ISSN 2616-5775 Vol. 9, Issue 4: 77-85, DOI: 10.25236/AJCIS.2026.090410

Lesion Region Segmentation of Endometrial Cancer Based on an Improved YOLOv8

Qianqian Zhang[1,a], Jie Ying[1,b,*], Yu Wang[1,c], Le Fu[2,d]

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
[2]First Maternity and Infant Hospital, Tongji University , Shanghai, China
[a]2335050707@st.usst.edu.cn, [b]yingjsh@163.com, [c]2335061722@st.usst.edu.cn, [d]fule0125@qq.com
**Corresponding author*

Abstract: *To address the segmentation challenges in endometrial cancer MR images, including blurred lesion boundaries, irregular morphology, and large scale variation, an improved YOLOv8 model was proposed. The original YOLOv8-seg served as the baseline, and three key algorithmic improvements were introduced. An efficient multi-scale attention module based on cross-spatial learning was embedded into the backbone network to enhance the discriminative ability for lesion regions. The SPPF structure in the backbone network was replaced with a focal modulation module to achieve adaptive modeling of multi-scale contextual information. In the neck network, a dual-path feature fusion module that integrates the re-parameterization concept with an improved CSP (Cross Stage Partial) structure was designed to strengthen the collaborative representation of local details and high-level semantic information. The model was trained and evaluated on a proprietary dataset consisting of 803 endometrial cancer MRI slices. Experimental results show that the model achieved Recall, IoU, Precision, and DSC values of 92.2%, 80.1%, 96%, and 88%, respectively, which verifies the effectiveness and advancement of the proposed method in endometrial cancer lesion segmentation.*

Keywords: *Deep Learning, MRI; YOLOv8, Endometrial Cancer, Image Segmentation*

1. Introduction

Endometrial cancer (EC) is a common malignant tumor of the female reproductive system with an increasing worldwide incidence[1]. Due to superior soft tissue contrast and multiparametric imaging, magnetic resonance imaging (MRI) is a key modality for preoperative EC staging and myometrial invasion depth assessment[2]. Thus, accurate EC lesion segmentation is clinically vital for individualized treatment, surgical guidance, and prognosis. However, manual delineation is time-consuming, labor-intensive, and relies on physician experience, leading to poor reproducibility and unmet clinical demand.

U-Net[3] and its variants[4,5] use encoder-decoder architectures with skip connections to fuse high-level semantics and low-level spatial details, performing well in segmentation tasks. Despite this, inherent complexities like low resolution, ambiguous boundaries, small lesions, and complicated backgrounds remain challenging. Researchers continue exploring new architectures to improve accuracy. The Transformer[6], offering global context modeling, combined with U-Net forms models like TransUNet[7]. Kalantar et al.[2] proposed a deep learning framework with a multi-head dilated encoder for cervical cancer segmentation on multiparametric MRI. This framework uses independent encoders for modalities like T2-weighted (T2WI) and diffusion-weighted imaging (DWI) to alleviate misalignment, optimizing a residual U-Net baseline. Additionally, AESC-TransUnet[8] addresses limitations in low-resolution and small lesion segmentation via an attention-enhanced selective channel Transformer.

Recent studies show YOLOv8 excels in instance segmentation across multiple domains. However, applying YOLOv8 to EC MRI faces challenges: lesions vary in shape and size with indistinct boundaries, and surrounding organs with similar morphology complicate segmentation[1]. This study proposes an improved YOLOv8-seg method for EC MR image segmentation. By introducing advanced attention mechanisms and multi-scale feature fusion, the method enhances fine-grained feature capture and global context perception. Integrating medical image characteristics with YOLOv8 strengths, we developed a high-accuracy, efficient automatic model named EFF-YOLOv8 to provide reliable

ISSN 2616-5775 Vol. 9, Issue 4: 77-85, DOI: 10.25236/AJCIS.2026.090410

imaging evidence for clinical EC management.

2. EFF-YOLOv8 Model

2.1 EFF-YOLOv8 Model Architecture

To address the challenges of blurred lesion boundaries, scale variation, and irregular morphology in EC medical image segmentation, this paper proposes an efficient and accurate segmentation network named EFF-YOLOv8. The network is built upon the original YOLOv8 architecture and introduces innovative modifications to key modules in the backbone and neck, forming an end to end segmentation framework with strong feature perception, adaptive scale modeling, and efficient feature fusion capability. The overall structure is shown in Figure 1.

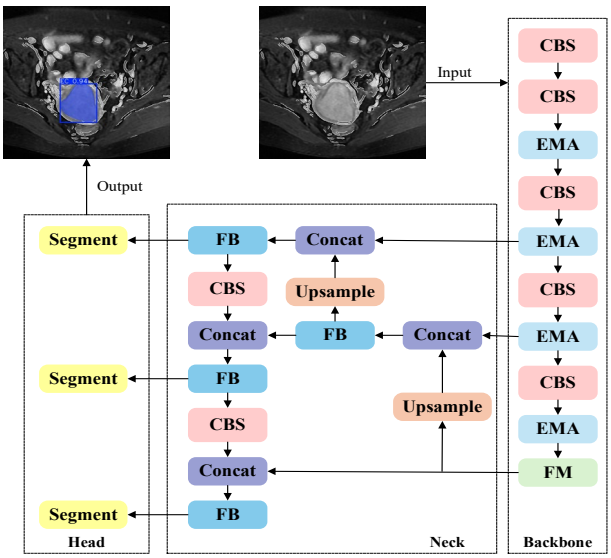

Figure 1 Network architecture of EFF-YOLOv8.

EFF-YOLOv8 follows a four stage pipeline consisting of input, feature extraction, feature fusion, and segmentation output. After preprocessing, the medical images are fed into the network and multi scale features are first extracted by the backbone. The original C2f module is enhanced by introducing an efficient multi scale attention module based on cross spatial learning, referred to as EMA, which enables accurate capture of critical lesion features. In addition, a focal modulation module, denoted as FM, replaces the original SPPF module to improve adaptive modeling capability for lesions with varying scales. The features generated by the backbone are then forwarded to the neck. A dual path feature fusion module, termed FB, is designed based on reparameterization and an improved CSP structure, which deeply integrates local details with global semantic information and produces high quality fused representations. Finally, the fused features are processed by the segmentation head to generate lesion segmentation results with the same spatial resolution as the input, providing accurate assistance for clinical diagnosis. Through collaborative optimization of these modules, the proposed network significantly improves segmentation performance for complex lesions while maintaining computational efficiency.

2.2 Efficient Multi Scale Attention Module Based on Cross Spatial Learning

In EC images, lesion regions often exhibit blurred boundaries and textures similar to surrounding tissues, which imposes higher requirements on discriminative capability during feature extraction. The backbone of the original YOLOv8 mainly relies on local convolution for feature modeling and is insufficient to capture critical spatial correlations within lesion regions. To address this issue, an efficient multi scale attention mechanism based on cross spatial learning, referred to as Efficient Multi Scale Attention (EMA)[9], is introduced in the backbone stage to improve the structure of the original C2f module. This module enhances feature perception for endometrial cancer lesions. It aims to dynamically strengthen responses to lesion regions through adaptive attention while suppressing background noise interference, thereby improving segmentation accuracy.

ISSN 2616-5775 Vol. 9, Issue 4: 77-85, DOI: 10.25236/AJCIS.2026.090410

EMA first groups the input feature along the channel dimension and divides it into M groups of sub feature maps:

$$\mathbf{X}=\left[\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_M\right],\mathbf{X}_i\in\mathbb{R}^{\frac{C}{M}\times H\times W} \tag{1}$$

where $\mathbf{X}\in\mathbb{R}^{C\times H\times W}$ denotes the input feature map, C represents the number of channels, and H and W denote the height and width, respectively. $\mathbf{X}_i$ denotes the i th sub feature map, M is the number of groups, and C/M represents the number of channels in each group. This grouping strategy facilitates learning diverse semantics and maintains balanced spatial semantic distribution for each sub feature group, which avoids information collapse caused by global pooling.

Subsequently, for each sub feature map $\mathbf{X}_i$,EMA adopts a multi branch parallel structure for attention modeling. As illustrated in Figure 2, one branch encodes direction aware global context information by applying one dimensional global average pooling along horizontal and vertical directions. For the i th sub feature map, the horizontal pooling output $\mathbf{z}_i^H(h)$ at vertical position h and the vertical pooling output $\mathbf{z}_i^W(w)$ at horizontal position w are computed as follows:

$$\mathbf{z}_i^H(h)=\frac{1}{W}\sum_{w=1}^{W}\mathbf{X}_i(h,w) \tag{2}$$

$$\mathbf{z}_i^W(w)=\frac{1}{H}\sum_{h=1}^{H}\mathbf{X}_i(h,w) \tag{3}$$

In Eq. (2), $h\in[1,H]$ denotes the vertical coordinate of the feature map. In Eq. (3), $w\in[1,W]$ denotes the horizontal coordinate of the feature map.

This process effectively preserves positional information and captures long range dependencies without introducing additional spatial compression. The features from the two directions are then concatenated and passed through a shared 1×1 convolution to model cross channel interactions. Afterward, a Sigmoid activation function generates direction aware attention weights, which are used to adaptively recalibrate the original features. To compensate for the limited local modeling capability of one dimensional pooling, EMA introduces another parallel branch based on 3×3 convolution to capture multi scale contextual information within local spatial neighborhoods, thereby enhancing sensitivity to lesion boundaries and texture variations. After branch level feature modeling, EMA further fuses the outputs of different branches through a cross spatial learning mechanism. Specifically, two dimensional global average pooling is first applied to the branch outputs to generate a global spatial descriptor vector $\mathbf{s}_i$, and the computation is defined as follows:

$$\mathbf{s}_i=\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{F}_i(h,w,:) \tag{4}$$

In Eq. (4), $\mathbf{F}_i\left(h,w,:\right)$ denotes the feature vector at spatial coordinate $\left(h,w\right)$ of the i th sub feature map after branch processing. Subsequently, Softmax is applied to generate a spatial weight distribution, and matrix multiplication is then employed to establish pixel wise pairwise relationships, enabling information interaction across spatial dimensions. This design simultaneously focuses on locally salient regions and global structural information, which effectively enhances the discriminative capability of feature representation.

Finally, the attention weighted results of all feature groups are activated by Sigmoid and remapped to the original feature space, followed by fusion along the channel dimension to obtain an output feature map with the same size as the input. Overall, the EMA module avoids channel dimension compression and achieves precise enhancement of critical lesion regions through grouped modeling, multi scale parallel structure, and cross spatial interaction learning, thereby providing more discriminative feature representations for subsequent segmentation.

ISSN 2616-5775 Vol. 9, Issue 4: 77-85, DOI: 10.25236/AJCIS.2026.090410

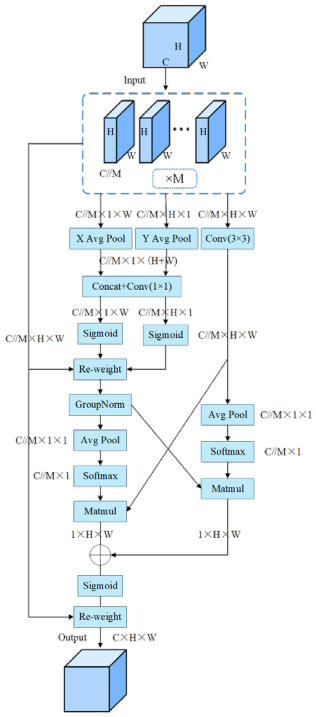

Figure 2 Architecture of EMA module.

2.3 Focal Modulation Module

EC lesion regions exhibit significant scale diversity, and different regions show distinct dependencies on local details and global contextual information. However, the SPPF structure adopted in YOLOv8 performs multi scale modeling in a fixed manner, which makes it difficult to achieve adaptive adjustment for different spatial locations. To enhance adaptive modeling capability for multi scale contextual information and address the rigid context capture and limited generalization caused by fixed receptive field structures when handling scale varying lesions, this study introduces a novel interaction module without self attention, referred to as Focal Modulation (FM)[10], to replace the original SPPF module in the YOLOv8 backbone.

The core idea of FM is that it does not aggregate contextual information by computing pairwise relationships among all tokens as in traditional self attention mechanisms. Instead, it adopts a more efficient and intuitive strategy that first aggregates and then modulates. Through multi level context modeling and a gating mechanism, FM adaptively aggregates spatial contextual information at different scales and generates modulation factors to perform position wise modulation on the original features. This design enhances representation capability of key regions while maintaining computational efficiency. The overall structure is illustrated in Figure 3.

Specifically, given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, the feature is first mapped to a new feature space through a linear projection layer. The projected initial feature map is denoted as $\mathbf{T}_0 \in \mathbb{R}^{C\times H\times W}$, where $f_z()$ represents the linear projection function. Subsequently, stacked depth-wise convolutions are employed for hierarchical contextualization. At the ℓ th focal level ($\ell \in \{1,2,...,L\}$, L denotes the number of focal levels), the output feature map $\mathbf{T}_\ell$ is expressed as follows:

$$\mathbf{T}_\ell = GeLU(DWConv_{k_\ell}(\mathbf{T}_{\ell-1})) \tag{5}$$

Here, $DWConv_{k_\ell}$ denotes the depth wise separable convolution with kernel size k_ℓ, where k_ℓ represents the convolution kernel size at the ℓ th level, and $GeLU()$ denotes the Gaussian Error Linear Unit activation function. By progressively enlarging the convolution kernel size, this process gradually captures visual contextual information at different granularities from local to global. Furthermore, to incorporate global context, global average pooling is applied to the highest level feature $\mathbf{T}_L$ to obtain a

ISSN 2616-5775 Vol. 9, Issue 4: 77-85, DOI: 10.25236/AJCIS.2026.090410

global feature T, which supplements the global contextual information.

Subsequently, in the gated aggregation stage, the model learns a spatial and level aware gating weight $\mathbf{G} = f_g(\mathbf{X}) \in \mathbb{R}^{H\times W\times(L+1)}$, where $f_g()$ denotes the gating weight learning function implemented by a linear layer, enabling adaptive fusion of contextual features from different levels. The final modulator M is generated through weighted summation followed by a subsequent linear transformation:

$$\mathbf{M} = h(\sum_{\ell=1}^{L+1} \mathbf{G}_\ell \odot \mathbf{T}_\ell) \tag{6}$$

Here, $\mathbf{G}_\ell$ denotes the slice of the gating weight G at the ℓ th layer, and corresponds to the adaptive weighting coefficient at the same layer. The operator $\odot$ represents element-wise multiplication, and $h()$ denotes a linear layer that promotes cross-channel interaction.

Finally, the original input X is projected through another linear pathway to obtain the query feature q. An element-wise affine transformation is then applied between q and the modulator M, enabling refined modulation for each token and producing the output Y. This design allows the network to adaptively focus on discriminative regions according to the input content while avoiding the high computational complexity of conventional self-attention mechanisms, thereby providing a more efficient and interpretable feature representation for medical image segmentation tasks.

2.4 Dual-Path Feature Fusion Module Based on Re-parameterization and Improved CSP Structure

The conventional YOLOv8 Neck performs multi-scale feature fusion via top-down and bottom-up pyramids. However, this structure shows limited capability in jointly modeling local fine-grained and high-level semantic information in medical imaging, especially when EC lesions exhibit unclear boundaries and significant morphological variations. To address this, a dual-path feature fusion module (FB) is introduced into the Neck, integrating re-parameterization with an improved CSP structure to enhance feature representation for medical image segmentation. Figure 4 illustrates the module architecture.

Following the Cross-Stage Partial Connection (CSP)[11] principle, the module introduces efficient processing units. A dual-branch structure first processes input features, where both branches adjust channels through a Conv1x1+BN+Act sequence. This design reorganizes features while reducing computational complexity. One branch serves as a shortcut to preserve original information, while the other forms an enhancement path to strengthen nonlinear representation.

Within the enhancement path, a re-parameterization-based convolutional structure incorporates multiple branches, including 3x3 and 1x1 convolutions with batch normalization, to increase feature diversity. This enables the network to learn spatial characteristics of EC lesion regions more effectively, improving modeling of irregular shapes and small lesions. Furthermore, an activation function and an additional 3x3 convolution enlarge the receptive field and strengthen local contextual modeling, helping the network distinguish lesions from background tissues. Outputs from both paths are concatenated along the channel dimension, followed by a 1x1 convolution for fusion and compression. This process generates high-quality features integrating local details with global semantics.

These improvements enable the YOLOv8 Neck to achieve sufficient information interaction. The FB module enhances representation for complex lesion structures and provides discriminative, robust features for the segmentation head, improving overall EC medical image segmentation performance.

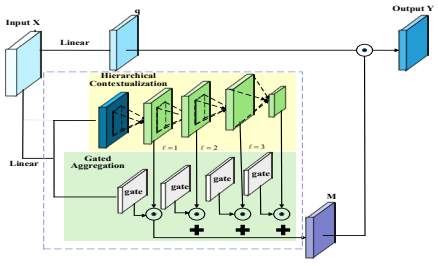

Figure 3 Architecture of FM module.

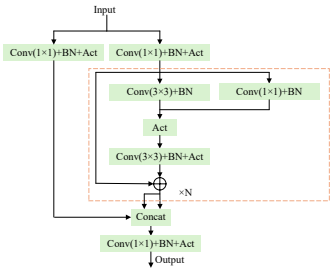

Figure 4 Architecture of FB module.

ISSN 2616-5775 Vol. 9, Issue 4: 77-85, DOI: 10.25236/AJCIS.2026.090410

3. Dataset construction and experimental environment

3.1 Data Construction

The data are provided by the First Maternity and Infant Hospital affiliated with Tongji University and include MR images from 207 patients. The images are acquired using a 1.5T (Optima MR360) magnetic resonance imaging system equipped with a phased-array coil. After screening, each patient contributes 2 to 13 valid slices, resulting in 803 two-dimensional MR images containing EC lesions for training and evaluation. All lesions are manually delineated slice by slice by experienced radiologists from Shanghai First Maternity and Infant Hospital and independently reviewed by senior radiologists. When discrepancies arise, the final annotation is determined through consensus to ensure reliable and consistent labeling.

3.2 Dataset Preprocessing

The region of interest (ROI) is first extracted from each image and uniformly resized to 512×512. The Canny algorithm is then applied for edge detection, and the results are superimposed on the ROI image, followed by grayscale normalization. Data augmentation is further performed using horizontal flipping and random rotation, as shown in Figure 5. To prevent data leakage, the dataset is strictly partitioned at the patient level, where all slices from the same patient belong to the same subset. The dataset is divided into training, validation, and test sets at a ratio of 8:1:1.

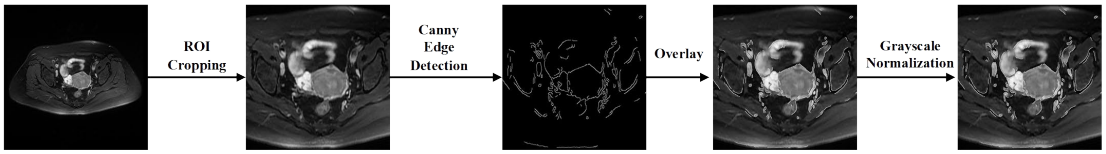

Figure 5 Flow of data preprocessing.

3.3 Experimental Setup

The experiments are conducted on a platform with an Intel Xeon Platinum 8352V CPU, 90 GB of system memory, and a vGPU with 32 GB of video memory. The operating system is Ubuntu 22.04, and GPU acceleration is implemented using PyTorch 2.8.0, Python 3.12, and CUDA 12.8. The model is trained for 150 epochs using the AdamW optimizer. The initial learning rate is 0.002, the momentum coefficient is 0.9, and the batch size is 16.

3.4 Evaluation Metrics

To comprehensively assess model performance, recall is adopted as a key indicator for measuring missed lesion detection. A higher recall indicates fewer missed lesions and more complete lesion coverage. In addition, three commonly used metrics, including Dice similarity coefficient (DSC), precision, and intersection over union (IoU), are employed. The corresponding formulas are as follows:

$$DSC = \frac{2 \times TP}{FP + 2 \times TP + FN} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$IOU = \frac{TP}{TP + FP + FN} \tag{10}$$

Here, TP, FP, and FN denote the numbers of true positive, false positive, and false negative pixels, respectively.

ISSN 2616-5775 Vol. 9, Issue 4: 77-85, DOI: 10.25236/AJCIS.2026.090410

4. Experimental results and analysis

4.1 Comparative Experiments

To comprehensively evaluate the segmentation performance of the proposed model, four classical medical segmentation models, including U-Net, SegNet[12], DeepLabv3+[13], and TransUNet, are selected for comparison. All models are trained and tested on the same EC dataset under identical conditions to ensure fairness. Quantitative results of each metric are presented in Table 1, the visual comparison is shown in Figure 6, and Recall curves across epochs are shown in Figure 7. Here, red indicates the ground truth lesion region, blue indicates the lesion region predicted by the model, and green indicates the overlapping area between the ground truth and predicted lesion regions.

Table 1.Comparative Experiments

Model	DSC/%	IOU/%	Precision/%	Recall/%
U-Net	67.5	52.1	70.7	67.1
TransUNet	73.55	61.68	73.85	75.32
SegNet	79.09	68	81.3	81.6
DeepLabv3+	84.6	74.6	85.7	85.4
EFF-YOLOv8(Ours)	**88**	**80.1**	**96**	**92.2**

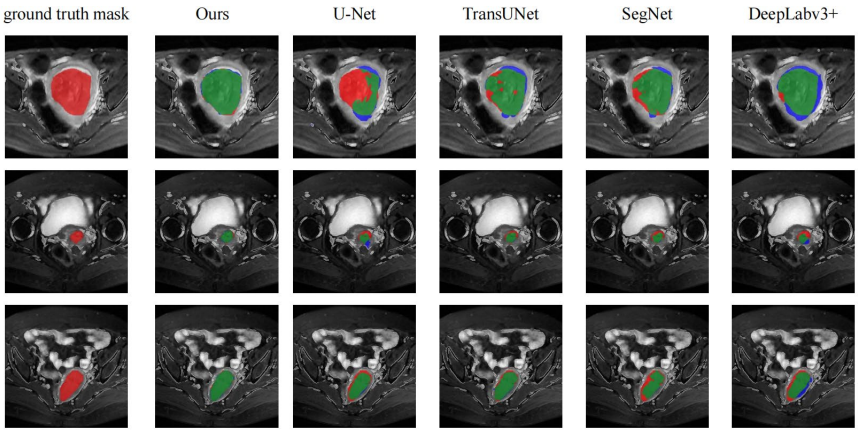

Figure 6 Segmentation performance comparison of different models on the dataset

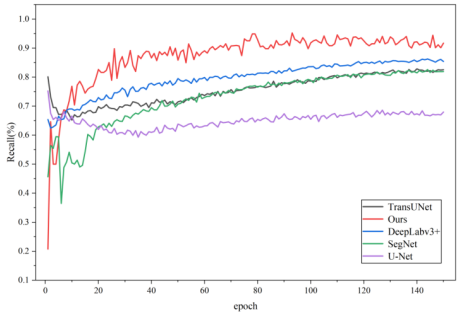

Figure 7 Recall versus epoch for different models.

The comparative results show that the proposed model achieves the best performance across all four evaluation metrics. Compared with the strongest baseline, DeepLabv3+, the DSC increases from 84.6% to 88%, improving by 3.4%, and the IoU rises from 74.6% to 80.1%, improving by 5.5%, indicating significantly better region overlap and segmentation consistency. Precision improves from 85.7% to 96% with a gain of 10.3%, while recall increases from 85.4% to 92.2% with a gain of 6.8%. Both false positives and missed detections are markedly reduced, and lesion boundary delineation becomes more accurate. U-Net, TransUNet, and SegNet show clear gaps from DeepLabv3+ in DSC and IoU, suggesting that conventional encoder-decoder architectures and hybrid Transformer structures achieve suboptimal segmentation on this dataset. The visual results in Figure 6 further confirm these findings. The comparison models exhibit blurred boundaries and missed lesion regions, whereas the proposed model produces results more closely aligned with the ground-truth masks and achieves more

ISSN 2616-5775 Vol. 9, Issue 4: 77-85, DOI: 10.25236/AJCIS.2026.090410

continuous and precise recognition of small lesions. Overall, the improvements introduced to YOLOv8 effectively enhance endometrial cancer image segmentation and achieve performance clearly superior to existing classical models.

4.2 Ablation experiments

To verify the effectiveness of each proposed module for EC segmentation, progressive ablation experiments are conducted using YOLOv8 as the baseline. Four configurations are evaluated. The first is the original YOLOv8 baseline model. The second incorporates the EMA module into YOLOv8. The third further introduces the FM module on this basis. The fourth additionally integrates the FB module, constituting the final model. All experiments are performed on the EC dataset. Quantitative results are reported in Table 2, performance improvements are illustrated in Figure 8, and Recall curves across epochs are shown in Figure 9

Table 2. Ablation experiment

Model	DSC/%	IOU/%	Precision/%	Recall/%
YOLOv8	83	74	87.3	84.5
YOLOv8+EMA	86	75	91.6	88.7
YOLOv8+EMA+FM	86.59	78.02	94.4	89.1
YOLOv8+EMA+FM+FB(Ours)	**88**	**80.1**	**96**	**92.2**

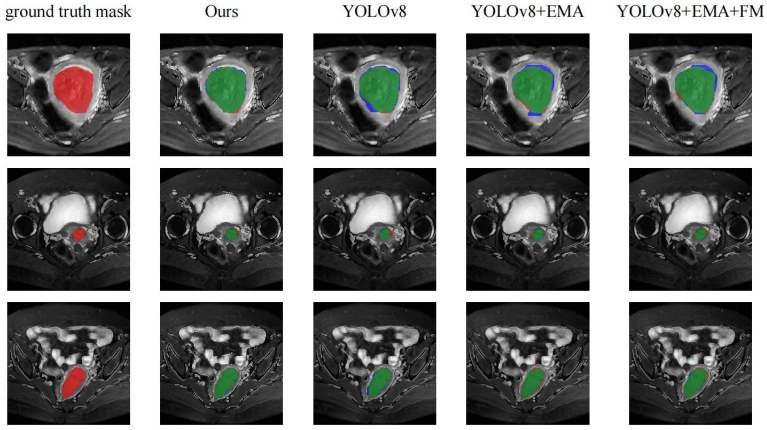

Figure 8: Segmentation results with the improved method.

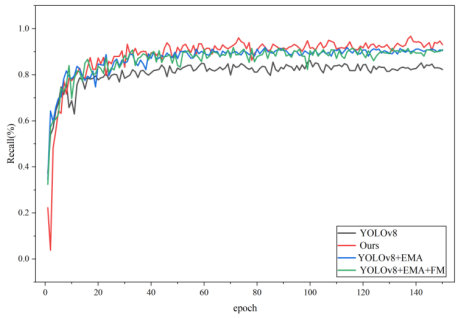

Figure 9 Recall comparison of different YOLO-based models.

The ablation results indicate that each proposed module brings consistent performance gains, showing a progressive improvement trend. The baseline YOLOv8 achieves a DSC of 83%. After introducing the EMA module, the DSC increases to 86%, improving by 3%, and recall rises from 84.5% to 88.7%, demonstrating enhanced lesion feature perception and fewer missed detections. With the FM module, the DSC further increases slightly to 86.59%, while the IoU improves from 75% to 78.02%, showing clearer gains in region overlap. After incorporating the FB module, all metrics reach optimal values. The DSC, IoU, precision, and recall achieve 88%, 80.1%, 96%, and 92.2%, respectively, corresponding to improvements of 5%, 6.1%, 8.7%, and 7.7% over the baseline. These results indicate that the collaboration of multiple modules enhances segmentation accuracy and region completeness. The visual results in Figure 8 are consistent with the quantitative trends in Table 2. As

ISSN 2616-5775 Vol. 9, Issue 4: 77-85, DOI: 10.25236/AJCIS.2026.090410

modules are progressively integrated, segmentation boundaries become smoother and false detections decrease. The recall curves in Figure 9 show that the final model converges more stably and maintains a higher overall level, further verifying the effectiveness of the improved architecture.

5. Discussion and Conclusion

This study addresses challenges in endometrial cancer MRI segmentation, including blurred lesion boundaries, scale variation, and irregular morphology, and proposes an efficient and accurate segmentation network, EFF-YOLOv8, based on an improved YOLOv8-seg architecture. Experiments on the EC dataset show that EFF-YOLOv8 outperforms the original YOLOv8 baseline, confirming the effectiveness and superiority of the proposed method. However, this work still has limitations. First, the model is validated only on single-center MRI data, and its generalization across devices, field strengths, and multimodal settings needs further evaluation. Second, the model generates two-dimensional slice-level segmentation results and has not been extended to three-dimensional voxel-level reconstruction, limiting support for spatial continuity and volumetric lesion analysis. Future work will further improve the endometrial cancer image segmentation algorithm to enhance segmentation performance and support computer-assisted clinical diagnosis.

References

[1] Ying Jie, Huang Wei, Fu Le, et al. Weakly supervised segmentation of uterus by scribble labeling on endometrial cancer MR images[J]. Computers in Biology and Medicine, 2023, 167: 107582.
[2] Kalantar R, Curcean S, Winfield J M, et al. Deep learning framework with multi-head dilated encoders for enhanced segmentation of cervical cancer on multiparametric magnetic resonance imaging[J]. Diagnostics, 2023, 13(21): 3381.
[3] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [C]//Proc of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer International Publishing, 2015: 234-241.
[4] Zhou Zongwei, Siddiquee M M R, Tajbakhsh N, et al. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation[J]. IEEE Transactions on Medical Imaging, 2019, 39(6): 1856-1867.
[5] Çiçek Ö, Abdulkadir A, Lienkamp S S, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation [C]//Proc of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer International Publishing, 2016: 424-432.
[6] Huang Huimin, Lin Lanfen, Tong Ruofeng, et al. Unet 3+: a full-scale connected unet for medical image segmentation [C]//Proc of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE Press, 2020: 1055-1059.
[7] Chen Jieneng, Mei Jieru, Li Xianhang, et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers[J]. Medical Image Analysis, 2024, 97: 103280.
[8] Huang Wenlei, Xiao Hongxiang. AESC-TransUnet: attention enhanced selective channel transformer U-Net for medical image segmentation: W. Huang, H. Xiao[J]. Signal, Image and Video Processing, 2025, 19(9): 710.
[9] Ouyang D, He S, Zhang G, et al. Efficient multi-scale attention module with cross-spatial learning [C]//Proc of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE Press, 2023: 1-5.
[10] Yang J, Li C, Dai X, et al. Focal modulation networks[J]. Advances in Neural Information Processing Systems, 2022, 35: 4203-4217.
[11] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN [C]//Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE Press, 2020: 390-391.
[12] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.
[13] Chen L C, Zhu Yukun, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]//Proc of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 801-818.