Online ISSN: 3104-9338
Print ISSN: 3104-932X

Research on Colorectal Polyp Semantic Segmentation Based on E-UNet Network

Zheming Zhao[1, *], Qianli Ma[2]

[1] Department of Electronics and Information Engineering, Changchun University of Science and Technology, Jilin, China

[2] Department of Internet Economics and Trade, Fujian University of Technology, Fujian, China

Received: November 24, 2025

Revised: November 26, 2025

Accepted: November 26, 2025

Published online: November 30, 2025

To appear in: *International Journal of Advanced AI Applications,* Vol. 1, No. 8 (December 2025)

* Corresponding Author: Zheming Zhao (zzm_626@163.com)

Abstract. This study proposes an enhanced computer network named E-UNet, designed to improve the semantic segmentation capability for colorectal polyps. Addressing challenges in colorectal polyp images such as immense scale variations, blurry boundaries, and color similarity to the background, E-UNet makes key improvements based on the UNet architecture. First, Pyramidal Convolution (PyConv) is adopted to replace standard convolution, solving the problem of multi-scale feature extraction without increasing computational cost. Second, Soft Pooling is introduced to replace Max Pooling, reducing information loss during down-sampling and preserving more key low-amplitude signals such as blurry boundaries and subtle textures. Finally, an improved CBAM (I-CBAM) attention mechanism is designed. By processing channel and spatial attention in parallel and optimizing the MLP structure, it dynamically focuses on polyp morphology and key features, thereby overcoming issues such as the similarity between polyp and background colors. Experimental results on the authoritative public dataset Kvasir-SEG show that E-UNet outperforms mainstream methods such as UNet, UNet++, and Attention UNet across all evaluation metrics, achieving an mIoU of 87.5%, an mRecall of 91.9%, and an mPrecision of 93.1%. Ablation studies further verify the effectiveness of the I-CBAM, PyConv, and Soft Pooling modules, with the complete model achieving a 3.7% mIoU improvement compared to the baseline UNet.

Keywords: *Colorectal Polyp; Semantic Segmentation; Pyramidal Convolution; Soft Pooling; Improved CBAM*

1. Introduction

Colorectal polyp semantic segmentation technology, by enabling precise, pixel-level identification of lesions in colonoscopy images, holds significant value in clinical medicine. This technology can substantially enhance the efficacy of early colorectal cancer screening[1]. Acting as a "second pair of eyes" for physicians, it effectively reduces the likelihood of missing easily overlooked lesions, such as diminutive or flat polyps, thereby significantly lowering the missed diagnosis rate of manual examinations and providing an objective basis for clinical decision-making[2]. In terms of optimizing diagnostic and treatment workflows, segmentation results can directly guide the precise implementation of automated polypectomy. When integrated with advanced endoscopic techniques like Narrow Band Imaging (NBI), it improves diagnostic confidence. Furthermore, it facilitates the construction of structured reporting systems, laying the foundation for telemedicine and tiered diagnosis and treatment. Moreover, this technology propels the development of clinical research and medical education. By establishing large-scale polyp segmentation databases, it not only aids epidemiological studies but also serves as a high-quality teaching resource for training endoscopists. As an ideal experimental scenario for validating computer vision algorithms in the medical field, breakthroughs in colorectal polyp semantic segmentation technology not only directly serve colorectal cancer prevention and control but also offer successful experiences that can be transferred to the analysis of other endoscopic images. Ultimately, this contributes to building an intelligent diagnostic and treatment system covering the entire process of screening, diagnosis, treatment, and follow-up.

Long et al.[3] first proposed Fully Convolutional Networks (FCNs) by replacing the fully connected layers at the end of classification networks with convolutional layers, enabling them to accept inputs of arbitrary sizes and output dense heatmaps. This work pioneered the end-to-end pixel-level prediction paradigm, laid a solid foundation for deep learning-based semantic segmentation, and marked the beginning of a new era in the field. Yu et al.[4] systematically introduced atrous convolution (also known as dilated convolution) into convolutional neural networks. By inserting zero values between convolutional kernel elements, this technique expands the receptive field without increasing the number of parameters. It effectively mitigates the resolution loss caused by pooling operations and has become a key component for many subsequent segmentation models, enabling them to capture multi-scale context information while maintaining feature map resolution. UNet, proposed by Ronneberger et al.[5], adopts a symmetric encoder-decoder structure and skip connections, enabling the efficient fusion of

shallow detailed features captured by the encoder path with deep semantic information recovered by the decoder path. Its elegant and efficient design has not only made it a benchmark model in the field of biomedical image segmentation but has also widely influenced the development of general semantic segmentation architectures. He et al.[6] introduced ResNet, where residual connections solve the degradation problem of very deep networks, enabling the construction of deeper and more expressive semantic segmentation encoders. Paszke et al.[7] proposed ENet, designing a highly asymmetric encoder-decoder structure that allocates more computational resources to the encoder while keeping the decoder very lightweight. This model heavily prioritizes inference speed and efficiency, offering a cost-effective solution for resource-constrained real-time applications. Badrinarayanan et al.[8] proposed SegNet, with a core innovation where the decoder utilizes pooling indices recorded during the encoder's max-pooling steps for non-linear upsampling. This method is more efficient than learned upsampling, improves boundary segmentation accuracy while maintaining low memory consumption, and provides an important design concept for real-time segmentation applications. Zhao et al.[9] proposed PSPNet, which utilizes a Pyramid Pooling Module (PPM) to aggregate contextual information from different sub-regions of the image. By employing pooling kernels of various sizes, this module captures multi-scale context ranging from local to global, effectively aiding the network in distinguishing confusing categories and enhancing segmentation consistency for large objects and the overall scene. Chen et al. [10]developed a network that processes multi-scale images via an attention mechanism. Its core attention mechanism module can adaptively learn and fuse the importance of feature maps at different scales. This approach achieves efficient and intelligent fusion of multi-scale features, significantly improving the model's segmentation accuracy for objects of varying sizes in complex scenes. Chen et al.[11] on the basis of DeepLabv3 added a lightweight decoder structure to construct DeepLabv3+ network, aiming to gradually recover object boundary details lost in the encoder path. This model combines the encoder-decoder structure's advantages in boundary recovery and atrous convolution's ability in multi-scale context modeling, becoming the culmination in this series that balances accuracy and detail. Lin et al.[12] proposed RefineNet, a classic encoder-decoder model based on Multi-Path Refinement, specifically designed for high-accuracy semantic segmentation. Rather than using simple skip connections directly, it employs complex RefineNet Blocks. Each block receives feature maps from the same resolution and all lower resolutions (i.e., more abstract levels) of the encoder as input. Internally, these blocks fuse multi-resolution features through residual convolutions and chained pooling (utilizing multiple pooling kernels of different sizes), thereby capturing richer contextual information.

Existing CNNs in feature extraction process mainly rely on their locality inductive bias, but because convolutional kernels' receptive field range is limited, cannot effectively capture global dependencies, leading to detail information lost or boundary blurred. At the same time this research aiming at colorectal polyp images' specific characteristics, improves UNet model's feature extraction and skip connection parts, to better capture target multi-scale information, enhance small target and boundary recognition accuracy, can fully use local features, and then improve segmentation performance.

2. UNet Semantic Segmentation Network

UNet[5] is a convolutional neural network with an encoder-decoder architecture, originally proposed by Olaf Ronneberger et al. in 2015, primarily for biomedical image segmentation tasks. It was later widely adopted in various semantic segmentation scenarios due to its excellent performance. Its core design revolves around a symmetric U-shaped structure, which merges high-resolution features from the encoder path with up-sampled features from the decoder path via skip connections, balancing both global semantic information and local spatial details.

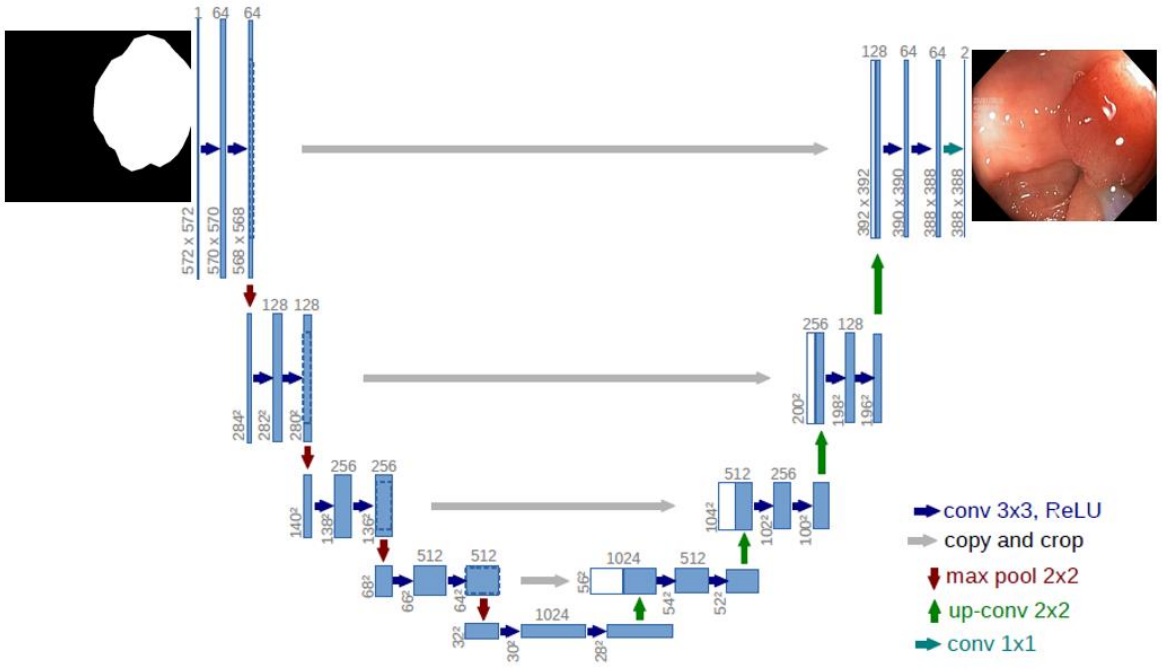

Figure 1. The UNet network model structure.

The encoder of UNet consists of multiple convolutional and pooling layers, progressively extracting multi-level features and reducing spatial resolution to capture the image's contextual semantic information. The decoder, conversely, gradually restores spatial resolution through transposed convolutions or up-sampling operations, and by combining feature maps from the

corresponding levels of the encoder, it achieves precise localization and detail reconstruction. Skip connections are the key innovation of UNet; they directly pass high-resolution features from each stage of the encoder to the corresponding layer of the decoder, effectively mitigating the loss of spatial information caused by down-sampling and significantly improving the accuracy of boundary segmentation. The network structure of UNet is shown in Figure 1.

3. E-UNet Colorectal Polyp Semantic Segmentation Network

This paper proposes an enhanced computer network named E-UNet, designed to improve the semantic segmentation capability for colorectal polyps. This network is optimized based on the UNet architecture, significantly enhancing its feature extraction and fusion abilities.

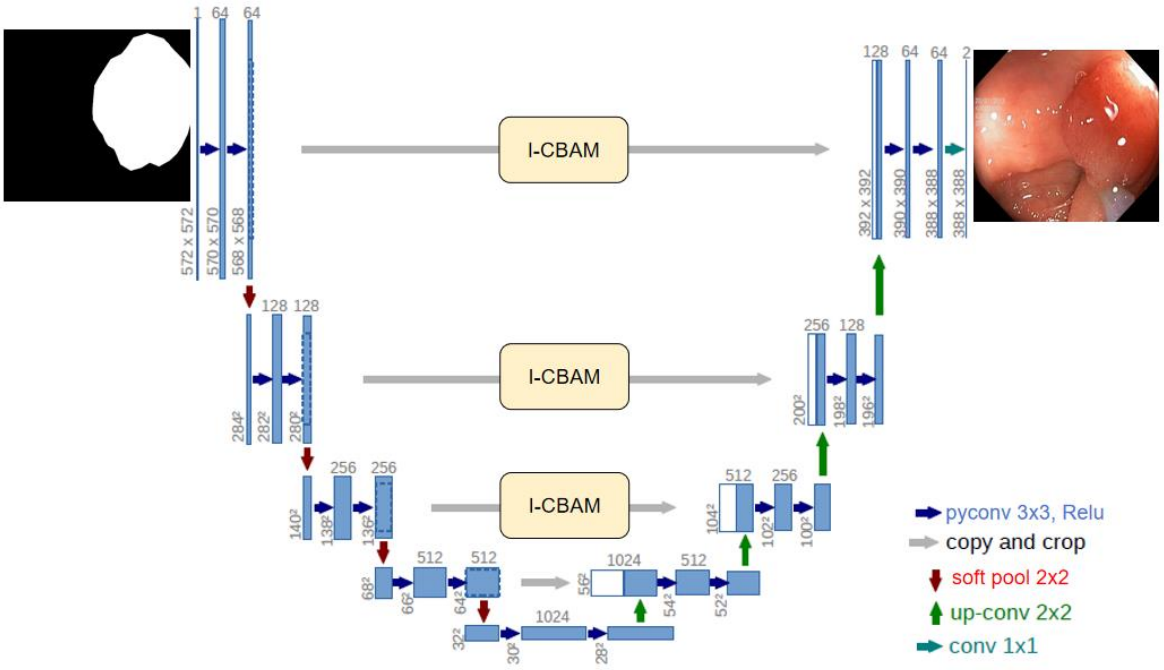

Figure 2. The E-UNet network model structure.

In this study, two key improvements are made to the UNet encoder to construct a feature extractor that is highly robust to both scale variations and fine details. The traditional UNet encoder, due to its fixed convolutional kernel sizes, can only observe the image with a single receptive field at each layer, making it difficult to cope with the immense scale variation challenge in colorectal polyps. Therefore, this study replaces standard convolution with Pyramidal Convolution (PyConv). Furthermore, it is crucial to ensure that information is not lost during the down-sampling process. Max Pooling crudely discards a large amount of key features, including blurry polyp boundaries and subtle textures. To ensure that the valuable details captured by PyConv can be passed to the deeper layers with high fidelity, this study replaces Max Pooling with Soft Pooling. Finally, to address challenges such as blurry polyp

edges and the similar color between polyps and the background, the network introduces an improved CBAM (I-CBAM) attention mechanism. This module can dynamically focus on the polyp's morphology and the key features across various dimensions. Figure 2 illustrates the overall architectural design of the E-UNet network for colorectal polyp semantic segmentation.

3.1. E-UNet Feature Extraction Design

The core idea of the UNet architecture is its use of a symmetrical encoder-decoder structure, which employs skip connections to fuse deep semantic information with shallow, high-resolution features. Along the encoder path, UNet relies on successive convolutional neural network (CNN) layers and pooling operations to extract local features in a hierarchical manner and progressively enlarge the receptive field.

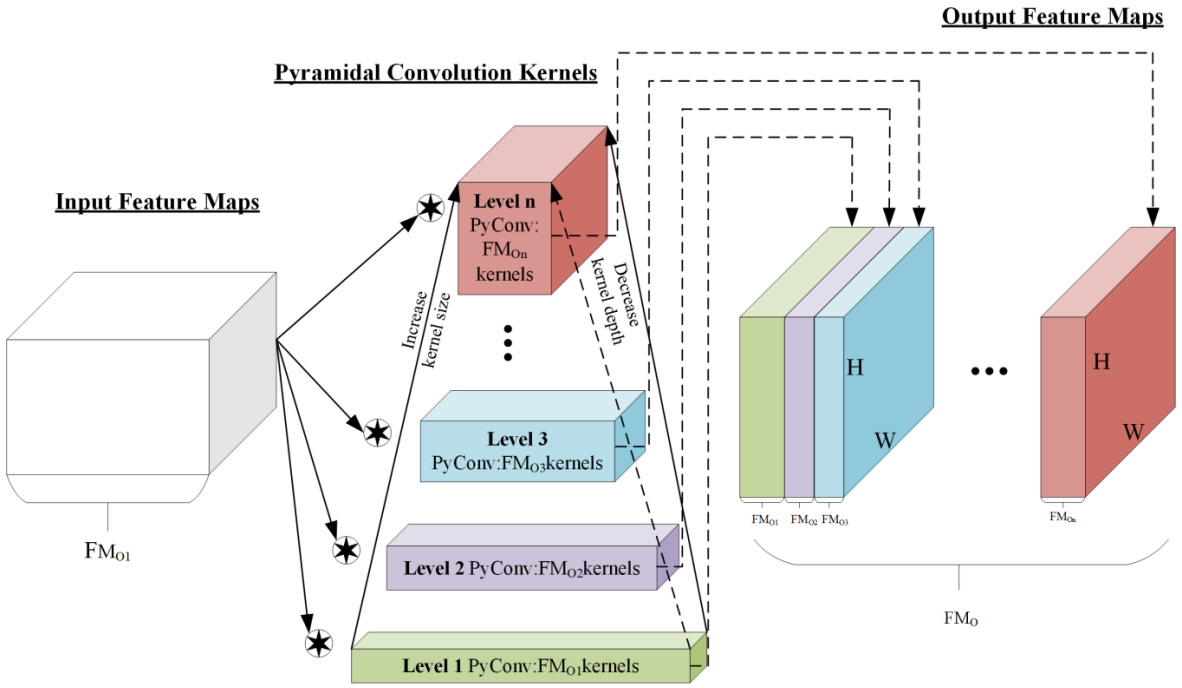

Figure 3. Pyramid convolution calculation process

However, when using original, standard convolutional operations as the encoder, especially when applied to the highly challenging task of colorectal polyp semantic segmentation, two significant limitations exist. First, the traditional UNet encoder utilizes fixed-size convolutional kernels (e.g., standard 3×3 convolutions) at each convolutional layer. This design results in a receptive field that is highly uniform and homogeneous at each layer. In the clinical endoscopic images characteristic of colorectal polyps, the lesions themselves exhibit extreme multi-scale properties: Polyps can range from tiny, flat lesions, only a few millimeters in diameter and difficult to discern against the mucosal background, to large pedunculated or sessile polyps that occupy a significant portion of the visual field. The borders of a polyp may be sharp and well-

defined, but they are frequently blurry and ill-defined, showing a gradual transition into the surrounding normal mucosa, particularly in the case of flat polyps. Although stacking and pooling allow the deeper layers of the network to acquire a larger receptive field for capturing the global structure of large polyps, feature extraction in the early and mid-stages of the encoder remains strictly confined to information at similar scales.

In this study, Pyramidal Convolution (PyConv)[13][14] with its computational process illustrated in Figure 3, is introduced to replace the original, fixed-size convolutional kernels in UNet. This is done precisely to address this problem of multi-scale feature extraction without increasing the model's computational cost or complexity.

In the UNet encoder path, Max Pooling is the standard operation for achieving down-sampling and expanding the receptive field. However, its "winner-takes-all" selection mechanism possesses severe drawbacks. In a typical 2×2 window, Max Pooling permanently discards 75% of the feature information, retaining only the peak activation in that region. In high-precision tasks like colorectal polyp segmentation, this discarded information often contains key low-amplitude signals such as the polyp's blurry boundaries, subtle mucosal textures, or tiny lesions. This irreversible information loss occurs early in the encoder, making it difficult for the decoder to reconstruct precise segmentation details, even with the aid of skip connections. Furthermore, during backpropagation, the gradient only flows to the selected maximum value, leading to a sparse gradient flow, which not only reduces training efficiency but also limits the model's ability to learn complex features.

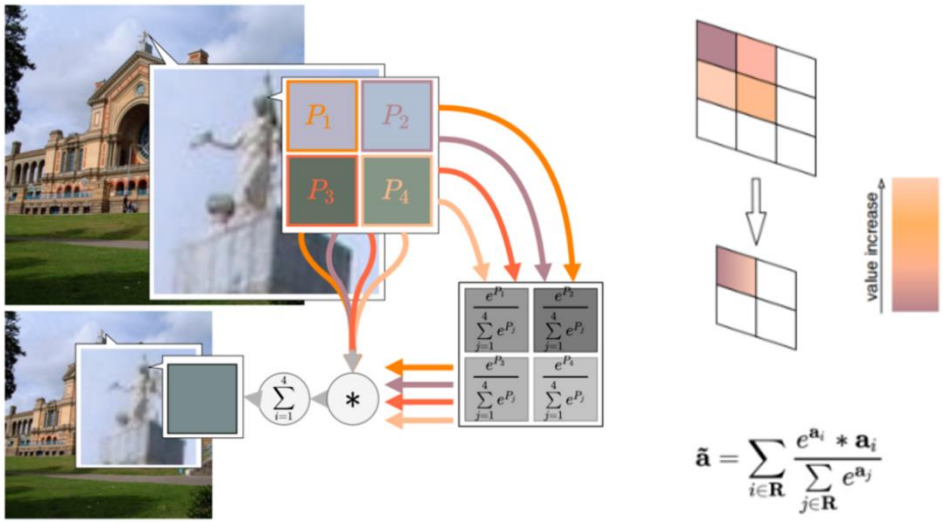

Figure 4. Soft pooling calculation process.

In contrast to the "hard" selection of Max Pooling, Soft Pooling[15] adopts a weighted-average down-sampling strategy. It does not discard any information; instead, it performs a

weighted sum of all activation values within the pooling window based on their importance (often calculated via Softmax). The core advantage of this mechanism is the complete retention of feature information: features with high activation values receive high weights, while secondary features with lower activation values (e.g., edges or texture details) are still preserved in the feature map with smaller weights, rather than being completely eliminated. Concurrently, because the pooled output is a smooth function of all its inputs, gradients can flow unimpeded to all neurons within the window during backpropagation. This dense and smooth gradient flow greatly enhances training stability and enables the network to learn richer and more robust feature representations.

Assuming R is a pooling region, a_i is the i-th activation value in that region R. First, the Softmax function is used to calculate the weight w_i, for each activation value a_i, within the region R. The weight w_i reflects the importance of a_i relative to the other values in that region, is shown in Equations (1):

$$w_i = \frac{e^{a_i}}{\sum_{j \in R} e^{a_j}} \tag{1}$$

Then, the calculated weight w_i is multiplied by the corresponding activation value a_i, and all the results are summed together to obtain the final pooled output ã, is shown in Equations (2).

$$\tilde{a} = \sum_{i \in R} w_i \cdot a_i = \sum_{i \in R} \frac{e^{a_i}}{\sum_{j \in R} e^{a_j}} \cdot a_i \tag{2}$$

3.2. E-UNet Skip Connection Design

The traditional Convolutional Block Attention Module (CBAM[16]) is a powerful and widely adopted mechanism designed to refine feature maps by inferring attention weights along both the channel ("what") and spatial ("where") dimensions. However, its standard architecture possesses two key limitations that hinder its performance on high-fidelity tasks like ancient script processing.

First, the Channel Attention Module (CAM) in a standard CBAM employs a shared Multilayer Perceptron (MLP) that follows a "dimension reduction followed by dimension increase" strategy. This architecture, designed for computational efficiency, creates an information bottleneck. By first compressing the channel information, the model risks irretrievably discarding subtle, low-amplitude signals that represent the very fine details crucial for interpreting degraded scripts. Second, the traditional CBAM framework is cascaded, meaning the spatial attention module operates on the feature map after it has already been re-

weighted by the channel attention module. This sequential dependency creates instability and mutual interference. An error or suboptimal weighting from the CAM (e.g., suppressing a channel that contained spatially critical information) is propagated and permanently baked into the input for the Spatial Attention Module (SAM), which cannot recover the lost information. This entanglement limits the expressive power of both modules.

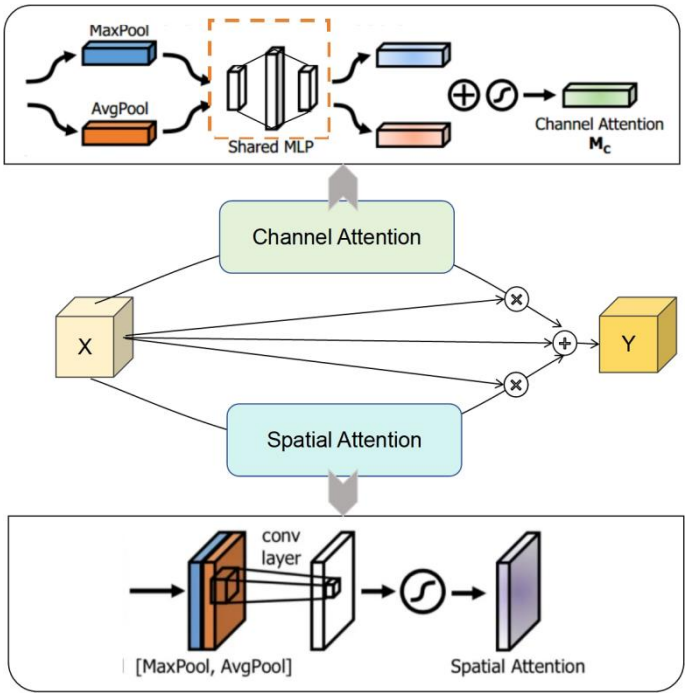

Figure 5. The improved CBAM (I-CBAM) model structure.

To address these critical shortcomings, this study proposes an improved CBAM (I-CBAM[17]), which introduces two fundamental architectural modifications specifically engineered to preserve fine details and enhance model stability. The I-CBAM reverses the processing order within the shared MLP of the CAM to an "dimension increase followed by dimension reduction" structure. By first expanding the channel dimensions, the module creates a higher-dimensional feature space. This allows the MLP to learn more complex and richer cross-channel relationships before compressing this information into the final channel attention weights. This "pre-amplification" of feature combinations ensures that subtle, nuanced details from the original image are effectively captured and preserved, rather than being prematurely discarded by a dimensional bottleneck. The I-CBAM abandons the traditional cascaded structure in favor of a parallel architecture. In this new design, both the CAM and the SAM receive the exact same original input feature map from the preceding convolutional layer. They operate independently and concurrently, allowing each module to learn its respective attention weights directly from the pristine, unmodified features. This decoupling is paramount: it

eliminates the mutual interference and error propagation inherent in the cascaded design. The channel attention can focus purely on "what" is important, and the spatial attention can focus purely on "where" it is important, without one's decisions compromising the other.

The final refined feature map is generated by combining the outputs of these two independent modules. As illustrated in Figure 5, which details the I-CBAM model structure, this approach of independently learning and then integrating channel and spatial priorities not only enhances the stability of the learning process but also significantly improves the overall effectiveness of feature extraction.

The input features (X) are processed by the Channel Attention Module (CAM) to obtain the weight F_{CAM}, while simultaneously being processed by the Spatial Attention Module (SAM) to obtain the weight F_{SAM}. The two resulting weights (F1 and F2) are then multiplied with the input features (X) respectively. Finally, the weighted features are combined with the original input features (X) through a residual connection to produce the output features (Y). The calculation process is shown in Equations (3), (4), and (5).

$$F_{CAM} = \sigma\left(MLP\left(AvgPool(x)\right) + MLP\left(MaxPool(x)\right)\right) \tag{3}$$

$$F_{SAM} = \sigma\left(f^{7\times7}([AvgPool(x); MaxPool(x)])\right) \tag{4}$$

$$Y = F_{CAM} \times X + F_{SAM} \times X + X \tag{5}$$

σ stands for Sigmoid activation operation, **MLP**$(*)$ stands for shared multilayer perceptron operation with first dimension up and then dimension down. $\mathbf{f^{7\times7}}(*)$ represents 7×7 convolution dimensionality reduction operation.

4. Experiments and results analysis

4.1. Dataset selection and processing

Kvasir-SEG[18] is a public, authoritative benchmark dataset specifically for colorectal polyp semantic segmentation. The dataset consists of 1,000 images from actual colonoscopies and their corresponding 1,000 pixel-level ground truth masks. These masks were all manually drawn and validated by experienced medical experts, ensuring extremely high annotation quality. The core value of Kvasir-SEG lies in its inclusion of various real-world clinical challenges: polyps exhibit immense scale variations (from tiny to large) and diverse morphologies (such as flat, pedunculated, and sessile types). Furthermore, many polyps have blurry boundaries that are indistinct from the surrounding normal mucosa and similar in color. The dataset also contains varying image resolutions and artifacts like reflections and blur.

To meet experimental requirements, all images were uniformly resized to a 256×256 resolution. The dataset was divided into a 70% training set, a 20% validation set, and a 10% test set.

4.2. Experimental environment and the parameter settings

All experiments in this study were conducted on a workstation equipped with an Intel® Core i7-14650HX CPU (base frequency: 2.2 GHz, max turbo frequency: 4.1 GHz), an RTX 4060 GPU, and the Windows 10 Home operating system. The experimental details are summarized in Table 1.

Table 1. Experimental parameters and values.

Parameter	Numerical Value
Initial learning rate	0.001
Learning rate Scheduler	PolynomiaILR
Decay rate	0.9
Network optimizer	Adam
Batch size	16
Number of iterations	100
Number of categories	2

4.3. Evaluating indicator

This experiment utilizes the Mean Intersection over Union (mIoU), Mean Precision (mPrecision), and Mean Recall (mRecall) metrics to evaluate the algorithm's performance. Precision measures the accuracy of the positive-class pixels predicted by the model. A high precision value indicates that the regions the model identified as positive have a high degree of confidence, and the false detection rate is low. Recall is used to evaluate the completeness of the model's coverage of actual positive-class pixels. A high recall value means the model can effectively detect the vast majority of target objects, with few missed detections. Intersection over Union (IoU) not only considers the correctness of pixel classification but also more strictly assesses the agreement between the predicted boundaries and the g round truth boundaries. If the IoU value is high, it signifies that the model is not only accurate in its pixel-level class judgments but also precise in its shape delineation, indicating high overall segmentation quality and balanced performance across classes. Their mathematical descriptions are as follows:

The mIoU coefficient is shown in Equation 6:

$$mIou = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP_k}{TP_k + FP_k + FN_k} \tag{6}$$

The mPrecision coefficient is shown in Equation 7:

$$mPrecision = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP_k}{TP_k + FP_k} \tag{7}$$

The mRecall coefficient is shown in Equation 8:

$$mRcall = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP_k}{TP_k + FN_k} \tag{8}$$

In the equation, TP represents true positives, FN represents false negatives, FP represents false positives, and k denotes the number of categories to be segmented in the dataset.

4.4. Analysis of experimental results

To evaluate the segmentation performance of the E-UNet network on the public colorectal polyp dataset Kvasir-SEG, UNet, UNet++[19], and Attention UNet[20] were selected and trained alongside the E-UNet model under the unified environment described in this study for quantitative comparison. The quantitative results on the Kvasir-SEG dataset, based on the evaluation metrics detailed in Section 4.3, are presented in Table 2. The best metrics in the table are indicated in bold, and the second-best metrics are underlined.

Table 2. Quantitative metric comparison of different models on the Kvasir-SEG dataset.

Module	mIou (%)	mPrecision (%)	mRcall (%)
UNet	83.8	89.1	87.2
UNet++	84.5	91.1	89.5
Attention UNet	86.2	92.9	91.2
E-UNet	87.5	93.1	91.9

As indicated in Table 2, the E-UNet network model proposed in this chapter outperforms other mainstream methods in overall experimental accuracy, achieving the highest scores across all evaluation metrics. On the Kvasir-SEG dataset, it attained an mIoU of 87.5%, an mRecall of 91.9%, and an mPrecision of 93.1%. Compared to the baseline UNet network, these represent improvements of 3.7% in mIoU, 4.7% in mRecall, and 4.0% in mPrecision.

Figure 6 is the above three comparison networks and E-Unet network's segmentation visualization effects on gastroscopy dataset Kvasir-SEG. It can be seen that segmentation networks with CNN as backbone when processing endoscopic lesions, consecutively use convolution, pooling operations, leading to weaker performance in capturing long-range context information in images, leading to segmentation result accuracy decline. This paper's network not only can better identify global information, but also further overcome the problem of polyp color being similar to background, and can detect polyp tissues of different shapes, sizes and colors, realizing clearer and more accurate region and boundary division, avoiding

missed detection situations.

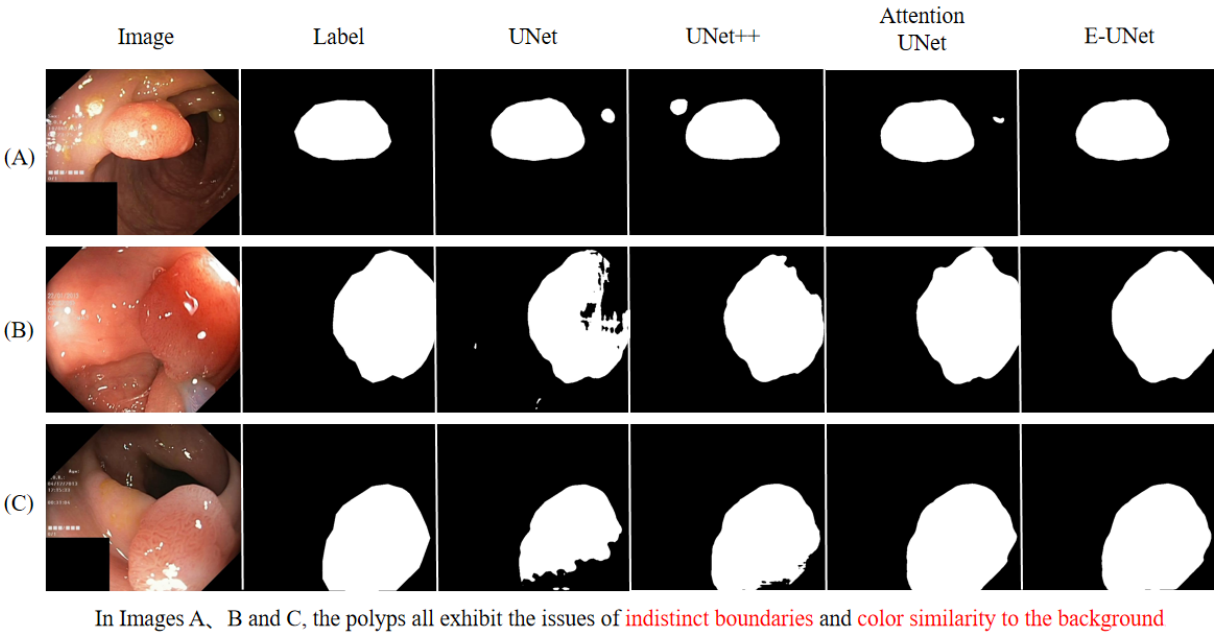

Figure 6. Structure segmentation visualization results of different methods on the Kvasir-SEG dataset.

To verify the optimization effect of the E-Unet network in terms of computational complexity, the UNet, UNet++, and Attention UNet models from the comparative experiments were selected for comparison. The evaluation metrics include FLOPs (Floating-Point Operations) and Parameters, aiming to assess the model's computational efficiency and inference performance. The comparative results of model parameters are presented in Table 3.

Table 3. Comparison of model parameter values.

Module	Parameters/M	FLOPs/G	mIou
UNet	7.76	14.21	83.8
UNet++	9.15	35.62	84.5
Attention UNet	56.6	133.45	86.2
E-UNet	7.09	17.01	87.5

Table 3 compares the UNet, UNet++, Attention UNet, and E-UNet models across three dimensions: parameter count, computational complexity, and segmentation performance. In terms of parameter count, E-UNet achieves a lightweight design with 7.09 M parameters, which is lower than the baseline UNet's 7.76 M, while Attention UNet has 56.6 M parameters, significantly increasing storage costs. Regarding computational complexity, the FLOPs of the models are approximately positively correlated with the parameter count. The baseline UNet (14.21 G) and E-UNet (17.01 G) exhibit lower computational costs, whereas UNet++ (35.62 G) and Attention UNet (133.45 G) show substantial increases in FLOPs. Notably, the computational complexity of Attention UNet is 9.4 times that of the baseline UNet. From the

perspective of performance-complexity trade-off, E-UNet achieves the optimal mIoU of 87.5 with parameter counts and computational complexity close to those of the baseline UNet. In contrast, although Attention UNet improves the mIoU to 86.2, its parameter count and FLOPs are 7.3 times and 9.4 times those of the baseline UNet, respectively, indicating a significant efficiency cost. Meanwhile, UNet++ achieves only a slight improvement in mIoU (+0.7) at the expense of increased complexity (parameter count +17.9%, FLOPs +150.7%). These results demonstrate that E-UNet achieves a better balance among parameter count, computational complexity, and segmentation performance, validating its dual advantages in computational efficiency and task performance.

The results of the model ablation study are detailed in Table 4. In experiment (a), we introduced the CBAM module into the UNet network. Experiment (b) replaced the CBAM module in experiment (a) with an improved CBAM (I-CBAM) module. Experiment (c) built upon experiment (b) by introducing PyConv to replace the standard convolutions in the UNet network. Finally, building on experiment (c), Soft Pooling was introduced to replace the max pooling in the UNet network, constructing the complete E-UNet network for colorectal polyp semantic segmentation.

Table 4. Experimental results of model validity verification.

Module	CBAM	I-CBAM	PyConv	Soft Pooling	mIou (%)
UNet					83.8
Experiment(a)	√				86.1
Experiment(b)		√			86.3
Experiment(c)		√	√		87.1
E-UNet		√	√	√	87.5

Table 3 presents the ablation study results of the E-UNet model, aiming to verify the effectiveness of each improved module. Using the original UNet as a baseline (mIoU 83.8%), the introduction of the standard CBAM module first significantly increased the mIoU by 2.3% to 86.1%. Subsequently, replacing the standard CBAM with the improved I-CBAM brought a further increase of 0.2% (mIoU 86.3%). Building on this, the introduction of PyConv further boosted performance by 0.8%, reaching 87.1%. Finally, by integrating Soft Pooling to construct the complete E-UNet model, the mIoU ultimately reached 87.5%. The entire ablation study shows that the complete model achieved a total performance improvement of 3.7% compared to the baseline, with each step demonstrating a positive gain. This fully validates the necessity and effectiveness of the I-CBAM, PyConv, and Soft Pooling modules in enhancing colorectal polyp segmentation performance.

5. Conclusion

Addressing the challenges present in colorectal polyp endoscopic images, such as large variations in lesion scale, blurry boundaries, and high color similarity to the mucosal background, this study proposes E-UNet, an improved high-precision semantic segmentation network based on UNet. First, Pyramidal Convolution (PyConv) is utilized to replace standard convolution, effectively enhancing the model's ability to capture multi-scale polyp features. Second, Soft Pooling is introduced to replace traditional Max Pooling, maximally preserving critical low-amplitude information, such as blurry boundaries and fine textures, during the down-sampling process. Finally, an improved attention mechanism (I-CBAM) is designed, which significantly enhances the model's ability to focus on polyp morphology and suppresses background noise interference through parallel spatial and channel attention modules and an optimized MLP structure.

Experimental results on the authoritative public dataset Kvasir-SEG indicate that the comprehensive performance of E-UNet is superior to mainstream networks such as UNet, UNet++, and Attention UNet, with its mIoU, mRecall, and mPrecision reaching 87.5%, 91.9%, and 93.1%, respectively. Ablation studies further verify the effectiveness and necessity of the three aforementioned improvement modules, with the complete model achieving a 3.7% improvement in segmentation accuracy over the baseline UNet. The E-UNet model proposed in this study can provide more precise pixel-level lesion identification and holds promise as an effective auxiliary tool for early clinical colorectal cancer screening, reducing missed diagnosis rates and optimizing diagnostic and treatment workflows.

References

[1] Jafar, A., Abidin, Z. U., Naqvi, R. A., & Lee, S. W. (2024). Unmasking colorectal cancer: A high-performance semantic network for polyp and surgical instrument segmentation. Engineering Applications of Artificial Intelligence, 138, 109292.

[2] Manan, M. A., Feng, J., Yaqub, M., Ahmed, S., Imran, S. M. A., Chuhan, I. S., & Khan, H. A. (2024). Multi-scale and multi-path cascaded convolutional network for semantic segmentation of colorectal polyps. Alexandria Engineering Journal, 105, 341-359.

[3] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[4] Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

[5] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: Springer international publishing.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[7] Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147.
[8] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.
[9] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).
[10] Chen, L. C., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3640-3649).
[11] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).
[12] Lin, G., Milan, A., Shen, C., & Reid, I. (2017). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1925-1934).
[13] Duta, I. C., Liu, L., Zhu, F., & Shao, L. (2020). Pyramidal convolution: Rethinking convolutional neural networks for visual recognition. arXiv preprint arXiv:2006.11538.
[14] Li, Z., Yang, S., & Zhao, Z. (2025, May). Oracle Bone Inscription Recognition Based on the S-MobileViT Network. In 2025 5th International Symposium on Computer Technology and Information Science (ISCTIS) (pp. 118-123). IEEE.
[15] Stergiou, A., Poppe, R., & Kalliatakis, G. (2021). Refining activation downsampling with SoftPool. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10357-10366).
[16] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
[17] Liu, K., Song, J., Zhao, Z., Liu, H., Qu, Z., & Wang, S. (2025, April). Research on the Recognition of Bamboo and Silk Scripts Based on the E-MobileViT Network. In Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area Education Digitalization and Computer Science International Conference (pp. 172-178).
[18] Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., De Lange, T., Johansen, D., & Johansen, H. D. (2019, December). Kvasir-seg: A segmented polyp dataset. In International conference on multimedia modeling (pp. 451-462). Cham: Springer International Publishing.
[19] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation[C]//International workshop on deep learning in medical image analysis. Cham: Springer International Publishing, 2018: 3-11.
[20] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.