algorithms

MDPI

Article

MSEANet: Multi-Scale Selective Edge Aware Network for Polyp Segmentation

Botao Liu [1], Changqi Shi [1] and Ming Zhao [2,*]

[1] School of Computer Science, Yangtze University, Jingzhou 434023, China; liubotao920@163.com (B.L.); shichangqi17@gmail.com (C.S.)
[2] School of Internet of Things Engineering, Wuxi University, Wuxi 214105, China
[*] Correspondence: hitmzhao@gmail.com

Abstract: The colonoscopy procedure heavily relies on the operator's expertise, underscoring the importance of automated polyp segmentation techniques in enhancing the efficiency and accuracy of colorectal cancer diagnosis. Nevertheless, achieving precise segmentation remains a significant challenge due to the high visual similarity between polyps and their backgrounds, blurred boundaries, and complex localization. To address these challenges, a Multi-scale Selective Edge-Aware Network has been proposed to facilitate polyp segmentation. The model consists of three key components: (1) an Edge Feature Extractor (EFE) that captures polyp edge features with precision during the initial encoding phase, (2) the Cross-layer Context Fusion (CCF) block designed to extract and integrate multi-scale contextual information from diverse receptive fields, and (3) the Selective Edge Aware (SEA) module that enhances sensitivity to high-frequency edge details during the decoding phase, thereby improving edge preservation and segmentation accuracy. The effectiveness of our model has been rigorously validated on the Kvasir-SEG, Kvasir-Sessile, and BKAI datasets, achieving mean Dice scores of 91.92%, 82.10%, and 92.24%, respectively, on the test sets.

Keywords: polyp segmentation; context fusion; edge aware; high-frequency information; deep learning

check for updates

Academic Editors: Francesc Pozo and Antonio Sarasa-Cabezuelo

Received: 6 December 2024
Revised: 7 January 2025
Accepted: 9 January 2025
Published: 12 January 2025

Citation: Liu, B.; Shi, C.; Zhao, M. MSEANet: Multi-Scale Selective Edge Aware Network for Polyp Segmentation. *Algorithms* **2025**, *18*, 42. https://doi.org/10.3390/a18010042

1. Introduction

Colorectal cancer (CRC) ranks as the third most prevalent and serious cancer worldwide [1] and is characterized by a high incidence and mortality rate. A strong correlation exists between polyps—especially adenomatous polyps—and the development of CRC, as these polyps represent a critical precursor stage. Consequently, effective prevention of CRC necessitates early detection through screening tests such as colonoscopy. Colonoscopy remains the gold standard for CRC screening due to its high diagnostic accuracy; however, it is heavily dependent on the operator's expertise and the procedural environment. Despite advancements, challenges such as the complex intestinal topology, inconsistent lighting conditions, and continuous organ deformation contribute to a substantial polyp miss rate. Studies reveal a missed detection rate of 26.8% for polyps in the right colon and 21.4% for polyps in the left colon [2,3]. These limitations highlight the urgent need for automated polyp segmentation systems to assist clinicians in improving diagnostic reliability and reducing error rates.

Deep learning has emerged as a transformative approach in medical image analysis, offering robust solutions for developing next-generation imaging applications [4]. The introduction of U-Net by Ronneberger et al. in 2015 [5] was a significant milestone, establishing deep learning as a cornerstone in medical image segmentation. Nevertheless, polyp segmentation presents unique challenges, including an imbalance between foreground and background classes, as well as blurred and indistinct edges. Previously, a potential solution came from the statistical region-based segmentation method developed by Slabaugh et al., which improves segmentation by modeling image regions statistically, particularly for ultrasound images with complex backgrounds [6]. While this method was not initially designed for polyp segmentation, its approach to handling low-contrast regions can be beneficial in our context as well.

Polyp boundaries, influenced by variations in shape, size, light, and texture, can make polyp boundaries difficult to distinguish and works like U-Net++ [7]. ResUnet++ [8] aims to address these complexities. However, U-Net and its variants usually use a symmetric encoder-decoder structure, which is able to process the global information of the image, but there is no specialized mechanism to enhance the edge details, and the local feature maps may lose some critical edge information in the decoding stage, which may affect the accurate segmentation of edge details. In addition, although the U-like structure of the network improves the fusion of features to a certain extent by skip connection, the capture of multi-scale edge information is still limited. Recently, some methods have been proposed to try to solve these problems. PraNet [9] uses a reverse attention module to mine boundary clues, establish the relationship between region and boundary clues, and calibrate misaligned predictions through a cyclic cooperation mechanism between regions and boundaries to improve segmentation accuracy. CPFNet [10] employs a context pyramid fusion network to provide multiple levels of global context to the decoder through reconstructed skip connections. TGA-Net [11] leverages a text-guided attention mechanism to enhance polyp segmentation by integrating natural language descriptions of polyp size and quantity with image features. However, the lack of standardized text information limits its effectiveness.

Attention mechanisms have gained prominence across various domains in recent years. While these mechanisms show potential for improving polyp segmentation, challenges persist due to foreground-background imbalance and edge blurring. Overemphasis on attention mechanisms may introduce irrelevant noise, adversely impacting model accuracy. Consequently, accurately distinguishing the polyp foreground from its edges within a similar background remains a significant challenge. This is particularly critical in extracting high-frequency edge details and seamlessly integrating them with contextual features to enhance segmentation performance.

To address the aforementioned challenges, we propose the Multi-Scale Selective Edge-Aware Network (MSEANet), a novel framework that selectively fuses edge information and multi-scale contextual information via an edge-aware module for polyp segmentation. The proposed method effectively integrates global contextual information with high-frequency edge features of polyps, thereby enhancing segmentation performance. Furthermore, MSEANet operates in a fully end-to-end manner, eliminating the need for additional manual annotations or separate training stages. The main contributions are summarized as follows:

(1) An enhanced Edge Feature Extractor (EFE) is introduced to capture high-frequency edge information of polyps during the early stages of the encoder, ensuring precise delineation of polyp boundaries.

(2) The Cross-layer Context Fusion (CCF) block is designed to effectively merge local structural features with global contextual information, improving the model's ability to understand target characteristics and accurately localize polyps in complex scenes.

(3) Through the proposed Selective Edge Aware (SEA) module, edge information and contextual features extracted by the CCF block are integrated and preserved across multiple scales. This design significantly enhances segmentation accuracy by maintaining the fidelity of polyp boundaries.

(4) Our proposed MSEANet was rigorously tested on the Kvasir-SEG [12], Kvasir-Sessile [13], and BKAI [14] datasets, achieving mean Dice scores of 91.92%, 80.63%, and 91.50%, respectively. These results demonstrate the effectiveness of our method in diverse datasets.

In the following sections, we provide a detailed description of the proposed algorithm, including related work, methods, experiments, and conclusion.

2. Related Work

2.1. Encoder-Decoder Model

The advent of convolutional neural networks (CNNs) has revolutionized medical image segmentation, providing the foundation for the development of sophisticated models. Among these, U-Net [5] stands out as a seminal architecture that introduced an encoder-decoder structure, enabling the efficient extraction of rich image features across multiple hierarchical levels, making U-Net a widely adopted solution in medical imaging tasks. Its skip connections effectively bridge the semantic gap between low- and high-level features. Building upon U-Net’s success, the DeepLab family of networks [15–17] has pushed the boundaries of segmentation accuracy by incorporating innovations like dilated convolutions and Atrous Spatial Pyramid Pooling (ASPP). These techniques expand the receptive field without increasing the computational burden, allowing the network to capture global contextual information while preserving spatial resolution. These advancements have laid the groundwork for automated polyp segmentation, demonstrating the potential of encoder-decoder models in addressing complex medical image analysis tasks.

However, despite their achievements, traditional encoder-decoder architectures that rely solely on convolutional operations face notable limitations. On the one hand, global contextual information is transmitted from deeper stages to shallower stages, and it may be diluted due to the weak feature extraction ability of individual stages. On the other hand, the skip connection in each stage ignores global information and is an indiscriminate combination of local information, which will introduce irrelevant clutter and lead to misclassification of pixels. In polyp segmentation, challenges such as blurred edges and the inability to effectively integrate global and local contextual information often arise. These shortcomings hinder the precise delineation of polyp boundaries and the comprehensive understanding of intricate patterns within complex scenes. Addressing these limitations requires innovative designs that go beyond conventional convolutional paradigms to enhance edge sensitivity and multi-scale contextual understanding. For example, multidimensional signal analysis techniques, such as wavelet transform and Fourier transform, can be leveraged to capture features at multiple scales. These methods have been shown to enhance feature fusion and improve boundary detection by explicitly modeling multi-scale information.

2.2. Parallel Attention Model

Attention mechanisms have garnered significant interest for their ability to enhance feature representation, making them a valuable component in image segmentation tasks. Self-attention [18], in particular, enables models to prioritize critical features while filtering out irrelevant information. For instance, a stepped network [19] has been introduced for real-time polyp segmentation in colonoscopy images. This network employs four blocks for spatial feature extraction, integrating a dual attention module within each block and utilizing a multi-scale fusion module to consolidate features across scales. However, such models often suffer from slow processing speeds and limited ability to effectively integrate global contextual information. Recently, models incorporating parallel attention structures have demonstrated promising results. For example, the Parallel Reverse Attention Network (PraNet) [9] leverages a reverse attention mechanism to focus on polyp regions while employing a recursive feature aggregation module to capture global contextual information effectively. Similarly, PRAPNet [20] is an improved deep learning model designed to further enhance polyp segmentation performance through its Parallel Residual Atrous Pyramid structure.

The combination of parallel structures and attention mechanisms not only accelerates network processing but also enhances focus on critical aspects such as edge features and global contextual information. This capability is particularly advantageous in scenarios where polyps exhibit high visual similarity to the background or have blurred boundaries. However, efficiently extracting edge features and seamlessly fusing them with contextual information remain significant challenges. To address this, this study proposes a network model based on a parallel attention structure, leveraging its robust contextual fusion capabilities and flexible module integration to achieve improved polyp segmentation.

3. Methods

MSEANet is designed to accurately capture complex boundary details while improving segmentation performance across diverse clinical scenarios. The overall network architecture is illustrated in Figure 1a. Built on an encoder-decoder framework, MSEANet comprises a feature encoder, an EFE module, three CCF modules, and four SEA modules. The functionality and design of each of these components are detailed in the following sections.

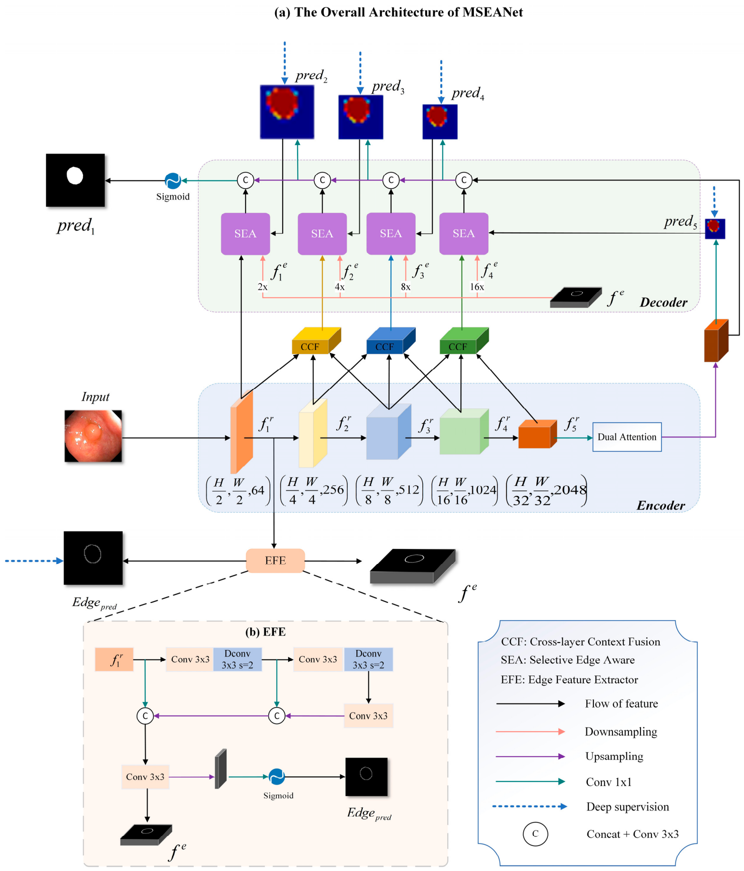

Figure 1. (**a**) Overall architecture of the proposed MSEANet. (**b**) Architecture of EFE; Dconv denotes depthwise separable convolution.

3.1. Feature Encoder

ResNet-50 [21] can take advantage of its natural residual structure to accelerate the convergence of the network and also effectively mitigate the gradient vanishing problem. The feature encoder utilizes a pre-trained ResNet-50 [21] as its backbone network to extract feature maps $\{f_i^r, \text{i} \in 1,2,3,4,5\}$ at various stages and scales. These feature maps encapsulate diverse spatial and semantic information throughout the encoding process. In the early stages, feature maps generally exhibit higher spatial resolution, containing richer spatial details critical for edge detection.

Initially, the feature maps f_1^r with the most spatial information are passed through the EFE module to generate polyp edge feature maps and prediction maps. Subsequently, feature maps $\{f_1^r, f_2^r, f_3^r\}$, $\{f_2^r, f_3^r, f_4^r\}$, $\{f_3^r, f_4^r, f_5^r\}$ extracted from the backbone network are grouped across layers and passed to the CCF module, which is responsible for fusing the context information. Finally, in order to obtain the enhanced feature representation of the feature map f_5^r in different dimensions, we introduce Dual Attention [22] and apply it. Dual Attention [22] includes Position Attention and Channel Attention, which are two mechanisms that can enhance feature representation in different dimensions and are especially effective in segmentation tasks, helping to suppress redundant features that are irrelevant to the segmentation task, enhancing the model's attention to the polyp itself, and reducing the possibility of mis-segmentation. To be compatible with the segmentation task, we remove the last average pooling layer and the fully connected layer of ResNet-50 [21].

3.2. Edge Feature Extractor

The edges of polyps are often blurred and challenging to distinguish. Inspired by Edge-Prioritized Polyp Segmentation (EPPS) [23], this paper improves the EFE to accurately capture high-frequency features of polyp edges. This module extracts and fuses edge information through multi-level convolution and upsampling operations, as illustrated in Figure 1b. The design of EFE focuses on efficiently extracting edge-related high-frequency information from the feature maps generated at the initial stage of the encoder. Specifically, the input feature map undergoes successive 3 × 3 convolutions to extract base features. Atrous Separable Convolution is employed for downsampling, which helps to preserve spatial information typically lost with traditional max-pooling operations. After two downsampling operations, the resolution of the feature map is progressively reduced while its semantic information becomes more enriched. The downsampled feature maps are then upsampled using bilinear interpolation, spliced with the preceding layer's feature maps processed by 1 × 1 convolution, and further fused through a 3 × 3 convolution.

This process is conducted at two scales to ensure the fine-grained edge information is fully integrated with the deep semantic features. The upsampled feature maps are passed through a convolution layer and a sigmoid function to generate edge prediction maps $Edge_{pred}$. Simultaneously, the EFE module outputs fused edge-enhanced features f^e for further processing during the subsequent decoding stage.

Manual annotation is not required for ground truth edge labeling. Instead, the Canny operator is used to extract edge ground truth from polyp masks. This approach provides greater accuracy and eliminates the variability and uncertainty often associated with manual labeling.

3.3. Cross-Layer Context Fusion

The encoder in a segmentation network learns global context information, including the surroundings and category characteristics of objects [24,25]. However, in complex scenarios, the encoder often struggles to adequately capture the edges, shapes, and structures of polyps, and it has limitations in extracting multi-scale contextual information. To address these challenges, we designed the CCF module, which enhances the network's understanding of complex polyp morphology by fusing feature maps from different layers and combining contextual information across multiple scales and receptive fields. As illustrated in Figure 2a, the workflow of the CCF module consists of four main steps: feature maps $\{x_1, x_2, x_3\}$ from three different scales are passed through a standard 3×3 convolution operation to standardize them to the same number of channels as x_1. The feature maps x_2 and x_3 are scaled to the same spatial resolution as x_1 using 2× downsampling and 2× upsampling, respectively. The three adjusted feature maps are concatenated along the channel dimension, generating a richer multi-scale feature representation. On the concatenated feature map, four parallel dilated convolutions with dilation rates of {1, 6, 12, 18} are applied. These dilation rates capture multi-scale contextual information from varying receptive fields, enriching the feature map's contextual expression and diversity. Mathematically, this process can be summarized as:

$$C_i \ = Concat\left\{Dsample\left(f_i^r\right), f_{i+1}^r, Upsample\left(f_{i+2}^r\right)\right\},\ i = 1,2,3 \tag{1}$$

where C_i represents the feature extracted from the ith stage of the encoder. By fusing feature maps from adjacent scales, the feature map C_i with enhanced contextual expression is obtained. Finally, the MSFBlock [26] is introduced to fuse C_i into f_i^c, as shown in Figure 2b. The MSFBlock processes feature maps from varying receptive fields and performs hierarchical fusion, integrating multi-scale contextual information and enhancing the feature maps' expressive power.

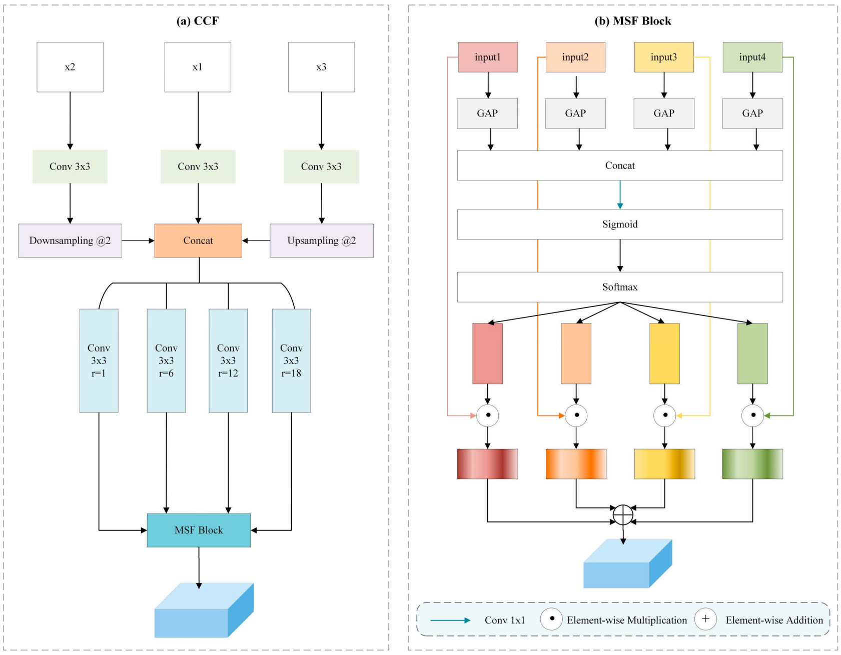

Figure 2. (**a**) Structure of Cross-layer Context Fusion (CCF) module. (**b**) Structure of Multi-scale Selective Fusion (MSF).

In MSEANet, three CCF modules are strategically placed between the encoder and decoder. These modules guide high-level global semantic information to different feature extraction stages, significantly improving the network's ability to extract multi-scale contextual information and understand the complex edges and morphology of polyps. This design enhances the accuracy and robustness of polyp segmentation, particularly in challenging scenarios.

3.4. Selective Edge Aware

The structure of the SEA module is illustrated in Figure 3. During the decoding stage, the SEA module robustly fuses edge information across multiple scales, significantly enhancing the model's ability to delineate complex polyp boundaries. In polyp segmentation tasks, precise edge information is critical for improving segmentation performance. The SEA module leverages a reverse attention mechanism in conjunction with edge feature fusion, increasing the model's sensitivity to the target boundaries.

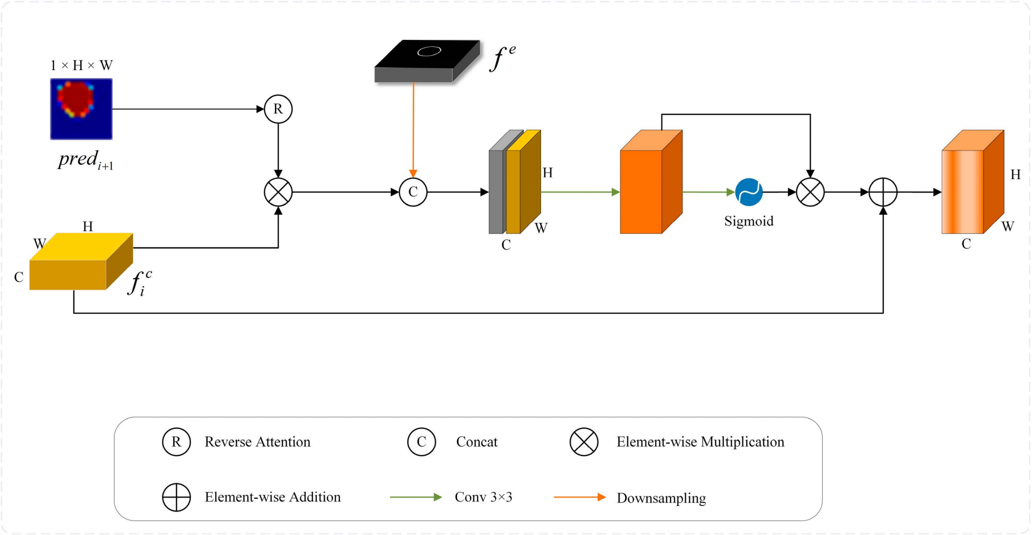

Figure 3. Structure of Selective Edge Aware (SEA) module.

Each SEA module processes the predicted map $pred_{i+1}$ generated by the deep supervision output of the previous layer, activating it with a sigmoid function to produce a probability map of salient regions. This probability map is inverted to generate a background attention map, which improves the model's focus on background areas. Subsequently, the background feature map is progressively fused with the global contextual feature map f_i^c from the CCF module, creating a refined feature map where salient regions are removed. To further enhance edge information, the SEA module integrates edge features f^e extracted by the EFE module. These edge features are concatenated with the background feature map, and the concatenated output is processed through a 3 × 3 convolution operation to fully integrate multi-scale information, resulting in an enriched fused feature map f_i^b. The mathematical calculations can be summarized as follows:

$$f_i^b = F_{conv}\left[Concat\left(\left(\left(1-\sigma\left(pred_{i+1}\right)\right)\otimes f_i^c\right), f^e\right)\right] \tag{2}$$

where f_i^b is the fused feature map, F_{conv} represents the 3 × 3 convolution operation, σ is sigmoid function, and $Concat$ denotes the concatenation operation. Subsequently, a simple gating mechanism is employed to adaptively weight the fused features, amplifying the model's response in critical areas and improving the localization of edge and shape information. A residual connection mechanism is then applied, where the weighted features are added element-wise to the original input features. This residual design preserves the global semantic information from the original features while enhancing local edge details. The calculation can be expressed as:

$$\hat{f}_i^b = \left(f_i^b \otimes \sigma\left(f_i^b\right)\right) \bigoplus f_i^c \tag{3}$$

where $\hat{f}_i^b$ is the final output feature map, f_i^b represents the fused feature map, $\otimes$ means element-wise multiplication, $\bigoplus$ indicates element-by-element addition, and f_i^c denotes the original input features.

By organically combining edge features, deep prediction features, and global contextual information, the SEA module can accurately capture the boundaries and shapes of polyps. This design effectively addresses the challenge of insufficient preservation of high-frequency information, thereby improving segmentation accuracy and robustness in complex scenarios.

3.5. Loss Function

We employed the DiceBCE Loss $L(\cdot)$ to optimize the model training process, ensuring improved performance in the polyp segmentation task. DiceBCE Loss is a combination of Dice Loss and Binary Cross Entropy Loss. This loss function helps the model to better capture complex boundary areas while ensuring pixel-level classification accuracy.

For each different scale of the output feature map $pred_i$, we compute its loss with respect to the label $mask_i$, which is mathematically expressed as:

$$loss_i = L(pred_i, mask_i) \tag{4}$$

where $mask_i$ is the labeled map obtained by sampling operation of ground truth.

We compute the difference between the edge prediction map and the true edge map by separate edge loss function. The formula is:

$$loss_e = L\left(Edge_{pred}, Edge_{gt}\right) \tag{5}$$

where $loss_e$ represents the loss between the edge prediction map $Edge_{pred}$ output by the EFE module and the ground truth of the edge $Edge_{gt}$.

The final total loss is the weighted sum of the sum of the individual scale losses and the edge loss, which is expressed as:

$$loss_{total} = \sum_{i=1}^{5} loss_i + loss_e \tag{6}$$

This total loss $loss_{total}$ is the optimization objective of the model, which aims to improve the accuracy of the multi-scale segmentation effect and edge details by optimizing this loss function.

4. Experiments

In this section, we present a qualitative comparison of MSEANet against state-of-the-art methods using widely recognized polyp segmentation datasets. The evaluation highlights the strengths and performance of MSEANet in comparison with existing techniques. This paper employed five widely used metrics, including mean Intersection over Union (mIoU), mean Dice coefficient (mDice), Recall, Precision, and F2-score. IoU measures the similarity between predicted segmentation results and actual segmentation results. In the following formula, $Prediction$ is the prediction area of the model, $Ground\ Truth$ is the true label area, $\cap$ represents the intersection, and $\cup$ represents the union.

$$IoU = \frac{|Prediction \cap Ground\ Truth|}{|Prediction \cup Ground\ Truth|} \tag{7}$$

Dice is used to measure the similarity between two sets, emphasizing the overlapping parts and being sensitive to small object segmentation. The formula is as follows:

$$Dice = \frac{2 \cdot |Prediction \cap Ground\ Truth|}{|Prediction| + |Ground\ Truth|} \tag{8}$$

Recall quantifies the proportion of positive examples recognized by the model and represents the proportion of correctly predicted pixel points to the true positive example pixel points, with the formula:

$$Recall = \frac{|Prediction \cap Ground\ Truth|}{|Ground\ Truth|} \tag{9}$$

Precision evaluates the proportion of pixels predicted to be positive instances that are actually positive instances, indicating the accuracy of the prediction, and is calculated by the formula:

$$Precision = \frac{|Prediction \cap Ground\ Truth|}{|Prediction|} \tag{10}$$

F2-score is a weighted form of Dice and Recall, with more emphasis on Recall for tasks requiring higher sensitivity, and is calculated as:

$$F2 = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{11}$$

where β is a weighting factor indicating the relatively higher importance of Recall. In this paper, β is equal to 2.

4.1. Datasets

To evaluate the performance of the proposed MSEANet, we conducted extensive tests on three publicly available polyp segmentation benchmark datasets: Kvasir-SEG [12], Kvasir-Sessile [13], and BKAI [14]. Sample images from these datasets, along with their corresponding edge prediction maps, are presented in Figure 4.

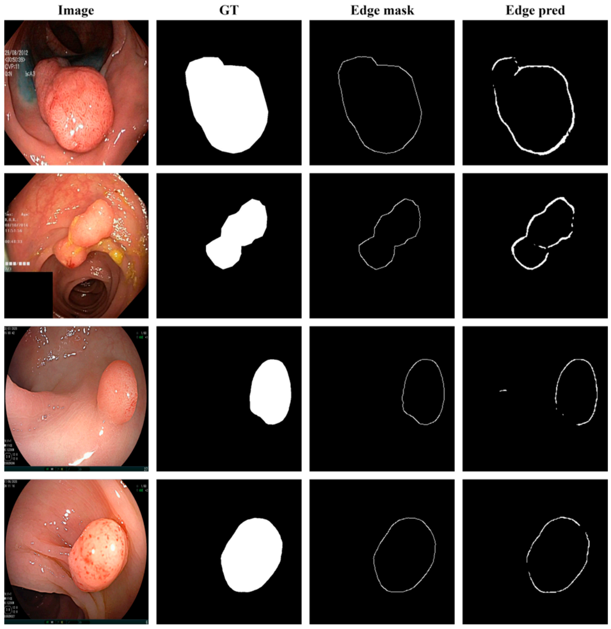

Figure 4. Image refers to sample images from the Kvasir-SEG and BKAI datasets. GT represents the ground truth annotations of polyps; Edge mask is the ground truth of polyp edges generated using the Canny operator; Edge pred denotes the predicted polyp edge maps generated by the EFE module.

Kvasir-SEG [12]: Kvasir-SEG is a publicly available endoscopic image dataset specifically designed for polyp segmentation tasks. It consists of 1000 annotated polyp images with a variety of shapes, sizes, positions, and appearances. The image resolutions range from 332 × 487 to 1920 × 1072 pixels, offering diverse challenges for segmentation models. All images are derived from endoscopic examination videos and have been meticulously annotated by experts from the University Hospital of Oslo, Norway. These reliable annotations serve as high-quality labels for model training and evaluation, making Kvasir-SEG a widely used benchmark for assessing segmentation performance.

Kvasir-Sessile [13]: Kvasir-Sessile is an extended dataset developed by researchers at the Norwegian University of Science and Technology (NTNU) to target the segmentation of flat polyps, a challenging and less prominent type of polyp. These polyps often feature ambiguous boundaries and irregular shapes, complicating the segmentation task. The dataset's images are annotated by medical experts, including trained endoscopists, ensuring precise labeling of polyp boundaries and morphology. This high-quality dataset supports model development and testing for more challenging segmentation scenarios.

BKAI [14]: The BKAI dataset, curated by the BKAI laboratory in Vietnam, is a diverse collection of endoscopic images for polyp segmentation. It includes images of various types of polyps and represents a range of complex clinical scenarios, including variations in image quality, lighting conditions, and background complexities. The dataset's annotations are manually generated by experienced endoscopists, ensuring accurate boundary localization and precise labeling. This dataset provides a robust foundation for training and evaluating models in challenging clinical contexts.

4.2. Experimental Preparation

The proposed model is implemented using the PyTorch 2.0.0 framework, with all experiments conducted on an NVIDIA GeForce RTX 3090 GPU. The Kvasir-SEG [12] is divided into training and testing sets in the official ratio of 880:120, while the Kvasir-Sessile [13] is split in an 8:1:1 ratio for training, validation, and testing. For the BKAI [14], since no test data is provided, we divided the training data into an 8:1:1 ratio for training, validation, and testing. All images are uniformly resized to 256 × 256 pixels, and data augmentation techniques, including random rotation, vertical and horizontal flipping, and random erasing, are applied to enhance the training dataset's diversity. The model is trained with a learning rate of 1×10^{-4} using the Adam optimizer [27] and a batch size of 16. We tried different sets of hyperparameters to determine the optimal configuration of the model. To prevent overfitting, an early stopping mechanism is implemented, which halts training if the validation loss does not improve after 40 epochs. Additionally, a ReduceLROnPlateau learning rate scheduler is used to adapt the learning rate during training. The training process usually takes about 35 min, and the model converges after about 50 epochs.

4.3. Results

As shown in Figure 5, in situations where lighting is uneven and polyps are small and difficult to discern, the model performs exceptionally well. This is primarily due to its dedicated EFE module, which ensures the enhancement of edge information capture even under challenging lighting conditions. In contrast, other methods, such as DeepLabv3+, may be adversely affected by such conditions, failing to accurately capture the boundary information. Notably, in images containing multiple polyps, MSEANet accurately segments each polyp region. This success stems from the model's robust understanding of global contextual information and its effective integration of edge and global context features during the decoding stage. The use of multi-scale features effectively compensates for limitations in the receptive field, ensuring that even small polyps are accurately segmented.

Furthermore, during the decoding phase, the model effectively fuses multi-scale information, allowing it to accurately segment multiple polyps. This capability ensures that the model does not miss any polyps. This design of MSEANet not only significantly expands the model's receptive field, but also enhances its capability to accurately segment small polyps and identify multi-polyp regions.

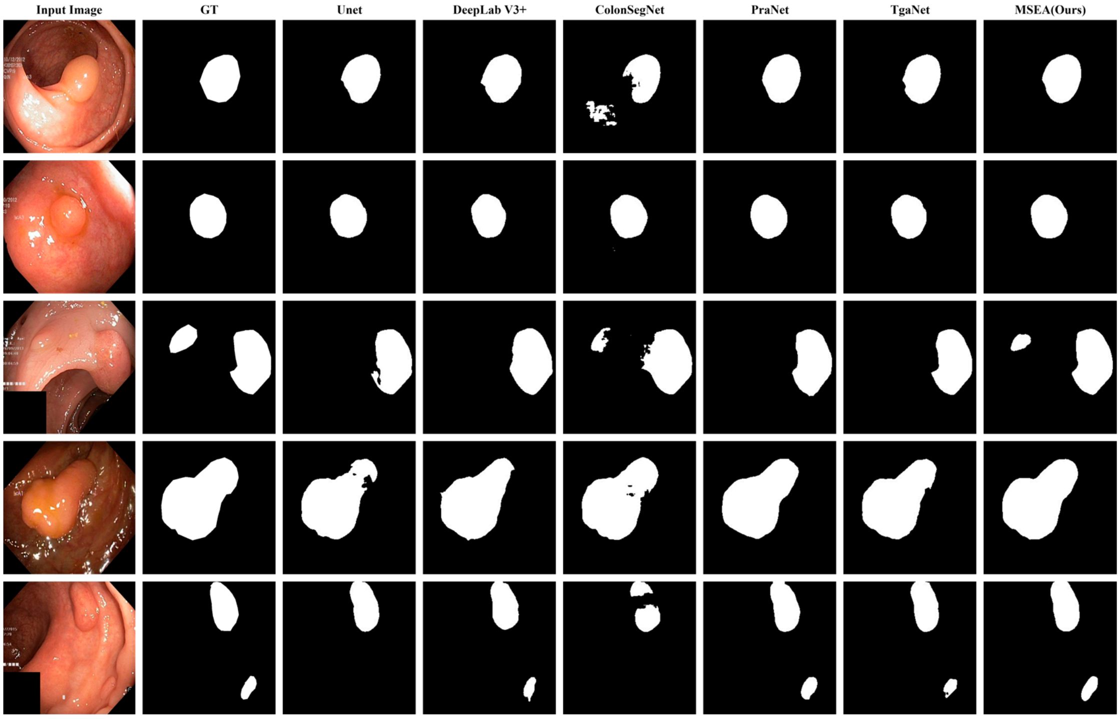

Figure 5. Comparison results of MSEANet with other advanced models on the Kvasir-SEG.

The quantitative results are summarized in Tables 1 and 2. In this study, MSEANet was compared with several state-of-the-art methods commonly used in polyp segmentation, including U-Net [5], ColonSegNet [28], Deeplabv3+ [17], PraNet [9], and TGA-Net [11]. These algorithms hold substantial importance in the field.

Comparison on Kvasir-SEG: As shown in Table 1, MSEANet outperforms other state-of-the-art methods on the Kvasir-SEG dataset across all metrics. Notably, it achieves a mIoU of 86.91% and a mDice of 91.92%, critical performance indicators in the medical domain. Additionally, other metrics such as Recall, Precision, and F2-score also reflect the model's advanced capabilities.

Comparison on Kvasir-Sessile: The Kvasir-Sessile dataset holds substantial clinical relevance as it includes challenging flat and sessile polyps [11]. As presented in Table 1, MSEANet surpasses all other methods, achieving a mIoU of 72.88%. Compared to PraNet [9], the mIoU increased by nearly 6%. These results underscore MSEANet's robustness and effectiveness in handling clinically significant yet difficult polyp types.

Comparison on BKAI: Table 2 showcases the comparative results on the BKAI dataset, where MSEANet achieves a mIoU of 87.55% and a mDice of 92.24%, along with an F2-score of 91.79%. These metrics highlight the model's ability to excel even in diverse and complex clinical scenarios. The high performance validates the model's ability to segment polyp regions accurately while significantly reducing the false negative rate. By extracting and effectively fusing features across multiple scales and receptive fields, MSEANet demonstrates a clear advantage over other advanced methods.

Model Result Analysis: The results demonstrate that MSEANet achieves excellent performance across multiple datasets, which can be attributed to its unique architecture.

Specifically, MSEANet incorporates dedicated modules for processing edge information and multi-scale contextual features. During the decoder phase, these features are efficiently fused through specialized mechanisms, allowing the model to capture both fine-grained details and global context. As shown in Figure 5, MSEANet excels in handling challenging scenarios, including blurry edges, multiple polyps, and complex backgrounds, thanks to its edge-aware and multi-scale information fusion capabilities. In contrast, other methods may struggle in these specific situations due to the absence of specialized edge-awareness modules or inadequate integration of global context. These architectural differences highlight why MSEANet outperforms other advanced models.

Table 1. Comparisons with other advanced methods on the Kvasir-SEG and Kvasir-Sessile.

Method	Kvasir-SEG					Kvasir-Sessile				
	mIoU	*mDice*	*Rec*	*Prec*	*F2*	*mIoU*	*mDice*	*Rec*	*Prec*	*F2*
U-Net [5]	76.82	84.01	86.38	88.14	84.69	24.72	36.88	72.37	32.64	46.35
ColonSegNet [28]	70.92	80.14	83.08	84.32	80.73	21.13	32.78	52.34	33.36	38.68
DeepLabV3+ [17]	80.44	87.48	88.74	90.18	87.68	59.27	70.78	70.85	82.25	70.09
PraNet [9]	83.02	89.8	90.60	91.64	90.09	66.71	77.36	80.69	82.44	78.71
TGA-Net [11]	83.30	89.82	91.32	91.23	90.29	69.23	79.78	79.35	**84.88**	79.89
MSEANet (Ours)	**86.91**	**91.92**	**92.45**	**93.87**	**91.69**	**72.88**	**82.10**	**90.92**	76.84	**86.76**

Table 2. Comparisons with other advanced methods on the BKAI.

Method	BKAI				
	mIoU	*mDice*	*Rec*	*Prec*	*F2*
U-Net [5]	75.99	82.86	82.95	89.89	82.64
ColonSegNet [28]	68.81	77.48	78.52	87.11	77.46
DeepLabV3+ [17]	83.14	89.37	88.70	92.30	88.82
PraNet [9]	82.64	89.04	89.01	92.47	88.85
TGA-Net [11]	84.09	90.23	90.26	92.08	90.02
MSEANet (Ours)	**87.55**	**92.24**	**91.81**	**95.21**	**91.79**

4.4. Ablation Study

To evaluate the effectiveness and importance of the proposed components, we conducted four ablation experiments on the Kvasir-SEG dataset. The results, summarized in Table 3, provide insights into the contribution of each module to the overall performance of MSEANet.

Table 3. Ablation study of MSEANet on Kvasir-SEG.

Method	*mIoU*	*mDice*	*Rec*	*Prec*	*F2*
Baseline	82.92	88.56	91.80	89.56	89.89
+CCF	84.74	90.23	92.02	91.94	91.60
+EFE&SEA	85.34	90.49	91.33	93.0	90.45
+ALL	**86.91**	**91.92**	**92.45**	**93.87**	**91.69**

The baseline achieved a mIoU of 82.92% and a mDice of 88.56%. When the CCF module was added, the performance improved significantly, with the mIoU increasing to 84.74%. This improvement highlights the CCF module's ability to integrate multi-scale contextual information effectively, enhancing the model's global understanding of polyp morphology.

The inclusion of the EFE further boosted performance, with the mIoU reaching 85.34%. This result demonstrates the EFE module's role in capturing high-frequency edge details, enabling better delineation of polyp boundaries.

Similarly, the addition of the SEA improved performance, showcasing its capacity to fuse edge and global context information. The combined effect of the EFE and SEA modules reflects the model's increased sensitivity to edge features and its ability to enhance segmentation accuracy, particularly in challenging cases. Finally, when all modules—CCF, EFE, and SEA—were integrated, the model achieved its highest performance, with a mIoU of 86.91%, a mDice of 91.92%, and an F2-score of 91.69%. These results validate the complementary contributions of each module and demonstrate their synergistic effect in achieving superior segmentation outcomes.

5. Conclusions

In this study, we proposed the Cross-layer Fusion and Edge-Aware network (MSEANet) to address challenges in polyp segmentation, including insufficient extraction of contextual information and insensitivity to high-frequency edge details across multiple scales. To tackle these issues, the EFE module was designed to prioritize the extraction of edge information from features rich in spatial details during the early stages of the encoder. Concurrently, the CCF module was introduced to integrate global context information across multiple scales and receptive fields, enhancing the model's understanding of polyp morphology.

The features extracted by the EFE and CCF modules were subsequently passed to the SEA module, where high-frequency edge features were effectively fused with global context features. This fusion improved the model's sensitivity to edge information, addressing a critical limitation in existing polyp segmentation methods.

Comprehensive experiments on three benchmark datasets—Kvasir-SEG, Kvasir-Sessile, and BKAI—validated the effectiveness of MSEANet. The results demonstrated significant improvements in segmentation performance across all key evaluation metrics, highlighting MSEANet's potential for clinical applications in colonoscopy. These findings underscore the value of combining edge-aware strategies with multi-scale context fusion for advancing polyp segmentation technology.

Although our model performs well on multiple datasets, real-time performance and more complex samples are required in clinical environments. Therefore, our future work will explore model lightweighting and multimodal data fusion to improve model robustness and real-time performance so it can perform better in diverse and complex clinical environments.

Author Contributions: Conceptualization, B.L. and C.S.; Formal analysis, M.Z.; Methodology, C.S.; Resources, B.L.; Software, C.S.; Supervision, B.L. and M.Z.; Validation, M.Z. and C.S.; Writing—original draft, B.L. and C.S.; Writing—review & editing, B.L. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This study received no external funding.

Data Availability Statement: The data involved in the experiments were downloaded from their respective official websites.

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

References

1. Gupta, M.; Mishra, A. A systematic review of deep learning based image segmentation to detect polyp. *Artif. Intell. Rev.* **2024**, *57*, 7. [CrossRef]
2. Kim, N.H.; Jung, Y.S.; Jeong, W.S.; Yang, H.J.; Park, S.K.; Choi, K.; Park, D.I. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intest. Res.* **2017**, *15*, 411–418. [CrossRef] [PubMed]
3. Rex, D.K.; Cutler, C.S.; Lemmel, G.T.; Rahmani, E.Y.; Clark, D.W.; Helper, D.J.; Lehman, G.A.; Mark, D.G. Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies. *Gastroenterology* **1997**, *112*, 24–28. [CrossRef] [PubMed]

4. Alzahrani, Y.; Boufama, B. Biomedical image segmentation: A survey. *SN Comput. Sci.* **2021**, *2*, 310. [CrossRef]
5. 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 International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
6. Slabaugh, G.; Unal, G.; Wels, M.; Fang, T.; Rao, B. Statistical region-based segmentation of ultrasound images. *Ultrasound Med. Biol.* **2009**, *35*, 781–795. [CrossRef] [PubMed]
7. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; Proceedings 4. Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 3–7.
8. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; De Lange, T.; Halvorsen, P.; Johansen, H.D. Resunet++: An advanced architecture for medical image segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 225–2255.
9. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 263–273.
10. Feng, S.; Zhao, H.; Shi, F.; Cheng, X.; Wang, M.; Ma, Y.; Xiang, D.; Zhu, W.; Chen, X. CPFNet: Context pyramid fusion network for medical image segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 3008–3018. [CrossRef] [PubMed]
11. Tomar, N.K.; Jha, D.; Bagci, U.; Ali, S. TGANet: Text-guided attention for improved polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 151–160.
12. 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 International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 451–462.
13. Jha, D.; Smedsrud, P.H.; Johansen, D.; De Lange, T.; Johansen, H.D.; Halvorsen, P.; Riegler, M.A. A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2029–2040. [CrossRef] [PubMed]
14. Ngoc Lan, P.; An, N.S.; Hang, D.V.; Long, D.V.; Trung, T.Q.; Thuy, N.T.; Sang, D.V. Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In Proceedings of the Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual, 4–6 October 2021; Proceedings, Part II; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 15–28.
15. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
16. Chen, L.C. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
17. 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.
18. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**. [CrossRef]
19. Feng, R.; Lei, B.; Wang, W.; Chen, T.; Chen, J.; Chen, D.Z.; Wu, J. SSN: A stair-shape network for real-time polyp segmentation in colonoscopy images. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 225–229.
20. Han, J.; Xu, C.; An, Z.; Qian, K.; Tan, W.; Wang, D.; Fang, Q. PRAPNet: A Parallel Residual Atrous Pyramid Network for Polyp Segmentation. *Sensors* **2022**, *22*, 4658. [CrossRef] [PubMed]
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
22. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
23. Lei, M.; Wang, X. EPPS: Advanced Polyp Segmentation via Edge Information Injection and Selective Feature Decoupling. *arXiv* **2024**, arXiv:2405.11846.
24. 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]
25. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.

26. Xie, L.; Li, C.; Wang, Z.; Zhang, X.; Chen, B.; Shen, Q.; Wu, Z. Shisrcnet: Super-resolution and classification network for low-resolution breast cancer histopathology image. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, 8–12 October 2023; Springer Nature Switzerland: Cham, Switzerland, 2023; pp. 23–32.
27. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
28. Jha, D.; Ali, S.; Tomar, N.K.; Johansen, H.D.; Johansen, D.; Rittscher, J.; Riegler, M.A.; Halvorsen, P. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *IEEE Access* **2021**, *9*, 40496–40510. [CrossRef] [PubMed]

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.