

Colorectal Polyp Image Segmentation Based on DeepLab-RCP

Shaosai Wang^{1,*}, Xuewen Ding^{1,2}, Hemao Jiang¹, Yanlong Jia¹

¹ School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China

² Tianjin Yunzhitong Technology Co., Ltd., Tianjin 300350, China

* Corresponding author: Shaosai Wang (Email: 1282161216@qq.com)

Abstract: Colorectal cancer early screening and polyp resection are of great significance for prevention. With the development of deep learning technologies, computer-aided detection has become a mainstream method for polyp segmentation. However, challenges such as variations in polyp size, complex shapes, and blurred boundaries remain. To address these issues, this paper proposes an improved polyp segmentation model based on the DeepLabv3+ framework, named DeepLab-RCP (DeepLab with ResNeXt, Cross-Dense Fusion, and Parallel Swin-ASPP). The model designs a Cross-Dense Fusion (CDF) module to tackle size variations and a Parallel Swin Transformer and Atrous Spatial Pyramid Pooling (PS-ASPP) module to handle shape complexity. Additionally, ResNeXt-50 and Depthwise Separable Convolution (DSConv) are employed to optimize boundary segmentation. Experimental results demonstrate that the model achieves mean Dice coefficients and mean intersection over union of 92.8% and 86.9%, respectively, on the Kvasir-SEG dataset, and 95.5% and 91.6% on the CVC-ClinicDB dataset, validating its effectiveness.

Keywords: Polyp Segmentation; Deeplab; Multi-scale Feature Fusion; ResNeXt50; DSConv.

1. Introduction

In recent years, deep learning technologies have achieved breakthrough progress in numerous fields of computer vision, owing to their powerful feature learning capabilities and end-to-end optimization characteristics[1]. In the task of colorectal polyp segmentation, deep learning-based methods have demonstrated significant advantages over traditional approaches[2]. Firstly, they can automatically learn discriminative multi-level features from raw images, eliminating the need for complex manual feature design, thereby enhancing the robustness and richness[3] of feature representation. Secondly, through end-to-end training, models can simultaneously optimize both feature extraction and segmentation decision-making[4], significantly improving segmentation accuracy and processing efficiency. Lastly, leveraging the acceleration capabilities of hardware such as GPUs[5], deep learning models enable rapid training and efficient inference, providing technical feasibility for real-time clinical auxiliary diagnosis.

Considering the evolution of core model architectures, existing deep learning methods can be broadly categorized into two classes: those based on convolutional neural networks (CNNs) and those based on Transformers[6]. In 2015, the fully convolutional network proposed by Long et al. first realized end-to-end semantic segmentation[7], laying a crucial foundation for pixel-level prediction tasks. In the same year, Ronneberger et al. developed the U-Net architecture tailored for biomedical imaging[8], which effectively combines low-level detail information with high-level semantic information through its encoder-decoder structure and skip connections, exerting a profound influence in the field of medical image segmentation[9]. Subsequently, numerous research efforts have focused on improving U-Net. For instance, U-Net++, proposed by Zhou et al., enhances multi-scale feature fusion through dense skip connections. Models such as R2U-Net[10], MultiResUNet, and DoubleU-

Net further strengthen the foundational architecture from perspectives including recurrent convolution, multi-resolution analysis[11], and dual U-shaped structures, respectively.

Beyond the U-Net series, models represented by the DeepLab series have introduced the Atrous Spatial Pyramid Pooling (ASPP) module[12], which captures multi-scale contextual information through parallel convolutional layers with different dilation rates[13]. This effectively expands the receptive field without sacrificing resolution, demonstrating excellent performance when segmenting polyps with significant size variations[14]. DeepLabv3+ further incorporates an encoder-decoder structure, leveraging the rich semantic information extracted by ASPP and the spatial details provided by the encoder to achieve more precise boundary segmentation.

However, these methods typically employ simple addition or channel concatenation for feature fusion[15], which can lead to issues of feature redundancy and information interference. To address this[16], MSNet innovatively utilizes feature subtraction operations to highlight differences between adjacent scale features[17], thereby reducing redundant information. UACANet, on the other hand, divides initial segmentation regions into foreground, background[18], and uncertain areas, and processes them separately using a contextual attention mechanism, significantly improving the model's segmentation performance at low-contrast boundary locations[19].

Simultaneously, the rise of Vision Transformer (ViT) has introduced new research paradigms for polyp segmentation tasks. Transformers[20], with their powerful global context modeling capabilities, can effectively capture long-range dependencies within images. Researchers have attempted to combine the local feature extraction capabilities of CNNs with the global relationship modeling advantages of Transformers, primarily forming three integration approaches: The first is a serial structure[21], such as TransUnet, which

first uses CNNs to extract local features and then employs Transformers to establish global contextual relationships. The second is a parallel structure[22], like Transfuse, which integrates features extracted by CNNs and Transformers through carefully designed fusion modules[23]. The third approach uses pure Transformer encoders, adopting hierarchical Vision Transformers such as PVT and Swin Transformer as the backbone network to directly generate multi-scale feature sequences[24]. Methods like Polyp-PVT, SegFormer, and ColonFormer have demonstrated outstanding performance in polyp segmentation tasks.

2. DeepLab-RCP Model Design

2.1. Overall Architecture Design

The DeepLab-RCP model is built upon the DeepLabv3+ framework, integrating multiple specialized modules to address key challenges in polyp segmentation: significant size

variation, complex shapes, and low boundary contrast. The model follows an encoder-decoder structure, utilizing ResNeXt-50 as its backbone network. This backbone employs a "split-transform-merge" strategy with grouped convolutions to extract rich, discriminative multi-scale features. Figure 1.

A Cross-layer Dense Fusion (CDF) module is incorporated after the backbone, using dense connections and channel attention to effectively aggregate and refine multi-level features. The core Parallel Swin-ASPP (PS-ASPP) module combines multi-scale atrous convolutions with a multi-window self-attention-based Transformer branch in parallel. This dual-path design enables the simultaneous capture of local details and global contextual information. The decoder is constructed using depthwise separable convolutions (DSConv), creating a lightweight yet powerful network for feature fusion that maintains performance while reducing computational complexity.

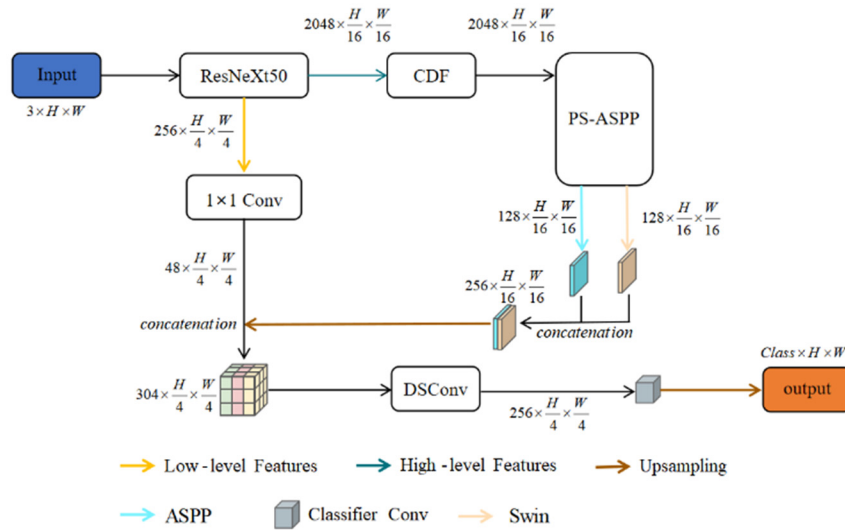


Figure 1. DeepLab-RCP

During processing, the input image passes through the ResNeXt-50 backbone to generate multi-scale features. Shallow features preserve spatial details, while deep features are enhanced by the CDF module. The refined high-level features then enter the PS-ASPP module, where the two parallel branches extract complementary representations. Their outputs are fused into a context-rich feature map. In the decoder, this feature map is upsampled and concatenated with shallow features from the encoder. The DSConv-based fusion network progressively restores spatial details and refines boundaries, culminating in a final convolutional layer that produces the pixel-level segmentation output. This integrated design allows the model to effectively leverage features at

different levels, achieving accurate polyp segmentation with maintained computational efficiency.

2.2. Cross-Dense Fusion

The Cross-layer Dense Fusion (CDF) module is a core innovative component in the DeepLab-RCP model, designed to address the challenge of multi-scale feature fusion caused by significant polyp size variations. Inspired by the dense connections in DenseNet, the CDF module employs a densely connected network structure to achieve adaptive aggregation and refinement of features from different layers, significantly enhancing the model's ability to represent polyps of varying sizes. Figure 2.

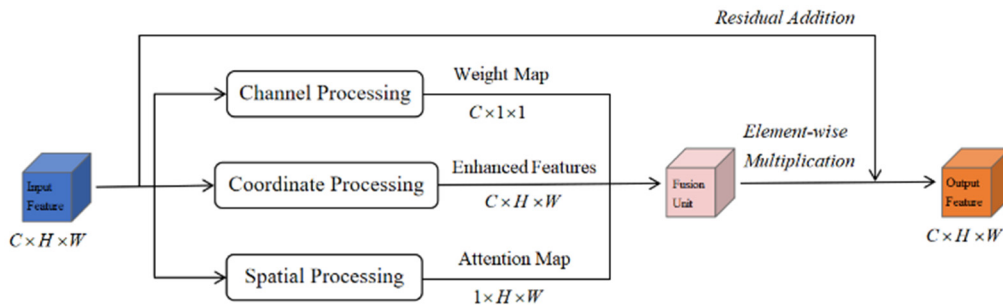


Figure 2. CDF

The module follows a three-stage "Growth-Aggregation-Calibration" process. First, multiple dense layers progressively expand the feature dimensions. Then, the outputs from all layers are concatenated and fused. Finally, a channel attention mechanism adaptively recalibrates the fused features, producing a feature representation that is both rich in detail and highly discriminative.

Mathematically, let the input feature map be $x \in \mathbb{R}^{C \times H \times W}$. The forward propagation of the CDF module can be described as follows:

For the l -th dense layer $l=1,2,\dots,L$, the input is the concatenation of the outputs from all previous layers and the original input:

$$\text{input}_l = [x, F_1, F_2, \dots, F_{l-1}] \quad (1)$$

Where $[\]$ denotes concatenation along the channel dimension, and F_i represents the output of the i -th layer. The transformation of each dense layer is defined as:

$$F_l = H_l(\text{input}_l) = \text{ReLU}(\text{BN}(\text{Conv}_{3 \times 3}(\text{input}_l))) \quad (2)$$

After L dense layers, all features are concatenated:

$$F_{\text{fused}} = [x, F_1, F_2, \dots, F_L] \quad (3)$$

Subsequently, a channel attention mechanism performs a squeeze-and-excitation operation to calibrate the fused features. Global average pooling is first applied to obtain channel statistics:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_{\text{fused}}^{(c)}(i, j) \quad (4)$$

The channel weights are then learned through two fully connected layers:

$$s = \sigma(W_2 \delta(W_1 z)) \quad (5)$$

Where δ is the ReLU activation function, σ is the Sigmoid function, W_1 and W_2 are learnable parameters, and r is the reduction ratio. The refined features are obtained as:

$$F_{\text{refined}} = s \cdot F_{\text{fused}} \quad (6)$$

Finally, an adaptive convolution adjusts the channel dimension back to the original size:

$$F_{\text{out}} = \text{Conv}_{1 \times 1}(F_{\text{refined}}) \quad (7)$$

This design enables the network to leverage both low-level details and high-level semantic features. The dense connections facilitate feature reuse, while the channel attention mechanism adaptively emphasizes important feature channels, effectively enhancing the model's capability to represent polyps of diverse sizes.

2.3. Parallel Swin Transformer and Atrous Spatial Pyramid Pooling

The PS-ASPP module is a key component of the DeepLab-RCP model, designed to synergistically combine the efficient local feature extraction capability of convolutional neural networks with the global context modeling strength of Transformers. This enables effective handling of the challenge posed by the complex and variable shapes of polyps. As illustrated in Figure 3, for an input feature map X the forward propagation of the PS-ASPP module proceeds through two parallel branches: the ASPP branch and the Swin Attention branch. Figure 3.

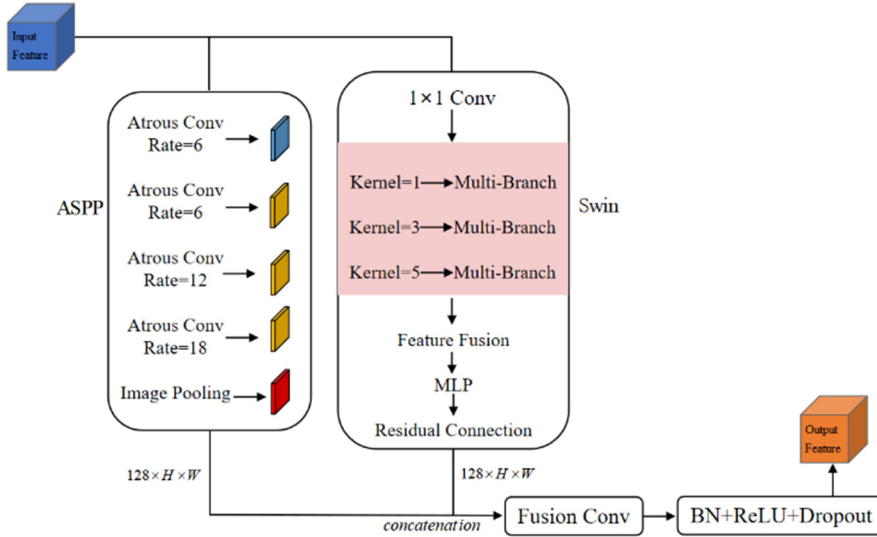


Figure 3. PS-ASPP

The ASPP branch adopts the standard Atrous Spatial Pyramid Pooling structure, consisting of five parallel paths: a 1×1 convolution branch, three 3×3 dilated convolutions with and an image-level feature branch:

$$Y_1 = \text{Conv}_{1 \times 1}(X) \quad (8)$$

$$Y_2 = \text{Conv}_{3 \times 3}^{\text{dilation}=r}(X) \quad (9)$$

$$Y_3 = \text{Conv}_{3 \times 3}^{\text{dilation}=2r}(X) \quad (10)$$

$$Y_4 = \text{Conv}_{3 \times 3}^{\text{dilation}=3r}(X) \quad (11)$$

$$Y_5 = \text{Upsample}(\text{Conv}_{1 \times 1}(\text{GAP}(X))) \quad (12)$$

where GAP denotes global average pooling and Upsample

restores the feature map to its original spatial dimensions. The final output of the ASPP branch is given by:

$$F_{\text{aspp}} = \text{Conv}_{1 \times 1}(\text{Concat}(Y_1, Y_2, Y_3, Y_4, Y_5)) \quad (13)$$

Subsequently, this output is fed into the multi-window attention module. For each attention head, the input feature is $Z \in \mathbb{R}^{C_k \times H \times W}$. The QKV triplet is generated via convolutional operations:

$$Q, K, V = \text{Split}(\text{Conv}_{1 \times 1}(Z + P)) \quad (14)$$

where P denotes the learnable positional encoding. The scaled dot-product attention is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q \otimes K^*}{\sqrt{d_k}}\right) \otimes V \quad (15)$$

where $\otimes$ represents matrix multiplication and d_k is the dimensionality of the Key. Multi-window attention is implemented via parallel paths with different convolutional kernel sizes. For the i -th pathway, the computation can be expressed as $A_i = \text{DWConv}_{k_i \times k_i}(X)$ where k represents the kernel size and DWConv indicates depthwise separable convolution. The outputs from all pathways are then concatenated and processed through a fusion layer, followed by feature transformation via an MLP to generate the final Swin branch output.

In the feature fusion stage, the outputs from both the ASPP and Swin branches are concatenated and refined through a convolutional layer:

$$F_{\text{out}} = \text{Conv}(F_{\text{concat}}) \quad (16)$$

3. Experiments and Analysis

3.1. Experimental Environment

The experiments in this study were conducted using the PyTorch 1.10.0 deep learning framework within a Python 3.8 environment, running on the Ubuntu 20.04 operating system. The computational platform utilized an NVIDIA RTX 3090 GPU (24GB VRAM) coupled with the CUDA 11.3 parallel computing architecture, providing powerful computational support for model training. The CPU was configured as an Intel Xeon Gold 6330 processor (2.00GHz, 14 vCPU) to ensure stable execution of data preprocessing tasks.

Regarding training parameter settings, the Adam optimizer was employed for model optimization with an initial learning rate set to 0.0001. All input images were uniformly resized to a resolution of 512×512 pixels, with a batch size set to 8, and the models were trained for 100 epochs. This configuration ensured training stability while fully leveraging GPU memory resources, guaranteeing both the efficiency of the experimental process and the reproducibility of the results.

3.2. Datasets

This study employs two public datasets, Kvasir-SEG and CVC-ClinicDB, as the primary sources for training and validation data. To ensure experimental standardization and result comparability, both datasets are uniformly and randomly divided into training, validation, and test sets following an 8:1:1 ratio. Specifically, the training set is used for model parameter learning and optimization, the validation set facilitates hyperparameter tuning and model selection during the training process, and the test set is reserved for the final evaluation of the model's generalization performance.

3.3. Evaluation Metrics

The evaluation metrics for colorectal polyp segmentation employ the mean Dice coefficient (mDice) and mean intersection over union (mIoU). These metrics are calculated based on four-pixel classification categories: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The Dice coefficient assesses the overlap between predicted regions and ground truth annotations, while the intersection over union quantifies boundary consistency through the ratio of intersection to union. As the Dice coefficient approaches 1, it indicates a high degree of overlap between the segmented area and the ground truth,

demonstrating robust performance particularly along complex edges. An increase in the IoU value reflects improved alignment accuracy between predicted and actual boundaries, reducing segmentation ambiguity in blurred regions.

3.4. Comparative Experimental Results

To comprehensively evaluate the advancement of the DeepLab-RCP model, this section compares DeepLab-RCP with six mainstream segmentation methods on two public datasets, Kvasir-SEG and CVC-ClinicDB. The models involved in the comparison include UNet, UNet++, PSPNet, PraNet, CaraNet, and DeepLabV3+ (with Xception as its backbone network). All models were trained and tested under the same experimental environment, using identical training strategies and dataset partitioning to ensure fairness in comparison. The evaluation metrics employed were the mean Dice coefficient (mDice) and mean intersection over union (mIoU). The detailed comparison results are presented in Table 1.

Table 1. Performance Comparison with Mainstream Models

Model	Kvasir-SEG		CVC-ClinicDB	
	mDice	mIoU	mDice	mIoU
UNet	70.8	71.9	60.4	61.8
UNet++	86.3	79.0	80.6	77.0
PSPNet	83.1	77.7	89.4	87.4
ParNet	81.3	75.8	78.8	81.1
CaraNet	69.8	61.0	50.0	46.8
DeepLabV3+	45.3	41.4	47.7	45.0
DeepLab-RCP	92.8	86.9	95.5	91.6

From the experimental results, it can be observed that the proposed DeepLab-RCP model achieved the best performance on both datasets. On the Kvasir-SEG dataset, DeepLab-RCP reached mDice and mIoU scores of 93.0% and 87.3%, respectively, significantly outperforming other compared methods. Compared to the second-best performing model, UNet++, these values represent improvements of 6.7 and 8.3 percentage points, respectively. On the CVC-ClinicDB dataset, the advantage of DeepLab-RCP was even more pronounced, with mDice and mIoU scores reaching 95.5% and 91.6%, respectively. These represent improvements of 6.1 and 4.2 percentage points compared to PSPNet.

It is noting that the baseline model, DeepLabV3+, performed suboptimally on both datasets. The analysis suggests that the Xception backbone network it employs may have limitations in extracting polyp features, and the standard ASPP module may be inadequate for handling the complex morphology and low-contrast boundaries in polyp images. In contrast, DeepLab-RCP, by incorporating the ResNeXt50 backbone, the Cross-layer Dense Fusion (CDF) module, and the Parallel Swin-ASPP (PS-ASPP) module, significantly enhances the model's ability to represent multi-scale polyp features and improves segmentation accuracy for complex boundaries. These improvements enable DeepLab-RCP to maintain high segmentation consistency while better adapting to the distribution characteristics of different datasets, thereby demonstrating excellent generalization performance and clinical practical value.

3.5. Ablation Study

To validate the effectiveness of each core module in the DeepLab-RCP model, a systematic ablation study was conducted on the CVC-ClinicDB dataset. The experiments used DeepLabv3+ with ResNeXt50 as the backbone as the baseline model. The Cross-layer Dense Fusion (CDF) module, the Parallel Swin-ASPP (PS-ASPP) module, and Depthwise Separable Convolutions (DSConv) were incrementally added to evaluate the impact of different module combinations on polyp segmentation performance. The evaluation metrics employed were the mean Dice coefficient (mDice) and mean intersection over union (mIoU). The experimental results are presented in Table 2.

Table 2. Quantitative results of eight ablation experiments on the CVC-ClinicDB dataset

Method	CVC-ClinicDB	
	mDice	mIoU
Baseline	93.9	88.9
Baseline+CDF	94.4	89.7
Baseline+PS-ASPP	94.5	90.0
Baseline+DSConv	94.3	89.6
Baseline+CDF+PS-ASPP	94.6	90.1
Baseline+CDF+DSConv	94.9	90.5
Baseline+PS-ASPP+DSConv	94.9	90.6
Baseline+CDF+PS-ASPP+DSConv	95.5	91.6

The results indicate that adding any single module improves the performance of the baseline model. When only the CDF module was added, mDice and mIoU reached 94.4% and 89.7%, respectively, representing improvements of 0.5 and 0.8 percentage points over the baseline. This demonstrates that the CDF module effectively enhances multi-scale feature fusion capability through its dense connections and channel attention mechanism. Adding only the PS-ASPP module increased mDice and mIoU to 94.5% and 90.0%, respectively, proving that the parallel Swin-ASPP structure can synergistically utilize the multi-scale context capture ability of atrous convolutions and the global modeling advantage of self-attention. Using the DSConv module alone also provided stable improvements (mDice 94.3%, mIoU 89.6%), indicating that depthwise separable convolution optimizes model efficiency by reducing the number of parameters while maintaining feature extraction capability.

Furthermore, experiments combining modules demonstrated the complementary nature of the components. When both the CDF and PS-ASPP modules were added, performance improved significantly compared to the baseline (mDice 94.6%, mIoU 90.1%). Combinations of CDF with DSConv (mDice 94.9%, mIoU 90.5%) and PS-ASPP with DSConv (mDice 94.9%, mIoU 90.6%) achieved even better results, indicating that combining feature enhancement with lightweight design can further optimize segmentation accuracy. Finally, when all modules were integrated, the model achieved the best performance (mDice 95.5%, mIoU 91.6%), representing improvements of 1.6 and 2.7 percentage points over the baseline, respectively. This fully validates the necessity of the combined action of the CDF, PS-ASPP, and DSConv modules for improving polyp segmentation effectiveness. Their synergistic mechanism effectively addresses multiple challenges, including size variation,

complex shapes, and low contrast.

3.6. Comparison of Segmentation Results

To visually demonstrate the segmentation performance of the DeepLab-RCP model, this section selects several representative polyp cases for a comparative visual analysis, comparing the segmentation results of our method with other mainstream segmentation models. Figure 1 displays the segmentation results of various models on typical polyp images, including the original image, ground truth, and the segmentation results from UNet, UNet++, PSPNet, PraNet, CaraNet, DeepLabV3+, and DeepLab-RCP. As shown in the first and second columns of Figure 4.

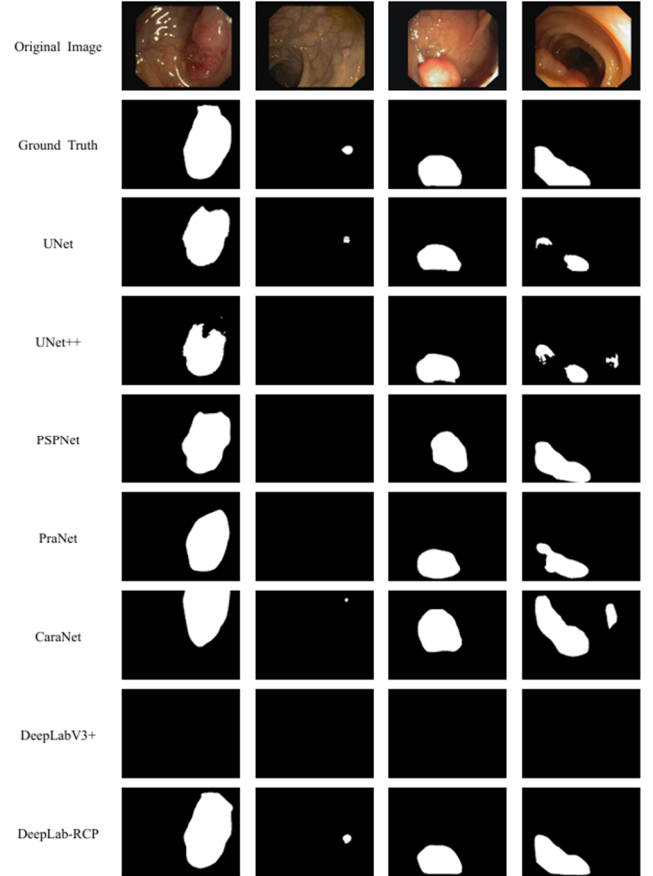


Figure 4. Comparison of Segmentation

From the visual results, it is evident that when handling polyps with significant size variations, DeepLab-RCP, leveraging the feature fusion mechanism of the CDF module, enables the model to capture both the detailed features of small polyps and the global structure of large polyps. For large or prominently featured polyps, the predicted regions by DeepLab-RCP exhibit complete contours without significant "gaps" or hollows. For tiny polyps in the images, models like UNet++ and PSPNet failed to detect them, whereas DeepLab-RCP, utilizing the multi-scale feature calibration capability of the CDF module, achieved precise segmentation of these small polyps.

Secondly, the PS-ASPP module plays a key role in segmenting polyps with complex and variable shapes. As shown in the fourth column of Figure 4, when faced with irregularly shaped polyps, traditional methods like DeepLabV3+ produced segmentation results with smoothed boundaries and loss of detail. In contrast, DeepLab-RCP, through the multi-window self-attention mechanism in the parallel Swin-ASPP, effectively captures the complex spatial

dependencies of the polyp, thereby generating a segmentation contour that more closely aligns with the true boundary.

Finally, in handling low-contrast boundaries, reflective highlights, and blurred edges, as shown in the second and third columns of Figure 4, when the contrast between the polyp and the surrounding mucosal tissue is low, models like UNet++ and PraNet produced results with blurred boundaries, as well as false positives and missed detections. DeepLab-RCP, benefiting from the enhanced feature extraction capability of ResNeXt-50 and the boundary optimization effect of depthwise separable convolutions, produces clear and complete segmentation edges.

4. Conclusion

This paper addresses the challenges in colorectal polyp segmentation, such as significant size variation, complex morphological diversity, and blurred boundaries, by proposing an improved model named DeepLab-RCP based on the DeepLabv3+ framework. The model enhances performance through three core modules: first, the Cross-layer Dense Fusion module leverages dense connections and a channel attention mechanism to strengthen multi-scale feature fusion capability; second, the Parallel Swin-ASPP module integrates the multi-scale feature extraction of atrous convolutions with the global context modeling capacity of Swin Transformer; finally, the ResNeXt-50 backbone network and depthwise separable convolutions are adopted to maintain feature extraction power while optimizing model efficiency.

Experimental results on the Kvasir-SEG and CVC-ClinicDB datasets demonstrate that DeepLab-RCP achieves optimal performance across multiple evaluation metrics. Ablation studies further validate the effectiveness and synergistic effects of each module. Visual comparisons reveal the model's significant advantages in handling small polyps, complex shapes, and low-contrast boundaries. This research provides an effective solution for precise colorectal polyp segmentation and holds substantial application value for clinical auxiliary diagnosis.

References

- [1] Y. S. B. A. R. and M. A., "Deep learning applications in medical image segmentation: overview, approaches, and challenges," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. Hoboken, NJ: John Wiley & Sons, 2025, pp. 15–64.
- [2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 3431–3440.
- [3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in **Proc. 18th Int. Conf. Medical Image Computing and Computer-Assisted Intervention**, Munich, Germany, 2015, pp. 234–241.
- [4] K. Wang, Y. T. Lu, J. Liu, et al., "GAML: Geometry-aware mutual learning for polyp segmentation," *Biomedical Signal Processing and Control*, vol. 108, p. 107965, 2025.
- [5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
- [6] T. K. Dutta, et al., "SAM-Mamba: Mamba guided SAM architecture for generalized zero-shot polyp segmentation," in *Proc. IEEE Winter Conf. Applications of Computer Vision*, Waikoloa, HI, 2025, pp. 4655–4664.
- [7] H. Zhao, J. Shi, X. Qi, et al., "Pyramid scene parsing network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 2881–2890.
- [8] L. C. Chen, Y. Zhu, G. Papandreou, et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. European Conf. Computer Vision*, Munich, Germany, 2018, pp. 801–818.
- [9] M. Yang, K. Yu, C. Zhang, et al., "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 3684–3692.
- [10] C. H. Huang, H. Y. Wu, and Y. L. Lin, "HarDNet-MSEG: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
- [11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 7132–7141.
- [12] S. Woo, J. Park, J. Y. Lee, et al., "CBAM: Convolutional block attention module," in *Proc. European Conf. Computer Vision*, Munich, Germany, 2018, pp. 3–19.
- [13] A. Srivastava, D. Jha, S. Chanda, et al., "MSRF-Net: A multi-scale residual fusion network for biomedical image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 5, pp. 2252–2263, May 2021.
- [14] D. Jha, P. H. Smedsrud, M. A. Riegler, et al., "Kvasir-SEG: A segmented polyp dataset," in *Proc. 26th Int. Conf. MultiMedia Modeling*, Daejeon, South Korea, 2020, pp. 451–462.
- [15] D. Vázquez, J. Bernal, F. J. Sánchez, et al., "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthcare Eng.*, vol. 2017, p. 4037190, 2017.
- [16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, et al., "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Granada, Spain, 2018, pp. 3–11.
- [17] H. Zhao, J. Shi, X. Qi, et al., "Pyramid scene parsing network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 2881–2890.
- [18] T. K. Dutta, et al., "SAM-Mamba: Mamba guided SAM architecture for generalized zero-shot polyp segmentation," in *Proc. IEEE Winter Conf. Applications of Computer Vision*, Waikoloa, HI, 2025, pp. 4655–4664.
- [19] L. C. Chen, Y. Zhu, G. Papandreou, et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. European Conf. Computer Vision*, Munich, Germany, 2018, pp. 801–818.
- [20] D. Fan, Q. Yang, X. Zhang, et al., "PraNet: Parallel reverse attention network for polyp segmentation," in **Proc. 23rd Int. Conf. Medical Image Computing and Computer-Assisted Intervention**, Lima, Peru, 2020, pp. 555–563.
- [21] J. Liu, K. Wang, Y. T. Lu, et al., "CaraNet: Contextual axial reverse attention network for small medical object segmentation," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 8, pp. 3608–3617, Aug. 2023.
- [22] D. Jha, P. H. Smedsrud, M. A. Riegler, et al., "Kvasir-SEG: A segmented polyp dataset," in *Proc. 26th Int. Conf. MultiMedia Modeling*, Daejeon, South Korea, 2020, pp. 451–462.
- [23] D. Vázquez, J. Bernal, F. J. Sánchez, et al., "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthcare Eng.*, vol. 2017, p. 4037190, 2017.
- [24] J. Moorthy and U. D. Gandhi, "A survey on medical image segmentation based on deep learning techniques," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 117, 2022.