

QCFSNet: A Quaternion-Guided Inter-Channel and Frequency-Spatial Feature Interaction Strategy for Endoscopic Image Segmentation

Penghui Huang¹, Guoheng Huang^{1,*}, Guo Zhong^{2,*}, Shiqiang Ma³, Meixu Zhu⁴ and Qingjian Ye^{5,*}

¹ School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China

² School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, China

³ Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China

⁴ School of Biomedical and Pharmaceutical Sciences, Guangdong University of Technology, Guangzhou 510006, China

⁵ Department of Gynecology, The Third Affiliated Hospital of Sun Yat-Sen University, Guangzhou 510630, China

* Correspondence: kevinwong@gdut.edu.cn (G.H.); yb77410@um.edu.mo (G.Z.); yeqjian@mail.sysu.edu.cn (Q.Y.)

How To Cite: Huang, P.; Huang, G.; Zhong, G.; et al. QCFSNet: A Quaternion-Guided Inter-Channel and Frequency-Spatial Feature Interaction Strategy for Endoscopic Image Segmentation. *Artificial Intelligence and Emerging Technologies* **2026**, *3*(4), 14. <https://doi.org/10.53941/aiet.2026.100014>

Received: 9 April 2026

Revised: 10 August 2026

Accepted: 28 September 2026

Published: 8 October 2026

Abstract: In the field of medical consumer electronics, particularly in cervical cancer diagnostics, endoscopic imaging is often affected by uneven illumination and the high similarity in textures and features between lesion regions and surrounding tissues, which severely hinders accurate boundary perception and lesion segmentation. To address these challenges, we propose QCFSNet, a Quaternion-Guided Channel-Frequency-Spatial Feature Interaction framework for segmentation in endoscopic images. This approach incorporates a frequency representation mechanism based on quaternion convolution and wavelet transform to extract high-dimensional and highly discriminative features while alleviating the impact of illumination inconsistency on feature modeling. In addition, a boundary-adaptive attention module is introduced to suppress redundant responses and enhance the contrast between lesion and background regions, while a frequency-guided spatial-channel fusion mechanism strengthens semantic integration across multiple domains. Experiments on a private dataset (CervicalEndoDB) and public benchmarks (CVC-300 and UW-Sinus-Surgery-C/L) demonstrate that the proposed method holds great promise in improving segmentation perception in electronic endoscopic imaging, thereby better supporting downstream clinical tasks such as physician intervention planning.

Keywords: semantic segmentation; endoscopic image analysis; quaternion convolutions; wavelet transform; frequency domain attention

1. Introduction

The rapid development of deep learning has significantly accelerated the integration of artificial intelligence into electronic healthcare systems. In particular, electronic endoscopic imaging has become a crucial tool for physicians in acquiring visual information of lesion areas and assisting clinical decision-making [1], especially in the diagnosis of diseases such as cervical cancer. Accurate segmentation of lesion regions in endoscopic images not only improves diagnostic precision and efficiency, but also provides reliable references for treatment planning and prognosis assessment [2,3]. However, inconsistent lighting conditions during endoscopic acquisition often lead to unstable imaging quality. Additionally, lesion regions typically exhibit discrete distributions and irregular shapes, and share highly similar textures and colors with surrounding anatomical structures such as the squamocolumnar junction, vascular networks, and Nabothian cysts [4]. These challenges make it difficult to achieve precise perception and segmentation of lesion areas in cervical endoscopic images. The core difficulties include: (1) how to accurately



localize lesion regions that are sparsely distributed and morphologically irregular, which requires models to capture multi-scale contextual features while preserving local details; and (2) how to enhance boundary clarity between lesions and background, especially under conditions of visual similarity and uneven illumination, where deep networks often lose critical high-frequency structures, resulting in blurred edges. Therefore, designing segmentation strategies that are robust to illumination variation, capable of restoring high-frequency details, and effective in enhancing boundary perception remains a key scientific challenge in electronic endoscopic image segmentation.

To tackle these challenges, numerous researchers have developed advanced deep learning models [5–15], which have been extensively applied in medical image segmentation. U-Net initially proposed an encoder-decoder architecture incorporating skip connections to extract multi-scale features from images [16], achieving significant success and widespread application in medical image segmentation. Building upon U-Net, several methods and architectures have been developed. To highlight lesion foreground representation under multi-scale features, PraNet introduced a reverse attention mechanism [7], enhancing the distinction between foreground and background to improve lesion localization. However, it remains susceptible to interference from background information. To address this, RA-DENet refined the reverse attention mechanism by incorporating a context-aware interference elimination module [11], effectively reducing false positives and further improving lesion localization accuracy. CFA-Net reinforced boundary representation by integrating boundary information with semantic features across multiple scales [12]. Similarly, UMFNet adopts a multi-branch network to further enhance the comprehensiveness of feature extraction [14], thereby improving overall segmentation performance. However, these methods predominantly rely on spatial domain features. Although attention mechanisms are introduced to enhance feature learning, spatial attention remains insufficient in capturing boundaries within regions of similar color and texture, as demonstrated in Figure 1a. Therefore, how to supplement spatial representations with frequency-domain cues while retaining accurate localization and preserving fine-grained edges remains an unavoidable challenge for high-precision endoscopic segmentation in intelligent imaging systems.

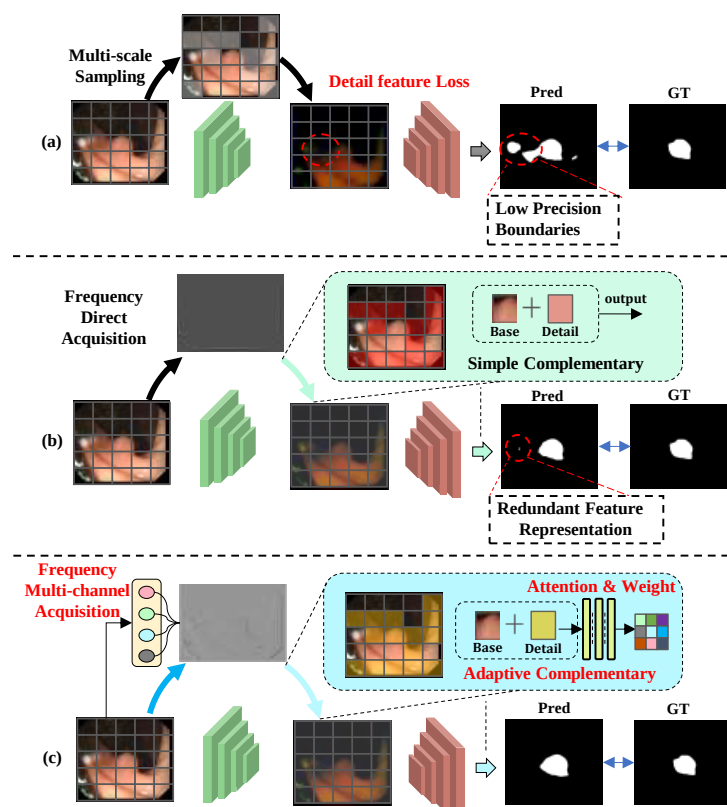


Figure 1. Different strategies for endoscopic segmentation. (a) The loss of high-frequency details caused by continuous convolutions and downsampling in the spatial domain leads to misclassification issues. (b) Simple frequency domain enhancement strategies can improve boundary details but may introduce redundant noise. (c) Integrated flexible frequency and spatial domain information, adaptive retention of high-frequency details and suppression of redundant information.

Consequently, numerous researchers have begun incorporating frequency-domain information to address the limitations of spatial-domain features. Under the premise of integrating multi-scale features, MEGANet employs a

Laplacian filter to extract high-frequency details [13], thereby enhancing boundary perception of lesion regions. Similarly, XNet leverages wavelet transform for high- and low-frequency feature fusion, facilitating the learning of boundary semantic information [17]. SW/SE-CNN proposes various wavelet transform techniques to enhance the semantic richness of features and improve foreground-background contrast [18]. However, while filter-based methods effectively capture high-frequency information, their flexibility is often insufficient in complex scenarios, such as bleeding or cysts. Conversely, wavelet-based approaches frequently exhibit limited boundary extraction capability and weak adaptability in challenging conditions characterized by low contrast, similar boundary regions, and high-frequency noise interference. These shortcomings can lead to redundant feature predictions, as depicted in Figure 1b. Therefore, a more adaptable and robust frequency-domain feature enhancement strategy is imperative to address the segmentation of image in such intricate scenarios.

Driven by the need for more flexible and effective feature extraction and integration, this paper introduces a novel frequency extraction method based on quaternion channel interaction and wavelet transform, combined with an adaptive learning mechanism for frequency and spatial features, as illustrated in Figure 1c. To implement this approach, an end-to-end boundary-enhanced segmentation network, termed QCFSNet, is proposed. QCFSNet consists of three core modules: Quaternion-Driven Frequency Enhancement Module (QFEM), the Boundary Adaptive Module (BAM), and the Spatial-Frequency Channel Interaction Module (SFI). The network uses ResNet-34 as the backbone for multi-scale feature extraction [19]. In QFEM, wavelet transform is applied to extract high-frequency and low-frequency features from multiple channels, and quaternion convolution is used to model these features, flexibly capturing frequency details and anti-interference information, thereby improving segmentation performance. Unlike traditional frequency domain methods that capture a single type of feature, QFEM also considers channel interaction, making the captured frequency information more flexible, enhancing detail expression, and reducing the impact of interference. Then, BAM combines and learns high/low-frequency information with multi-scale features through feature mask attention and gating mechanisms, enhancing focus on key regions, bridging the semantic gap between the spatial and frequency domains, and suppressing redundant information interference. Finally, in the decoding stage, SFI combines channel attention from CBAM with Fourier-based frequency domain attention [20], strengthening global feature representation after each decoder layer and improving boundary clarity. We validate our method through experiments on the private CervicalEndoDB dataset and several public endoscopic polyp datasets. The results show that our method significantly outperforms existing state-of-the-art methods in various metrics, such as Dice coefficient and IoU. Additionally, the visualized frequency features in the experiments in Section 4 further demonstrate the effectiveness and generalization ability of QCFSNet.

The main contributions of this paper are as follows:

- Our QFEM leverages wavelet transform and quaternion convolution to model the multi-channel integrated high and low-frequency features, enabling more flexible capture of effective frequency feature representations to alleviate detail loss in spatial domain operations.
- Our BAM combines feature mask attention and adaptive gating mechanisms to flexibly integrate frequency features from QFEM, enhancing frequency domain information learning, bridging the semantic gap between spatial and frequency domains, strengthening boundary feature expression, and suppressing redundant information interference.
- SFI introduces feature interaction with a combination of channel attention and Fourier transform-based frequency domain operations, efficiently capturing global correlations, enhancing pixel relationships, mitigating blur effects, and improving the representation of key regions.

2. Related Work

2.1. Quaternion Neural Networks

In the field of deep learning, increasing research efforts have focused on exploring effective representations for multidimensional inputs. The development of hypercomplex neural networks was initially motivated by complex-valued deep learning. Trabelsi et al. introduced Deep Complex Networks, establishing fundamental operations for complex-valued neural architectures and demonstrating the potential of extending neural networks beyond real-valued representations [21]. Subsequently, Zhu et al. proposed Quaternion Convolutional Neural Networks, which extended convolution operations into quaternion domains and enabled structured multi-channel feature representation through quaternion algebra [22]. Quaternions provide a unified representation for multiple input data, enabling effective modeling of interdependencies among inputs and overcoming the limitations of traditional real-valued methods in processing multidimensional data. Compared with independent channel processing strategies, such as group convolution, quaternion convolution explicitly models the correlations among multiple components through quaternion algebra. Meanwhile, unlike attention-based fusion methods that rely on additional weighting

mechanisms, quaternion convolution captures intrinsic inter-channel relationships with structured parameter sharing. Therefore, quaternion convolution provides a compact and effective representation for multi-channel feature learning. Recent studies have demonstrated the advantages of quaternions in various tasks, including color image recognition, classification, segmentation, 3D acoustic signal processing, and speech recognition [23–27]. Recent studies have further explored the combination of hypercomplex representations and frequency-domain analysis, particularly for medical image representation. Sigillo et al. investigated quaternion wavelet networks by integrating quaternion algebra with wavelet decomposition, demonstrating the potential of hypercomplex frequency-aware learning for generalizing single- and multi-channel medical inputs [28]. Grassucci et al. further proposed a task- and model-agnostic wavelet-driven framework that incorporates wavelet processing into deep networks for medical imaging applications, highlighting the effectiveness of frequency decomposition for extracting complementary representations [29]. These developments provide a foundation for integrating quaternion-based modeling with frequency-domain feature learning.

Inspired by these advancements, we incorporate quaternion convolutions into QCFSNet to enhance multi-channel frequency representation. Specifically, the Quaternion-Driven Frequency Enhancement Module (QFEM) leverages quaternion convolutions to jointly process high- and low-frequency features extracted via wavelet transforms together with gradient cues. By exploiting cross-channel correlations through quaternion algebra, QFEM enables adaptive fusion of complementary frequency components and enhances lesion boundary representation for accurate segmentation.

2.2. Frequency Domain Learning

Frequency domain information has been widely applied in image segmentation, particularly for capturing boundary details and texture features. Traditional frequency feature extraction methods include Discrete Cosine Transform (DCT), filter decomposition, and wavelet transform. As a classical frequency analysis tool, DCT has been introduced into CNNs to enhance image processing performance [30,31]. However, due to its fixed frequency basis, DCT lacks adaptability to local image features and noise variations, leading to limitations in handling complex backgrounds and fine-grained details. To this end, filter decomposition methods have been adopted to reduce noise interference and improve boundary threshold recognition accuracy through flexible frequency component extraction, as seen in [32,33]. Additionally, MEGANet further introduced Laplacian filters to capture high-frequency boundary features [13], enhancing the contrast between the foreground and background, thus improving boundary segmentation accuracy. Although filters are effective in extracting frequency features, their inherent preset thresholds poorly adapt to images, especially when processing different types of endoscopic images, where they may not effectively handle the challenges of complex backgrounds and noise variations. To overcome this, recent studies have started to explore combining wavelet transforms with neural networks to achieve dynamic, learnable frequency feature extraction methods. Wavelet transforms, through multi-scale decomposition, effectively capture both high and low-frequency information in images, possessing strong localization capabilities, making them flexible in handling the varied details of images. Models that combine wavelet transforms with deep networks for optimizing frequency feature extraction have been proposed in several works, demonstrating significant results in image restoration, super-resolution, and downsampling tasks [18,34–36]. Furthermore, wavelet-based learning strategies have been investigated for medical image analysis. Grassucci et al. proposed a task-agnostic wavelet-driven framework that integrates wavelet processing into deep networks, demonstrating the effectiveness of frequency decomposition for extracting complementary representations in medical imaging. Although these methods perform excellently in simpler backgrounds, they still struggle to flexibly handle the interference from various background information, particularly in complex backgrounds and highly similar tissue textures.

Motivated by these observations, QCFSNet combines quaternion-guided frequency modeling with adaptive spatial-frequency feature refinement. Specifically, QFEM jointly models multi-channel high- and low-frequency components together with gradient cues, while BAM adaptively integrates these frequency representations with spatial features to suppress redundant responses and strengthen boundary-related information. This design improves the utilization of complementary frequency and spatial cues under complex backgrounds and highly similar tissue textures.

2.3. Feature Interaction Learning

Feature interaction learning is critical in image processing tasks as it facilitates the network's ability to learn more discriminative semantic features, thereby improving performance in segmentation and classification tasks. Recently, various methods have been proposed to enhance feature interactions within networks. For instance, BCFNet introduces a boundary-guided semantic cross-fusion strategy to enhance multi-scale feature fusion [37],

while dual cross-attention mechanisms have also been explored to improve semantic feature integration [38]. In addition, channel prior attention [39] and multi-scale channel-spatial attention networks [40] have been designed to promote feature interaction by integrating spatial and channel attention in a unified framework. These approaches effectively highlight critical regions and improve the identification accuracy of lesion areas. However, in low-contrast or blurry images, these methods often face challenges of redundant recognition, particularly in complex backgrounds where the ability to discern key regions is limited. Thus, enhancing pixel-level correlations and mitigating the blurring effects caused by sampling remains essential. To address this, Kong et al. introduced frequency-domain element-wise multiplication within the Fourier transform to optimize feature interaction in the spatial domain, enabling more efficient correlation learning at reduced computational complexity [41]. Building on this concept, we designed the SFI module in QCFSNet. This module integrates channel attention and frequency-domain attention mechanisms to enable feature interaction across spatial and frequency domains via Fourier-based frequency-domain attention. By leveraging frequency-domain element-wise multiplication to capture spatial-domain correlations, the SFI significantly enhances pixel-level associations, alleviates ambiguity effects, and improves the representation of key regions. Consequently, it substantially strengthens the network's ability to distinguish lesion regions.

3. Materials and Methods

3.1. QCFSNet Structure

The structure of QCFSNet is shown in Figure 2, designed as an encoder-decoder segmentation network. Prior to the encoding stage, the input image is processed through the QFEM module for multi-channel frequency domain analysis to extract high and low-frequency information maps f_i^{fre} , which serve as supplementary frequency domain features.

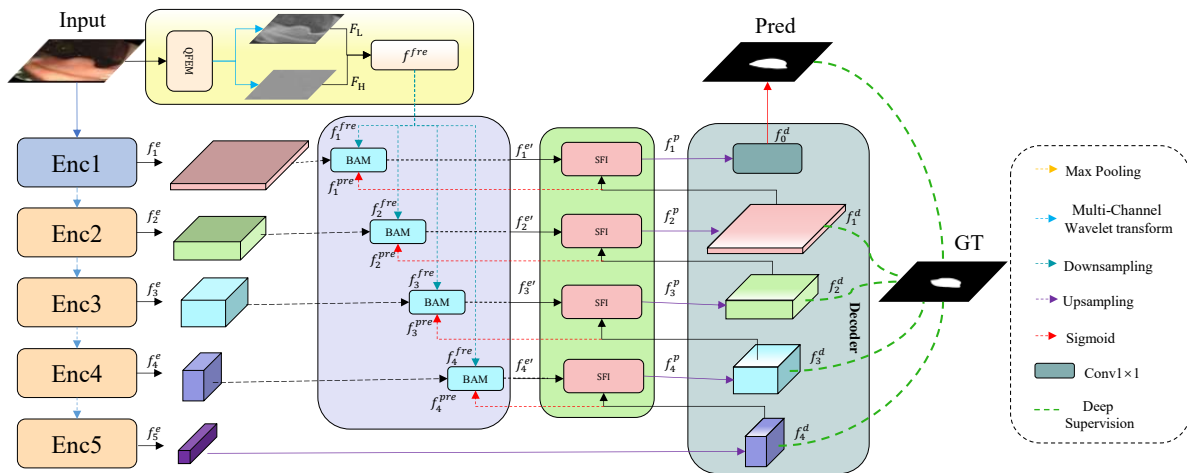


Figure 2. An overview of the QCFSNet network. The first step involves efficient frequency information extraction through QFEM, followed by feature modeling and adaptive learning via the multi-scale BAM. Finally, feature interaction through SFI after each decoder layer enhances the boundary feature representation and segmentation robustness.

In the encoding stage, ResNet-34 [19] is used as the backbone to extract multi-scale contextual and high-level features. The input image $x \in \mathbb{R}^{3 \times H \times W}$ is progressively downsampled by the encoder to generate multi-scale features $f_i^e \in \mathbb{R}^{C_i \times \frac{H}{s_i} \times \frac{W}{s_i}}$, where $i \in \{1, 2, 3, 4\}$, C_i denotes the channel dimension of the i -th encoder stage, and $s_i \in \{2, 4, 8, 16\}$ denotes the corresponding downsampling factor. To align the frequency-domain representations with the encoder features at different scales, we define $F_{D_i}(\cdot)$ as the frequency-domain projection and downsampling operation associated with the i -th encoder stage. It progressively adjusts the spatial resolution and channel dimension of the high- and low-frequency features extracted by QFEM. The aligned frequency-domain features are therefore represented as:

$$f_i^{\text{fre}} = (f_i^h, f_i^l) = F_{D_i}(f^h, f^l), \quad i \in \{1, 2, 3, 4\}. \quad (1)$$

where F_{D_i} denotes the scale-specific projection and downsampling operation used to align the high- and low-frequency representations with the i -th encoder stage.

During the decoding stage, the network recovers the image by progressively upsampling and incorporating enhanced features. The output feature at each layer is $f_i^d \in \mathbb{R}^{C_i \times \frac{H}{s_i} \times \frac{W}{s_i}}$, ($i = 1, 2, 3, 4$), with the smallest four scales' features generating prediction heatmap features that are passed to the corresponding BAM for inverse attention feature learning. The final decoded feature f_0^d is passed through the segmentation head to generate the

mask prediction.

At each decoding stage, the encoder feature f_i^e , the corresponding frequency-domain feature f_i^{fre} , and the prediction feature derived from the decoder feature f_i^d are jointly fed into the corresponding BAM. By integrating spatial and frequency-domain information, BAM enhances boundary-related representations while suppressing redundant responses, producing the enhanced encoder feature $f_i^{e'}$, expressed as:

$$f_i^{e'} = \text{BAM}(f_i^e, f_i^{\text{fre}}, \delta(f_i^d)), \quad i \in \{1, 2, 3, 4\}, \quad (2)$$

where $\delta(\cdot)$ denotes the prediction operation that maps the decoder feature f_i^d to the corresponding prediction heatmap f_i^{pre} .

To further improve image reconstruction quality and enhance boundary features, the SFI module is used to fuse the decoded output with the BAM-enhanced features $f_i^{e'}$ before passing them to the next decoding layer, optimizing the feature representation as:

$$f_i^{d'} = \text{SFI}(f_i^{e'}, f_i^d), \quad (3)$$

By combining low-level decoded features f_i^d with boundary-enhanced features $f_i^{e'}$, and applying channel and frequency domain attention, the reconstructed image quality is improved, leading to clearer feature representations and enhanced segmentation accuracy.

QCFSNet is designed as a general architecture that supports various encoder backbones and can perform feature analysis and detail enhancement across multiple domains, improving segmentation accuracy and generalization capability. The following sections will provide a detailed description of the three key modules in QCFSNet: the Quaternion-Driven Frequency Enhancement Module (QFEM), the Boundary Adaptive Module (BAM), and the Spatial-Frequency Channel Interaction Module (SFI).

3.2. Quaternion-Driven Frequency Enhancement Module (QFEM)

Filters can effectively extract different frequency information from images. However, fixed operators have limitations when dealing with most images, especially when the background exhibits similar color and texture features. Although conventional methods such as wavelet transform can adapt flexibly when combined with deep learning, their ability to capture such information remains insufficient. Therefore, the challenge still lies in how to flexibly extract effective frequency features to address the interference of similar backgrounds and uneven illumination. To this end, we have designed a multi-channel frequency extraction module based on quaternion convolution and Haar wavelet transform, the structure of which is shown in Figure 3a.

For the input variable $X \in \mathbb{R}^{3 \times H \times W}$, we first decompose it into RGB channels: $X_R \in \mathbb{R}^{1 \times H \times W}$, $X_G \in \mathbb{R}^{1 \times H \times W}$, $X_B \in \mathbb{R}^{1 \times H \times W}$. Next, we perform Haar wavelet transform on each channel separately, obtaining the high-frequency component X_H and low-frequency component X_L :

$$(A_C, H_C, V_C, D_C) = F_{\text{Haar}}(X_C) \quad (4)$$

$$X_{H_C} = F_{\text{Haar}}^{-1}(A_0, H_C, V_C, D_C) \quad (5)$$

$$X_{L_C} = X_C - X_{H_C}, \quad C \in \{R, G, B\} \quad (6)$$

where $F_{\text{Haar}}(\cdot)$ denotes the Haar wavelet decomposition operation, and $F_{\text{Haar}}^{-1}(\cdot)$ denotes the inverse Haar wavelet transform. The (A_C, H_C, V_C, D_C) represent the low-frequency component and the horizontal, vertical, and diagonal high-frequency components obtained from the wavelet transform of the channel feature X_C . A_0 represents a zero tensor of the same size as (A_C) , and the pure high-frequency component X_H is obtained by performing the inverse wavelet transform on (A_0, H_C, V_C, D_C) . The low-frequency component X_L is then derived by subtracting X_H from the corresponding channel input feature X_C .

Simultaneously, the input variable X is processed by gradient computation and grayscale transformation, represented as:

$$X_{\text{Grad}} = \text{Gradient}(X), \quad X_{\text{Gray}} = \text{Conv}_{1 \times 1}(X) \quad (7)$$

where $\text{Gradient}()$ computes the gradient of the image to obtain the gradient feature X_{Grad} , and X_{Gray} represents the grayscale image feature obtained by a 1×1 convolution.

After wavelet decomposition, the high- and low-frequency components of RGB channels are denoted as $\mathbf{H}_R$, $\mathbf{H}_G$, $\mathbf{H}_B$ and $\mathbf{L}_R$, $\mathbf{L}_G$, $\mathbf{L}_B$, respectively, where R , G , and B indicate the color channels. By combining these frequency components with X_{Grad} and X_{Gray} , the high- and low-frequency matrices are constructed as:

$$\mathbf{F}_H = \begin{bmatrix} \mathbf{X}_{\text{Grad}} \\ \mathbf{H}_R \\ \mathbf{H}_G \\ \mathbf{H}_B \end{bmatrix}, \quad \mathbf{F}_L = \begin{bmatrix} \mathbf{X}_{\text{Gray}} \\ \mathbf{L}_R \\ \mathbf{L}_G \\ \mathbf{L}_B \end{bmatrix}. \quad (8)$$

where the bracket notation denotes channel-wise stacking rather than conventional matrix multiplication. Each component belongs to $\mathbb{R}^{1 \times H \times W}$, and therefore $\mathbf{F}_H, \mathbf{F}_L \in \mathbb{R}^{4 \times H \times W}$.

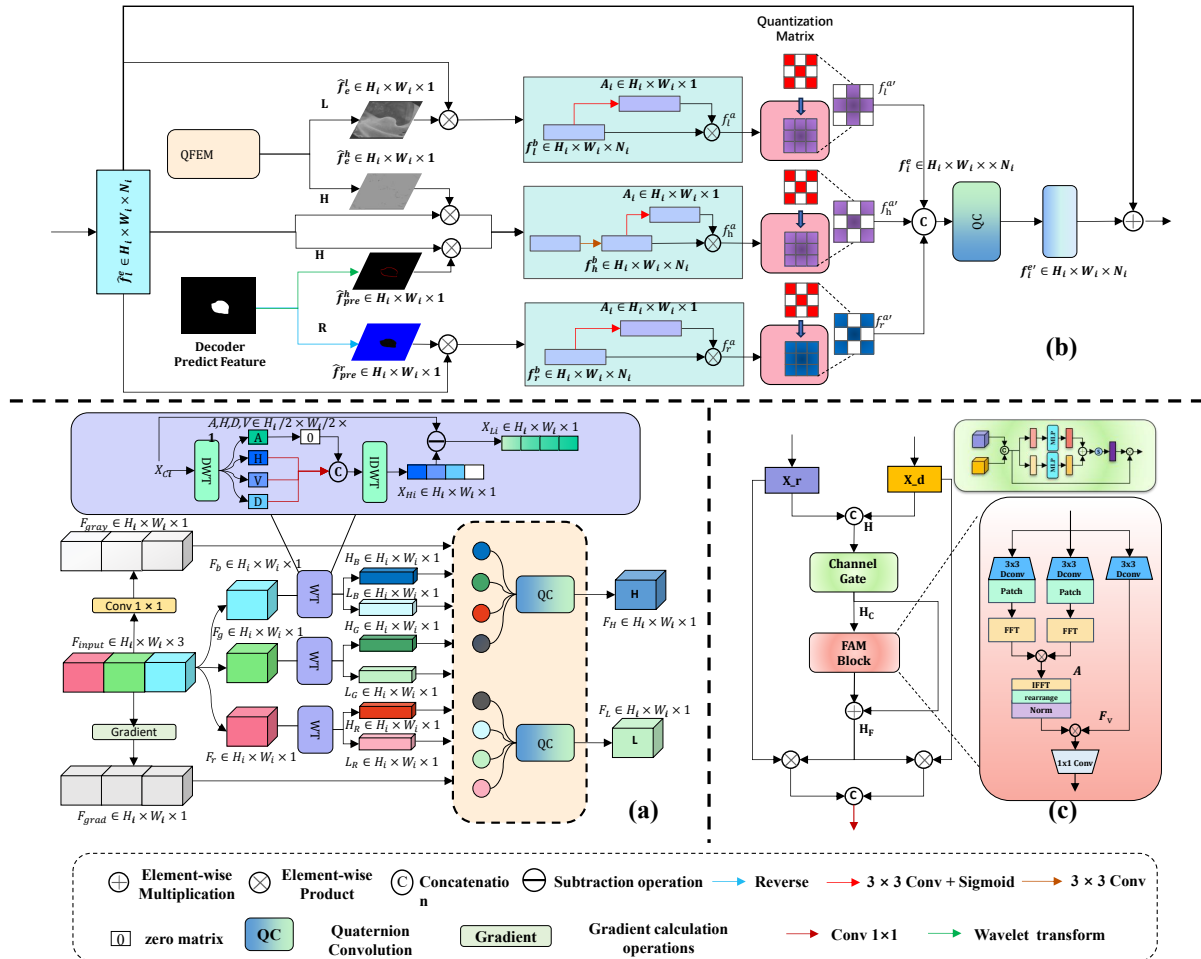


Figure 3. Detailed visualization of key modules in QCFSNet: (a) Quaternion-Driven Frequency Enhancement Module (QFEM) utilizing wavelet transform and quaternion convolution for effective extraction of high- and low-frequency information; (b) Boundary Adaptive Module (BAM) with adaptive gating and mask attention mechanisms to highlight lesion boundaries and suppress redundant features; (c) Spatial-Frequency Interaction (SFI) module integrating channel and frequency-domain attention to enhance global feature representation and feature interaction.

We then introduce quaternion convolution to perform feature interaction learning. The quaternion convolution kernel is defined as:

$$\mathbf{W} = \mathbf{W}_e + \mathbf{W}_r i + \mathbf{W}_g j + \mathbf{W}_b k \quad (9)$$

After quaternion integration, we obtain the high-frequency and low-frequency detail features via:

$$\mathbf{W} \otimes \mathbf{F}_H = \begin{bmatrix} \mathbf{W}_{\text{grad}} & -\mathbf{W}_r & -\mathbf{W}_g & -\mathbf{W}_b \\ \mathbf{W}_r & \mathbf{W}_{\text{grad}} & -\mathbf{W}_b & \mathbf{W}_g \\ \mathbf{W}_g & \mathbf{W}_b & \mathbf{W}_{\text{grad}} & -\mathbf{W}_r \\ \mathbf{W}_b & -\mathbf{W}_g & \mathbf{W}_r & \mathbf{W}_{\text{grad}} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{\text{grad}} \\ \mathbf{H}_R \\ \mathbf{H}_G \\ \mathbf{H}_B \end{bmatrix} \quad (10)$$

$$\mathbf{W} \otimes \mathbf{F}_L = \begin{bmatrix} \mathbf{W}_{\text{gray}} & -\mathbf{W}_r & -\mathbf{W}_g & -\mathbf{W}_b \\ \mathbf{W}_r & \mathbf{W}_{\text{gray}} & -\mathbf{W}_b & \mathbf{W}_g \\ \mathbf{W}_g & \mathbf{W}_b & \mathbf{W}_{\text{gray}} & -\mathbf{W}_r \\ \mathbf{W}_b & -\mathbf{W}_g & \mathbf{W}_r & \mathbf{W}_{\text{gray}} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{\text{gray}} \\ \mathbf{L}_R \\ \mathbf{L}_G \\ \mathbf{L}_B \end{bmatrix} \quad (11)$$

Finally, the frequency-discriminative feature map $f_i^{\text{fre}} = (f_i^h, f_i^l)$ is obtained at each encoder layer, where both f_i^h and f_i^l are derived by applying a 3×3 convolution-based downsampling to the high-frequency $\mathbf{F}_H$ and low-frequency $\mathbf{F}_L$ components, respectively.

3.3. Boundary Adaptive Module (BAM)

To effectively leverage the frequency information captured by the QFEM, this study designs a boundary-adaptive module (BAM) based on wavelet transform. BAM semantically integrates frequency information with spatial features from the encoder and employs an adaptive gating mechanism to selectively and effectively weight the enhanced features, thereby improving feature contrast, suppressing redundant feature interference, and enhancing the focus on key regions such as boundaries, as shown in Figure 3b. Specifically, the input to the i -th layer BAM module consists of the following three components: the visual features from the encoder $f_i^e \in \mathbb{R}^{H_i \times W_i \times C_i}$, the predicted heatmap features from the decoder output after applying the Sigmoid function $f_i^{\text{pre}} \in \mathbb{R}^{H_i \times W_i \times C_i}$, and the high- and low-frequency feature maps extracted by the QFEM $f_i^{\text{fre}} \in \mathbb{R}^{H_i \times W_i \times 1}$.

The BAM architecture mainly consists of three branches: the high-frequency branch, the low-frequency branch, and the reverse attention branch. These three branches optimize the encoder output from different perspectives: fine detail supplementation, global information retention, and foreground-background contrast enhancement. In each BAM, first, for the high-frequency branch, the input encoder feature $f_i^e \in \mathbb{R}^{H_i \times W_i \times C_i}$ is element-wise multiplied with the high-frequency feature $f_i^h \in \mathbb{R}^{H_i \times W_i \times 1}$ obtained from the QFEM module and the high-frequency feature $f_i^{\text{pre-h}} \in \mathbb{R}^{H_i \times W_i \times 1}$ from the decoder's prediction after wavelet transform decomposition. Then, the results are passed through a 3×3 convolution to obtain the high-frequency enhanced feature f_h^b , which is represented as:

$$f_h^b = \text{Conv}_{3 \times 3} \left((f_i^e \otimes f_i^h), (f_i^e \otimes f_i^{\text{pre-h}}) \right) \quad (12)$$

where $\otimes$ denotes element-wise multiplication. By analyzing the combined high-frequency features from the predicted map and the initial captured high-frequency features, the network performs constraint attention based on the differences between the front and back, which helps better localize adaptive boundary and other high-frequency regions.

The low-frequency branch multiplies the input encoder features $f_i^e \in \mathbb{R}^{H_i \times W_i \times C_i}$ with the low-frequency feature map $f_i^l \in \mathbb{R}^{H_i \times W_i \times 1}$ obtained from the QFEM module, and applies a 1×1 convolution to extract the low-frequency attention component f_l^b , represented as:

$$f_l^b = \text{Conv}_{1 \times 1} (f_i^e \otimes f_i^l) \quad (13)$$

For the reverse attention branch, the predicted heatmap f_i^{pre} from the decoder output is processed in reverse to obtain the background feature $f_i^{\text{pre-r}} \in \mathbb{R}^{H_i \times W_i \times 1}$. Then, it is element-wise multiplied with the encoder feature f_i^e and passed through a 1×1 convolution to extract the background attention component f_r^b :

$$f_r^b = \text{Conv}_{1 \times 1} (f_i^e \otimes f_i^{\text{pre-r}}), \quad f_i^{\text{pre-r}} = 1 - f_i^{\text{pre}} \quad (14)$$

After obtaining the feature components from these three branches, an attention mask $\mathcal{A}_i$ is introduced to further enhance the module's learning of the target region. This mask is generated using the Sigmoid function σ , which increases the weight of key regions to guide the network's focus on important areas from each branch. The calculation of the attention mask is given by:

$$f_i^a = f_i^b + (f_i^b \otimes \mathcal{A}_i), \quad \mathcal{A}_i = \sigma (\text{Conv} (f_i^b)) \quad (15)$$

where $i \in \{h, l, r\}$, f_i^b represents the feature component in different branches, and f_i^a represents the enhanced feature after applying region-based weight optimization.

Before the three-branch feature fusion, an adaptive gating mechanism is designed to reduce semantic differences between different components and prevent redundant information interference between branches. A learnable quantization matrix $\mathbf{W}_{H_i \times W_i}$ is introduced to apply element-wise weighting to the enhanced features from each branch, and through backpropagation, the features are correlated and integrated, enabling adaptive learning of the features. The computation is as follows:

$$f_i^{a'} = \mathbf{W}_{H_i \times W_i} \cdot f_i^a \quad (16)$$

where $\mathbf{W}_{H_i \times W_i}$ is a matrix of size $H_i \times W_i$ initialized to zero, and is used as a learnable parameter that is dynamically updated during training. m and n correspond to the height H_i and width W_i of the feature map. After

obtaining the key attention features $f_l^{a'}$, $f_h^{a'}$, and $f_r^{a'}$ from each branch, the region-enhanced features $f_i^{a'}$ from each branch are concatenated together, and a quaternion convolution is applied to perform channel-wise feature learning, integrating them into an adaptive enhanced feature f_i^{en} . Finally, to stabilize the feature learning of the network and ensure more effective mapping of region weights, the result is added to the original encoder features f_i^e via a skip connection to output the final region-enhanced feature $f_i^{e'}$, which optimizes the boundary perception for the mask and supports improved segmentation accuracy. The mathematical representation of $f_i^{e'}$ is:

$$f_i^{e'} = F_{QC} \left(\text{Cat}(f_l^{a'}, f_h^{a'}, f_r^{a'}) \right) \quad (17)$$

where F_{QC} represents the quaternion convolution operation, and Cat represents the concatenation operation.

3.4. Spatial-Frequency Channel Interaction Module (SFI)

In feature fusion and interaction tasks, traditional encoder-decoder frameworks primarily utilize spatial attention to enhance feature representations, but they often struggle to capture fine-grained boundary details and model global dependencies between distant pixels. Moreover, directly fusing the frequency-refined output of BAM with low-level decoder features may introduce noise due to the spatial blurriness of the latter, thereby undermining the advantages of frequency-domain enhancements. To address these issues, we propose the Spatial-Frequency Channel Interaction (SFI) module, as illustrated in Figure 3c, which integrates channel attention and frequency-domain modeling in a unified framework. Specifically, concatenated input features are first refined by a channel attention gate to emphasize high-dimensional representations. These features are then passed into a Frequency Attention Module (FAM), where Fast Fourier Transform (FFT) is used to perform pixel-wise interaction in the frequency domain, allowing the model to capture both low- and high-frequency components. The enhanced features are subsequently transformed back to the spatial domain, enabling joint enhancement of global context and boundary detail. The effectiveness of this design in preserving global consistency and improving structural sharpness is further evaluated through frequency-response analysis in Section 4.3.3.

Specifically, the input to the SFI module includes two feature components: the region-enhanced features $f_i^{e'}$ from BAM and the decoder features f_i^d at the corresponding scale. These are denoted as X_r and X_d , respectively. First, they are concatenated along the channel dimension to form a joint feature H , which is refined using the channel attention mechanism from CBAM [20], yielding:

$$H_C = \mathbf{F}_{CG}(\text{Cat}(X_r, X_d)) \quad (18)$$

where $\mathbf{F}_{CG}$ denotes channel attention and Cat represents channel-wise concatenation. H_C serves as the preliminary fused representation. To further capture global dependencies and enhance fine-grained pixel details while suppressing noise from low-level decoder features, H_C is passed to the Frequency Attention Module (FAM). FAM leverages Fast Fourier Transform (FFT) for efficient frequency-domain interaction modeling. Unlike conventional spatial-domain attention, FFT-based operations use element-wise multiplication, offering lower computational cost and better pixel-level correlation modeling. In FAM, H_C is first divided into $n \times n$ patches, and a 3×3 depthwise separable convolution is applied to extract local features, generating the query F_q , key F_k , and value F_v . FFT is then applied to F_q and F_k to obtain their frequency representations, and the pixel-level attention map A is computed as:

$$A = \mathcal{F}^{-1}(\mathcal{F}(F_q) \cdot \overline{\mathcal{F}(F_k)}) \quad (19)$$

where $\mathcal{F}()$ represents the Fourier transform, $\mathcal{F}^{-1}()$ represents the inverse Fourier transform, $\overline{\mathcal{F}()}$ represents the conjugate transpose operation, and $\odot$ denotes element-wise multiplication. This frequency domain point multiplication allows the network to focus on pixel-level fine granularity and correlations with lower computational complexity. The frequency domain attention features are then mapped back via the inverse Fourier transform, reshaped, and weighted by matrix multiplication with the spatial domain value features F_v . The final aggregated feature H_F is obtained via a residual operation as follows:

$$H_F = H_C + \mathbf{R}(A) \cdot F_v \quad (20)$$

where $\mathbf{R}$ denotes reshaping and normalization operations. Through a series of operations, the aggregated feature H_F enables both fine-grained pixel-level details and global dependency modeling. Additionally, to better fuse the two feature vectors and reduce semantic differences between the two domains, we introduce an interaction gating mechanism. The aggregated feature H_F is multiplied with the input features X_r and X_d respectively for interaction. The interacted features are then re-concatenated and passed through a convolution operation to extract the final

output fused feature, as represented by:

$$f_i^{d'} = \text{Conv}_{1 \times 1} (\text{Cat} (H_F \odot X_r, H_F \odot X_d)). \quad (21)$$

Through the synergistic effects of the channel attention mechanism, FAM module, and interaction gating, the SFI module effectively combines frequency and spatial domain operations for feature interaction, enhancing feature clarity and pixel-level correlations. This emphasizes the representation of key regions while reducing the blur effects of low-resolution features, enabling the next decoding layer to analyze the mask segmentation regions in a more discriminative manner.

3.5. Loss Function and Evaluation Metrics

The loss function employed in this paper is a hybrid formulation, combining binary cross-entropy loss L_{bce} and Dice loss L_{dice} , defined as follows:

$$L_{\text{bce}} = -\frac{1}{N} \sum_{n=1}^N \left[y^{(n)} \log(\hat{y}^{(n)}) + (1 - y^{(n)}) \log(1 - \hat{y}^{(n)}) \right] \quad (22)$$

$$L_{\text{dice}} = 1 - \frac{2 \sum_{n=1}^N y^{(n)} \hat{y}^{(n)} + \epsilon}{\sum_{n=1}^N y^{(n)} + \sum_{n=1}^N \hat{y}^{(n)} + \epsilon} \quad (23)$$

$$L = L_{\text{bce}} + L_{\text{dice}} \quad (24)$$

where N denotes the number of pixels, $y^{(n)}$ and $\hat{y}^{(n)}$ denote the ground-truth label and predicted foreground probability, respectively, and ϵ is a small constant to ensure numerical stability. During training, deep supervision is applied to the decoder outputs at different stages. The same hybrid BCE–Dice loss is calculated for each decoder prediction, and the total training loss is defined as:

$$L_{\text{total}} = \sum_{i=0}^4 L_i, \quad (25)$$

where L_i denotes the hybrid loss at the i -th supervised decoder output, with $i = 0$ corresponding to the final prediction. All supervised outputs are equally weighted.

To quantitatively evaluate segmentation performance on both the private CervicalEndoDB dataset and public benchmarks, we employ six metrics: mean Dice (mDice), mean IoU [42], weighted F-measure F_{β}^w [42], structural measure S_{α} [43], enhanced alignment measure $E_{\phi}^{\max}$ [44], and mean absolute error (MAE). These metrics are widely used for segmentation performance evaluation [7, 13].

4. Experiment

4.1. Datasets and Implementation Details

To evaluate the effectiveness of the proposed framework, we conducted experiments on a private cervical endoscopy dataset, CervicalEndoDB, and several public endoscopic segmentation benchmarks. CervicalEndoDB, collected from the Third Affiliated Hospital of Sun Yat-sen University, comprises 1022 cervical endoscopic images with pixel-level annotations, approved by the institutional ethics committee (Approval ID: “ZYTH-III-2024-105-01”). The dataset presents various clinically challenging structures, including squamous-columnar junctions, vascular networks, Nabothian cysts, and bleeding areas. All images were acquired from cervical endoscopic examinations and annotated with pixel-level masks by experienced clinicians. The annotations were subsequently reviewed to ensure consistency and reliability. The dataset was randomly divided into training and validation subsets with an 8:2 ratio, and no overlap existed between the two subsets to prevent potential data leakage. For public datasets, we used Kvasir-SEG [45] (900 training, 100 testing) and CVC-ClinicDB [46] (550 training, 62 testing) as seen-domain training sources. Generalization was further evaluated on CVC-ColonDB [47] (380 images), CVC-300 [48] (60 images), and ETIS [49] (196 images), which collectively provide diverse anatomical and imaging scenarios. To assess robustness under challenging endoscopic imaging conditions, we further evaluated QCFSNet on the UW-SINUS-SURGERY-C/L dataset [50], which contains cadaveric and live sinus surgery scenes with uneven illumination, specular reflections, and motion blur. In particular, the presence of illumination variations and specular highlights introduces unstable appearance changes and local intensity disturbances, making this dataset suitable for evaluating the robustness of segmentation models under degraded visual conditions. Therefore, this dataset

provides a complementary evaluation of QCFSNet's ability to preserve structural and boundary information under illumination-related interference.

Our framework was implemented in PyTorch and trained on an NVIDIA RTX 4090 GPU. Following the training strategy in ACSNet [51], we used stochastic gradient descent (momentum = 0.9, weight decay = 1×10^{-5}) with a polynomial learning rate decay:

$$lr = init_lr \times \left(1 - \frac{epoch}{nEpoch}\right)^{power}, \quad (26)$$

where $init_lr = 1 \times 10^{-3}$, $nEpoch = 200$, and $power = 0.9$. Input images were resized to 352×352 (also tested at 384×384 for the private dataset), and results were evaluated at the original resolution. Standard data augmentations, including random flipping, rotation, and multi-scale training ($\{0.75, 1.0, 1.25\}$), were applied to improve model generality.

4.2. Performance Comparison

To comprehensively evaluate the performance of QCFSNet in the cervical endoscopic lesion segmentation task and verify its effectiveness and generalization ability on endoscopic datasets, we conducted quantitative and qualitative comparisons with several widely used methods on both the private dataset CervicalEndoDB and multiple public datasets. These methods include: U-Net [16], U-Net++ [6], SFA [5], PraNet [7], SANet [8], MSNet [2], PEFNet [52], M²UNet [10], RA-DENet [11], CFANet [12], and MEGANet-34 [13]. It is noteworthy that the results for U-Net are taken from the MEGANet paper, and the results for PEFNet were adjusted based on M²UNet; other methods' results are directly quoted from their respective original publications.

4.2.1. Quantitative Analysis

QCFSNet exhibits consistently superior segmentation performance across both private and public endoscopic datasets, as shown in Tables 1 and 2. On the private dataset CervicalEndoDB, QCFSNet outperforms the classical U-Net by 16.1% in mDice and 15.8% in mIoU, highlighting its enhanced capability in regional perception and boundary delineation for complex cervical anatomical structures. When compared with the recently proposed MEGANet-34, QCFSNet achieves noticeable improvements on public datasets: it boosts mDice/mIoU by 0.7%/1.2% on Kvasir-SEG and 0.9%/1.0% on CVC-300. On the ETIS dataset, where lesion regions are relatively small and structurally simple, complex architectures like RA-DENet tend to overfit. In contrast, QCFSNet maintains stable segmentation performance, benefiting from its adaptive learning strategy and frequency-aware design. Furthermore, as shown in Table 3, compared with RA-DENet, QCFSNet reduces parameters by 9.1M and achieves a relative mDice improvement of 2.3%. Compared with MEGANet, QCFSNet uses 0.4M fewer parameters and improves mDice by 1.3%. These results demonstrate the superior balance of QCFSNet between segmentation accuracy and model efficiency. These quantitative results collectively validate the adaptability, stability, and generalization ability of QCFSNet across diverse endoscopic scenarios.

Table 1. Experimental results from the private cervical endoscopy dataset CervicalEndoDB. The best result for each metric is highlighted in **bold**.

Methods	mDice ↑	mIoU ↑	S_α ↑	MAE ↓
U-Net [16]	59.7	54.5	72.7	3.8
U-Net++ [6]	63.3	55.2	75.6	3.5
PraNet [7]	68.9	56.9	77.7	3.2
SANet [8]	73.6	65.3	81.3	1.8
MSNet [2]	74.2	66.8	82.4	1.9
PEFNet [52]	70.2	60.3	79.3	2.0
RA-DENet [11]	74.8	69.2	85.1	1.6
CFANet [12]	73.6	67.5	83.2	1.7
MEGANet-34 [13]	74.2	68.1	83.4	1.7
Ours	75.8	70.3	84.7	1.6

Furthermore, to evaluate the practical computational characteristics of QCFSNet, we report additional efficiency metrics on the ClinicDB dataset, including model parameters, FLOPs, and peak GPU memory consumption, as

summarized in Table 4. QCFSNet contains 28.9 M trainable parameters, requires 59.8 G FLOPs, and consumes 2.3 GB peak GPU memory during inference. Compared with U-Net, QCFSNet reduces the parameter count by approximately 6.8% and achieves relative improvements of approximately 13.4% and 17.3% in mDice and mIoU, respectively, with only moderate increases of approximately 5.5% in FLOPs and 9.5% in peak inference memory. Compared with MEGANet, QCFSNet improves mDice and mIoU by approximately 1.3% and 1.4%, respectively, while reducing the parameter count, FLOPs, and peak inference memory by approximately 1.4%, 2.4%, and 4.2%, respectively. These results demonstrate that QCFSNet achieves a favorable balance between segmentation accuracy, parameter efficiency, and computational complexity.

Table 2. Comparison of QCFSNet with other methods across datasets. Best results are in **bold**; all metrics are percentages (%). $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively.

Methods	Kvasir-SEG (Seen)						CVC-300 (EndoScene) (Unseen)					
	mDice $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{\max} \uparrow$	MAE $\downarrow$	mDice $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{\max} \uparrow$	MAE $\downarrow$
U-Net	81.8	74.6	79.4	85.8	89.3	5.5	71.0	62.7	68.4	84.3	87.6	2.2
U-Net++	82.1	74.3	80.8	86.2	89.8	4.8	70.7	62.4	68.7	83.9	89.8	1.8
SFA	72.3	61.1	67.0	78.2	84.9	7.5	46.7	32.9	34.1	64.0	81.7	6.5
PraNet	89.8	84.0	88.5	91.5	94.8	3.0	87.1	79.7	84.3	92.5	97.2	1.0
SANet	90.4	84.7	89.2	91.5	95.3	2.8	88.8	81.5	85.9	92.8	97.2	0.8
MSNet	90.7	86.2	89.3	92.2	94.4	2.8	86.9	80.7	84.9	92.5	94.3	1.0
PEFNet	89.2	83.8	—	—	—	2.9	87.1	79.7	—	—	—	1.0
M ² UNet	90.7	85.5	—	—	—	2.5	89.0	81.9	—	—	—	0.7
RA-DENet	91.8	87.0	90.9	92.9	96.6	2.1	89.1	82.1	87.1	93.4	97.5	0.7
CFANet	91.5	86.1	90.3	92.4	96.2	2.3	89.3	82.7	87.5	93.8	97.8	0.8
MEGANet-34	91.1	85.9	90.4	91.6	95.4	2.6	88.7	81.8	86.3	92.4	95.9	0.9
Ours	91.8	87.1	91.1	92.8	96.7	2.1	89.6	82.8	87.7	93.5	97.6	0.7

Methods	ColonDB (Unseen)						ETIS (Unseen)					
	mDice $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{\max} \uparrow$	MAE $\downarrow$	mDice $\uparrow$	mIoU $\uparrow$	$F_{\beta}^w \uparrow$	$S_{\alpha} \uparrow$	$E_{\phi}^{\max} \uparrow$	MAE $\downarrow$
U-Net	51.2	44.4	49.8	71.2	77.6	6.1	39.8	33.5	36.6	68.4	74.0	3.6
U-Net++	48.3	41.0	46.7	69.1	76.0	6.4	40.1	34.4	39.0	68.3	77.6	3.5
SFA	46.9	34.7	37.9	63.4	76.5	9.4	29.7	21.7	23.1	55.7	63.3	10.9
PraNet	70.9	64.0	69.6	81.9	86.9	4.5	62.8	56.7	60.0	79.4	84.1	3.1
SANet	75.3	67.0	72.6	83.7	87.8	4.3	75.0	65.4	68.5	84.9	89.7	1.5
MSNet	75.5	67.8	73.7	83.6	88.3	4.1	71.9	66.4	67.8	84.0	83.0	2.0
PEFNet	71.0	63.8	—	—	—	3.6	63.6	57.2	—	—	—	1.9
M ² UNet	76.7	68.4	—	—	—	3.6	67.0	59.5	—	—	—	2.4
RA-DENet	76.8	68.6	74.3	85.4	90.3	3.2	76.0	68.8	71.4	85.6	90.3	1.9
CFANet	74.3	66.5	72.8	83.5	89.8	3.9	73.2	65.5	69.3	84.5	89.2	1.4
MEGANet-34	78.1	70.6	76.6	84.5	89.9	3.8	78.9	70.9	75.3	86.6	91.5	1.5
Ours	78.9	71.1	77.6	84.9	89.3	3.3	79.0	71.2	75.4	86.8	90.4	1.6

Table 3. Comparison of the performance and network efficiency of QCFSNet with other existing SOTA methods on the ClinicDB dataset. The highest values are shown in **bold**. All metrics are in (%).

Methods	ClinicDB (Seen)			Params (M)
	mDice $\uparrow$	mIoU $\uparrow$	MAE $\downarrow$	
PraNet [7]	89.9	84.9	0.9	32.6
M ² UNet [10]	90.1	85.3	0.8	28.8
RA-DENet [11]	92.1	86.5	1.1	38.0
MEGANet (ResNet-34) [13]	93.0	88.5	0.8	29.3
Ours	94.2	89.7	0.6	28.9

Table 4. Segmentation performance and computational complexity on ClinicDB. All complexity metrics were measured at an input resolution of 352×352 under the same settings. Memory denotes peak GPU memory during inference. The best result for each metric is highlighted in **bold**.

Method	Input	mDice (%) $\uparrow$	mIoU (%) $\uparrow$	Params (M) $\downarrow$	FLOPs (G) $\downarrow$	Memory (GB) $\downarrow$
U-Net	352^2	83.1	76.5	31.0	56.7	2.1
MEGANet (ResNet-34)	352^2	93.0	88.5	29.3	61.3	2.4
QCFSNet	352^2	textbf94.2	89.7	28.9	59.8	2.3

4.2.2. Qualitative Analysis

To further validate the performance and generalization capability of QCFSNet in lesion segmentation, we conducted a visual comparison across both private and public datasets, as shown in Figure 4. Panels (a)–(d) illustrate the segmentation results of QCFSNet on the private CervicalEndoDB dataset, involving typical clinical challenges such as squamocolumnar junctions, Nabothian cysts, and blood occlusion. Compared to U-Net, PraNet, and MEGANet,

QCFSNet exhibits superior performance in boundary localization and structural consistency, maintaining stable and accurate segmentation even under complex backgrounds as shown in panels (b)–(d). Panels (e)–(i) present the results on public datasets, including Kvasir-SEG, CVC-300, ColonDB, ETIS, and ClinicDB. QCFSNet demonstrates strong robustness and stability across diverse datasets and lesion types. In particular, under imaging disturbances such as illumination variation and specular interference, QCFSNet maintains clearer lesion boundaries and more complete structural contours. This benefit can be attributed to the complementary effects of QFEM and BAM, where frequency-domain enhancement provides robust structural cues and boundary-adaptive learning improves contour localization. Whether dealing with background interference in panels (e) and (g), or addressing the size and shape variations of polyps in panels (f), (h), and (i), QCFSNet consistently maintains sharp boundaries and complete structural contours. These results validate its generalization across various scenarios and its overall applicability in endoscopic clinical tasks. Furthermore, to better assess the model's capability in boundary modeling, we visualize the activation heatmaps before the decoder stage, as shown in Figure 5a. Compared to the traditional spatial-domain method (Spa), the flexible frequency-domain enhancement method (Q-Fre) used in QCFSNet shows stronger activation near lesion boundaries, effectively highlighting high-frequency edge information and demonstrating superior boundary awareness. This provides more reliable feature support for the quantitative analysis of lesion regions. In contrast, although the Spa method generates certain activations within lesion interiors, it exhibits weaker responses near the edges, making it difficult to accurately delineate lesion contours and failing to meet clinical measurement accuracy requirements.

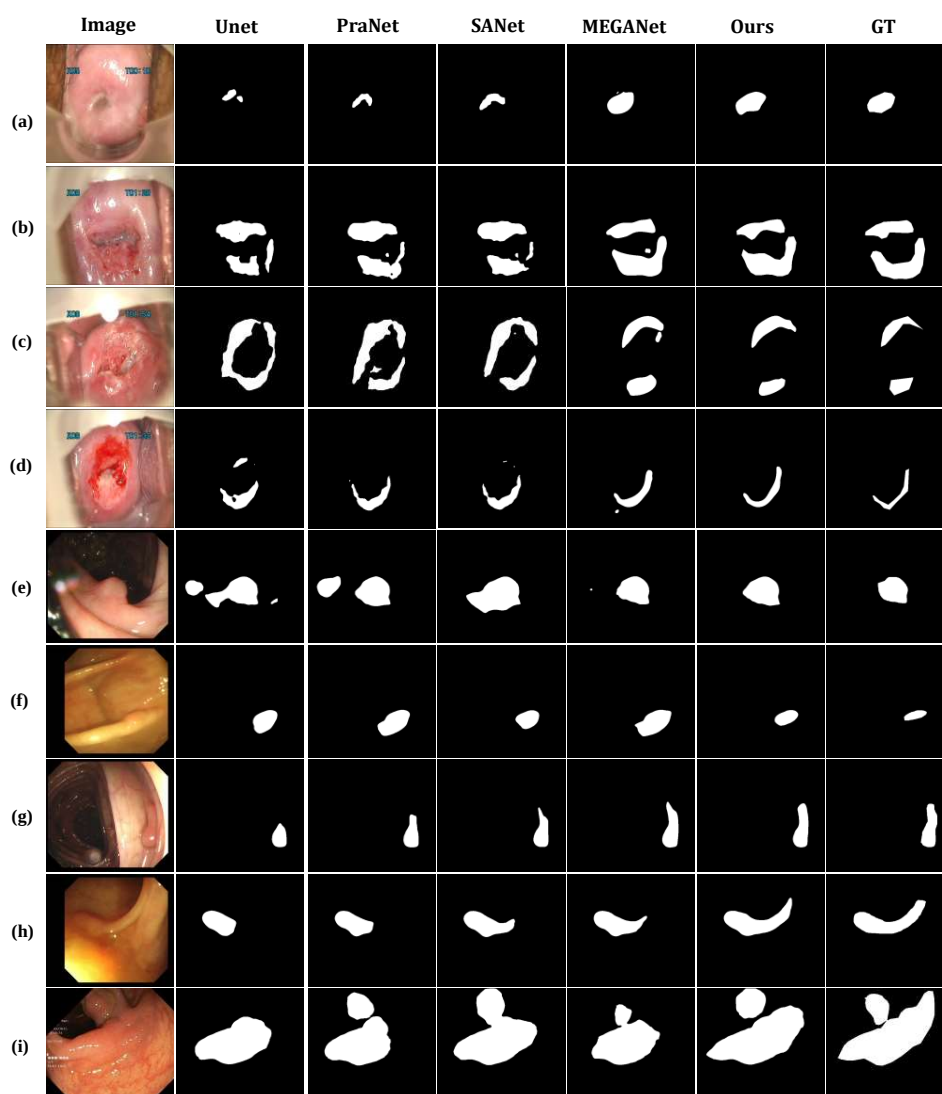


Figure 4. Qualitative segmentation results of QCFSNet compared with other state-of-the-art methods on private and public datasets. (a–d) Results on the private CervicalEndoDB dataset involving complex clinical challenges such as squamocolumnar junctions, Nabothian cysts, and blood occlusion. (e–i) Results on public datasets, specifically (e,i) from CVC-300, (f,h) from ColonDB, and (g) from Kvasir-SEG, demonstrating the effectiveness and generalization of QCFSNet across diverse endoscopic scenarios.

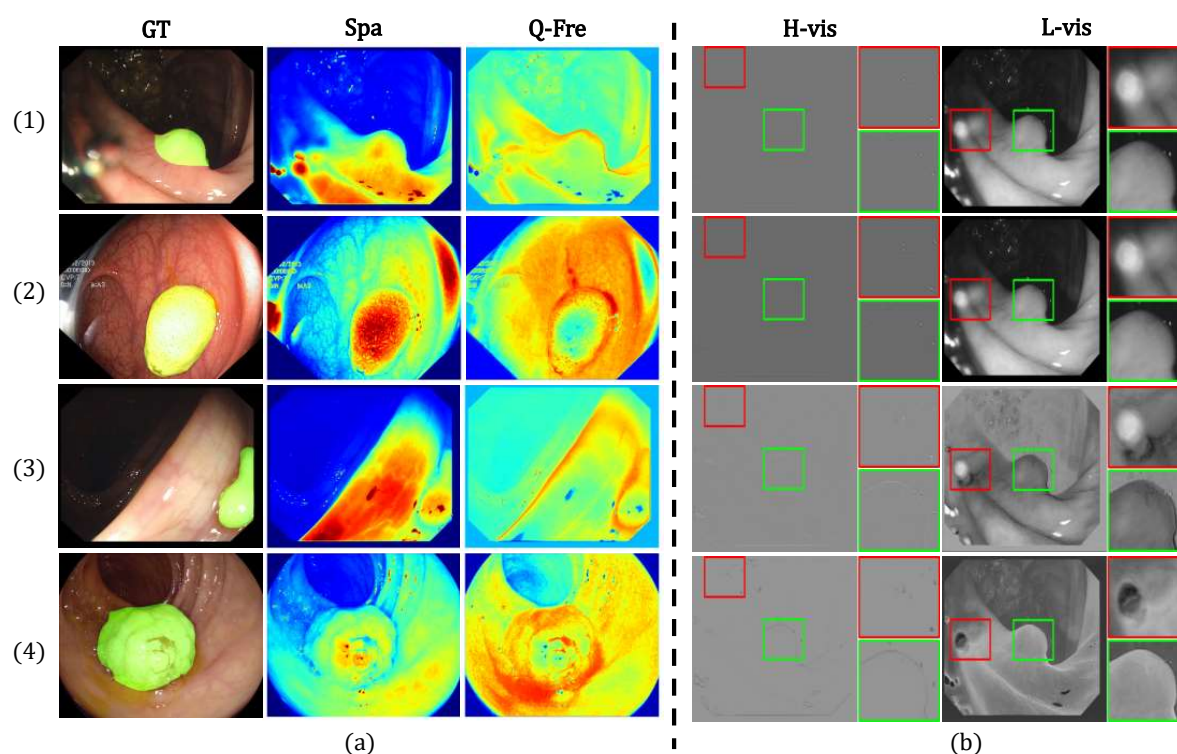


Figure 5. Visualization of frequency-domain feature effectiveness. **(a)** Comparison between a spatial-domain method (Spa) and the proposed frequency-enhanced approach (Q-Fre), demonstrating stronger boundary responses and improved lesion edge clarity with Q-Fre. **(b)** Ablation visualization of the QFEM module: high-frequency (H-vis) and low-frequency (L-vis) features extracted by different configurations, highlighting the superior boundary discrimination and noise suppression capabilities of the multi-channel quaternion convolution (MCQ).

4.3. Ablation Study

To evaluate the independent roles and collaborative contributions of each module in QCFSNet, we conducted a component-level ablation study on the CervicalEndoDB dataset, as shown in Table 5. Based on the encoder-decoder network structure, QFEM, BAM, and SFI modules were incrementally integrated and analyzed. In the second row, frequency features extracted by QFEM are multiplied with encoder features and fused with decoder features. The third row applies BAM directly on encoder features to model attention masks and gating mechanisms. The fourth row feeds skip connection features into the SFI module, where spatial-frequency channel interaction is performed before upsampling. Each module independently enhances performance, verifying the validity of their respective designs. Rows 5 to 7 further assess the collaborative optimization among modules, especially highlighting the combined contribution of QFEM and BAM to boundary and texture discrimination under similar backgrounds. When all three modules are integrated (row 8), the model achieves the best performance across all metrics, demonstrating the superiority of QCFSNet in multi-scale feature enhancement and boundary refinement. In the following subsections, we further analyze the internal mechanisms and effectiveness of QFEM, BAM, and SFI, respectively. In addition to the quantitative analysis on CervicalEndoDB, we conducted a visual ablation study on the *UW-SINUS-SURGERY-C* dataset to further evaluate the effectiveness of each module under realistic clinical imaging conditions. As illustrated in Figure 6, segmentation performance improves progressively as QFEM, BAM, and SFI modules are integrated into the baseline. Specifically, QFEM enhances frequency-based feature representation, BAM improves boundary localization in the presence of anatomical ambiguity, and SFI reinforces spatial-frequency channel interaction to cope with illumination interference and visual clutter. The complete QCFSNet model produces more accurate and stable segmentation results, particularly under common disturbances in electronic endoscopic images, such as specular highlights, blood occlusion, and motion blur. These findings validate the individual contributions of each module and demonstrate the practical value of QCFSNet in improving segmentation reliability under complex electronic endoscopic imaging conditions.

Table 5. Ablation study of QCFSNet modules conducted on the CervicalEndoDB datasets. The best result for each metric is highlighted in **bold**.

QFEM	BAM	SFI	mDice ↑	mIoU ↑	S_{α} ↑	MAE ↓
✓	✓	✓	67.3	61.2	77.8	2.5
			71.8	65.3	81.6	2.2
			69.8	63.1	78.8	2.5
✓	✓	✓	72.1	67.3	82.3	2.1
			74.3	69.1	83.1	1.7
			73.1	67.9	83.6	1.9
✓	✓	✓	73.5	68.3	82.8	1.8
			75.8	70.3	84.7	1.6

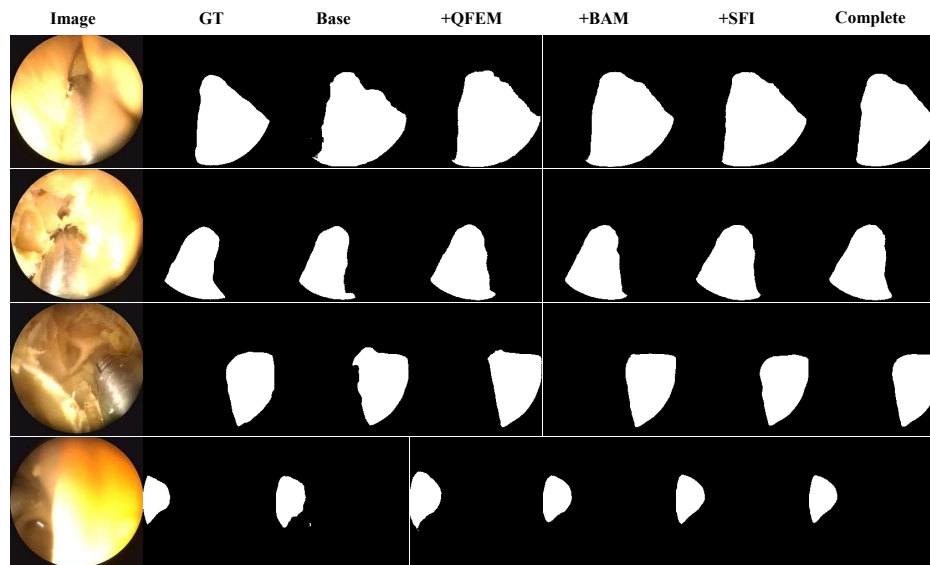


Figure 6. Ablation study visualization on the UW-SINUS-SURGERY-C dataset. From left to right: input images, ground truth (GT), baseline model (Base), baseline with QFEM (+QFEM), with BAM (+BAM), with SFI (+SFI), and the complete model (Complete). Each module is progressively added based on the previous configuration.

4.3.1. Efficiency of QFEM

The overall ablation results demonstrate that frequency information supplementation is beneficial for boundary refinement and global context modeling. To validate the internal design of QFEM, we conducted a detailed ablation on the CVC-300 dataset, as shown in Table 6. Four variants were tested: (1) DWT without channel separation, (2) Laplacian filtering (LF), (3) multi-channel DWT with standard convolution (MCC), and (4) multi-channel DWT with quaternion convolution (MCQ). Results show that introducing multi-channel wavelet decomposition and quaternion convolution significantly improves segmentation performance, confirming their effectiveness in extracting more discriminative and structure-aware frequency features. Figure 5b provides a qualitative comparison. Compared with other variants, MCQ produces clearer high-frequency boundaries and stronger low-frequency global consistency, while effectively suppressing background noise.

Table 6. Ablation study of QFEM modules conducted on the CVC-300 dataset. The best result for each metric is highlighted in **bold**. DWT: wavelet transform, LF: Laplace filter, MCC: standard convolution, MCQ: quaternion convolution.

DWT	LF	MCC	MCQ	mDice ↑	mIoU ↑	S_{α} ↑	MAE ↓
✓	✓	✓	✓	87.1	80.2	91.6	1.1
				88.7	81.6	92.4	0.9
✓				89.1	81.9	93.1	0.8
✓				89.6	82.8	93.5	0.7

4.3.2. Efficiency of BAM

To validate the internal mechanism of the Boundary Adaptive Module (BAM), we conducted component-wise ablation experiments on the CervicalEndoDB dataset, as summarized in Table 7. BAM is composed of high-frequency, low-frequency, reverse attention, and adaptive gating branches. By removing each branch individually (BAM_H , BAM_L , BAM_R , BAM_{AG}), we observe consistent performance degradation, especially when the high-frequency or adaptive gating branch is excluded. These results highlight the essential role of high-frequency components in enhancing texture and boundary cues, as well as the effectiveness of adaptive weighting in suppressing redundant background information and promoting lesion-focused feature enhancement.

Table 7. Ablation study of BAM modules conducted on the CervicalEndoDB datasets. mDice, mIoU, S_α , and MAE are used as evaluation metrics. The best result for each metric is highlighted in **bold**.

Method	mDice $\uparrow$	mIoU $\uparrow$	S_α $\uparrow$	MAE $\downarrow$
BAM_H	72.1	65.8	81.9	2.1
BAM_L	74.8	68.2	83.8	1.7
BAM_R	73.6	67.3	82.7	1.9
BAM_{AG}	75.1	68.6	84.3	1.7
BAM	75.8	70.3	84.7	1.6

4.3.3. Efficiency of SFI

To evaluate the effectiveness of the Spatial-Frequency Channel Interaction (SFI) module, we compare model variants with and without SFI in the first part of Table 5. The integration of SFI consistently improves mDice and mIoU, confirming its role in enhancing feature-level interaction between spatial and frequency domains. Moreover, we analyze the frequency-domain modeling capability of the SFI module through bar-chart visualizations of FFT responses, as shown in Figure 7. Compared to its No-FFT counterpart, the SFI-enhanced version exhibits significantly higher magnitude responses across both low- and high-frequency ranges. In particular, the frequency bands near the low-frequency index 0 and the high-frequency index 65,000 show marked increases, indicating improved structural preservation and enhanced edge-texture representations. These observations validate that the SFI module effectively integrates global contextual structures while highlighting fine-grained boundary details in the frequency domain.

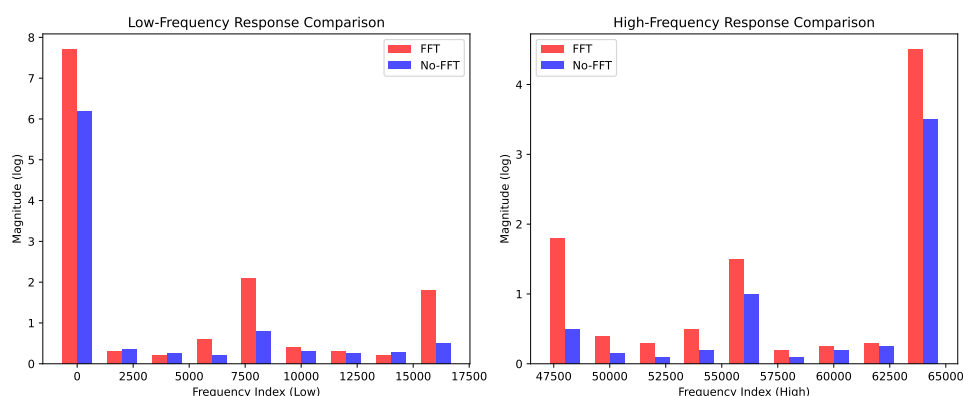


Figure 7. Comparison of frequency responses with and without FFT enhancement. The **left subfigure** illustrates the distribution of low-frequency components, while the **right subfigure** highlights high-frequency components. The X-axis indicates the frequency index range, and the Y-axis shows the log-scaled magnitude response.

5. Discussion

Accurately delineating cervical-lesion boundaries is the prerequisite for the perception layer of intelligent image-guided medical robots. High-quality masks serve as spatial priors underpinning automated diagnosis, tool–tissue interaction monitoring, path planning, and closed-loop visual servoing. Without reliable pixel-level perception, the downstream control stack loses critical geometric context, and robotic assistance risks remaining

reactive rather than truly autonomous. Resolving long-standing issues such as blood occlusion, texture heterogeneity, and uneven illumination will therefore accelerate the development of robot-based intelligent platforms for cervical therapy.

QCFSNet addresses these challenges by explicitly integrating spatial- and frequency-domain information. QFEM combines quaternion convolution with wavelet decomposition to capture high-frequency boundary cues and low-frequency contextual information. BAM adaptively integrates these frequency features with encoder representations to enhance lesion-related regions and suppress background interference, while SFI further strengthens long-range semantic interaction during decoding. Experiments on the private CervicalEndoDB dataset and multiple public benchmarks demonstrate improved Dice and IoU scores and clearer boundary delineation across diverse endoscopic scenarios.

Despite these improvements, severe specular highlights and abrupt illumination changes may still affect frequency representation. Moreover, quaternion operations, wavelet transformations, and spatial-frequency interactions introduce additional computational and memory costs, while the limited native support for quaternion convolution in current deep-learning libraries increases implementation complexity. Future work will focus on illumination-adaptive feature modeling, lightweight frequency processing, efficient quaternion operators, and hardware-aware optimization for real-time deployment on resource-constrained medical devices.

6. Conclusions

This paper presents QCFSNet, a frequency-enhanced segmentation framework for electronic endoscopic imaging. By integrating multi-channel frequency extraction, adaptive attention, and spatial–frequency interaction, QCFSNet effectively handles common challenges such as blood occlusion, texture variability, and uneven illumination, leading to improved semantic representation and boundary precision. Extensive experiments on the private CervicalEndoDB dataset and multiple public benchmarks demonstrate its superior segmentation performance and strong generalization over state-of-the-art methods. Ablation studies further validate the individual and joint contributions of each module. The high-quality lesion masks generated by QCFSNet have potential to support early-stage cervical disease screening and assistive visual analysis in portable or embedded diagnostic systems. Future work will focus on more lightweight frequency modeling and real-time optimization to facilitate deployment on resource-constrained endoscopic platforms.

Author Contributions

P.H.: methodology, experimental design, and writing—original draft preparation; G.H.: writing—review and editing, supervision, and funding acquisition; G.Z.: writing—review and editing; S.M.: writing—review and editing; M.Z.: manuscript revision; Q.Y.: data support. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Key Areas Research and Development Program of Guangzhou under Grant 2023B01J0029, the Guangdong Provincial Key Laboratory of Cyber-Physical System under Grant 2020B1212060069, and the Beijing Xisike Clinical Oncology Research Foundation under Grant Y-2019AZQN-1049.

Institutional Review Board Statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Ethics Committee of the Third Affiliated Hospital of Sun Yat-sen University. The private CervicalEndoDB dataset used in this study comprises 1022 cervical endoscopic images with pixel-level annotations collected from the Third Affiliated Hospital of Sun Yat-sen University (Approval ID: ZYTH-III-2024-105-01).

Informed Consent Statement

Not applicable.

Data Availability Statement

The public datasets used in this study are available from the polyp segmentation benchmark repository on Zenodo (<https://zenodo.org/records/5579392>), which contains the standardized training and testing splits for Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS-Larib, and CVC-300. The UW-Sinus-Surgery-C/L dataset is available at <https://github.com/SURA23/Sinus-Surgery-Endoscopic-Image-Datasets>.

Acknowledgments

In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used ChatGPT to assist in drafting and refining sections of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Jin, Z.; Huang, G.; Zhang, F.; et al. SQ-ViT: A Multi-Scale Vision Transformer With Quaternion for Endoscopic Images Classification. *IEEE Trans. Consum. Electron.* **2025**, *71*, 828–838. <https://doi.org/10.1109/tce.2024.3518755>.
2. Zhao, X.; Zhang, L.; Lu, H. Automatic Polyp Segmentation via Multi-scale Subtraction Network. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*; Springer International Publishing: Cham, Switzerland, 2021; pp. 120–130. https://doi.org/10.1007/978-3-030-87193-2_12.
3. Tang, D.; Wang, L.; Ling, T.; et al. Development and validation of a real-time artificial intelligence-assisted system for detecting early gastric cancer: A multicentre retrospective diagnostic study. *eBioMedicine* **2020**, *62*, 103146. <https://doi.org/10.1016/j.ebiom.2020.103146>.
4. Domgue, J.F.; Dille, I.; Gnangnon, F.H.R.; et al. Utility of colposcopy for the screening and management of cervical cancer in Africa: a cross-sectional analysis of providers' training and practices. *BMC Health Serv. Res.* **2024**, *24*, 1619. <https://doi.org/10.1186/s12913-024-11982-1>.
5. Fang, Y.; Chen, C.; Yuan, Y.; et al. Selective Feature Aggregation Network with Area-Boundary Constraints for Polyp Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*; Springer International Publishing: Cham, Switzerland, 2019; pp. 302–310. https://doi.org/10.1007/978-3-030-32239-7_34.
6. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; et al. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. <https://doi.org/10.1109/tmi.2019.2959609>.
7. Fan, D.P.; Ji, G.P.; Zhou, T.; et al. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. *arXiv* **2020**, arXiv:2006.11392. <http://arxiv.org/abs/2006.11392v4>.
8. Wei, J.; Hu, Y.; Zhang, R.; et al. Shallow Attention Network for Polyp Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*; Springer International Publishing: Cham, Switzerland, 2021; pp. 699–708. https://doi.org/10.1007/978-3-030-87193-2_66.
9. Cong, R.; Zhang, Y.; Yang, N.; et al. Boundary Guided Semantic Learning for Real-Time COVID-19 Lung Infection Segmentation System. *IEEE Trans. Consum. Electron.* **2022**, *68*, 376–386. <https://doi.org/10.1109/tce.2022.3205376>.
10. Trinh, Q.H.; Bui, N.T.; Nguyen-Mau, T.H.; et al. M2UNet: MetaFormer Multi-Scale Upsampling Network for Polyp Segmentation. In Proceedings of the 2023 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4–8 September 2023; pp. 1115–1119. <https://doi.org/10.23919/eusipco58844.2023.10290110>.
11. Wang, K.; Liu, L.; Fu, X.; et al. RA-DENet: Reverse Attention and Distractions Elimination Network for polyp segmentation. *Comput. Biol. Med.* **2023**, *155*, 106704. <https://doi.org/10.1016/j.compbiomed.2023.106704>.
12. Zhou, T.; Zhou, Y.; He, K.; et al. Cross-level Feature Aggregation Network for Polyp Segmentation. *Pattern Recognit.* **2023**, *140*, 109555. <https://doi.org/10.1016/j.patcog.2023.109555>.
13. Bui, N.T.; Hoang, D.H.; Nguyen, Q.T.; et al. MEGANet: Multi-Scale Edge-Guided Attention Network for Weak Boundary Polyp Segmentation. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 7970–7979. <https://doi.org/10.1109/wacv57701.2024.00780>.
14. Wan, Y.; Zhou, D.; Wang, C. UMF-Net: A UNet-based multi-branch feature fusion network for colon polyp segmentation. *Biomed. Signal Process. Control.* **2025**, *99*, 106851. <https://doi.org/10.1016/j.bspc.2024.106851>.
15. Zhang, L.; Xiong, W.; Wu, D.; et al. Mul2UNet: Deformable Multi-Scale UNet With Decoupled Channel Fusion for Oral Squamous Cell Carcinoma Segmentation. *IEEE Trans. Consum. Electron.* **2025**, *71*, 8477–8491. <https://doi.org/10.1109/tce.2025.3582774>.
16. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597v1. <http://arxiv.org/abs/1505.04597v1>.
17. Zhou, Y.; Huang, J.; Wang, C.; et al. XNet: Wavelet-Based Low and High Frequency Fusion Networks for Fully- and Semi-Supervised Semantic Segmentation of Biomedical Images. In Proceedings of the 2023 IEEE/CVF

- International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 21028–21039. <https://doi.org/10.1109/iccv51070.2023.01928>.
18. Esteki, S.; Naghsh-Nilchi, A.R. SW/SE-CNN: semi-wavelet and specific image edge extractor CNN for Gaussian image denoising. *Neural Comput. Appl.* **2024**, *36*, 5447–5469. <https://doi.org/10.1007/s00521-023-09314-1>. Correction in *Neural Comput. Appl.* **2024**, *36*, 5471.
 19. He, K.; Zhang, X.; Ren, S.; et al. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. <https://doi.org/10.1109/cvpr.2016.90>.
 20. Woo, S.; Park, J.; Lee, J.Y.; et al. CBAM: Convolutional Block Attention Module. *arXiv* **2018**. <https://doi.org/10.48550/arxiv.1807.06521>.
 21. Trabelsi, C.; Bilaniuk, O.; Zhang, Y.; et al. Deep Complex Networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017; Volume 3.
 22. Zhu, X.; Xu, Y.; Xu, H.; et al. Quaternion Convolutional Neural Networks. In *Computer Vision—ECCV 2018*; Springer International Publishing: Cham, Switzerland, 2018; pp. 645–661. https://doi.org/10.1007/978-3-030-01237-3_39.
 23. Wang, J.; Huang, G.; Zhong, G.; et al. QGD-Net: A Lightweight Model Utilizing Pixels of Affinity in Feature Layer for Dermoscopic Lesion Segmentation. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 5982–5993. <https://doi.org/10.1109/jbhi.2023.3320953>.
 24. Gai, S.; Huang, X.L. Reduced Biquaternion Convolutional Neural Network for Color Image Processing. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1061–1075. <https://doi.org/10.1109/tcsvt.2021.3073363>.
 25. Singh, S.; Tripathi, B.K.; Rawat, S.S. Deep quaternion convolutional neural networks for breast Cancer classification. *Multimed. Tools Appl.* **2023**, *82*, 31285–31308. <https://doi.org/10.1007/s11042-023-14688-4>.
 26. Communiello, D.; Lella, M.; Scardapane, S.; et al. Quaternion Convolutional Neural Networks for Detection and Localization of 3D Sound Events. In Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8533–8537. <https://doi.org/10.1109/icassp.2019.8682711>.
 27. Parcollet, T.; Zhang, Y.; Morchid, M.; et al. Quaternion Convolutional Neural Networks for End-to-End Automatic Speech Recognition. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 22–26. <https://doi.org/10.21437/interspeech.2018-1898>.
 28. Sigillo, L.; Grassucci, E.; Uncini, A.; et al. Generalizing medical image representations via quaternion wavelet networks. *Neurocomputing* **2025**, *638*, 130195. <https://doi.org/10.1016/j.neucom.2025.130195>.
 29. Grassucci, E.; Sigillo, L.; Uncini, A.; et al. GROUSE: A Task and Model Agnostic Wavelet-Driven Framework for Medical Imaging. *IEEE Signal Process. Lett.* **2023**, *30*, 1397–1401. <https://doi.org/10.1109/lsp.2023.3321554>.
 30. Ehrlich, M.; Davis, L. Deep Residual Learning in the JPEG Transform Domain. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3483–3492. <https://doi.org/10.1109/iccv.2019.00358>.
 31. Chang, H.Y.; Lin, Y.C. Deep Learning Approach to the Quality Restoration for JPEG Images. In Proceedings of the International Conference on Computer and Communications Management, Okayama, Japan, 29–31 July 2022.
 32. Jumiawi, W.A.H.; El-Zaart, A. A Boosted Minimum Cross Entropy Thresholding for Medical Images Segmentation Based on Heterogeneous Mean Filters Approaches. *J. Imaging* **2022**, *8*, 43. <https://doi.org/10.3390/jimaging8020043>.
 33. Peng, Y.; Pan, L.; Luan, P.; et al. Curvilinear object segmentation in medical images based on ODoS filter and deep learning network. *Appl. Intell.* **2023**, *53*, 23470–23481. <https://doi.org/10.1007/s10489-023-04773-4>.
 34. Ma, J.; Han, X.; Zhang, X.; et al. MLWAN: Multi-Scale Learning Wavelet Attention Module Network for Image Super Resolution. *Sensors* **2022**, *22*, 9110. <https://doi.org/10.3390/s22239110>.
 35. Huang, Y.; Huang, J.; Liu, J.; et al. WaveDM: Wavelet-Based Diffusion Models for Image Restoration. *IEEE Trans. Multimed.* **2024**, *26*, 7058–7073. <https://doi.org/10.1109/tmm.2024.3359769>.
 36. Xu, G.; Liao, W.; Zhang, X.; et al. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognit.* **2023**, *143*, 109819. <https://doi.org/10.1016/j.patcog.2023.109819>.
 37. Hu, T.; Qi, Y.; Wang, H.; et al. BCFNET: Boundary-Guided Semantic Cross Fusion for Polyp Segmentation. In Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, 27–30 May 2024; pp. 1–4. <https://doi.org/10.1109/isbi56570.2024.10635517>.
 38. Ates, G.C.; Mohan, P.; Celik, E. Dual Cross-Attention for medical image segmentation. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107139. <https://doi.org/10.1016/j.engappai.2023.107139>.
 39. Huang, H.; Chen, Z.; Zou, Y.; et al. Channel prior convolutional attention for medical image segmentation. *Comput. Biol. Med.* **2024**, *178*, 108784. <https://doi.org/10.1016/j.compbiomed.2024.108784>.
 40. Liu, K.H.; Lin, B.Y. MSCSA-Net: Multi-Scale Channel Spatial Attention Network for Semantic Segmentation of Remote Sensing Images. *Appl. Sci.* **2023**, *13*, 9491. <https://doi.org/10.3390/app13179491>.
 41. Kong, L.; Dong, J.; Ge, J.; et al. Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 5886–5895. <https://doi.org/10.1109/cvpr52729.2023.00570>.

42. Margolin, R.; Zelnik-Manor, L.; Tal, A. How to Evaluate Foreground Maps. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 248–255. <https://doi.org/10.1109/cvpr.2014.39>.
43. Cheng, M.M.; Fan, D.P.; et al. Structure-Measure: A New Way to Evaluate Foreground Maps. *Int. J. Comput. Vis.* **2021**, *129*, 2622–2638. <https://doi.org/10.1007/s11263-021-01490-8>.
44. Fan, D.P.; Gong, C.; Cao, Y.; et al. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence {IJCAI-18}, Stockholm, Sweden, 13–19 July 2018; pp. 698–704. <https://doi.org/10.24963/ijcai.2018/97>.
45. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; et al. Kvasir-SEG: A Segmented Polyp Dataset. In *Multimedia Modeling*; Springer: Cham, Switzerland, 2020; pp. 451–462. https://doi.org/10.1007/978-3-030-37734-2_37.
46. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; et al. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. <https://doi.org/10.1016/j.compmedimag.2015.02.007>.
47. Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE Trans. Med. Imaging* **2016**, *35*, 630–644. <https://doi.org/10.1109/tmi.2015.2487997>.
48. Vázquez, D.; Bernal, J.; Sánchez, F.J.; et al. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *J. Healthc. Eng.* **2017**, *2017*, 1–9. <https://doi.org/10.1155/2017/4037190>.
49. Silva, J.; Histace, A.; Romain, O.; et al. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.* **2014**, *9*, 283–293. <https://doi.org/10.1007/s11548-013-0926-3>.
50. Qin, F.; Lin, S.; Li, Y.; et al. Towards Better Surgical Instrument Segmentation in Endoscopic Vision: Multi-Angle Feature Aggregation and Contour Supervision. *IEEE Robot. Autom. Lett.* **2020**, *5*, 6639–6646. <https://doi.org/10.1109/lra.2020.3009073>.
51. Zhang, R.; Li, G.; Li, Z.; et al. Adaptive Context Selection for Polyp Segmentation. *arXiv* **2023**, arXiv:2301.04799v1. <http://arxiv.org/abs/2301.04799v1>.
52. Nguyen-Mau, T.H.; Trinh, Q.H.; Bui, N.T.; et al. PEFNet: Positional Embedding Feature for Polyp Segmentation. In *MultiMedia Modeling*; Springer: Cham, Switzerland, 2023; pp. 240–251. https://doi.org/10.1007/978-3-031-27818-1_20.