Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A MobileNetV2-Based Boundary-Aware Improved DeepLabv3+ Network for Gastrointestinal Endoscopic Lesion Segmentation

Yan Hua
Wuxi Ninth People's Hospital Affiliated to Soochow University
Ziyu Zhang
Kangda College of Nanjing Medical University
Jiawei Cao
Wuxi Ninth People's Hospital Affiliated to Soochow University
Leyao Zhang
zhangleyao@suda.edu.cn

Wuxi Ninth People's Hospital Affiliated to Soochow University

Research Article

Keywords: gastrointestinal endoscopy, lesion segmentation, DeepLabv3+, MobileNetV2, cascaded ASPP, boundary attention

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11148776/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objective: Gastrointestinal endoscopic lesion segmentation is challenging because lesions often exhibit substantial variations in morphology and size, blurred boundaries, small foreground regions, and complex mucosal background interference. This study proposed an improved DeepLabv3+ segmentation model that integrates a lightweight backbone, cascaded multi-scale contextual modeling, and boundary-aware auxiliary supervision to improve automatic lesion segmentation in endoscopic images.

Methods: DeepLabv3+ was used as the baseline framework. The original feature extraction backbone was replaced with MobileNetV2 to reduce model complexity while extracting both low-level spatial details and high-level semantic features. The ASPP module was redesigned as a cascaded multi-scale atrous convolution module to enable feature reuse among branches with different dilation rates. In the low-level feature pathway, a boundary attention module was introduced to enhance contour-related responses according to a predicted boundary probability map. During training, boundary targets were automatically generated from ground-truth masks, and the model was optimized using a combination of cross-entropy loss, Dice loss, and boundary BCE loss. The dataset included 1,445 gastrointestinal endoscopic images with pixel-level labels, of which 1,300 were used for training and 145 for testing. F1-score, sensitivity, specificity, accuracy, and IoU were used for evaluation.

Results: The complete model achieved an F1-score of 0.878, sensitivity of 0.860, specificity of 0.995, accuracy of 0.990, and IoU of 0.784 on the test set. Compared with the original DeepLabv3+ baseline, IoU increased from 0.755 to 0.784. Compared with SegNet, the best-performing competing model, the proposed model improved IoU by 0.040. Ablation experiments demonstrated that both the cascaded ASPP module and the boundary attention branch contributed to performance improvement.

Conclusion: The proposed improved DeepLabv3+ model effectively integrates multi-scale semantic information and low-level boundary details, achieving favorable segmentation performance for gastrointestinal endoscopic lesions. This method has potential value for lesion localization, contour delineation, and quantitative analysis in computer-assisted endoscopy.

1. Introduction

1.1 Background

Gastrointestinal tumors and their precancerous lesions represent important health burdens. Gastrointestinal endoscopy allows direct visualization of mucosal morphology and is an essential tool for detecting polyps, adenomas, sessile serrated lesions, and other mucosal abnormalities. Accurate recognition, localization, and boundary delineation of endoscopic lesions affect diagnostic decisions, biopsy sampling, assessment of endoscopic resection margins, and follow-up strategies.

However, gastrointestinal endoscopic images are complex and variable. Image quality is frequently affected by uneven illumination, specular reflection, mucus, motion blur, mucosal folds, and residual fluids. Lesions vary considerably in color, shape, size, elevation, and boundary clarity. Some flat or small lesions exhibit low contrast against surrounding normal mucosa, making them difficult to detect and delineate. Moreover, lesion contours are often irregular, and manual annotation may be subjective and inconsistent among annotators.

Image segmentation is a fundamental task in medical image analysis. Unlike classification, which only determines whether a lesion exists, segmentation provides pixel-level lesion location, size, and boundary information. Such information is more suitable for clinical localization and quantitative analysis. With the development of deep learning, convolutional neural networks and their variants have achieved remarkable progress in medical image segmentation.

1.2 Significance

This study has both clinical and technical significance. Clinically, an automatic segmentation model can help physicians rapidly localize lesion regions, reduce the burden of manual delineation, and improve the objectivity of lesion measurement and documentation. For lesions with unclear or irregular boundaries, segmentation models can serve as auxiliary tools to highlight suspicious regions. Technically, endoscopic lesion segmentation requires strong multi-scale perception, boundary detail recovery, and lightweight deployment capability, making it a meaningful scenario for evaluating medical image segmentation methods.

Although existing deep learning segmentation models have achieved promising results, they still face limitations, including insufficient multi-scale feature fusion, inaccurate boundary recovery, and high model complexity. To address these issues, this study used DeepLabv3 + as the basic framework and introduced a MobileNetV2 backbone, a cascaded ASPP module, and boundary attention-assisted supervision to improve the segmentation of lesion bodies and boundaries.

1.3 Related Work

Early medical image segmentation methods mainly relied on thresholding, region growing, edge detection, active contour models, and handcrafted features. These methods can be effective when image quality is stable and target boundaries are clear, but their robustness is limited for endoscopic images with complex backgrounds, illumination changes, and heterogeneous lesion morphology.

FCN first transformed classification networks into end-to-end pixel-wise prediction networks and laid the foundation for deep semantic segmentation [1]. U-Net introduced an encoder-decoder architecture with skip connections to fuse low-level details and high-level semantics, becoming a widely used framework in medical image segmentation [2]. SegNet used pooling indices for upsampling [3]. UNet + + reduced the semantic gap between encoder and decoder through nested skip connections [4]. Attention U-Net introduced attention gates to emphasize target-related regions [5]. MultiResUNet and UNet 3 + further enhanced U-Net through multi-resolution blocks and full-scale skip connections, respectively [22, 23].

The DeepLab family expands the receptive field through atrous convolution and uses ASPP to capture multi-scale contextual information [15, 16]. DeepLabv3 + further combines ASPP with an encoder-decoder structure and low-level feature fusion to improve boundary recovery [6]. Recently, Transformer-based models have also been introduced into segmentation tasks. For example, SegFormer uses a hierarchical Transformer encoder and a lightweight decoder [7, 14]. However, Transformer models usually depend on relatively large-scale training data, and their generalization on small medical datasets should be carefully evaluated. Feature pyramid networks and pyramid scene parsing networks provide additional approaches to multi-scale feature representation and context aggregation [18, 24].

In gastrointestinal endoscopic lesion segmentation, public datasets such as Kvasir-SEG have promoted standardized research [8]. PraNet uses a parallel partial decoder and reverse attention to mine polyp boundary regions [9]. HarDNet-MSEG, Polyp-PVT, and CaraNet further advanced this field from the perspectives of lightweight design, multi-scale fusion, Transformer modeling, and axial reverse attention [19–21]. Overall, balancing segmentation accuracy, boundary quality, and deployment efficiency under limited data remains a major challenge.

1.4 Contributions

This study proposes an improved DeepLabv3 + model based on MobileNetV2 and boundary-aware learning for gastrointestinal endoscopic lesion segmentation. The model uses an encoder-decoder structure. MobileNetV2 is used as the lightweight backbone to extract low-level detail features and high-level semantic features. A cascaded ASPP module is designed to enable feature reuse among atrous convolution branches with different dilation rates. A boundary attention branch is introduced into the low-level feature pathway to enhance contour-related responses, and boundary auxiliary supervision is incorporated during training.

The main contributions are as follows: (1) MobileNetV2 is adopted as a lightweight backbone for DeepLabv3+; (2) the ASPP module is redesigned into a cascaded feature-reuse structure; (3) a boundary attention branch is constructed to emphasize lesion contours; and (4) boundary auxiliary loss is introduced to strengthen boundary learning during training.

2. Materials and Methods

2.1 Dataset and Task Definition

This study used gastrointestinal endoscopic images and corresponding pixel-level segmentation masks. The dataset contained 1,445 images, including 1,300 training images and 145 testing images. The task was formulated as binary semantic segmentation. For an input image $I \in R(H \times W \times 3)$, the model outputs a prediction map $P \in R(H \times W \times C)$, where $C = 2$ corresponds to background and lesion classes. The dataset split is shown in Table 1.

*[*Table 1 *about here]*

Table 1
Dataset split.

Subset	Number of images	Purpose	Description
Training set	1300	Model learning	Used for backpropagation and data augmentation
Test set	145	Model evaluation	Used to evaluate segmentation performance
Total	1445	All images and masks	Binary lesion segmentation task

2.2 Image Preprocessing and Data Augmentation

All images were resized to 512×512 pixels during training and testing. During testing, aspect-ratio-preserving resizing and padding were used to reduce geometric distortion. Segmentation masks were resized using nearest-neighbor interpolation to avoid generating non-integer class labels. During training, random scaling, aspect-ratio jittering, horizontal flipping, Gaussian blurring, random rotation, and HSV color perturbation were applied to improve robustness to variations in imaging distance, viewing angle, illumination, and mild blur.

2.3 Overall Network Architecture

The proposed model is based on DeepLabv3 + and follows an encoder-decoder framework. The encoder extracts high-level semantic features and multi-scale contextual information, while the decoder fuses low-level spatial details and restores the spatial resolution of the prediction map. Compared with the original DeepLabv3+, the proposed model replaces the backbone with MobileNetV2, modifies the standard ASPP into a cascaded ASPP module, and introduces a boundary attention branch with auxiliary boundary supervision. The overall architecture is shown in Fig. 1.

*[*Figure 1 *about here]*

2.4 MobileNetV2 Backbone

MobileNetV2 is a lightweight convolutional neural network designed for mobile and efficient vision applications. Its core components include depthwise separable convolution, inverted residual blocks, and linear bottlenecks [10, 17]. Standard convolution performs spatial filtering and channel mixing simultaneously, whereas depthwise separable convolution decomposes them into depthwise convolution and pointwise convolution, thereby reducing computational cost.

$$Cost_{standard} = Dk \times Dk \times Min \times Mout \times H \times W \quad (2-1)$$

$$Cost_{depthwise} = Dk \times Dk \times Min \times H \times W + Min \times Mout \times H \times W \quad (2-2)$$

Here, Dk denotes the kernel size, Min and Mout denote the input and output channels, and H and W denote the spatial dimensions of the feature map.

In the proposed model, MobileNetV2 extracts both low-level and high-level features. Low-level features retain edge, texture, and color details, whereas high-level features provide semantic cues for lesion recognition.

2.5 Cascaded ASPP Module

ASPP is a core component of DeepLabv3 + and captures contextual information at multiple scales using atrous convolutions with different dilation rates [6, 13]. Atrous convolution expands the receptive field without substantially increasing the number of parameters.

$y(i) = \Sigma_k x(i + r \cdot k)\, w(k)$ (2–3)

In this equation, x is the input feature, w is the convolution kernel, and r is the dilation rate.

The improved ASPP module contains five branches: a 1×1 convolution branch, 3×3 atrous convolution branches with dilation rates of 6, 12, and 18, and an image-pooling branch. Unlike the standard parallel ASPP, the rate-12 branch receives the concatenation of the original high-level feature and the output of the rate-6 branch, and the rate-18 branch further fuses the original feature, the rate-6 output, and the rate-12 output. The structure of the cascaded ASPP module is shown in Fig. 2.

$F6 = Conv3\times3, r = 6(F)$ (2–4)

$F12 = Conv3\times3, r = 12(Concat(F, F6))$ (2–5)

$F18 = Conv3\times3, r = 18(Concat(F, F6, F12))$ (2–6)

$FASPP = Conv1\times1(Concat(F1\times1, F6, F12, F18, Fgap))$ (2–7)

*[*Figure 2 *about here]*

2.6 Boundary Attention Branch

Boundary regions are among the most challenging areas in gastrointestinal endoscopic lesion segmentation. Lesion boundaries may have similar colors to surrounding mucosa and can be affected by reflections, folds, and viewing angles. To enhance contour awareness, a boundary attention branch was introduced into the low-level feature pathway. The boundary attention branch and auxiliary supervision are illustrated in Fig. 3.

$B = Sigmoid(Conv1\times1(\delta(BN(Conv3\times3(Flow)))))$ (2–8)

$Fboundary = Flow \times (1 + B)$ (2–9)

Here, Flow denotes the low-level feature, B denotes the boundary probability map, and Fboundary denotes the boundary-enhanced low-level feature.

*[*Figure 3 *about here]*

2.7 Decoder and Feature Fusion

The decoder fuses high-level semantic features with low-level boundary-enhanced details and restores spatial resolution. The cascaded ASPP output is upsampled by bilinear interpolation to match the spatial resolution of the low-level feature. The boundary-enhanced low-level feature is compressed by a 1×1 convolution. The two features are concatenated along the channel dimension and further fused by 3×3 convolutions to generate the final binary segmentation map.

2.8 Loss Function

The training objective consists of the main segmentation loss and boundary auxiliary loss. The main segmentation loss includes cross-entropy loss and Dice loss [12]. Cross-entropy loss supervises pixel-wise classification, whereas Dice loss alleviates class imbalance caused by small foreground lesion regions.

$$L_{CE} = -\frac{1}{N}\sum_i \sum_c y_{i,c} \log(p_{i,c}) \quad (2–10)$$

$$L_{Dice} = 1 - \frac{2|P \cap G| + \varepsilon}{|P| + |G| + \varepsilon} \quad (2–11)$$

The boundary auxiliary loss supervises the boundary attention branch. Boundary targets are generated automatically from the ground-truth mask by identifying adjacent pixels with different labels and then expanding boundary regions using max pooling.

$$L_{boundary} = -\frac{1}{N}\sum_i [b_i \log(q_i) + (1-b_i) \log(1-q_i)] \quad (2–12)$$

$$L_{total} = L_{CE} + L_{Dice} + \lambda L_{boundary}, \lambda = 0.2 \quad (2–13)$$

2.9 Evaluation Metrics and Experimental Settings

F1-score, sensitivity, specificity, accuracy, and IoU were used to evaluate segmentation performance. Let TP, TN, FP, and FN denote true-positive, true-negative, false-positive, and false-negative pixels, respectively. The main experimental settings are summarized in Table 2. The cosine-decay schedule is related to stochastic optimization schedules described in SGDR [25].

$$SEN = \frac{TP}{TP + FN} \quad (2–14)$$

$$SPE = \frac{TN}{TN + FP} \quad (2–15)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (2–16)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (2–17)$$

IoU = TP / (TP + FP + FN) (2–18)

*[*Table 2 *about here]*

Table 2
Main experimental settings.

Parameter	Setting	Description
Input size	512×512	Unified image input size
Number of classes	2	Background and lesion
Training epochs	100	Complete training schedule
Batch size	8	Training batch size
Optimizer	Adam	Parameter optimization
Initial learning rate	5×10^-4	Used with cosine decay
Main loss	CE + Dice	Pixel classification and region overlap constraint
Boundary loss weight	0.2	Auxiliary boundary supervision

3. Results and Analysis

3.1 Evaluation Protocol

The effectiveness of the proposed model was evaluated through quantitative comparisons, ablation experiments, backbone comparisons, and qualitative visualization. The reported results were obtained on the 145-image test set in the final Chinese manuscript.

3.2 Quantitative Comparison with Existing Models

Table 3 compares the proposed model with several representative segmentation networks. The proposed model achieved the best or near-best performance in F1-score, sensitivity, accuracy, and IoU, with an F1-score of 0.878 and an IoU of 0.784.

*[*Table 3 *about here]*

Table 3
Quantitative comparison of different segmentation models.

Model	F1	SEN	SPE	ACC	IoU
U-Net	0.842	0.824	0.994	0.987	0.728
SegNet	0.853	0.812	0.996	0.988	0.744
UNet++	0.846	0.793	0.996	0.988	0.733
Att-UNet	0.836	0.802	0.995	0.987	0.718
FCN	0.833	0.804	0.994	0.986	0.714
SegFormer	0.728	0.703	0.990	0.978	0.573
Proposed model	0.878	0.860	0.995	0.990	0.784

Compared with SegNet, the best-performing competing model, the proposed model improved IoU by 0.040. Compared with U-Net, IoU increased by 0.056. These results indicate that the proposed model better handles lesion scale variation and boundary ambiguity through cascaded ASPP and boundary-aware feature enhancement.

3.3 Ablation Study

Ablation experiments were conducted to evaluate the contributions of the cascaded ASPP module and the boundary attention branch. MobileNetV2 was used as the unified backbone. The ablation results are summarized in Table 4.

*[*Table 4 *about here]*

Table 4
Ablation study of the proposed modules.

Configuration	F1	SEN	SPE	ACC	IoU
Original DeepLabv3 + + MobileNetV2	0.871	0.848	0.995	0.989	0.772
Cascaded ASPP + MobileNetV2	0.877	0.857	0.995	0.990	0.781
Cascaded ASPP + Boundary Attention + MobileNetV2	0.878	0.860	0.995	0.990	0.784

ΔIoU_ASPP = 0.781 − 0.772 = 0.009 (3 − 1)

ΔIoU_Boundary = 0.784 − 0.781 = 0.003 (3 − 2)

The cascaded ASPP module was the main source of performance gain, indicating that feature reuse among atrous convolution branches strengthens multi-scale representation. The boundary attention branch further improved sensitivity and IoU, suggesting that explicit contour enhancement helps recover lesion boundaries.

3.4 Comparison of Different Backbones

The performance of different backbones under the complete improved architecture is shown in Table 5. MobileNetV2 achieved the best overall performance in this dataset. The compared ResNet backbones are based on residual learning [11].

*[*Table 5 *about here]*

Table 5
Results of the complete model with different backbones.

Backbone	**F1**	**SEN**	**SPE**	**ACC**	**IoU**
ResNet18	0.867	0.840	0.995	0.989	0.766
ResNet34	0.832	0.824	0.993	0.986	0.713
ResNet50	0.864	0.841	0.995	0.989	0.760
ResNet101	0.839	0.820	0.994	0.987	0.723
ResNet152	0.853	0.848	0.994	0.988	0.744
MobileNetV2	0.878	0.860	0.995	0.990	0.784

MobileNetV2 outperformed deeper ResNet backbones in F1-score and IoU. This suggests that, under the current dataset scale, a lightweight backbone may provide better training stability and generalization. Deeper networks may require larger datasets to avoid overfitting.

3.5 Qualitative Visualization

Figure 4 shows qualitative comparisons among different models. Green regions indicate ground-truth masks, and red regions indicate model predictions.

*[*Figure 4 *about here]*

The visualization results show that most models can roughly localize large lesion bodies, but their predictions differ in boundary, low-contrast, and narrow-connection regions. The proposed model produced predictions that were more consistent with the ground truth, indicating the benefit of multi-scale semantic fusion and boundary attention. Figure 5 presents the qualitative visualization of the ablation experiments.

*[*Figure 5 *about here]*

The ablation visualization further demonstrates that the cascaded ASPP module improves lesion body coverage, while boundary attention refines contour consistency. These qualitative observations are consistent with the quantitative improvements.

4. Discussion

The proposed MobileNetV2-cascaded ASPP-boundary attention model achieved favorable performance in gastrointestinal endoscopic lesion segmentation. The improvement can be attributed to three aspects: lightweight feature extraction, multi-scale contextual fusion, and boundary detail enhancement. MobileNetV2 provided stable feature extraction under the current dataset scale. Cascaded ASPP improved information interaction among different receptive fields. Boundary attention strengthened the learning of ambiguous contour regions.

Clinically, accurate lesion segmentation can assist physicians in localizing lesion boundaries and may support lesion area estimation, follow-up comparison, and planning of endoscopic therapy. Because the model uses a lightweight backbone, it has potential for future deployment in real-time computer-assisted endoscopy systems.

This study has several limitations. First, the dataset was derived from a single project source, and external multicenter validation was not performed. Second, model complexity, FLOPs, and inference speed were not systematically reported. Third, the current task was binary segmentation and did not distinguish pathological subtypes. Future studies should include larger multicenter datasets, external validation, complexity analysis, and video-level stability evaluation.

5. Conclusion

This study proposed an improved DeepLabv3 + network based on MobileNetV2 and boundary-aware learning for gastrointestinal endoscopic lesion segmentation. MobileNetV2 was used as a lightweight backbone, cascaded ASPP was introduced to enhance multi-scale contextual representation, and boundary attention with auxiliary boundary loss was used to strengthen lesion contour learning.

The proposed model achieved an F1-score of 0.878, sensitivity of 0.860, specificity of 0.995, accuracy of 0.990, and IoU of 0.784 on the test set, outperforming several representative segmentation models. Ablation experiments confirmed the contributions of cascaded ASPP and boundary attention. Overall, the proposed model has potential for computer-assisted lesion localization, contour delineation, and quantitative analysis in gastrointestinal endoscopy.

Declarations

Ethics approval: This study was approved by the Ethics Committee of Wuxi Ninth People's Hospital Affiliated to Soochow University (approval No. KS2026-108-01). All endoscopic images were anonymized before analysis. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Data availability: The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Competing interests: The authors declare that they have no competing interests. This statement should be confirmed by all authors before submission.

Funding: This work was supported by Wuxi Science and Technology Association's Soft Project (KX-25-B61).

Author contributions: Leyao Zhang and Yan Hua contributed to study design, data organization, and manuscript drafting. Ziyu Zhang contributed to case screening, clinical data verification, image annotation, and manuscript drafting. Jiawei Cao contributed to image annotation, model adjustment, and data analysis. Leyao Zhang contributed to study conception, quality control, and manuscript revision. All authors reviewed and approved the final manuscript.

References

1. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431–3440.
2. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Med Image Comput Computer-Assisted Intervention, 2015: 234–41.
3. Badrinarayanan V, Kendall A, Cipolla R, SegNet:. A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(12):2481–95.
4. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J, UNet++:. A nested U-Net architecture for medical image segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018: 3–11.
5. Oktay O, Schlemper J, Folgoc LL et al. Attention U-Net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999, 2018.
6. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. European Conference on Computer Vision, 2018: 801–818.
7. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst. 2021;34:12077–90.
8. Jha D, Smedsrud PH, Riegler MA et al. Kvasir-SEG: A segmented polyp dataset. International Conference on Multimedia Modeling, 2020: 451–462.
9. Fan DP, Ji GP, Zhou T et al. PraNet: Parallel reverse attention network for polyp segmentation. Med Image Comput Comput Assist Intervention, 2020: 263–73.
10. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510–4520.

11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.
12. Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. International Conference on 3D Vision, 2016: 565–571.
13. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. International Conference on Learning Representations, 2016.
14. Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. Adv Neural Inf Process Syst, 2017: 5998–6008.
15. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL, DeepLab. Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell. 2018;40(4):834–48.
16. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.
17. Howard AG, Zhu M, Chen B et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.
18. Lin TY, Dollár P, Girshick R et al. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117–2125.
19. Huang CH, Wu HY, Lin YL, HarDNet-MSEG. A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 FPS. arXiv preprint arXiv:2101.07172, 2021.
20. Dong B, Wang W, Fan DP et al. Polyp-PVT: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932, 2021.
21. Lou A, Guan S, Loew MH, CaraNet. Context axial reverse attention network for segmentation of small medical objects. Medical Imaging 2022: Image Processing, 2022.
22. Ibtehaz N, Rahman MS, MultiResUNet. Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw. 2020;121:74–87.
23. Huang H, Lin L, Tong R et al. UNet 3+: A full-scale connected UNet for medical image segmentation. IEEE International Conference on Acoustics, Speech and Signal Processing, 2020: 1055–1059.
24. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2881–2890.
25. Loshchilov I, Hutter F, SGDR. Stochastic gradient descent with warm restarts. International Conference on Learning Representations, 2017.

Figures

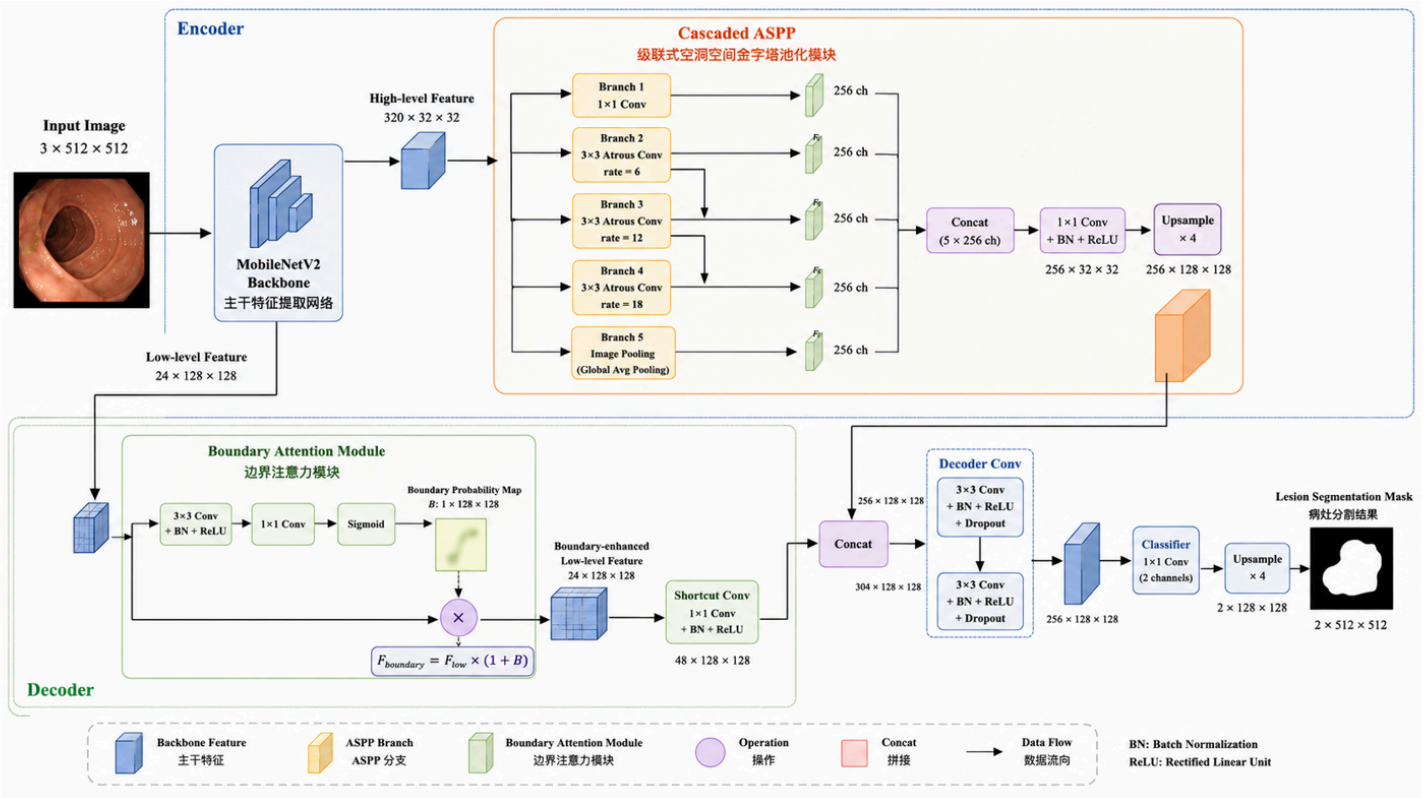


Figure 1

Overall architecture of the improved DeepLabv3+ model.

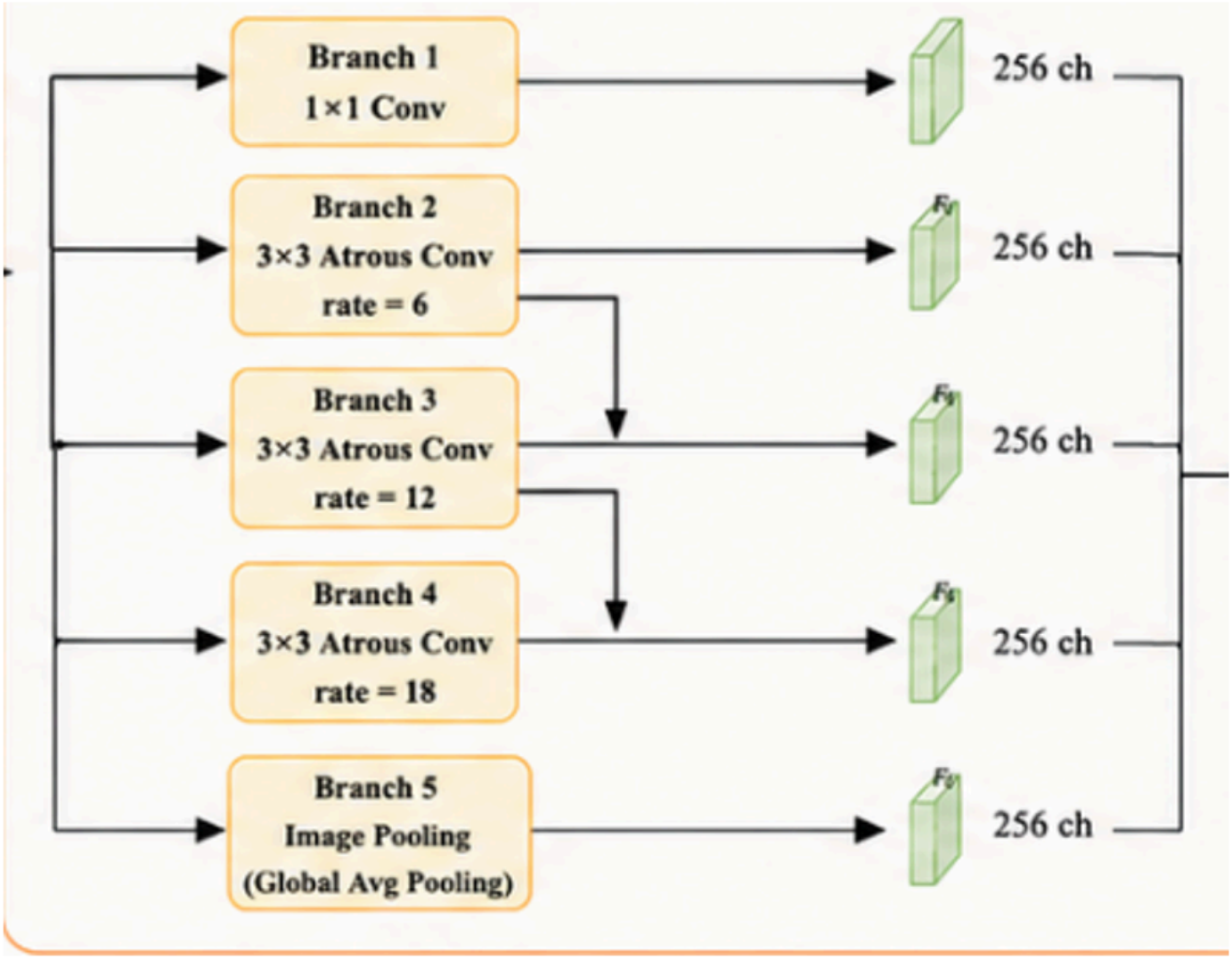


Figure 2

Structure of the cascaded ASPP module.

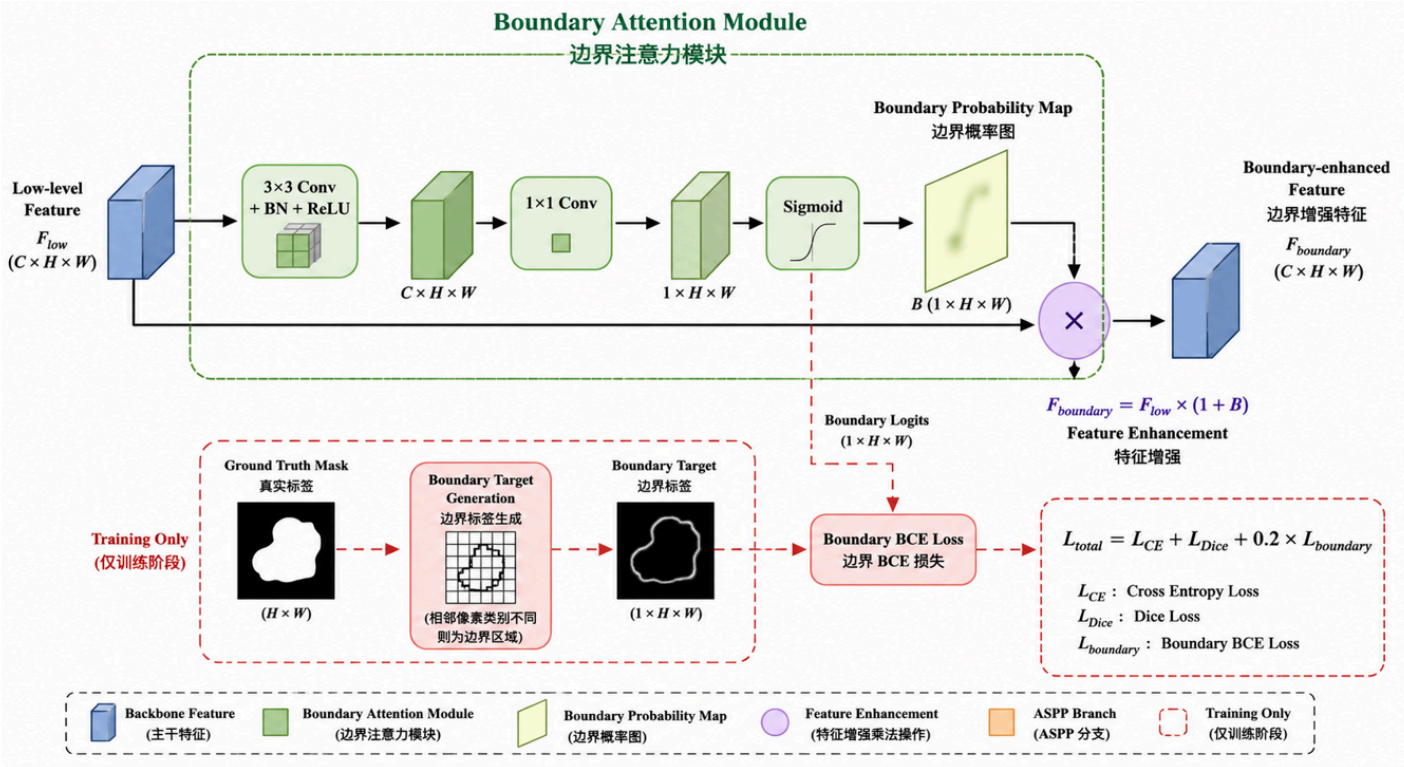


Figure 3

Boundary attention branch and auxiliary boundary supervision.

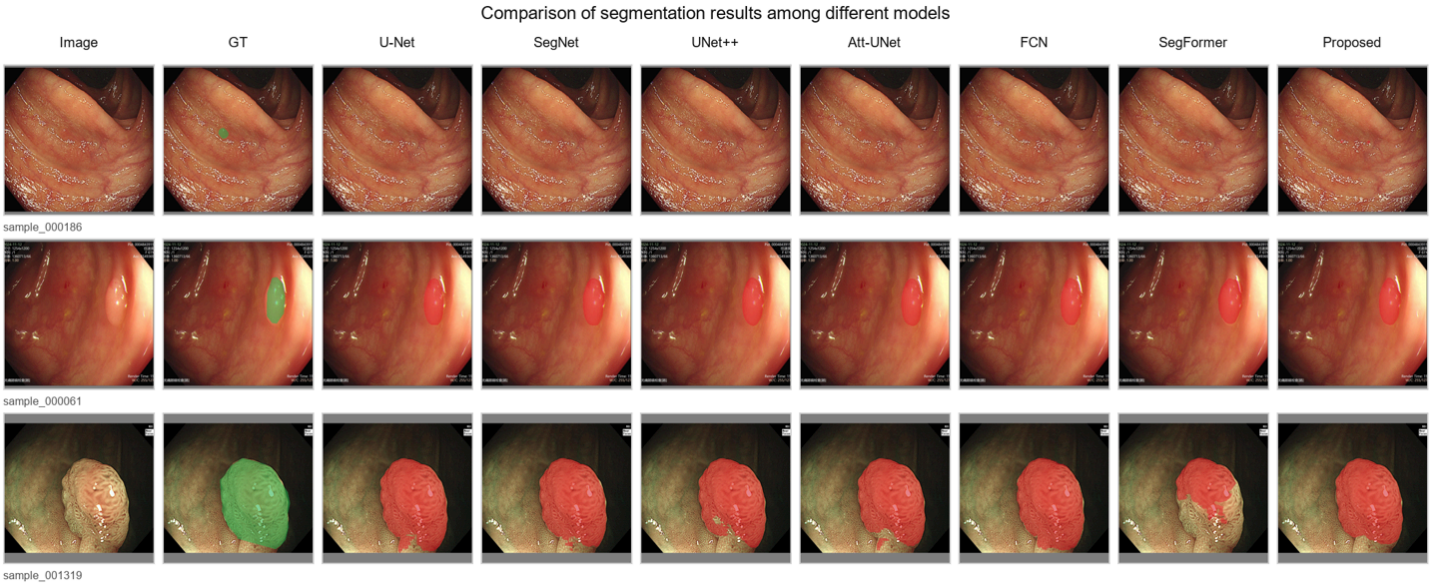


Figure 4

Qualitative comparison of different segmentation models.

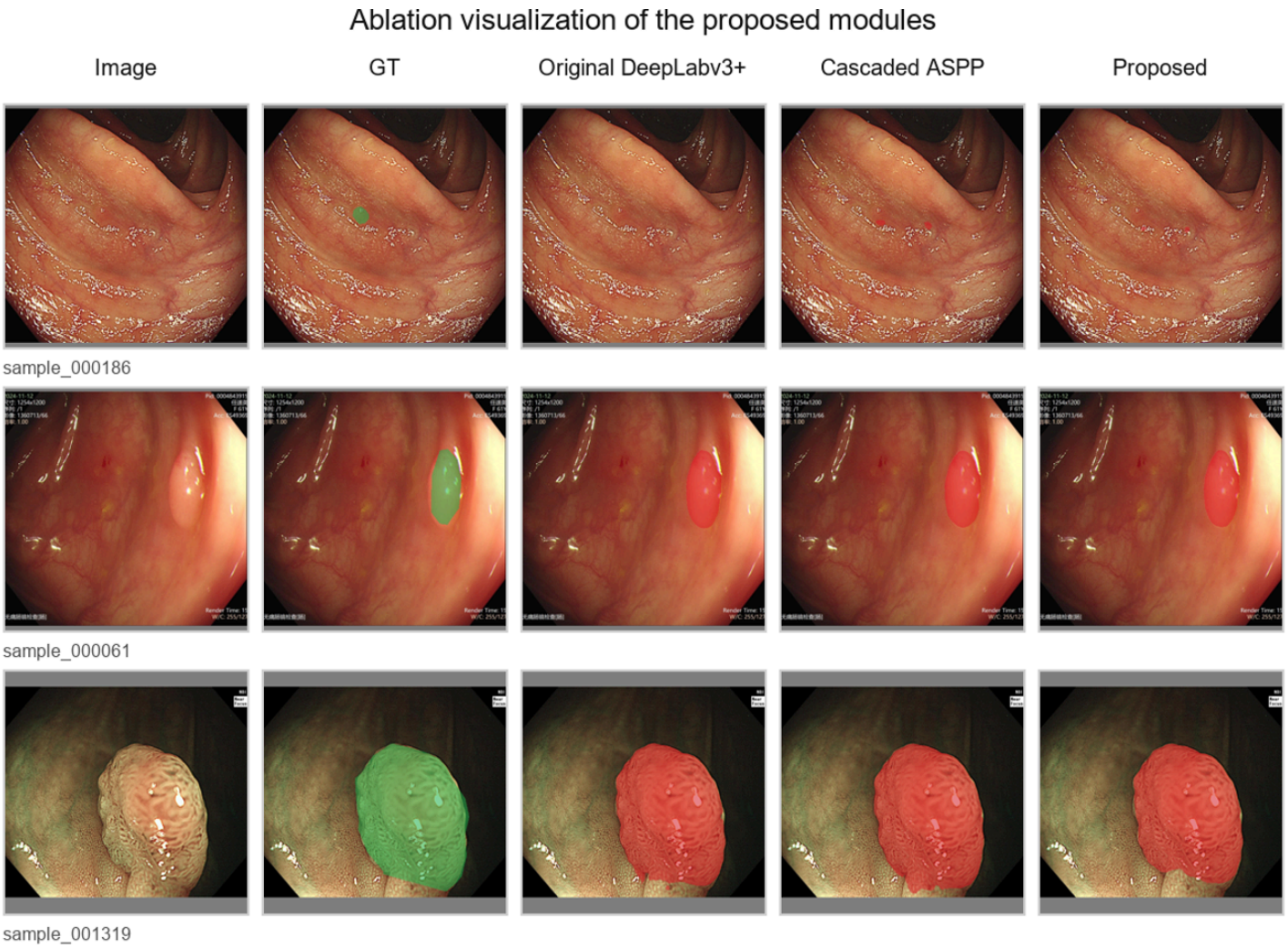


Figure 5

Qualitative visualization of ablation experiments.