Multi-lesion Segmentation of Fundus Images using Improved UNet++

Haoyan Jiang, Ji Zhao*

***Abstract*—Diabetic Retinopathy is a common microvascular complication of diabetes, and early and accurate diagnosis is crucial for minimizing its impact on vision. To address the complexity and diversity of lesions in diabetic retinopathy, as well as the presence of numerous small-scale lesions, this study proposes a multi-lesion segmentation framework based on an improved UNet++ architecture. Utilizing ResNet50 as the backbone network for feature extraction, we integrated a hybrid attention module into the residual block to enhance the model's feature extraction capability in handling the complexity of lesions. To address the information loss of small lesions during feature extraction, we introduced and adapted Across Feature Map Attention as an auxiliary branch, which enhances the segmentation accuracy of small lesions. Furthermore, considering the insufficient feature extraction capability for DR lesions in shallow network layers, the model abandoned the deep supervision structure of traditional UNet++. Experiments employed a weighted hybrid loss function. Evaluations conducted on IDRiD and DDR segmentation datasets demonstrated effective segmentation of four typical Diabetic Retinopathy lesions. Results indicated that compared with other research methods, our approach achieved superior performance in Dice Coefficient and IoU metrics.**

***Index Terms*—Diabetic retinopathy, Convolutional Neural Network, Semantic Segmentation, Attention Mechanism.**

I. Introduction

DIABETIC Retinopathy (DR) is a common chronic complication of diabetes. It is a series of typical pathological changes caused by retinal microvascular damage caused by diabetes, which affects vision and even causes blindness. DR patients will have different pathological characteristics at various stages of the disease, such as soft exudates (SEs), hard exudates (EXs), microaneurysms (MAs), hemorrhages (HEs), etc. According to the disease progression, diabetic retinopathy can be divided into two stages: nonproliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR).The PDR stage poses a significant threat to vision. Timely identification of NPDR lesions constitutes the optimal approach for delaying diabetic retinopathy progression and preserving visual acuity [1]. In clinical applications, ophthalmologists screen by manually observing the lesions in color fundus images. However, this screening method is not only affected by the subjective factors of doctors but also has a large workload. Consequently, developing automated lesion segmentation systems becomes imperative for enhancing DR diagnostic workflows.

Manuscript received April 1, 2024; revised April 16, 2025.

This work was supported by Research Project of Department of Education of Liaoning Province.

Haoyan Jiang is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: hy_j123456@163.com).

Ji Zhao is a professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (Corresponding author, e-mail: 319973500069@ustl.edu.cn).

As computer vision methodologies and neural network architectures continue to grow, automatic lesion segmentation methods are gradually emerging in DR screening. In recent years, much research has been based on Convolutional Neural Networks (CNNs), and some pixel-level lesion annotation databases have been published. These models can automatically extract the features of specific lesions and perform accurate segmentation in the image by learning annotated fundus images. In contrast to conventional image processing techniques, deep learning has shown better performance in processing complex fundus images. Although these works have made significant progress in the automatic segmentation of DR lesions, they are still full of significant challenges. Firstly, the structure of DR lesions is complex, and there are differences in size, shape, color, brightness, and other aspects among various lesions. Secondly, there are many small and medium-sized lesions in DR lesions. In the IDRiD dataset, the lesion size of images with a resolution of 4288 x 2848 is counted, and 50% of lesions are less than 269 pixels [2]. The small lesion size presents a significant challenge for CNN-based segmentation methods in capturing discriminative features with adequate spatial information. In addition, the chromatic characteristics, morphological profiles, and textural patterns of retinal structures closely resemble those observed in abnormal tissues, thereby potentially causing erroneous positive diagnoses during medical imaging analysis[3].

To address the challenges posed by the abundance of small-scale lesions and the structural complexity of DR, this paper proposed an improved segmentation model based on UNet++. The proposed model abandoned the traditional deep supervision training approach and employed ResNet50 as the backbone network. Additionally, a hybrid attention mechanism was integrated into the residual blocks to enhance feature extraction for lesion regions. To further enhance segmentation precision, we integrated an adjusted Across Feature Map Attention (AFMA) as a dedicated auxiliary branch. The effectiveness of the proposed model was validated through experiments on the IDRiD and DDR segmentation datasets.

This paper is organized as follows: In section 2, we review existing research on DR lesion segmentation. Section 3 presents a comprehensive explanation of our enhanced model architecture. In section 4, we conducted ablation experiments on our model and performed a comparative analysis with other mainstream medical image segmentation methods, which validated its superior performance. In conclusion, the study summarizes the findings and suggests potential avenues for further investigation.

II. Related Work

DR lesion segmentation is generally achieved by analyzing color fundus images. Lesion segmentation approaches

broadly fall into two primary groups: traditional methods-based and deep learning-based. The following will introduce these two types of methods separately.

A. Traditional methods

Early-stage diabetic retinopathy lesion segmentation methods primarily rely on digital image processing and machine learning-based approaches, categorized into morphology-based lesion segmentation, clustering-based lesion segmentation, and region growing-based lesion segmentation. However, the effectiveness of these methods is frequently constrained by suboptimal brightness and contrast in fundus imaging, resulting in compromised robustness, reduced segmentation accuracy, and failing to meet clinical screening requirements.

B. Deep learning methods

In recent years, deep neural networks demonstrating notable efficacy have been increasingly adopted for diabetic retinopathy lesion segmentation. In 2017, Tan et al. [4] first utilized a 10-layer CNN to simultaneously segment multiple lesions, including exudates, hemorrhages, and microaneurysms. They evaluated the output at the pixel level, demonstrating the feasibility of using a single CNN structure for segmenting multiple lesions simultaneously. In 2018, Playout et al. [5] proposed an extension to U-Net that could simultaneously segment red and bright lesions. Their decoder incorporated new architectures such as residual convolution, global convolution, and mixed pooling, employing two identical decoders, each dedicated to a specific lesion category. In 2019, Guo et al. [6] introduced a small object segmentation network, L-seg, capable of simultaneously segmenting four types of lesions: EX, SE, MA and HE. It uses VGG16 as the backbone network, incorporating multi-scale feature extraction (from low-level details to high-level semantics) to boost the ability to identify lesions of different sizes. Subsequently, Yan et al. [7] proposed a novel cascaded architecture to address the computational burden of high-resolution DR color images and the poor global background capture resulting from image tiling. The model was composed of three key components: GlobalNet, LocalNet, and the Fusion module. GlobalNet took downsampled image features as input and produced coarse segmentation maps with the same dimensions as the original image. LocalNet, on the other hand, processed cropped patches of the image to generate segmentation maps at the original resolution. The Fusion module combined feature maps from GlobalNet and incorporated them into LocalNet, empowering the framework to capture both global and local information simultaneously. Guo et al.[3] proposed a dual-input segmentation network architecture named DARNet. The framework employs ResNet101 and ResNet50 as backbone networks for feature extraction from the two input modalities, respectively. To integrate multi-level feature information, an Attention Refinement Module (ARM) is designed to dynamically fuse features across different hierarchical layers. Addressing the scale variation of different DR lesions, Liu et al. [2] modified the upsampling and downsampling parts of the convolutional neural network, designing a universal multi-to-multi feature recombination network (M2MRF) to segment them.

Fig. 1: Image Cropping

A marked increase in segmentation accuracy for microscopic DR lesions was thereby achieved.

This study develops a DR lesion segmentation method to address limitations in existing approaches. To handle the challenges of minute lesion sizes and complex feature variations in DR, we propose an improved UNet++ architecture for automated multi-lesion segmentation in retinal images.

III. Methods

UNet++ [8] is a widely used network architecture that introduces a series of nested and skip connections to better capture multi-scale feature information in images, reduce feature loss problems, and improve the model's perception ability. At the same time, it provides more contextual and detailed information, enabling it to handle better complex situations such as target boundaries and small structures. However, when processing DR fundus images, UNet++ still has certain limitations, as traditional encoder structures cannot solve the problems of slight target information loss and significant sample differences in feature extraction. Further improvements to the algorithm and model structure are required.

A. Image Preprocessing

In the IDRiD segmentation dataset, certain fundus images exhibit extensive black background regions along the periphery. These non-informative areas contain no ocular-relevant information, failing to provide effective learning features for models while simultaneously causing computational resource waste and compromising training efficiency. This paper proposes an adaptive cropping methodology that initially employs the Canny edge detection algorithm to precisely localize the main retinal structure within the IDRiD segmentation dataset, ensuring complete preservation of critical lesion areas throughout the cropping process while eliminating irrelevant background information. Subsequently, the minimum bounding rectangle technique is implemented to achieve adaptive trimming of black backgrounds, thereby enhancing image focus on retinal regions and improving overall data quality. The cropping process is shown in Figure 1.

B. Framework Design

To tackle the challenges of automatic segmentation in DR, we introduce an innovative medical image segmentation architecture adapted from the UNet++ framework. We employed ResNet50 [9] as the backbone network and incorporated a hybrid attention module into the residual blocks to gain the model's feature extraction capability.To address information loss during feature extraction for small-scale lesions in DR fundus images, AFMA was introduced

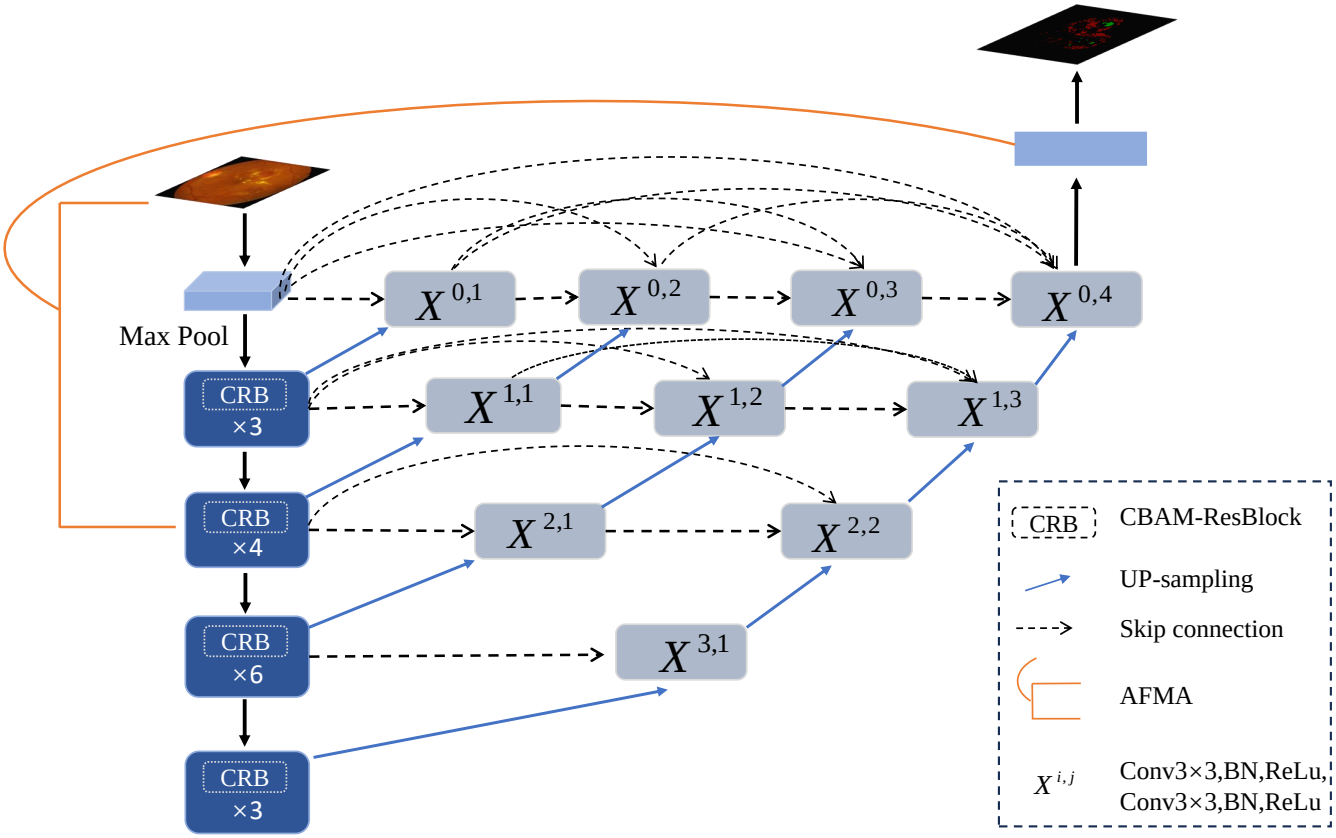

Fig. 2: Network structure

and adjusted as an auxiliary branch. Considering the high proportion of small lesions in DR samples and the insufficient representation of lesion features in the shallow layers of the model, we removed the deep supervision mechanism from the UNet++ architecture. Figure 2 delineates the complete framework of our model.

C. Encoder Design

1) Backbone Network: To improve the model's capability to adapt to different lesion types and stages in DR fundus images, more advanced feature extraction is required. UNet++ primarily utilizes basic convolution and pooling operations, limiting its feature representation. While deeper networks can improve fitting, excessive depth may lead to degradation. ResNet mitigates this issue through residual connections, enabling certain layers to bypass others and reducing dependency strength. With increased depth, residual blocks refine feature representation, enhancing learning capacity. Balancing computational efficiency with hierarchical feature acquisition requirements, our implementation adopted ResNet50 as the backbone network.

2) CBAM-ResBlock: To guide the model's emphasis primarily onto pathological areas rather than non-lesion content within retinal scans, this paper integrates the Convolutional Block Attention Module (CBAM) into the residual blocks of ResNet, enabling the network to extract key lesion features more effectively and enhancing the segmentation capability of the model.

As depicted in Figure 3, the CBAM framework [10] integrates dual attention components: Channel Attention Module (CAM) and Spatial Attention Module (SAM). CAM captures dependencies by modeling relationships across feature map channels, while SAM focuses on spatial correlations to enhance feature distribution understanding.

$$CAM(x) = Sigmoid(MLP(AvgPool(x)) + MLP(MaxPool(x))) \quad (1)$$

As shown in Equation 1, CAM adjusts the spatial dimensions of the feature map using both average and max pooling layers, maintaining the same number of channels. After passing through the MLP module, the results from the two layers are added together, and finally, the output result is obtained through a sigmoid activation function.

$$SAM(x) = Sigmoid(f^{7\times7}([AvgPool(x); MaxPool(x)])) \quad (2)$$

The SAM architecture processes input features through parallel max-pooling and average-pooling operations, yielding individual single-channel representations. These features are concatenated across the channel axis and subjected to a convolution layer for channel reduction. The resultant tensor undergoes sigmoid activation to produce the final spatial attention weights, with the complete computational workflow formalized in Equation 2.

Specifically, the CBAM was embedded into the bottleneck block by positioning it after the second 1×1 convolutional layer in the bottleneck block, forming the CBAM_ResBlock. This integration leverages both channel and spatial attention mechanisms to enhance the representation of lesion features. By combining CBAM with residual blocks, the model more effectively learns interdependencies between features, further strengthening its feature extraction capability. This design makes the architecture particularly suitable for extracting features from multiple co-occurring lesions in DR retinal

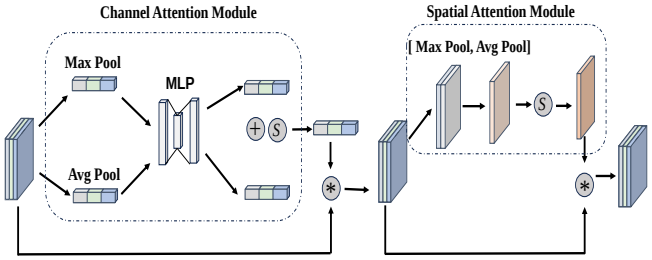

Fig. 3: CBAM structure

images. Ablation experiments confirmed that the CBAM module significantly improved lesion segmentation accuracy on DR images.

D. Information Loss Compensation

In DR fundus images, the presence of numerous microlesions poses significant challenges to feature extraction and segmentation tasks. While traditional convolution and pooling operations effectively extract deep-level features, they inevitably reduce image resolution, leading to progressive loss of microlesion information during hierarchical feature propagation. Given that these microlesions often serve as critical biomarkers for diseases, such information loss diminishes the model's sensitivity to lesion regions, thereby complicating the precise recovery of small-target features from low-resolution feature maps. To resolve this issue, this paper introduced and adapted the AFMA module as an auxiliary branch. The AFMA module compensates for information loss in convolution and pooling processes by capturing cross-hierarchical feature correlations between small and large objects within the same category, then integrating these associations into the final feature maps[11].

The auxiliary branch for information loss compensation operates in two phases. In the encoder phase, the model processes the input image $img \in \mathbb{R}^{H\times W\times C}$ and the Stage-3 output feature $F_3 \in \mathbb{R}^{H_3\times W_3\times C_3}$ from the backbone network through convolutional layers, aligning their channel dimensions with the segmentation category count N_C to obtain R_{img} and R_3. Channel-wise processing is then applied to R_{img} and R_3 : the feature map of each channel is partitioned into $d \times d$-sized patches, each flattened into a vector of length d^2. These flattened vectors are spatially ordered and combined into a block matrix of size $N \times d^2$, denoted as P^i_{img} and P^i_3 (where i is the channel index, and N denotes the number of patches). Next, the matrix multiplication between P^i_{img} and the transpose of P^i_3 is performed to obtain the relationship matrix A^i for the current channel.Finally, all channel-wise relationship matrices are concatenated to form the complete feature map A. In the second phase, the relationship feature map A is employed to optimize the decoder's predicted features. First, the decoder output mask $M_{mask} \in \mathbb{R}^{H\times W\times N_c}$ is resized to spatially align with R_3, yielding R_{mask}. For each channel in R_{mask}, the same patch partitioning, flattening, and block matrix construction are performed. All channel results are concatenated to form P_{mask}. Channel-wise matrix multiplication between A and P_{mask} produces P^i_{end} , the results from all channels are concatenated to form P_{end}. Finally, P_{end} is reshaped to match M_{mask}'s dimensions and multiplied element-wise with M_{mask} ,resulting in $O_{mask} \in \mathbb{R}^{H\times W\times N_c}$. The compensated output $Pre \in \mathbb{R}^{H\times W\times N_c}$ is obtained by summing O_{mask} and M_{mask}, thereby enhancing segmentation performance, the complete workflow is illustrated in Figure 4 [11].

E. Training module

1) Deep supervision: The deep supervision structure in UNet++ facilitates model pruning and reduces the number of model parameters. However, the difficulty of extracting model features due to the small scale of DR lesions and significant differences between samples makes it difficult, and the shallow features of the model contain less information. For pruning mode, the segmentation accuracy of subnetwork output feature maps is not high, and pruning will cause a decrease in segmentation accuracy. For ensemble mode, collecting the segmentation results of all segmentation branches and taking their average will cause information loss in profound network segmentation results, leading to a decrease in model segmentation accuracy. Therefore, our model abandoned the deep supervision structure and used the last layer of upsampled feature maps as output. Ablation experiment results demonstrated that removal of the deep supervision structure enhanced model segmentation accuracy in diabetic retinopathy cases.

2) Loss Function: The overall loss function of the model consist of two components to ensure optimization at different levels. First, the primary loss is the main segmentation loss, which seeks to reduce the discrepancy between the model's predicted segmentation of retinal lesions and the ground truth annotations. This ensures that the model accurately identifies lesion areas and improves segmentation precision.Second, an auxiliary branch loss was introduced to effectively supervise the generated feature map A, enhancing its quality and representational capability.

For the primary segmentation loss, this paper adopts the widely-used Cross-Entropy loss L_{CE} in deep learning. For mitigating the category disproportion across lesion versus background pixels within the DR dataset, we additionally introduce the Dice loss L_{Dice}. These two losses are combined in a weighted manner to form the primary segmentation loss, as formally defined in Equation 3, λ_1 and λ_2 represent the balance weights for L_{CE} and L_{Dice}, respectively.

$$L_{seg} = \lambda_1 L_{CE} + \lambda_2 L_{Dice} \tag{3}$$

For the auxiliary branch loss, this paper employs the mean squared error (MSE) loss function from reference [11] to supervise the feature map A.

$$L_{afma}=\frac{1}{C\cdot l_1\cdot l_2}\sum_{c=1}^{C}\sum_{i=1}^{l_1}\sum_{j=1}^{l_2}\left[A<c,i,j>-A_{gt}<c,i,j>\right]^2 \tag{4}$$

In the equation, A represents the relationship feature map obtained in the first stage of the auxiliary branch, where l_1 and l_2 represent the height and width of A, respectively. $A_{gt}$is derived from the segmentation label G. First, G is resized to match the dimensions of R_3, denoted as R_{gt} . For each channel i, G^i and R^i_{gt} undergo $d \times d$-patch partitioning and flattening to construct matrices P^i_G and P^i_{gt} ,Next, a matrix multiplication is performed between P^i_G and the transpose of

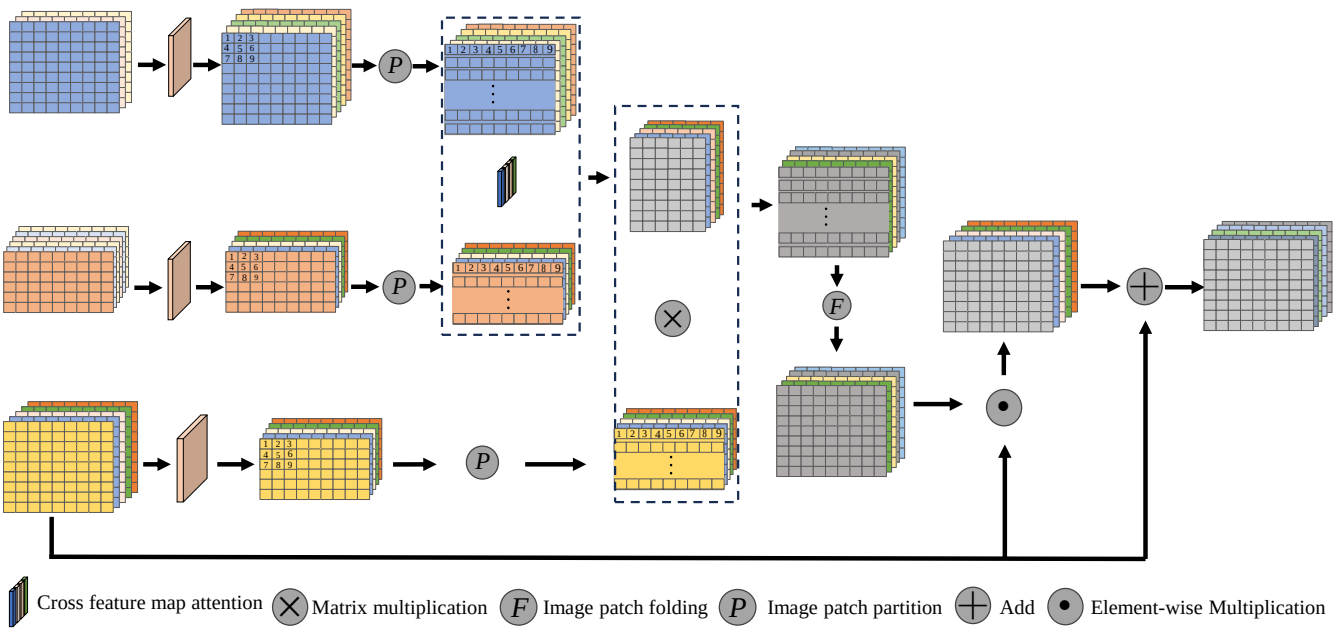

Fig. 4: Loss information compensation process

P_{gt}^{i} to obtain A_{gt}^{i}. Finally, the results from all channels are concatenated to construct the complete relationship feature map A_{gt}. Thus, the overall loss can be represented as:

$$L = \alpha L_{seg} + \beta L_{afma} \tag{5}$$

IV. Experiments and Analysis

A. Dataset

This paper evaluated the segmentation model's performance using the IDRiD and DDR segmentation datasets. The key characteristics of these datasets are summarized below:

1) IDRiD: The IDRiD segmentation dataset consists of 81 retinal images from India, with 54 images designated for the training set and 27 assigned to the test set. Each image in the dataset maintains a resolution of 4288 × 2848 pixels. Each image is annotated with pixel-level labels[12]. In the experiments of this study, we subdivided the training set the training set, dividing the 54 training images into 46 for training and 8 for validation.

2) DDR: The CFP images in the DDR segmentation dataset were captured using different retinal cameras. This dataset contains 757 color retinal images from Chinese individuals, each labeled with four types of lesions. The image resolution in the data set varies. In the course of the experiments, a total of 383 images were designated for the training phase, 149 images were set aside for validation, and 225 images were used for testing purposes[13].

B. Experimental environment

All experiments in this section were carried out on a high-performance computing server featuring an NVIDIA V100 GPU. The experimental framework was built using PyTorch with the following configurations: input image resolution of 1024×1024 pixels, batch size of 2, and 650 training epochs. The learning rate was set to 1e-4, with the Adam optimizer, momentum set to 0.9, and weight decay set to 0. In the AFMA module, the patch size was established at 10×10 pixels. The loss function weighting parameters were configured as λ_1=λ_2=1, α=1, and β=0.2. For identifying the best-performing network hyperparameters, we adopted the minimum validation loss criterion for final weight selection, followed by comprehensive performance evaluation on the test set image database.

C. Evaluation Criteria

To assess our method's lesion segmentation performance on the IDRiD and DDR segmentation datasets, two widely recognized evaluation metrics in semantic segmentation—Dice Coefficient (Dice) and Intersection over Union (IoU)-were employed for performance quantification. Their mathematical formulations are expressed as:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{6}$$

$$IOU = \frac{TP}{TP + FN + FP} \tag{7}$$

Where TP: Correctly classified positive-class instances, FP: Negative-class instances misclassified as positive, FN: Positive-class instances erroneously assigned to negative.

D. Ablation Experiments

To investigate the contribution of the ResNet50 backbone network, the CBAM module, the AFMA module, and deep supervision learning to the DR lesion segmentation performance, systematic experimental analysis was carried out using the IDRiD segmentation dataset. Using UNet++ with deep supervision as the baseline model, we compared the performance improvements brought by different modules.

The results in Table 1 show that UNet++ without deep supervision outperforms the version with deep supervision in terms of lesion segmentation accuracy. Furthermore, the integration of other structures significantly enhanced the

TABLE I: Ablation study results on IDRiD segmentation dataset

Model	Encoder	DS	CBAM	AFMA	mIou(%)	mDice (%)
Baseline	UNet++	√	-	-	36.49	51.56
(a)	UNet++	-	-	-	41.31	57.14
(b)	ResNet50	-	-	-	43.26	59.02
(c)	ResNet50	-	√	-	43.89	59.62
(d)	ResNet50	-	√	√	47.83	63.57

model's segmentation ability, with the AFMA module showing the most notable improvement, indicating its crucial role in enhancing the segmentation accuracy.

E. Comparative Experiments

To thoroughly assess the proposed method, we performed comparative experiments against mainstream semantic segmentation models using the DDR and IDRiD segmentation datasets. Experimental results from IDRiD and DDR datasets are separately presented and summarized in Table 2, where the best-performing outcomes per dataset are highlighted in bold.

As demonstrated in the comparative data of Table 2. On the IDRiD dataset,the proposed model gained a mean Dice coefficient (mDice) of 63.57%, demonstrating the highest average Dice coefficient that substantially surpassed UNet (55.01%), DeepLabv3+(56.44%), and UNeXt (51.29%). Compared with TransUNet, our approach exhibited significant superiority, particularly in HE lesion segmentation. Meanwhile, the model exhibited outstanding performance in the IoU metric, outperforming other models in overall segmentation as well as in the segmentation of EX, HE, and SE lesions. On the DDR dataset, our framework maintained competitive lesion segmentation capabilities, with overall segmentation results significantly surpassing those of UNet and UNeXt. Compared to TransUNet, the proposed model demonstrates superior performance in segmenting SE and MA lesion types. The proposed architecture's effectiveness in lesion segmentation has been rigorously verified through extensive multi-dataset experimentation.

Figure 5 depicts the segmentation results of partial fundus images in IDRiD segmentation dataset and DDR segmentation dataset, where (a) shows the ground truth labels, (b) presents the segmentation outcomes from the UNet model, (c) displays the segmentation results produced by the TransUNet model, (d) highlights the segmentation outcomes generated by the UNeXt model, and (e) depicts the segmentation outcomes obtained from our model. In the segmentation images, red indicates EX lesions, green indicates HE lesions, yellow indicates SE lesions, and blue indicates MA lesions. From the visualization results in Figure 5, it can be observed that our model's predictions closely match the ground truth labels, further demonstrating the effectiveness and reliability of the proposed model.

V. Conclusion

DR constitutes primary contributor to blindness, making accurate detection and segmentation of DR lesions crucial. Recent years have witnessed substantial progress in diabetic retinopathy lesion segmentation through deep neural networks. However, challenges remain, such as the large variation in lesions between samples and the prevalence of small lesions. To address these issues, we improved the UNet++ architecture by employing ResNet50 as the primary feature extraction framework and incorporating attention modules to raise the model's feature extraction capability. Additionally, we incorporated the AFMA module as an auxiliary branch to compensate for the loss of small lesion information during feature extraction. Furthermore, we discarded the traditional deep supervision structure and adopted a weighted hybrid loss function. While our model has improved segmentation accuracy for DR lesions, challenges remain, including missed and false detections, especially for small lesions like microaneurysms, owing to the insufficient quantity of precisely labeled DR lesion datasets. Future work will focus on deep learning-based medical image generation to mitigate data limitations. Additionally, we will refine our model to enhance small lesion segmentation accuracy and explore lightweight network designs without sacrificing performance.

References

[1] R. Chakrabarti, C. A. Harper, and J. E. Keeffe, "Diabetic retinopathy management guidelines," *Expert review of ophthalmology*, vol. 7, no. 5, pp. 417–439, 2012.

[2] Q. Liu, H. Liu, W. Ke, and Y. Liang, "Automated lesion segmentation in fundus images with many-to-many reassembly of features," *Pattern Recognition*, vol. 136, p. 109191, 2023.

[3] Y. Guo and Y. Peng, "Multiple lesion segmentation in diabetic retinopathy with dual-input attentive refinenet," *Applied Intelligence*, vol. 52, no. 12, pp. 14 440–14 464, 2022.

[4] J. H. Tan, H. Fujita, S. Sivaprasad, S. V. Bhandary, A. K. Rao, K. C. Chua, and U. R. Acharya, "Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network," *Information sciences*, vol. 420, pp. 66–76, 2017.

[5] C. Playout, R. Duval, and F. Cheriet, "A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 101–108.

[6] S. Guo, T. Li, H. Kang, N. Li, Y. Zhang, and K. Wang, "L-seg: An end-to-end unified framework for multi-lesion segmentation of fundus images," *Neurocomputing*, vol. 349, pp. 52–63, 2019.

[7] Z. Yan, X. Han, C. Wang, Y. Qiu, Z. Xiong, and S. Cui, "Learning mutually local-global u-nets for high-resolution retinal lesion segmentation in fundus images," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 597–600.

[8] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[11] S. Sang, Y. Zhou, M. T. Islam, and L. Xing, "Small-object sensitive segmentation using across feature map attention," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 5, pp. 6289–6306, 2022.

[12] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, 2018.

[13] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Information Sciences*, vol. 501, pp. 511–522, 2019.

TABLE II: Performance comparison with others for lesion segmentation.

Dataset	Methods	Dice (%)					IoU (%)				
		EX	HE	SE	MA	mDice	EX	HE	SE	MA	mIoU
DDR	**UNet [14]**	56.46	43.23	**43.85**	26.42	42.49	39.34	27.58	**28.09**	15.22	27.56
	TransUNet [15]	**57.38**	**50.39**	40.27	28.16	**44.05**	**40.24**	**33.68**	25.22	16.39	**28.88**
	UNeXt [16]	56.51	45.40	39.14	23.63	41.17	39.38	29.37	24.34	13.40	26.62
	Ours	56.64	47.74	42.86	**28.37**	43.90	39.51	31.36	27.28	**16.53**	28.67
IDRiD	**UNet [14]**	77.62	49.53	56.56	36.34	55.01	63.43	32.92	39.43	22.21	39.50
	DeepLabv3+ [17]	76.82	53.95	56.71	38.27	56.44	62.37	36.94	39.85	23.67	40.71
	TransUNet [15]	78.79	61.94	65.86	**45.80**	63.10	65.01	44.87	49.10	**29.71**	47.17
	UNeXt [16]	74.86	46.51	56.17	27.62	51.29	59.83	30.31	39.05	16.02	36.30
	Ours	**79.61**	**64.57**	**65.90**	44.20	**63.57**	**66.13**	**47.68**	**49.15**	28.37	**47.83**

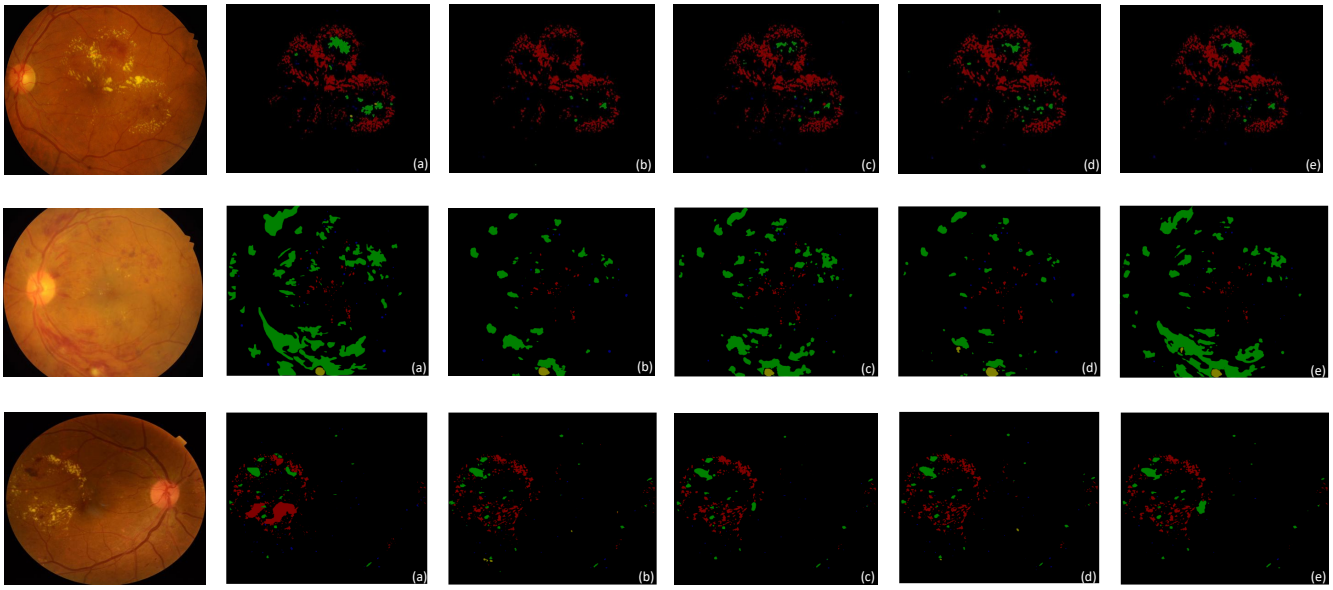

Fig. 5: Examples of Segmentation Results

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," 2021.

[16] J. M. J. Valanarasu and V. M. Patel, "Unext: Mlp-based rapid medical image segmentation network," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 23–33.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

Date of modification: January 6, 2025.

Brief description of the changes:

1) Fig.1: added Max Pool annotation explanation

2) Fig.3: changed the LN annotation to BN

3) Fig.5: corrected the direction of the arrow

4) Changed the value of the weight decay parameter from 0.1 to 0

Date of modification: March 12, 2025.

Brief description of the changes: The model structure has been adjusted, the experimental environment has been standardized, and the experimental data has been updated.

Date of modification: April 16, 2025.

Brief description of the changes:

Fig.5: the subfigures have been adjusted and subfigure labels have been corrected.