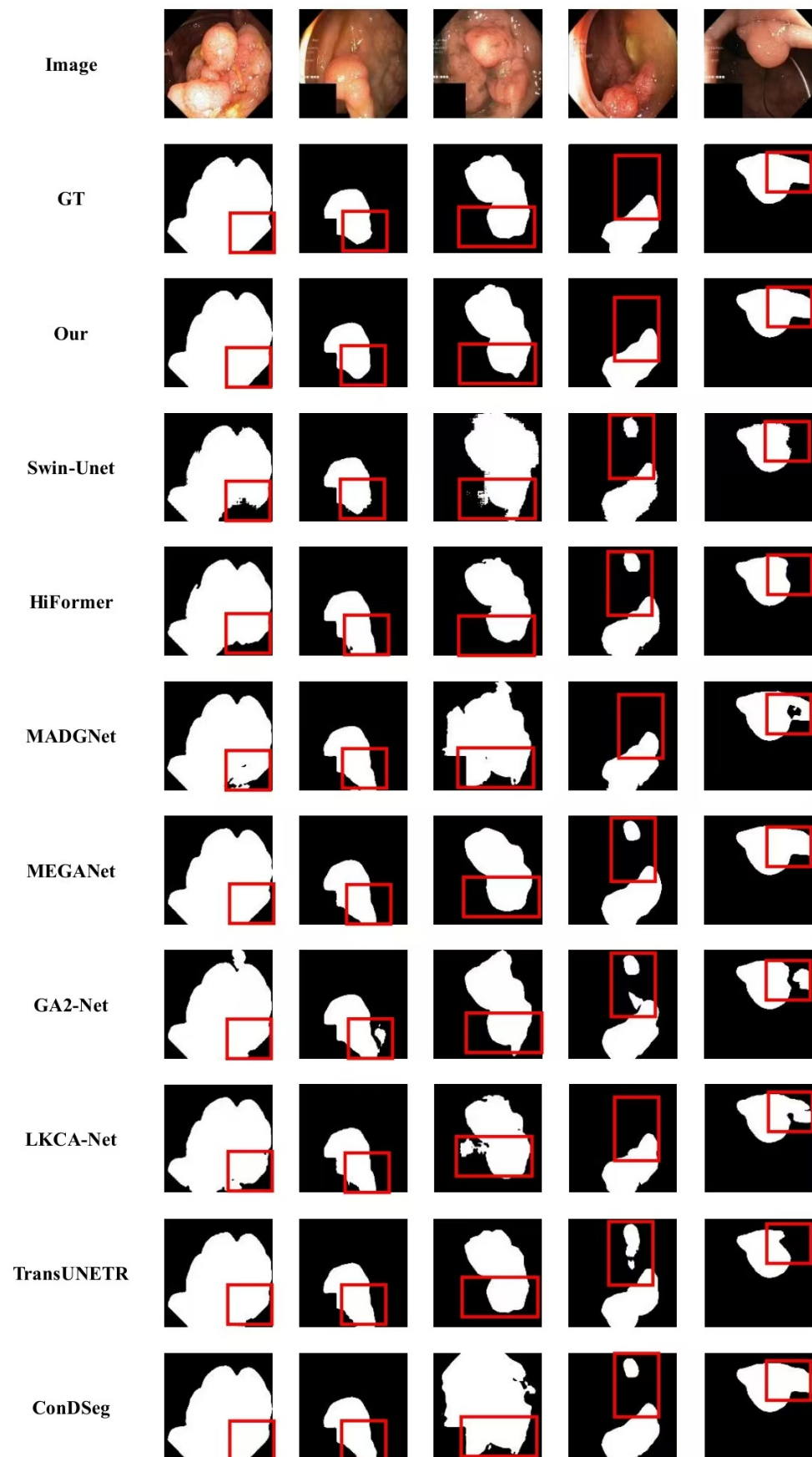


Figure 6. Qualitative results for the Kvasir dataset. Red bounding boxes highlight regions of performance differences.

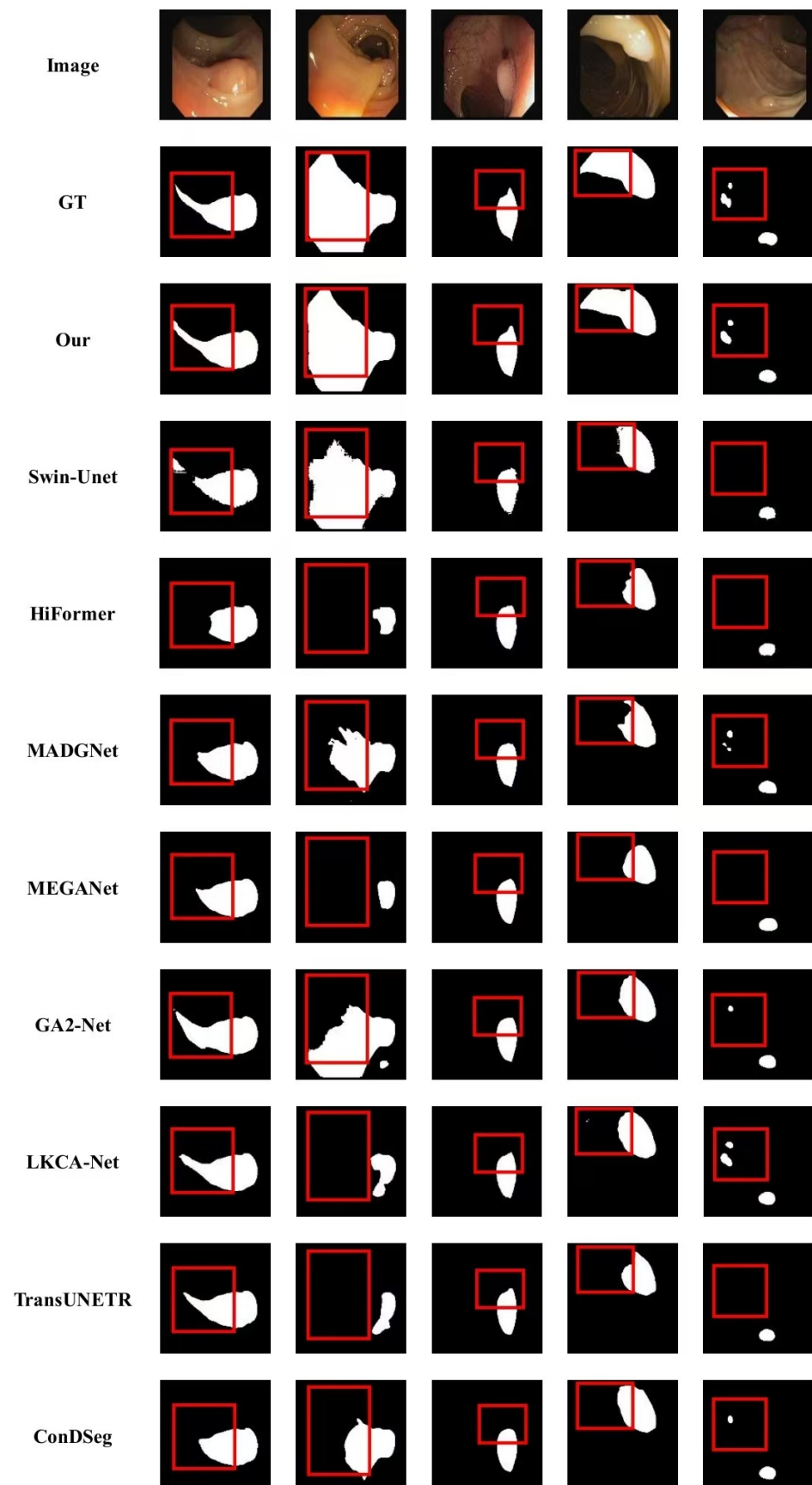


Figure 7. Qualitative results for the CVC-ClinicDB dataset. Red bounding boxes highlight regions of performance differences.

Figure 8 presents the qualitative results on the Glas dataset, further demonstrating the superiority of the proposed method. In the Glas dataset, the boundaries of target regions are extremely fragmented and intricate. Other methods exhibit issues such as over-connected within glandular structures, boundary discontinuities, or spurious segmentation results. In particular, MEGANet and MADGNet still exhibit excessive connectivity within glandular structures and blurred boundaries between glands when dealing with the densely packed and irregularly shaped glands in the Glas dataset. This limitation stems from their reliance on local frequency-domain decomposition, which makes it difficult to capture the global structural consistency required for complex topological structures. In contrast, our method generates results that exhibit greater similarity to the ground truth. It not only accurately preserves the internal lumen structures of the glands but also produces smooth and coherent boundaries. In the dataset of skin lesions shown in Figure 9, the challenge of the segmentation task lies in the low contrast between the lesions and the surrounding normal skin, as well as occlusion caused by complex artifacts such as hair. A qualitative comparison reveals that, when faced with such noisy scenarios, the feature representation capabilities of most models are significantly limited. Swin-UNet exhibits severe under-segmentation and local noise issues, while MEGANet and TransUNETR display significant morphological distortions in cases of hair occlusion and low-contrast lesion boundaries. In contrast, FBNet effectively suppresses interference from skin texture and hair noise, enabling precise lesion localization even under low-contrast conditions. In particular, for complex and irregular lesions, the segmentation masks generated by FBNet are closer to the ground truth labels thanks to the bidirectional guidance fusion module, which accurately captures fine edges.

To further evaluate the practical applicability of the proposed method, a systematic analysis of computational efficiency was conducted for all models, and the results are presented in Table 6. To ensure the rigor of the comparison, all models were evaluated under a unified experimental environment. The experimental results indicate that FBNet achieves a superior balance between model complexity and segmentation performance. Specifically, the computational complexity of FBNet is 20.09 GFLOPs, which is significantly lower than that of representative models such as MEGANet (26.34 GFLOPs) and GA2-Net (27.34 GFLOPs), and substantially lower than that of the computationally intensive model ConDSeg. Benefiting from the synergistic effect of the efficient PVT-v2-B2 backbone and the feature fusion module, FBNet is capable of extracting highly discriminative feature representations while maintaining high computational efficiency. These results demonstrate the potential of the proposed method in clinical medical image analysis scenarios.

Table 6. Computational cost comparison of different segmentation models.

Method	FLOPs (G)	Params (M)
Swin-UNet	25.54	41.34
HiFormer	17.75	34.14
MEGANet	26.34	29.27
MADGNet	25.82	33.37
GA2-Net	27.34	31.94
TransUNETR	12.48	10.25
ConDSeg	228.53	45.55
LKCA-Net	36.13	171.13
Ours	20.09	37.79

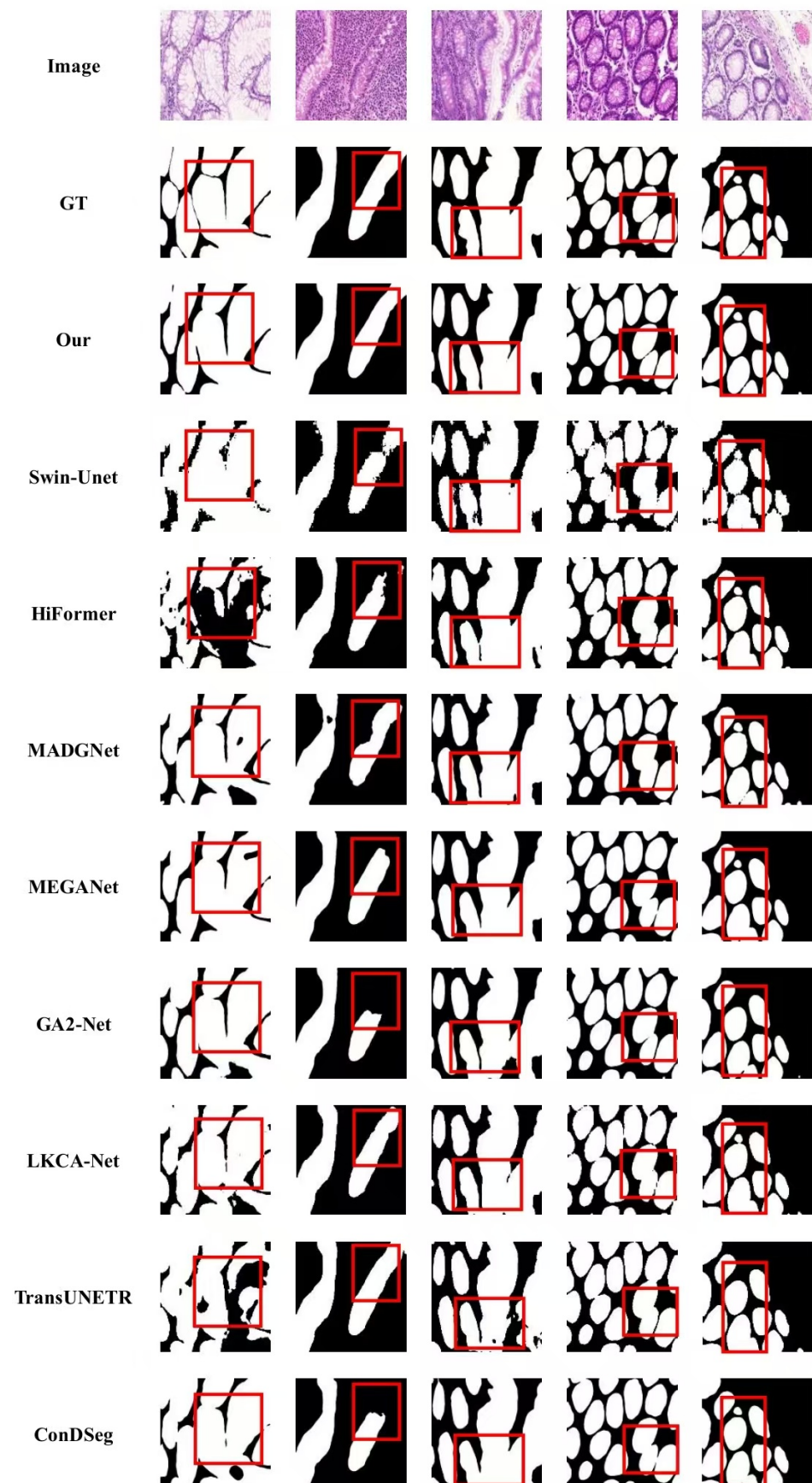


Figure 8. Qualitative results for the Glas dataset. Red bounding boxes highlight regions of performance differences.

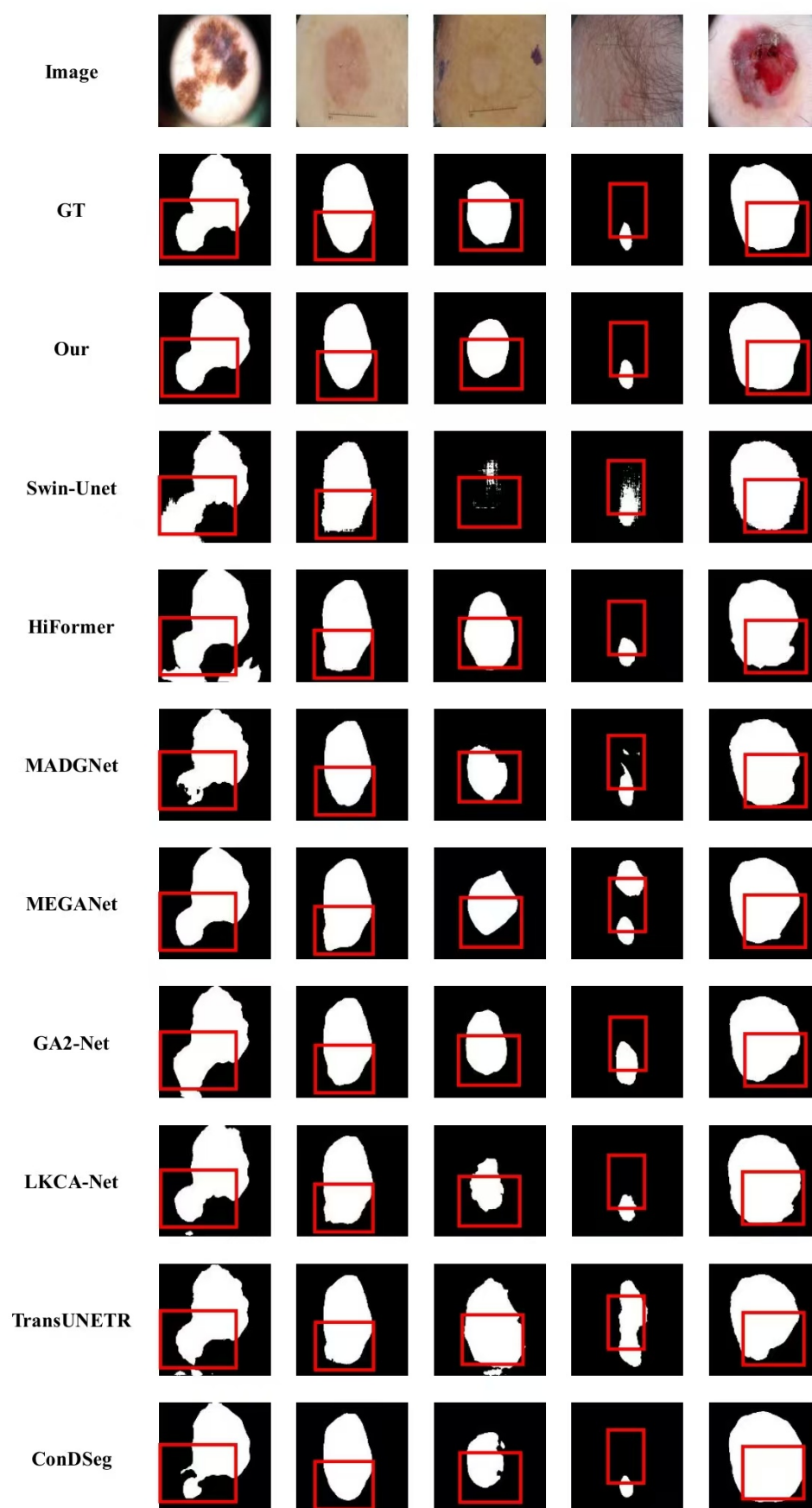


Figure 9. Qualitative results for the ISIC2018 dataset. Red bounding boxes highlight regions of performance differences.

4.5. Ablation Study

To assess the contribution of each key component, we conducted ablation experiments by progressively removing individual components while maintaining the same experimental settings. The quantitative results are summarized in Table 7.

Table 7. Ablation experiments for different modules.

FBK	Bidirectional Guidance Mechanism	SEA	mDice	mIoU	ACC	MAE↓
×	×	×	91.86	86.32	97.74	0.0228
✓	×	×	92.36	87.19	97.82	0.0220
✓	✓	×	92.44	87.40	98.10	0.0193
✓	✓	✓	92.82	87.92	98.18	0.0184

Note: × indicates that the module is not used; ✓ indicates that it is used; ↓ indicates lower values are better.

When all modules were removed, the model obtained its lowest performance, with mDice and mIoU at 91.86% and 86.32%, respectively. This indicates that relying solely on conventional spatial convolutions is insufficient for fully capturing subtle boundary information in medical images. After introducing the FBE module, mDice improved to 92.36%, validating the effectiveness of the frequency-domain enhancement strategy. By modulating the amplitude spectrum in the frequency domain, FBE effectively strengthens the edge response and texture detail representation in low-contrast regions, compensating for the limitations of spatial-domain operations in weak-boundary scenarios. Building upon this, the integration of the bidirectional feature guidance mechanism further improved mDice and mIoU to 92.44% and 87.40%, respectively. By establishing bidirectional interaction between shallow-level detail features and deep-level semantic features, this module effectively suppresses background noise and mitigates the semantic gap issue in multi-scale feature fusion. Ultimately, when the SEA module was combined, the model achieved optimal performance, with mDice, mIoU, and MAE reaching 92.82%, 87.92%, and 0.0184, respectively, significantly outperforming the baseline model. The SEA module explicitly enhances the model's sensitivity to structural changes and boundary regions through directional attention and variance attention mechanisms. The aforementioned ablation experiments validate the synergistic effectiveness of each module. FBE strengthens structural feature representation from the frequency-domain level, while BGF achieves refined modeling through cross-layer feature guidance. By integrating with the SEA module, the framework further focuses on local areas with high variation, thereby realizing a synergistic optimization of structural modeling and edge perception. Collectively, these three components enhance the model's capability to perceive structural and edge information, enabling more precise target segmentation.

Figure 10 presents the visualization results of segmentation under different module combinations. It is evident that the baseline model, lacking structural and boundary-aware guidance, struggles to accurately capture the morphology and contours of lesions. The introduction of the FBE module enhances feature contrast and structural representation capability, enabling the model to better capture structural characteristics of lesions. On the basis of this foundation, the BGF module was subsequently integrated and combined with cross-layer feature guidance and the SEA module. This strategy further strengthens the capability to perceive directional structures and boundary features. As a result, the structural integrity and boundary accuracy of the segmentation results are significantly improved.

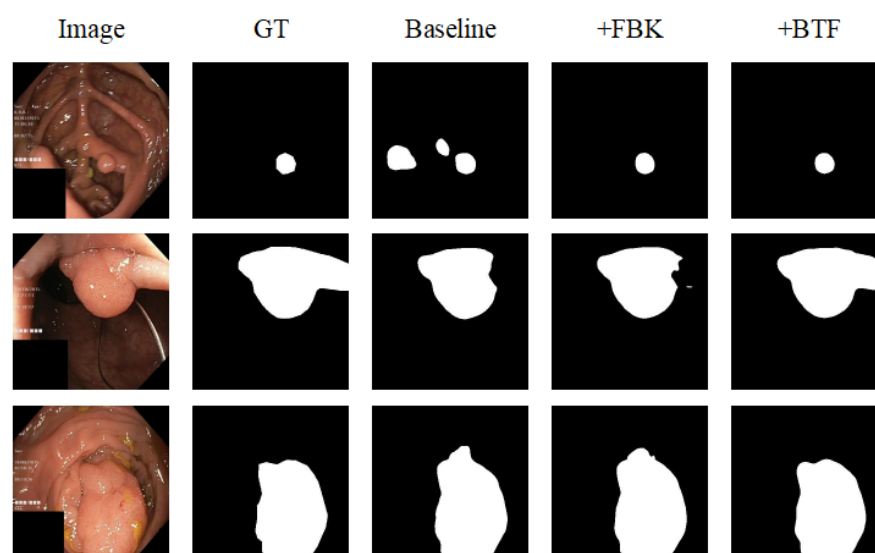


Figure 10. Visualization of segmentation results across different module combinations.

To investigate the effectiveness of different loss functions, the segmentation performance of Focal loss, Dice loss, BCE loss, and the combined Dice-BCE loss were all compared. The quantitative results are reported in Table 8. Although Focal loss exhibits certain advantages in handling class imbalance, it is marginally outperformed by the other methods in terms of overall regional consistency and boundary delineation capability. In contrast, the combined Dice and BCE loss achieves the best results across all evaluation metrics, including mDice, mIoU, Acc, and MAE, demonstrating more stable and comprehensive segmentation performance. These results indicate that integrating Dice loss with BCE loss provides superior optimization performance in balancing regional consistency and boundary detail preservation.

Table 8. Quantitative comparison of loss functions.

Loss Function	mDice	mIoU	ACC	MAE↓
Focal Loss	92.28	86.99	97.86	0.0214
Dice Loss	92.22	86.95	97.86	0.0214
BCE Loss	92.33	87.36	98.05	0.0202
Dice-BCE Loss	92.82	87.92	98.18	0.0184

Note: ↓ indicates lower values are better.

Additionally, the performance of different backbone networks was explored, as shown in Table 9. It can be observed that traditional convolutional backbone networks, ResNet30 and ResNet50 [44], achieve relatively stable performance in segmentation tasks. However, both exhibit high parameter counts and computational costs, with ResNet50 reaching 70.25 million parameters. In contrast, the PVT series of backbones demonstrates superior performance while maintaining similar computational efficiency. This advantage stems from their efficient global modeling and multi-scale feature extraction capabilities. Specifically, PVT-v2-B0, as a smaller-scale model, exhibits the lowest computational cost of 3.74 G FLOPs but shows noticeably lower mDice and mIoU scores compared with larger variants. Both PVT-v2-B4 and PVT-v2-B2 substantially enhance segmentation performance. Among them, PVT-v2-B2 achieves the optimal balance between performance and complexity, with mDice, mIoU, and MAE values of 92.82%, 87.92%, and 0.0184, outperforming other backbone networks. Therefore, PVT-v2-B2 was adopted as the backbone network

to perform feature extraction, enabling the model to obtain more discriminative feature representations while retaining high efficiency.

Table 9. Performance comparison of different backbone networks.

Backbone	FLOPs (G)	Params (M)	mDice	mIoU	ACC	MAE↓
ResNet30	18.24	32.29	91.40	85.70	97.47	0.0257
ResNet50	42.92	70.25	91.43	85.97	97.63	0.0241
PVT-v2-B0	3.74	6.66	86.32	78.73	95.82	0.0423
PVT-v2-B4	37.49	74.99	92.55	87.66	97.87	0.0216
PVT-v2-B2	20.09	37.79	92.82	87.92	98.18	0.0184

Note: ↓ indicates lower values are better.

5. Discussion

Beyond pixel-level segmentation precision (e.g., mDice and mIoU), the capability of computer vision algorithms to accurately distinguish and count individual targets is equally vital in practical clinical deployment [45,46]. Determining whether a specific region represents a single large lesion or is composed of multiple adjacent structures of varying sizes holds fundamental clinical significance. Although the research focus of FBNet is semantic segmentation, it demonstrates superior performance in boundary-aware modeling. By leveraging frequency-domain enhancement and synergistic guidance between deep and shallow features, FBNet effectively segments lesion boundaries characterized by complex topologies and dense distributions. This precise grasp of structural integrity not only enhances overall segmentation robustness but also provides more stable structural support for subsequent instance-level segmentation, object counting, and multi-system integration.

6. Conclusions

To address the challenges of weak boundaries, low contrast, and complex textures in medical image segmentation, a segmentation model named FBNet is proposed in this paper from a joint frequency-domain and spatial-domain perspective. The model enhances the perception of fine-grained features and structural consistency through two core designs. First, the frequency-based enhancement module (FBE) employs a content-aware gating mechanism to adaptively modulate amplitude spectra, strengthening feature details and contrast from a global perspective. Second, the Bidirectional Guided Feature Fusion module (BGF) achieves semantic complementarity between deep and shallow features. It employs the Directional and Variance Attention mechanism to synergistically optimize structural modeling and edge perception. Experimental results on multiple datasets demonstrate that FBNet outperforms existing mainstream methods in both segmentation accuracy and boundary consistency.

Despite the superior performance of FBNet, it still possesses certain limitations. The integration of frequency-domain transformations and feature fusion mechanisms inherently increases computational overhead, which, to some extent, constrains the model's deployment in resource-constrained scenarios. Therefore, future research will focus on model lightweighting strategies, such as leveraging knowledge distillation techniques. Our goal is to significantly reduce the memory footprint and inference latency while maintaining high segmentation precision with only minimal loss, thus facilitating the deployment of FBNet on edge devices and mobile healthcare platforms.

Author Contributions: Conceptualization, W.L.; Data Curation, W.L.; Formal Analysis, Y.L., W.L.; Funding Acquisition, Y.L.; Investigation, J.L.; Methodology, J.L., Y.L., W.L.; Project Administration, Y.L. and W.L.; Resources, Y.L.; Software, J.L.; Validation, Y.L.; Visualization, J.L.; Writing—Original

Draft, J.L.; Writing—Review and Editing, Y.L. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Project of the Hebei Education Department, grant number ZD2022102.

Data Availability Statement: The datasets that support the results of this study are publicly available datasets, and the use of these datasets in this work adheres to the licenses of these datasets. The Kvasir dataset is available at <https://github.com/DebeshJha/Kvasir-SEG> (accessed on 6 March 2026). The CVC-ClinicDB dataset is available at <https://www.kaggle.com/datasets/balraj98/cvcclinicdb> (accessed on 6 March 2026). The GLAS dataset is available at <https://github.com/wangminghui02/GLAS-dataset> (accessed on 6 March 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Statistical comparison on the Kvasir and CVC-ClinicDB datasets.

Method	Kvasir			CVC-ClinicDB		
	mDice (%)	Std	p-Value	mDice (%)	Std	p-Value
Swin-UNet	89.70	0.11	$p < 0.0001$	92.49	0.09	$p < 0.0001$
HiFormer	89.26	0.09	$p < 0.0001$	91.15	0.11	$p < 0.0001$
MEGANet	90.26	0.10	$p < 0.0001$	91.70	0.10	$p < 0.0001$
MADGNet	90.81	0.14	$p < 0.0001$	93.27	0.12	$p < 0.0001$
GA2-Net	91.28	0.12	$p < 0.0001$	93.86	0.23	$p < 0.0001$
TransUNETR	89.92	0.13	$p < 0.0001$	91.63	0.11	$p < 0.0001$
ConDSeg	90.13	0.09	$p < 0.0001$	92.10	0.11	$p < 0.0001$
LKCA-Net	90.13	0.18	$p < 0.0001$	93.60	0.13	$p < 0.0001$
Ours	92.82	0.17	–	94.97	0.11	–

Table A2. Statistical comparison on the Glas and ISIC2018 datasets.

Method	Glas			ISIC2018		
	mDice (%)	Std	p-Value	mDice (%)	Std	p-Value
Swin-UNet	90.66	0.10	$p < 0.0001$	89.61	0.07	$p < 0.0001$
HiFormer	91.53	0.11	$p < 0.0001$	90.75	0.07	$p < 0.0001$
MEGANet	93.19	0.16	$p < 0.0001$	90.92	0.14	$p < 0.0001$
MADGNet	93.01	0.06	$p < 0.0001$	89.33	0.05	$p < 0.0001$
GA2-Net	93.11	0.07	$p < 0.0001$	90.50	0.10	$p < 0.0001$
TransUNETR	91.77	0.11	$p < 0.0001$	90.87	0.11	$p < 0.0001$
ConDSeg	92.53	0.07	$p < 0.0001$	90.21	0.05	$p < 0.0001$
LKCA-Net	93.00	0.07	$p < 0.0001$	90.89	0.15	$p < 0.0001$
Ours	94.03	0.09	–	91.95	0.04	–

References

- Gu, R.; Wang, G.; Song, T.; Huang, R.; Aertsen, M.; Deprest, J.; Ourselin, S.; Vercauteren, T.; Zhang, S. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imaging* **2020**, *40*, 699–711. [CrossRef]
- Xia, Q.; Zheng, H.; Zou, H.; Luo, D.; Tang, H.; Li, L.; Jiang, B. A comprehensive review of deep learning for medical image segmentation. *Neurocomputing* **2025**, *613*, 128740. [CrossRef]
- Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2015; pp. 3431–3440.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

5. Wang, R.; Lei, T.; Cui, R.; Zhang, B.; Meng, H.; Nandi, A.K. Medical image segmentation using deep learning: A survey. *IET Image Process.* **2022**, *16*, 1243–1267. [[CrossRef](#)]
6. Xiao, H.; Li, L.; Liu, Q.; Zhu, X.; Zhang, Q. Transformers in medical image segmentation: A review. *Biomed. Signal Process. Control* **2023**, *84*, 104791. [[CrossRef](#)]
7. Heidari, M.; Kazerouni, A.; Soltany, M.; Azad, R.; Aghdam, E.K.; Cohen-Adad, J.; Merhof, D. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*; IEEE: Piscataway, NJ, USA, 2023; pp. 6202–6212.
8. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 14–24.
9. Liu, X.; Gao, P.; Yu, T.; Wang, F.; Yuan, R.-Y. CSWin-UNet: Transformer UNet with cross-shaped windows for medical image segmentation. *Inf. Fusion* **2025**, *113*, 102634. [[CrossRef](#)]
10. Fiaz, M.; Noman, M.; Cholakkal, H.; Anwer, R.M.; Hanna, J.; Khan, F.S. Guided-attention and gated-aggregation network for medical image segmentation. *Pattern Recognit.* **2024**, *156*, 110812. [[CrossRef](#)]
11. Wang, Y.; Liu, X.; Hu, J.; Xu, C.; Bai, W.; Zheng, K.; Dong, W. Gated partial self-attention network for medical image segmentation. *Biomed. Signal Process. Control* **2026**, *112*, 108515. [[CrossRef](#)]
12. Ji, Z.; Nie, M. LKCA-Net: Hybrid large kernel and clustering attention network for medical image segmentation. *Biomed. Signal Process. Control* **2026**, *112*, 108432. [[CrossRef](#)]
13. Zhu, S.; Yan, Y.; Wei, L.; Li, Y.; Mao, T.; Dai, X.; Du, R. SECA-Net: Squeezed-and-excitated contextual attention network for medical image segmentation. *Biomed. Signal Process. Control* **2024**, *97*, 106704. [[CrossRef](#)]
14. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Online, 22 February–1 March 2022; Volume 36, pp. 2441–2449. [[CrossRef](#)]
15. Yao, W.; Bai, J.; Liao, W.; Chen, Y.; Liu, M.; Xie, Y. From cnn to transformer: A review of medical image segmentation models. *J. Imaging Inform. Med.* **2024**, *37*, 1529–1547. [[CrossRef](#)]
16. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 568–578.
17. Yuan, F.; Zhang, Z.; Fang, Z. An effective CNN and Transformer complementary network for medical image segmentation. *Pattern Recognit.* **2023**, *136*, 109228. [[CrossRef](#)]
18. She, D.; Zhang, Y.; Zhang, Z.; Li, H.; Yan, Z.; Sun, X. Eoformer: Edge-oriented transformer for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer Nature: Cham, Switzerland, 2023; pp. 333–343.
19. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
20. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018*; Proceedings 4; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
21. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2020; pp. 1055–1059.
22. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–2111. [[CrossRef](#)]
23. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. Unetr: Transformers for 3d medical image segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*; IEEE: Piscataway, NJ, USA, 2022; pp. 574–584.
24. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 205–218.
25. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. [[CrossRef](#)]
26. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; Glocker, B.; Rueckert, D. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999. [[CrossRef](#)]

27. Zhao, X.; Jia, H.; Pang, Y.; Lv, L.; Tian, F.; Zhang, L.; Sun, W.; Lu, H. M²SNet: Multi-scale in multi-scale subtraction network for medical image segmentation. *arXiv* **2023**, arXiv:2303.10894.
28. Bui, N.-T.; Hoang, D.-H.; Nguyen, Q.-T.; Tran, M.-T.; Le, N. Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*; IEEE: Piscataway, NJ, USA, 2024; pp. 7985–7994.
29. Liu, Z.; Zheng, S.; Sun, X.; Zhu, Z.; Zhao, Y.; Yang, X.; Zhao, Y. The devil is in the boundary: Boundary-enhanced polyp segmentation. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 5414–5423. [[CrossRef](#)]
30. Rajasekar, D.; Theja, G.; Prusty, M.R.; Chinara, S. Efficient colorectal polyp segmentation using wavelet transformation and AdaptUNet: A hybrid U-Net. *Heliyon* **2024**, *10*, e33655. [[CrossRef](#)]
31. Nam, J.-H.; Syazwany, N.S.; Kim, S.J.; Lee, S.-C. Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 11480–11491.
32. Zou, F.; Liu, Y.; Chen, Z.; Zhanghao, K.; Jin, D. Fourier channel attention powered lightweight network for image segmentation. *IEEE J. Transl. Eng. Health Med.* **2023**, *11*, 252–260. [[CrossRef](#)]
33. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [[CrossRef](#)]
34. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2020; pp. 11534–11542.
35. Wei, J.; Wang, S.; Huang, Q. F³Net: Fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12321–12328. [[CrossRef](#)]
36. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-seg: A segmented polyp dataset. In *Proceedings of the MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, 5–8 January 2020*; *Proceedings, Part II 26*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 451–462.
37. Fan, X.; Zhou, J.; Jiang, X.; Xin, M.; Hou, L. CSAP-UNet: Convolution and self-attention paralleling network for medical image segmentation with edge enhancement. *Comput. Biol. Med.* **2024**, *172*, 108265. [[CrossRef](#)] [[PubMed](#)]
38. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. [[CrossRef](#)] [[PubMed](#)]
39. Silva, J.; Histace, A.; Romain, O.; Dray, X.; Granado, B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.* **2014**, *9*, 283–293. [[CrossRef](#)] [[PubMed](#)]
40. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; pp. 10347–10357.
41. He, Y.; Yi, Y.; Zheng, C.; Kong, J. BGF-Net: Boundary guided filter network for medical image segmentation. *Comput. Biol. Med.* **2024**, *171*, 108184. [[CrossRef](#)]
42. Xu, S.; Xiao, D.; Yuan, B.; Liu, Y.; Wang, X.; Li, N.; Shi, L.; Chen, J.; Zhang, J.X.; Wang, Y.; et al. Fafuse: A four-axis fusion framework of cnn and transformer for medical image segmentation. *Comput. Biol. Med.* **2023**, *166*, 107567. [[CrossRef](#)]
43. Lei, M.; Wu, H.; Lv, X.; Wang, X. Condseg: A general medical image segmentation framework via contrast-driven feature enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 4571–4579. [[CrossRef](#)]
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
45. Kristan, M.; Matas, J.; Tokmakov, P.; Felsberg, M.; Čehovin Zajc, L.; Lukežič, A.; Tran, K.-T.; Vu, X.-S.; Björklund, J.; Chang, H.J.; et al. The second visual object tracking and segmentation VOTS2024 challenge results. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2025; Volume 15629, pp. 357–383.
46. Laputin, F.; Sidorov, I.; Moshkin, A. Computer vision methods for assessing ovarian reserve. In *2024 International Russian Smart Industry Conference (SmartIndustryCon)*; IEEE: Piscataway, NJ, USA, 2024; pp. 656–661.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.