electronics

MDPI

Article

Redefining Contextual and Boundary Synergy: A Boundary-Guided Fusion Network for Medical Image Segmentation

Yu Chen [1,2], Yun Wu [1,2,*], Jiahua Wu [1,2], Xinxin Zhang [1,2], Dahan Wang [1,2] and Shunzhi Zhu [1,2]

[1] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China; 2222031098@s.xmut.edu.cn (Y.C.); wuplush@xmut.edu.cn (J.W.); zhangxinxin@xmut.edu.cn (X.Z.); wangdh@xmut.edu.cn (D.W.); szzhu@xmut.edu.cn (S.Z.)

[2] Fujian Key Laboratory of Pattern Recognition and Image Understanding, Xiamen 361024, China

* Correspondence: ywu@xmut.edu.cn

Abstract: Medical image segmentation plays a crucial role in medical image processing, focusing on the automated extraction of regions of interest (such as organs, lesions, etc.) from medical images. This process supports various clinical applications, including diagnosis, surgical planning, and treatment. In this paper, we introduce a Boundary-guided Context Fusion U-Net (BCF-UNet), a novel approach designed to tackle a critical shortcoming in current methods: the inability to effectively integrate boundary information with semantic context. The BCF-UNet introduces a Adaptive Multi-Frequency Encoder (AMFE), which uses multi-frequency analysis inspired by the Wavelet Transform (WT) to capture both local and global features efficiently. The Adaptive Multi-Frequency Encoder (AMFE) decomposes images into different frequency components and adapts more effectively to boundary texture information through a learnable activation function. Additionally, we introduce a new multi-scale feature fusion module, the Atten-kernel Adaptive Fusion Module (AKAFM), designed to integrate deep semantic information with shallow texture details, significantly bridging the gap between features at different scales. Furthermore, each layer of the encoder sub-network integrates a Boundary-aware Pyramid Module (BAPM), which utilizes a simple and effective method and combines it with a priori knowledge to extract multi-scale edge features to improve the accuracy of boundary segmentation. In BCF-UNet, semantic context is used to guide edge information extraction, enabling the model to more effectively comprehend and identify relationships among various organizational structures. Comprehensive experimental evaluations on two datasets demonstrate that the proposed BCF-UNet achieves superior performance compared to existing state-of-the-art methods.

Keywords: medical image segmentation; adaptive feature fusion; edge enhancement; U-Net

check for updates

Citation: Chen, Y.; Wu, Y.; Wu, J.; Zhang, X.; Wang, D.; Zhu, S. Redefining Contextual and Boundary Synergy: A Boundary-Guided Fusion Network for Medical Image Segmentation. *Electronics* **2024**, *13*, 4986. https://doi.org/10.3390/electronics13244986

Received: 31 October 2024
Revised: 13 December 2024
Accepted: 16 December 2024
Published: 18 December 2024

Copyright: © 2024 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Precise segmentation of medical images is essential in medical image analysis, aiming at detailed and pixel-level recognition of lesions [1]. Unlike natural images, medical images have special imaging equipment and principles. The details and contrast inside the medical images need to be particularly obvious to show subtle changes inside the body, such as tumors, blood vessels, etc. Moreover, the importance of medical images in medical analysis and diagnosis makes the edge detail information more worthy of attention, such as gradual changes in tissue density and indistinct boundaries between anatomical structures, as seen in Figure 1. With the advancement of modern intelligent healthcare, the demands for accuracy and efficiency in medical image segmentation are increasing, necessitating deep learning models with enhanced precision and greater adaptability [2].

Building on the advancements of deep learning in computer vision, Convolutional Neural Networks (CNNs) have demonstrated state-of-the-art performance across numerous segmentation tasks [3]. The U-Net [4], in particular, has emerged as a leading model for medical image segmentation, effectively utilizing its U-shaped architecture with skip

connections to mitigate the loss of spatial information [5]. However, CNNs are inherently limited by their restricted receptive fields, which hinders their ability to capture global context. This leads to the challenge of recognizing fuzzy boundaries and intricate structures in complex medical images, given the difficulty of extracting both local and global information. Moreover, the lack of contextual information limits CNN-based models from fully leveraging the relationship between spatial and structural features, which further complicates the accuracy of medical image segmentation. Recently, Transformer-based models have emerged as promising alternatives, as they possess self-attention mechanisms and have a high capability to capture long-range dependencies, which is a crucial factor in handling the global context of complex medical images [6]. Aiming to balance the capability between local and global feature extraction, hybrid networks [7,8] combine CNNs with the Transformers to enhance the performance of image segmentation. However, the increased model complexity, higher computational demands, and a need for large annotated datasets make hybrid networks impractical for many medical imaging scenarios. Considering the balance between efficiency and performance, the CNN-based model remains a practical and efficient choice for medical image segmentation [9].

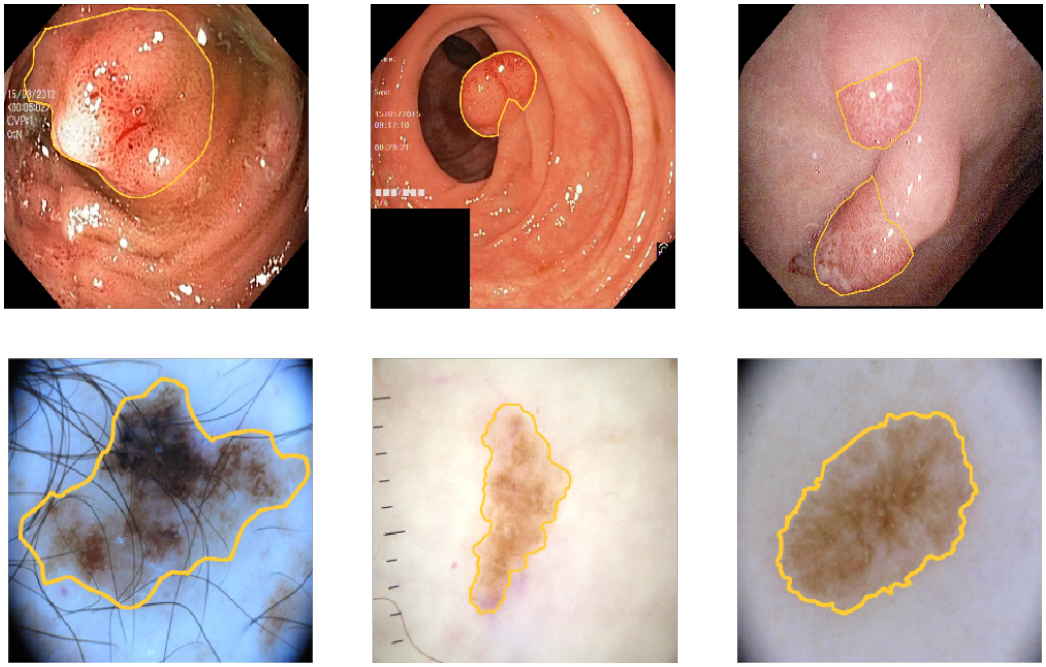


Figure 1. Two group examples of the skin lesion and colorectal polyp. The first row showcases images from colorectal polyp endoscopies, while the second row displays dermoscopic images of skin lesions. In each image, a solid yellow line delineates the boundary of the region of interest.

Medical images are often affected by noise and other artifacts, which may affect the diagnosis and analysis of diseases and lesions. Frequency domain image processing techniques can be applied to enhance the quality and medical images and the resolution of details. However, combining a CNN with a frequency filter often leads to vanishing and exploding gradient problems, because the single activation function causes the network difficulty in processing different frequency features. To utilize the advantage of the frequency filter and address the above problems, we propose a novel approach called the Adaptive Multi-Frequency Encoder (AMFE). Inspired by the Wavelet Transform (WT), the AMFE decomposes images into multiple frequency components to preserve crucial edge details and minimize noise. With the multi-frequency information, the AMFE can capture both local and global features to preserve essential details and global context simultaneously. To enhance the robustness of CNN in extracting the multi-frequency features, the AMFE proposes an adaptive activation strategy, enhancing its robustness to the highly nonlinear and intricate boundaries typical of medical images.

Additionally, existing U-Net variants often replace the original skip connections with alternative pathways, overlooking the critical relationship between boundary details and semantic context [10,11]. The general U-Net architecture struggles with accurately segmenting complex boundaries, particularly when the foreground target and the background tissue structures exhibit extremely high similarity, as illustrated in Figure 1. To overcome these limitations, we propose a Atten-kernel Adaptive Fusion Module (AKAFM) that integrates multi-level encoded features, enhancing semantic relationships and improving segmentation performance. In the AKAFM, a large, striped convolutional kernel and deformable convolution are used to obtain the dynamically weighted features, effectively enriching the contextual relationships of global channel features. Then, a Boundary-aware Pyramid Module (BAPM) is designed to improve the limited focus on boundary information in CNN-based segmentation methods. This module extracts multi-granularity edge information across sub-network layers by incorporating prior knowledge. As a result, BAPM significantly boosts the ability of the network to detect intricate edges.

In summary, we present a novel framework called Boundary-guided Context Fusion UNet (BCF-UNet), incorporating three innovative modules. First, the Adaptive Multi-scale Feature Extraction module (AMFE) is designed to effectively capture both local and global features. Next, the Atten-kernel Adaptive Fusion Module (AKAFM) is integrated into the skip connections to facilitate multi-scale feature fusion, enhancing the model's ability to understand complex anatomical structures. Finally, the Boundary-aware Pyramid Module (BAPM) is applied to the multi-scale decoders, strengthening their ability to extract fine-grained edge information across multiple scales. By combining these three modules, the framework effectively captures richer context and detailed features at each stage, leading to more precise segmentation results.

In summary, our main contributions are summarized as follows:

- We propose an AMFE module which decomposes images into multiple frequency components to preserve crucial edge details and minimize noise. To enhance the robustness of CNN in extracting the multi-frequency features, we design an adaptive activation strategy in AMFE to enhance the robustness of the highly nonlinear and intricate boundaries typical of medical images.
- We introduce a new multi-scale feature fusion module, the Atten-kernel Adaptive Fusion Module (AKAFM), designed to integrate low-level image features with high-level semantic information. This module leverages a combination of striped large kernel convolution and deformable convolution to capture rich contextual semantic information while maintaining edge details with high precision.
- We develop a straightforward yet efficient BAPM that leverages prior boundary information to supply complementary details, enabling more refined segmentation predictions.

2. Related Works

2.1. CNN-Based Segmentation Frameworks

In the past decade, Convolutional Neural Networks (CNNs) have become a cornerstone in the field of computer vision. Fully Convolutional Networks (FCNs) [12] have played a leading role in natural image semantic segmentation, particularly addressing challenges related to spatial resolution. The introduction of U-Net [4] marked a significant advancement in end-to-end semantic segmentation frameworks, demonstrating exceptional performance in medical image segmentation. Unlike FCNs, U-Net employs a U-shaped architecture with complementary upsampling and downsampling pathways, enhanced by skip connections that improve pixel-level localization and segmentation accuracy. This architecture effectively captures spatial hierarchies, leading to finer segmentation details [13]. TransUNet [7] is a hybrid architecture combining a CNN and Transformer. It enhances the ability of the model to deal with a large range of structures by introducing Transformer modules into the traditional U-Net network to capture long-range dependencies. It is especially suitable for medical images that require global context information. Similarly,

TransBTS [14] uses an architecture similar to 3D U-Net to model local and global information in spatial and slice/depth dimensions. It uses a 3D CNN backbone network in the encoder and decoder paths to capture local feature representations in spatial and depth dimensions. At the same time, TransBTS introduces the capabilities of Transformer by stacking ViT modules at bottleneck parts to take full advantage of its global modeling while maintaining efficient computation. However, Transformer models have high computational overhead and high demand for large-scale annotated data, which can be a challenge in the data-scarce medical field. Overall, the effective extraction of image context plays a critical role in enhancing segmentation performance.

2.2. Kolmogorov–Arnold Networks (KANs)

The Kolmogorov–Arnold theorem [15] provides the theoretical foundation for decomposing high-dimensional functions into multiple one-dimensional functions. Hornik et al. [16] proposed a continuous function of Multilayer Perceptrons (MLPs) with the Kolmogorov–Arnold theorem, which first introduced the Kolmogorov–Arnold theorem into neural network research. Recently, Liu et al. introduced a novel neural network architecture with the Kolmogorov–Arnold theorem, called Kolmogorov–Arnold Networks (KANs) [17]. KANs decompose complex multivariate functions into simpler one-dimensional functions, making them particularly effective for modeling intricate nonlinear relationships. In KANs, learnable B-splines are used as activation functions, excelling at capturing complex nonlinear relationships due to their flexibility and adaptability. The strong theoretical interpretability and explainability of KANs stem from their foundation in the Kolmogorov–Arnold theorem, which ensures that any multivariate continuous function can be represented as a superposition of one-dimensional functions and a finite number of summations. This inherent property allows the decomposition of complex functions into interpretable components, providing insights into how specific features contribute to the model's predictions. Furthermore, learnable B-splines add an additional layer of transparency, as their shape and behavior can be analyzed to understand feature interactions. Building on this theoretical foundation, Li et al. [18] first integrated KAN into U-Net by introducing dedicated KAN layers into the tokenized intermediate representation, offering improved performance in medical image segmentation. With the power capability of KANs, we propose an adaptive activation strategy which combines KAN with wavelet transform, which not only enhances the robustness of the feature extractor but also provides a pathway for improved explainability. The ability of wavelet transform to capture multi-frequency information complements KAN's decompositional nature, enabling a more interpretable feature extraction process.

3. Methodology

3.1. Overview of Our Proposed Architecture

Existing approaches, such as MultiResUNet [13] and U-Net++ [10], have shown that enhancing skip connections can improve model performance. However, these methods struggle to effectively capture contextual information and the intricate relationships between boundaries and semantic context. To overcome these limitations, we propose a novel medical image segmentation method, BCF-UNet, which addresses these challenges more effectively. In the BCF-UNet, We propose the AMFE, a robust feature extraction block to preserve essential edge details and reduce noise. Initially, it decomposes the image into different frequency components using wavelet transform. To enhance the robustness of CNN in extracting the multi-frequency features, a KANs-based adaptive activation strategy is proposed in AMFE. The extracted features are then processed by the AKAFM, which effectively uses semantic information to guide the segmentation of edge details. The deep semantic features are obtained from the feature map F_5. Outputs from the pooling operations are concatenated to enable multi-scale feature extraction, thereby enhancing the model's capacity to capture contextual information with greater accuracy. These fused features are subsequently forwarded to the BAPM, which employs a pyramid feature

extraction strategy alongside prior edge data to improve boundary detection. A visual summary of the proposed architecture can be found in Figure 2.

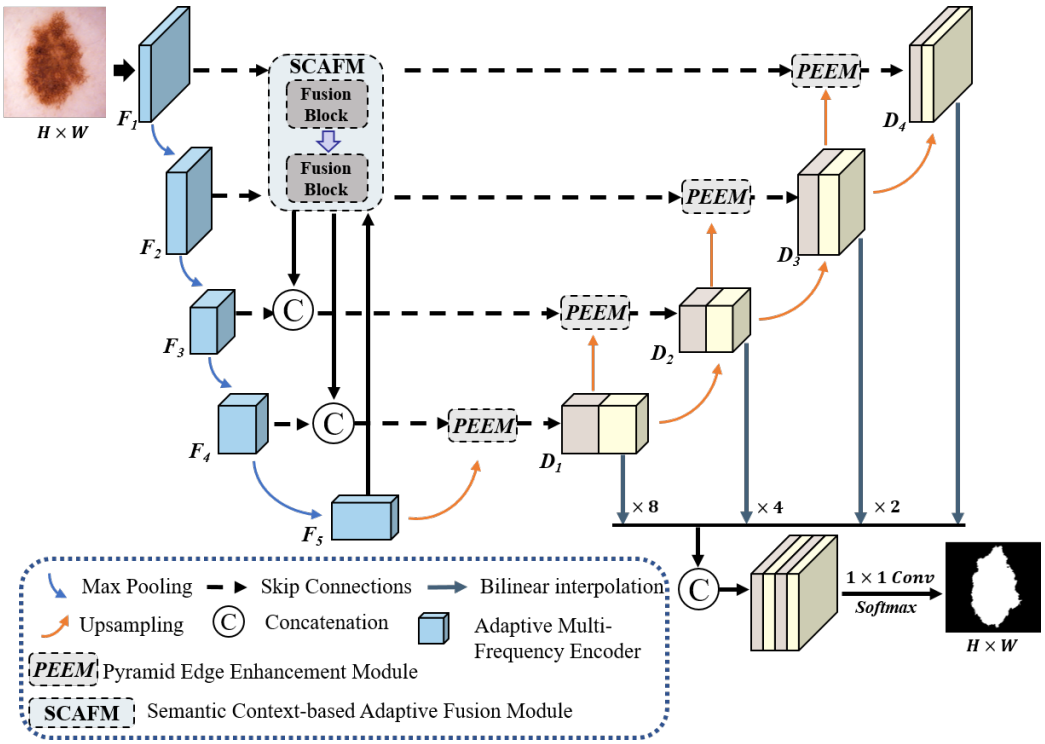


Figure 2. Illustration of the proposed BCF-UNet.

3.2. Adaptive Multi-Frequency Encoder

Multi-frequency image analysis is crucial in medical image segmentation because it can capture fine-grained details and broader contextual information. Traditional frequency domain methods like the Wavelet Transform (WT) have been widely adopted due to their capacity to decompose images into multi-frequency components, which is beneficial for preserving edge details and reducing noise. Hence, we combine WT with CNN to utilize the ability of WT to extract multi-frequency characteristics and the powerful spatial feature learning and nonlinear pattern recognition capability of CNN.

However, WT alone struggles when dealing with the complex nonlinear features commonly present in medical images, such as subtle texture variations and irregular, indistinct boundaries. This limitation arises because WT assumes a linear transformation framework that may not fully capture the intricate, high-dimensional characteristics of medical data. Meanwhile, activation functions like *ReLU*, *Sigmoid*, or *Tanh* are typically predefined and fixed for each neuron within CNNs or MLPs. This uniformity means the model cannot dynamically adapt its activation response based on data variations, which restricts its flexibility, particularly when handling diverse frequency domain features. However, we found experimentally that different frequency domain images obtained by wavelet transform may require different activation functions for effective activation. Because wavelet-transformed images contain diverse features across multiple scales and frequencies, each frequency may benefit from specialized activation functions that capture their unique characteristics. Therefore, a fixed activation function may limit the ability of the network to adapt to these variations and fully utilize the multi-frequency information provided by wavelet transform. To overcome these limitations, we propose introducing a flexible, learnable activation mechanism inspired by existing methods [17] that allow for adaptive function optimization. Unlike traditional fixed activation functions, this approach dynamically adjusts the activation behavior during training, based on the characteristics of the data. The activation functions are represented by flexible parameterized functions that are optimized as part of

the learning process. By incorporating adaptive activation functions, our model gains an enhanced capability to capture complex nonlinear relationships, making it more effective in processing fine details and challenging boundary features prevalent in medical images.

As shown in Figure 3, we firstly perform a WT on the given input feature F, decomposing it into four wavelet sub-bands. The WT utilizes a convolution operation with two kernels $L = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right], H = \left[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]$, followed by a standard downsampling operation. In AMFE, we compose the operation using the 2D Haar WT on both dimensions, resulting in a depth-wise convolution with a stride of 2, using the following set of four filters:

$$\begin{aligned} f_{LL} &= \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, f_{LH} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, \\ f_{HL} &= \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, f_{HH} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \end{aligned} \tag{1}$$

where f_{LL} is a low-pass filter used to extract the low-frequency components of an image. The remaining three filters, f_{LH}, f_{HL}, f_{HL}, are high-pass filters. For each input channel, the convolution output is given by:

$$[F_{LL}, F_{LH}, F_{HL}, F_{HH}] = WT(F) = Conv([f_{LL}, f_{LH}, f_{HL}, f_{HH}], F) \tag{2}$$

where F_{LL} represents the low-frequency components, which preserve the overall structure and primary information of the image, and F_{LH}, F_{HL}, F_{HL} represent the high-frequency components, depicting the details and edges of the image, in the horizontal, vertical, and diagonal directions, respectively.

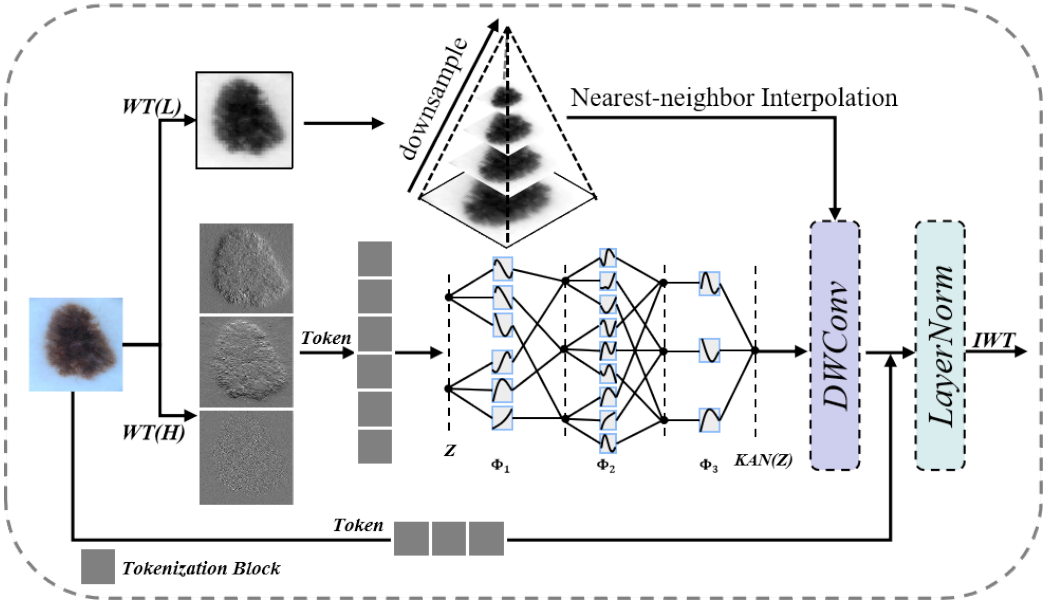


Figure 3. Adaptive Multi-Frequency Encoder.

Then, the inverse wavelet transform (IWT) is obtained by the transposed convolution of Equation (1):

$$\begin{aligned} F_{out} = IWT([F_{LL}, F_{LH}, F_{HL}, F_{HH}]) = Conv\text{-}transposed([f_{LL}, f_{LH}, f_{HL}, f_{HH}], \\ [F_{LL}, F_{LH}, F_{HL}, F_{HH}]) \end{aligned} \tag{3}$$

Subsequently, we handle these sub-bands independently to better utilize their unique properties for image segmentation. The low-frequency sub-band, characterized by smooth variations and large-scale structures, is well-suited for hierarchical analysis. To further

decompose and capture the multi-scale features of the low-frequency information, we transform it into a Gaussian pyramid [19]. The Gaussian pyramid provides a multi-resolution representation of the image, where each subsequent level is obtained by iteratively down-sampling and smoothing the previous level using a Gaussian kernel. As a result, we obtain a series of progressively lower-resolution images that effectively capture the global context at various scales. As shown in Figure 3, we take the original gray-scale image I_1 as the lowest level of the pyramid. Each subsequent level $I_y$of the Gaussian pyramid is generated by applying a Gaussian filter to the previous level $I_y - 1$ and then down-sampling it by a factor of 2. After the down-sampling operation, if the pixels in even rows and even columns of the image are removed, the gray value of the pixel at spatial position (i, j) in the y_{th} layer image can be computed using the following iterative algorithm:

$$I_y(i,j) = \sum_{m=-c}^{c} \sum_{n=-c}^{c} G(m,n) * I_{y-1}(2i-1-m, 2j-1-n) \tag{4}$$

where $*$ is the convolution operation, $y \in \{2, 3, \ldots, Y\}$, Y indicates the index of the Gaussian pyramid layer, and Y represents the total number of layers in the Gaussian pyramid. $G(m, n)$ is a $(2c + 1) \times (2c + 1)$ Gaussian window, which can be represented as follows:

$$G(m,n) = \frac{1}{2\pi\sigma^2} e^{-\frac{m^2+n^2}{2\sigma^2}} \tag{5}$$

where σ represents the standard deviation associated with the Gaussian filter. The series of images $\{I_1, I_2, \ldots, I_Y\}$, generated through the aforementioned convolution and down-sampling operations, forms the Gaussian pyramid.

Then, we tokenize the high-frequency features. First, we use a convolution operation to transform $F_{LL}, F_{LH}, F_{HL}, F_{HH}$ into features with an output size of $H_L \times W_L \times C_L$, where H_L and W_L are the height and width of the feature map, respectively, and C_L is the number of channels. Next, the convolution map is segmented into patches of size $P \times P$, resulting in a total number of patches $N = \frac{H_L \times W_L}{P^2}$. Each patch is then flattened into a one-dimensional vector of length $P^2 \times C_L$.

We subsequently use a trainable linear projection matrix E to map the flattened patch vectors into a D-dimensional embedding space. The projection matrix E has a size of $\mathbb{R}^{(P^2 \cdot C_L) \times D}$. All the flattened patch vectors X_L are projected through matrix E to obtain D-dimensional embedding vectors.

$$Z_0 = [X_L^1 E, X_L^2 E, \ldots, X_L^N E] \tag{6}$$

where X_L^1 represents the i_{th} flattened patch, and Z_0 is the collection of all D-dimensional embedding vectors after projection. With the embedding vectors, we pass them through three KAN layers to obtain multi-adaptive activation functions for each high-frequency domain. After that, different frequency information can be processed by a learnable activation function, and the network can filter for more confident frequency domain features. Subsequently, we add the original token features into the output of DWConv($DW(*)$) [20] for residual connection. Finally, a layer normalization (LN) and inverse wavelet are applied to restore the feature map to its original form. The i_{th} AMFE can be represented as follows:

$$Z_i = IWT\big(LN\big[Z_{i-1} + DW\big[I_y + KAN[WT(Z_{i-1})]\big]\big]\big) \tag{7}$$

where $Z_i \in \mathbb{R}^{H_i \times W_i \times D_i}$ represents the output feature map of the i_{th} layer. In our implementation, we set $L = 3$ and $i = 5$.

3.3. Atten-Kernel Adaptive Fusion Module

As shown in Figure 4, following the preliminary feature extraction, the two scales of features F_1 and F_2 are concatenated and passed through two distinct branches: the

Panoramic Channel Enhancement Module (PCEM) and the Deformable Spatial Attention Module (DSAM). These modules compute the channel context weight W_1 and spatial context weight W_2, respectively.

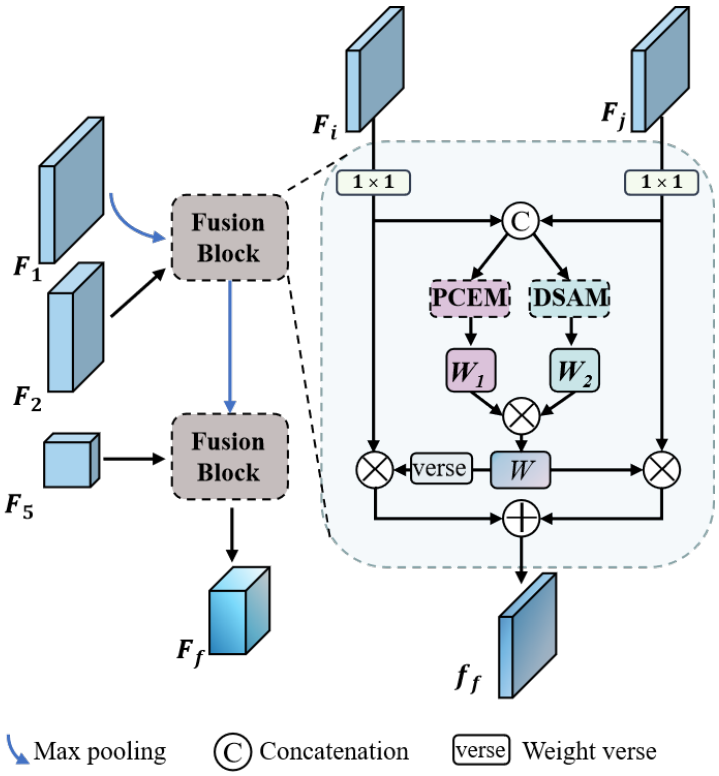


Figure 4. Atten-kernel Adaptive Fusion Module (AKAFM).

As illustrated in Figure 5, for the input features F, the PCEM generates the corresponding channel context weight W_1, which, combined with a striped large-kernel convolution, enables the PCEM to offer an extended receptive field, effectively capturing long-range dependencies in medical images. PCEM offers similar benefits compared to traditional large kernel convolutions but with significantly reduced parameters and computational cost.This efficiency is achieved by decomposing the deep convolution into three branches, each corresponding to different channel groups. Specifically, a convolution kernel of size $k \times k$ is decomposed into three distinct configurations: $k \times s$, $s \times k$ and $s \times s$. For instance, when $k = 13$ and $s = 3$, the parameter count for a traditional $k \times k$ convolution is 169. In contrast, the parameter count using our method is reduced to 87, representing an approximate 48.5% reduction compared to traditional $k \times k$ convolution. The computation of weight W_1 is as follows:

$$\begin{aligned} D^1 &= BN(f_{\mathrm{DW}}^{k\times s}(F)) \\ D^2 &= BN(f_{\mathrm{DW}}^{s\times k}(F)) \\ D^3 &= BN(f_{\mathrm{DW}}^{s\times s}(F)) \\ W_1 &= FC(C(F^1, F^2, F^3)) \end{aligned} \tag{8}$$

here, $D^{1,2,3}$ represents the feature maps obtained by passing the input feature F through three depth-wise convolutional layers, each followed by batch normalization (BN) [21]. The kernel size dimensions are denoted as $k = 13$ for the long dimension and $s = 3$ for the short dimension. The notation $C(*)$ refers to the concatenation.

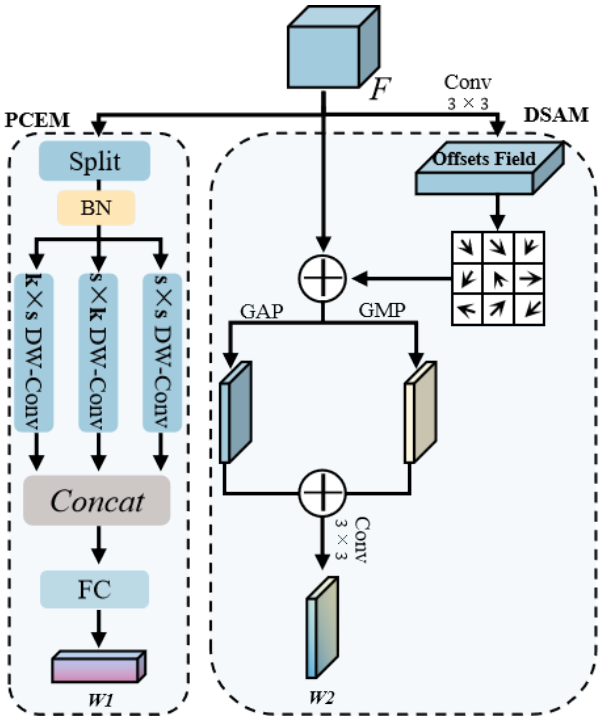


Figure 5. Panoramic Channel Enhancement Module (PCEM) and Deformable Spatial Attention Module (DSAM).

W_2 represents the weight of the DSAM, where deformable convolution enhances the network's robustness and sensitivity to irregular edges. In DSAM, the input feature F is processed by using Deform-DW convolution ($Conv_{D-DW}$). Depthwise convolution (DW Conv) reduces computational complexity by applying a separate filter to each input channel independently, instead of using a shared filter across all channels. This lowers the number of parameters and speeds up the computation. The key aspect of deformable convolution is the introduction of an offset field, which allows dynamic adjustment of the sampling positions of the convolutional kernel, rather than keeping them fixed. Specifically, the offset field, generated by a learnable subnetwork, adjusts the sampling points for each kernel to more precisely align with the complex edge structures in medical images. Afterward, two pooling operations are performed to obtain two feature maps, F'_{AP} and F'_{MP}. These feature maps are then concatenated, and the combined feature map is passed through a convolutional layer to compute the spatial-context weights W_2. The computation of W_2 can be expressed as:

$$\begin{aligned} W_2 &= Conv_{3\times3}(AveragePooling(F'); MaxPooling(F')]) \\ &= Conv_{3\times3}\left(\left[F'_{AP}; F'_{MP}\right]\right) \end{aligned} \tag{9}$$

where $F' = Conv_{D-DW}(F)$ and $Conv_{3\times3}$ refers to a standard convolution operation with a filter size of 3×3. Subsequently, the W_1 and W_2 are multiplied element-wise to compute the final weigh W, which can be expressed as:

$$W = \sigma(W_1 \times W_2) \tag{10}$$

here, $\sigma(*)$ denotes the sigmoid activation function, which is applied element-wise to the input. Using this fusion method, we can combine features from two different scales by first generating the final weight W and then applying it to the feature maps from the two scales. The feature fusion can be expressed as follows:

$$f_{i,j} = fus(F_i, F_j) = f_i \times (1 - W) + f_j \times W \tag{11}$$

where $fus(*)$ represents the entire fusion block operation, and $\times$ denotes channel-wise multiplication. The weight verse is denoted as $(1 - W)$, which is used to combine features with complementary emphasis. The final fused feature map F_f can be expressed as follows:

$$F_f = fus(F_5, f_{1,2}) \tag{12}$$

where (F_1, F_2, F_5) are the feature maps that are the outputs of the encoders at layers 1, 2, and 5 of the network, respectively.

3.4. Boundary-Aware Pyramid Module

In medical image segmentation, accurately segmenting lesion regions is often a significant challenge due to their complex shapes, varying sizes, and indistinct boundaries, as demonstrated in Figure 1, because some lesion regions have very complex edges. For example, in the setting of colorectal polyps, the similarity in appearance and color between the lesion area and the surrounding tissue makes it difficult to segment the lesion area. Therefore, improving the network's ability to capture edge details is crucial for accurately segmenting various lesion regions in medical images.

In this work, we propose a Boundary-aware Pyramid Module (BAPM) to address this challenge. To robustly augment the boundary information, we design a simple but effective pyramid feature extraction module. The module integrates multi-granularity information across different scales at each stage, ensuring that the network can accurately capture and enhance edges. The BAPM then combines the prior knowledge from the Sobel operator edge detection with the contextual information of surrounding pixels, guiding the network to focus more efficiently on the boundary regions. As illustrated in Figure 6, the feature map F_i with dimensions $H \times W$ is processed through pooling layers of different sizes, resulting in a reduced resolution of $\frac{H}{2} \times \frac{W}{2}$. Bilinear interpolation is then applied to upsample these features back to the original size of $H \times W$. This approach allows us to efficiently isolate the edge regions of the image, which correspond to the high-frequency information. The next step involves subtracting the feature F. The computational process is as follows:

$$F_{i,p}^k = F - Upsample(Pooling(F_i)), k \in \{1, \ldots, K\} \tag{13}$$

where $F_{i,p}^k$ denotes the edge features of the i-th stage obtained after applying the k-th pooling operations.

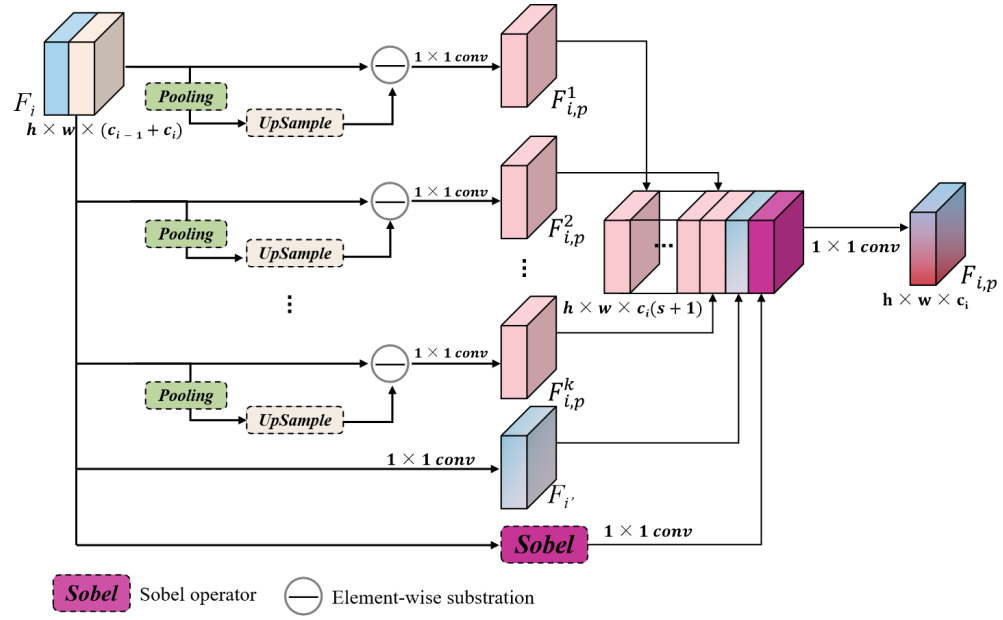


Figure 6. Structure of proposed Boundary-aware Pyramid Module (BAPM).

By making full use of these multi-granularity features, denoted as $F_{i,p}^1 \dots F_{i,p}^k$, we combine these features with the original ones, which are then further processed using prior knowledge provided by the Sobel operator, $S(F_i)$. The Sobel operator computes the partial derivatives in both the x and y directions within a 3×3 neighborhood centered around $f(x, y)$. To mitigate the influence of noise, the operator assigns higher weights to the central point in its 3×3 matrix [22]. The convolution kernel of the Sobel operator is defined as follows:

$$T_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad T_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{14}$$

We apply the Sobel operator for edge detection in the image by convolving it with both the horizontal template T_x and vertical template T_y. Afterward, the two gradient matrices are combined to compute the total gradient G [23]. Finally, the edge information obtained from the Sobel operator, denoted as $S(F_i)$, and the original features, F_i, which represent the edge information, is merged with the original feature map F_i through a 1×1 convolution operation:

$$F_{i,p} = Conv_{1\times1}\Big(C\Big(F_{i,p}^1, \dots, F_{i,p}^k, F_i, S(F_i)\Big); \theta_{i,p}\Big) \tag{15}$$

Here, $C(*)$ denotes the concatenation operation.

4. Experiments and Results

4.1. Dataset

In this paper, we utilized ISIC2018 [24] and Kvasir-SEG [25] to train and validate our proposed method and network. Detailed information about these datasets is as follows:

- The ISIC2018 dataset contains a diverse collection of skin lesion images, annotated with pixel-level segmentation masks. It is widely used for evaluating segmentation algorithms focused on skin lesion detection and classification, featuring images with various types of lesions, including melanomas, nevi, and basal cell carcinomas.
- The Kvasir-SEG dataset consists of gastrointestinal endoscopic images, primarily focused on colorectal polyp detection. It includes high-quality annotated images of polyps with segmentation masks, providing a valuable resource for evaluating segmentation models in the context of colonoscopy and endoscopy procedures.

4.2. Implementation Details and Metrics

All the experiments were implemented with PyTorch 1.12 on a single NVIDIA A100 GPU. The training process utilizeds the Adaptive Moment Estimation (Adam) optimizer with an initial learning rate of 0.0001, weight decay of 10^{-8}, a batch size of 16, and iteration length of 200 epochs. The learning rate was decayed by 0.5 every 256 epochs. The Dice loss function was employed for both training and evaluating each network. For final evaluation, five-fold cross-validation was applied to ensure robust performance metrics. In this paper, we evaluated the performance of our proposed method using four widely accepted metrics: the Dice similarity coefficient (Dice), Jaccard index (JA), accuracy (AC), and intersection over union (IoU).

$$\begin{aligned} \text{Dice} &= \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FN} + \text{FP}} \\ \text{JA} &= \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \\ \text{AC} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \\ \text{IoU} &= \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \end{aligned} \tag{16}$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives.

4.3. Ablation Studies

In this section, we conducted a series of ablation studies to thoroughly assess the contribution of each module within the proposed BCF-UNet framework, as shown in Table 1. Following this, the specific effects of various factors within the three modules are discussed in detail.

Table 1. Ablation experiments with proposed BCF-UNet. The bold font represents the best value in each metric.

Methods	Dice	IoU	Jaccard	Accuracy
U-Net	88.13	80.73	76.26	93.41
U-Net+AKAFM	92.89	87.34	79.31	94.79
U-Net+AMFE+AKAFM	93.06	87.68	79.38	94.80
U-Net+BAPM	90.49	83.85	77.26	94.62
U-Net+AMFE+BAPM	91.19	83.98	77.39	94.66
U-Net+AKAFM+BAPM	93.96	88.90	79.80	94.88
BCF-UNet (Full Model)	**94.09**	**88.99**	**79.86**	**94.95**

Ablation study of AMFE: To verify the effectiveness of our proposed module AMFE, we conducted a series of ablation experiments, as shown in Table 2. In these experiments, we replaced the KAN layer with MLP and tried three different activation functions, *ReLU*, *Sigmoid*, and *Tanh*, to evaluate the effectiveness of the adaptive activation function strategy. The superior performance of the original AMFE module can be attributed to its adaptive activation mechanism, which allows for more flexible and context-aware feature extraction compared to fixed activation functions like *ReLU*, *Sigmoid*, and *Tanh*. This adaptability is particularly advantageous in medical image segmentation tasks, where accurate boundary delineation and fine-grained feature representation are critical for achieving high segmentation performance.

Table 2. Ablation experiments with AMFE, replacing the original B-spline function with the activation function in parentheses. The bold font represents the best value in each metric.

Methods	Dice	IoU	Jaccard	Accuracy	Params (M)
AMFE (ReLU)	93.65	88.64	79.22	94.72	**8.00**
AMFE (Sigmoid)	93.33	88.59	78.99	94.69	**8.00**
AMFE (Tanh)	93.40	88.47	78.96	94.70	**8.00**
AMFE	**94.09**	**88.99**	**79.86**	**94.95**	11.51

Effects of different fusion methods: As shown in Table 3, three sets of comparative experiments were conducted on the AKAFM. In Table 3, AKAFM0 refers to the scenario where the encoders of the first, second, and fifth layers (F_1, F_2, F_5) were fed into AKAFM successively. In contrast, AKAFM1 involved only the encoders of the first and fifth layers (F_1, F_5). As indicated in the table, excluding the input F_2 led to a slight decrease in all metric scores for UNet. This reduction in performance was likely due to the substantial channel gap between F_1 and F_5, which may have resulted in significant loss of information. AKAFM2, which input all five layers into the AKAFM, showed similar experimental results to AKAFM0. However, the number of parameters in AKAFM2 was nearly seven times larger than that in AKAFM0. Taking into account both the parameter efficiency and performance, we ultimately chose AKAFM0 as our fusion strategy.

Table 3. Ablation experiments with different AKAFM linking methods. The bold font represents the best value in each metric.

Methods	Dice	IoU	Jaccard	Accuracy	Params (M)
U-Net	88.13	80.73	76.26	93.41	**3.20**
U-Net+AKAFM1	92.65	87.02	79.18	94.66	5.64
U-Net+AKAFM2	**92.90**	87.33	79.20	94.76	45.63
U-Net+AKAFM0	92.89	**87.34**	**79.31**	**94.79**	7.26

Effect of BAPM: As shown in Table 4, incorporating the BAPM resulted in improvements of 2.23% in Dice, 0.89% in IoU, and 1.18% in Accuracy. These results highlight the significant contribution of the BAPM in enhancing the model's ability to capture edge details. It is also noteworthy that removing the Sobel operator from the BAPM led to a slight reduction in performance, further emphasizing the importance of the Sobel operator in strengthening edge feature representation.

Table 4. Ablation Experiments with different BAPM methods. The bold font represents the best value in each metric.

Methods	Dice	IoU	Jaccard	Accuracy	Params (M)
U-Net	88.13	80.73	76.26	93.41	**3.20**
U-Net+BAPM (w/o Sobel)	90.36	81.62	77.19	94.59	3.68
U-Net+BAPM	**90.49**	**81.85**	**77.26**	**94.62**	4.25

4.4. Comparison with State-of-the-Art Methods

We evaluated our model against several state-of-the-art methods on two datasets, as shown in Table 5. On the ISIC2018 dataset, our model outperformed the current best-performing RA-Net [26] by a notable margin of 0.84% in the Dice score. When compared to transformer-based models, our CNN-based approach not only demonstrates superior performance, but also significantly reduces the number of parameters. While our model still has a higher parameter count than lightweight models like EGE-UNet [27], it strikes an optimal balance between performance and parameter efficiency. In terms of computational complexity, as measured by GFLOPs, our model demonstrates efficiency relative to its parameter size. Despite the increase in parameters from 3.20M (U-Net) to 11.51M, the GFLOPs increase only marginally from 7.16 to 8.12. This indicates that the additional complexity introduced by our model's advanced modules is well-optimized, resulting in a manageable computational cost.

Table 5. Comparison of different networks on ISIC2018 and Kvasir-SEG datasets. The bold font represents the best value in each metric.

		ISIC2018				Kvasir-SEG				Computational Complexity	
Networks	**Year**	**Dice (%)**	**IoU (%)**	**Jaccard (%)**	**Accuracy (%)**	**Dice (%)**	**IoU (%)**	**Jaccard (%)**	**Accuracy (%)**	**Params (M)**	**FLOPs (G)**
U-Net [4]	2015	88.13	80.73	76.26	93.41	85.59	77.73	78.52	78.52	3.20	7.16
R2U-Net [28]	2018	89.65	82.01	76.31	94.48	83.44	75.15	-	-	38.02	2.09
DeeplabV3+ [29]	2018	92.95	87.56	78.36	94.36	88.93	79.82	83.45	97.13	57.36	60.22
CE-Net [30]	2019	91.30	85.79	79.72	95.15	89.32	79.14	83.86	97.41	29.95	15.98
MultiResUNet [13]	2020	91.57	86.01	77.01	93.58	87.09	77.62	80.61	96.69	23.33	-
TransU-Net [7]	2021	92.30	86.29	77.29	94.51	90.25	79.86	84.20	97.62	63.88	60.75
SwinU-Net [31]	2021	91.03	84.02	76.93	**95.60**	89.24	79.10	83.82	97.40	36.13	33.39
CA-Net [32]	2021	92.67	87.16	79.42	94.85	90.73	80.21	84.43	97.45	2.79	6.33
MS RED [33]	2022	93.12	87.65	79.30	94.40	91.55	80.39	84.68	97.53	3.83	5.66
EGE-UNet [27]	2023	89.47	80.99	76.89	93.56	85.64	76.10	-	-	**0.053**	**0.07**
UNeXt [34]	2023	90.15	81.97	77.12	-	86.31	76.53	-	-	1.47	0.57
MFRA-Net [35]	2023	93.09	87.65	79.28	94.42	94.16	89.26	85.63	**97.85**	4.97	49.71
RA-Net [26]	2024	93.25	88.36	79.69	94.90	93.55	86.39	**85.97**	97.74	7.51	8.10
H-vmunet [36]	2024	88.65	79.49	76.03	94.37	-	-	-	-	8.97	0.74
VM-UNetV2 [37]	2024	89.73	81.37	77.36	94.65	91.34	84.15	83.47	97.52	5.11	1.56
Medsegdiff-v2 [38]	2024	93.29	85.33	79.52	94.02	-	-	-	-	46.00	983
BGDiffSeg [39]	2024	89.02	80.22	79.87	93.68	93.22	85.34	83.99	95.84	34.37	309.78
Ours	2024	**94.09**	**88.99**	**79.86**	94.95	**94.32**	**89.41**	85.59	97.80	11.51	8.12

4.5. Qualitative Analysis

To further validate the effectiveness of the two modules proposed for enhancing boundary differentiation, Figures 7 and 8 display the visualization results of our method compared to several state-of-the-art approaches. As seen in Figure 7, our method showed a distinct advantage in edge segmentation accuracy over existing techniques. Specifically, our model excelled on the skin lesion dataset, which contains significant inter-image variations, complex boundaries, and ambiguity. Similarly, the results in Figure 8 reveal that our method outperformed others, particularly in handling intricate edge details. These qualitative findings indicate that our approach can significantly enhance the accuracy of edge segmentation.

To assess the robustness of our model, we conducted additional training and testing on the colonoscopic polyp dataset. Table 4 presents our experimental results on the Kvasir-SEG dataset, comparing our method with the latest state-of-the-art techniques. The results demonstrate that BCF-UNet consistently outperformed current methods across most evaluation metrics. As shown in Figure 8, our approach notably excelled in processing detailed edges. Of particular significance is the model's strong performance even when the colon polyp foreground closely resembles the background. These experiments further underscore the robustness of BCF-UNet across different organs and a variety of medical segmentation tasks.

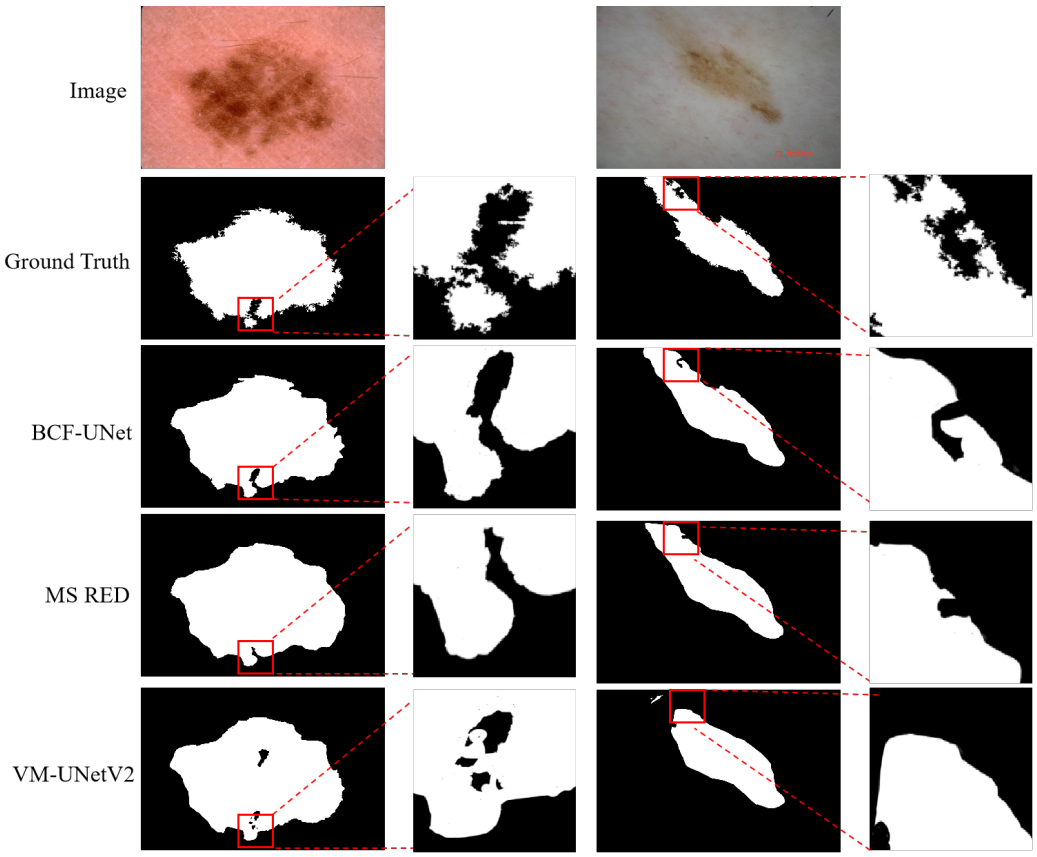


Figure 7. ISIC2018 Skin Lesion Dataset.

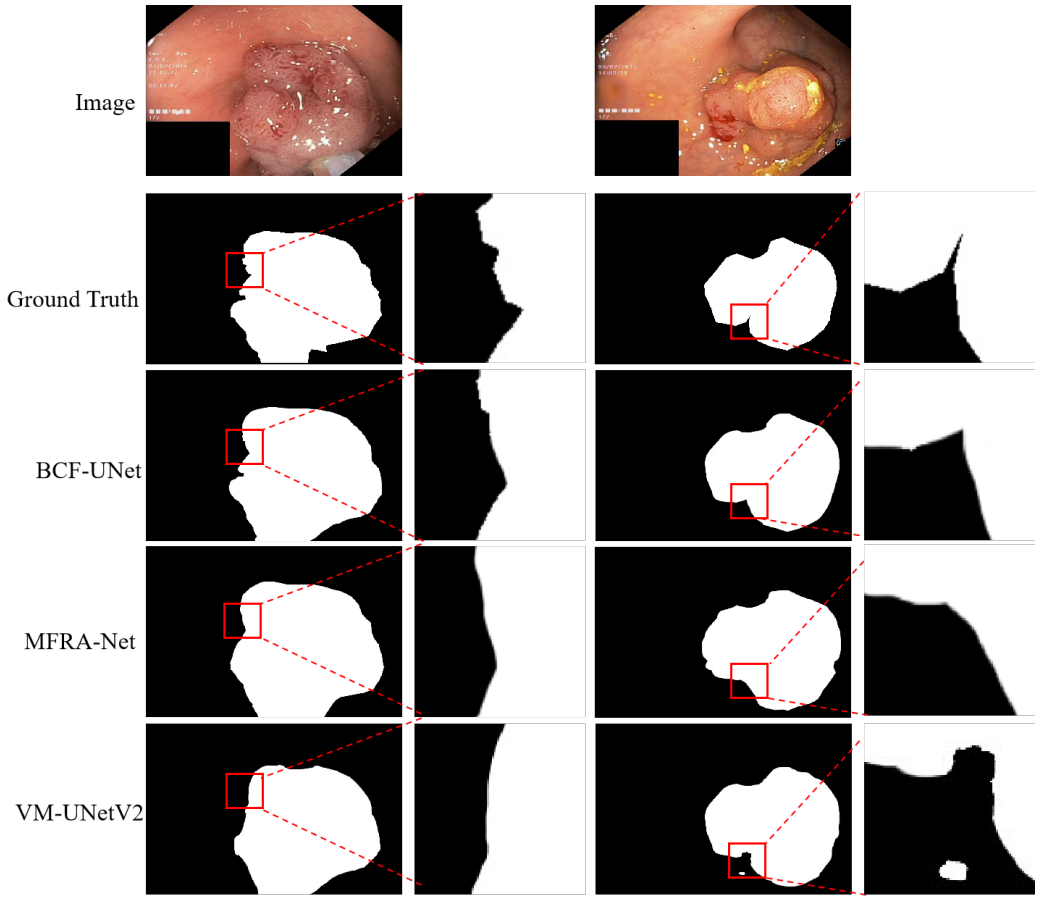


Figure 8. Kavasir -SEG polyp Dataset.

5. Conclusions

In this paper, we propose the Boundary-guided Context Fusion U-Net (BCF-UNet), introducing an innovative feature fusion strategy that enhances the traditional U-Net architecture by incorporating both global and local contextual information. A key component of our model is the Adaptive Multi-Frequency Encoder (AMFE), which, inspired by the Wavelet Transform and incorporating the adaptive activation strategy, decomposes images into multiple frequency components. This approach preserves crucial edge details while minimizing noise, effectively capturing both local and global features. Additionally, our Atten-kernel Adaptive Fusion Module (AKAFM) improves multi-scale feature fusion, strengthening the relationships between semantic and boundary information, while the Boundary-aware Pyramid Module (BAPM) enhances the ability of model to detect fine-grained edge information. The BCF-UNet has demonstrated superior performance on the ISIC2018 dataset and comparable results with leading methods for colon polyp segmentation.

Author Contributions: Methodology, Y.C.; Validation, S.Z.; Investigation, S.Z.; Writing—original draft, Y.C.; Writing—review & editing, Y.W. and J.W.; Visualization, X.Z.; Supervision, D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Natural Science Foundation of Xiamen (No. 3502Z202373058), the Unveiling and Leading Projects of Xiamen (No. 3502Z20241011), the Open Project of the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS2024101), the Science and Technology Planning Project617of Fujian Province (No. 2020Y9064), Fujian Key Technological Innovation and Industrialization Projects (2023XQ023) and Scientific Research Foundation for the Introduction of Talent at Xiamen University of Technology (YKJ23028R).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data cannot be made public to protect patient privacy.

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

References

1. Wang, R.; Chen, S.; Ji, C.; Fan, J.; Li, Y. Boundary-aware context neural network for medical image segmentation. *Med. Image Anal.* **2022**, *78*, 102395. [CrossRef] [PubMed]
2. Siddique, N.; Paheding, S.; Elkin, C.P.; Devabhaktuni, V. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access* **2021**, *9*, 82031–82057. [CrossRef]
3. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
4. 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; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
5. 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, Virtual, 22 February–1 March 2022; Volume 36, pp. 2441–2449.
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
7. 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.
8. 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 Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
9. Liu, X.; Peng, H.; Zheng, N.; Yang, Y.; Hu, H.; Yuan, Y. Efficientvit: Memory efficient vision transformer with cascaded group attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14420–14430.
10. 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: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
11. Ruan, J.; Xiang, S.; Xie, M.; Liu, T.; Fu, Y. MALUNet: A Multi-Attention and Light-weight UNet for Skin Lesion Segmentation. In Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 6–8 December 2022; pp. 1150–1156. [CrossRef]
12. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
13. Ibtehaz, N.; Rahman, M.S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* **2020**, *121*, 74–87. [CrossRef] [PubMed]
14. Wenxuan, W.; Chen, C.; Meng, D.; Hong, Y.; Sen, Z.; Jiangyun, L. Transbts: Multimodal brain tumor segmentation using transformer. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 109–119.
15. Kolmogorov, A.N. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. In *Doklady Akademii Nauk*; Russian Academy of Sciences: Moscow, Russia, 1957; Volume 114, pp. 953–956.
16. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
17. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. Kan: Kolmogorov-arnold networks. *arXiv* **2024**, arXiv:2404.19756.
18. Li, C.; Liu, X.; Li, W.; Wang, C.; Liu, H.; Yuan, Y. U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation. *arXiv* **2024**, arXiv:2406.02918.
19. Mertens, T.; Kautz, J.; Van Reeth, F. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2009; Volume 28, pp. 161–171.
20. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
21. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning. pmlr, Lille, France, 6–11 July 2015; pp. 448–456.
22. Chen, Y.; Wu, J.; Wang, D.H.; Zhang, X.; Zhu, S. Bridge the Gap of Semantic Context: A Boundary-Guided Context Fusion UNet for Medical Image Segmentation. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Urumqi, China, 18–20 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 165–179.

23. Jin-Yu, Z.; Yan, C.; Xian-Xiang, H. Edge detection of images based on improved Sobel operator and genetic algorithms. In Proceedings of the 2009 International Conference on Image Analysis and Signal Processing, Taizhou, China, 11–12 April 2009; IEEE: Piscataway Township, NJ, USA, 2009; pp. 31–35.
24. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv* **2019**, arXiv:1902.03368.
25. 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, Republic of Korea, 5–8 January 2020; Proceedings, Part II 26; Springer: Berlin/Heidelberg, Germany, 2020; pp. 451–462.
26. Naveed, A.; Naqvi, S.S.; Iqbal, S.; Razzak, I.; Khan, H.A.; Khan, T.M. RA-Net: Region-Aware Attention Network for Skin Lesion Segmentation. *Cogn. Comput.* **2024**, *16*, 2279–2296. [CrossRef]
27. Ruan, J.; Xie, M.; Gao, J.; Liu, T.; Fu, Y. Ege-unet: An efficient group enhanced unet for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 481–490.
28. Alom, M.Z.; Hasan, M.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv* **2018**, arXiv:1802.06955.
29. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
30. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Trans. Med Imaging* **2019**, *38*, 2281–2292. [CrossRef] [PubMed]
31. 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 Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 205–218.
32. 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] [PubMed]
33. Dai, D.; Dong, C.; Xu, S.; Yan, Q.; Li, Z.; Zhang, C.; Luo, N. Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Med. Image Anal.* **2022**, *75*, 102293. [CrossRef] [PubMed]
34. Valanarasu, J.M.J.; Patel, V.M. Unext: Mlp-based rapid medical image segmentation network. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 23–33.
35. Haider, A.; Arsalan, M.; Nam, S.H.; Hong, J.S.; Sultan, H.; Park, K.R. Multi-scale feature retention and aggregation for colorectal cancer diagnosis using gastrointestinal images. *Eng. Appl. Artif. Intell.* **2023**, *125*, 106749. [CrossRef]
36. Wu, R.; Liu, Y.; Liang, P.; Chang, Q. H-vmunet: High-order vision mamba unet for medical image segmentation. *arXiv* **2024**, arXiv:2403.13642.
37. Zhang, M.; Yu, Y.; Jin, S.; Gu, L.; Ling, T.; Tao, X. VM-UNET-V2: Rethinking vision mamba UNet for medical image segmentation. In Proceedings of the International Symposium on Bioinformatics Research and Applications, Kunming, China, 19–21 July 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 335–346.
38. Wu, J.; Ji, W.; Fu, H.; Xu, M.; Jin, Y.; Xu, Y. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 20–27 February 2024; Volume 38, pp. 6030–6038.
39. Guo, Y.; Cai, Q. BGDiffSeg: A Fast Diffusion Model for Skin Lesion Segmentation via Boundary Enhancement and Global Recognition Guidance. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Marrakesh, Morocco, 6–10 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 150–159.

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.