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

Haoyan Jiang, Ji Zhao*

***Abstract*—Diabetes retinopathy (DR) is one of the complications of diabetes. Early diagnosis of retinopathy is helpful to avoid vision loss or blindness. The difficulty of this task lies in the significant differences in the size and shape of lesions between different DR samples, with a higher proportion of small lesions. We propose a new multi-disease segmentation method based on UNet++ to improve the segmentation accuracy of DR lesions. We chose Resnet50 as the backbone network and introduced a new hybrid residual module to replace the original residual module. At the same time, to compensate for the loss of information in DR small lesions during the feature extraction process, we introduce the Across Feature Map Attention (AFMA) is an auxiliary branch that enhances the segmentation accuracy of small-scale lesions. Finally, in response to the difficulty in extracting DR lesions in shallow models, the model abandoned the deep supervision structure in UNet++. In addition, we use a weighted mixed loss function to train the model. We conducted experiments on IDRID and DDR public datasets, simultaneously segmenting four typical DR lesions. The results on intersection over union (IOU) and dice similarity coefficient (Dice) showed that our method achieved competitive performance compared to other research methods.**

***Index Terms*—Diabets retinopathy, Convolutional Neural Network, semantic segmentation, Attention mechanism.**

I. Introduction

DIABETS 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 exudation (SE), hard exudation (EX), microaneurysm (MA), bleeding point (HE), etc. According to the occurrence time, formation reason, and distribution characteristics of different lesions, diabetes retinopathy can be divided into two stages: the early stage of DR, known as nonproliferative diabetes retinopathy (NPDR). Currently, there is no evident focus on the patient's fundus. Early diagnosis of NPDR can help patients understand the disease status and promptly predict the disease's development. The second stage is the proliferative diabetes retinal stage (PDR). At this time, the patient's fundus appears to have severe retinal ischemia and new capillaries, and vision begins to decline or even become blind. Early screening of these NPDR lesions is the most effective method to slow down DR progress and prevent vision loss [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. Therefore, it is essential to create an automatic focus segmentation method for DR screening.

With the rapid development of computer vision and deep learning technology, automatic lesion segmentation methods are gradually emerging in DR screening. In recent years, much research has been based on Convolutional Neural Networks (CNN), 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. Compared to traditional image processing methods, 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]. Such a small lesion size poses an excellent challenge for CNN-based segmentation methods in learning discriminative representations with sufficient spatial information. In addition, the color, contour, and texture of tissues on the retina (such as blood vessels and optic disc) are similar to those of lesions, which can easily lead to false positive results.

We propose an improved model based on UNet++ to address the issues of multiple small-scale lesions, complex structures of various lesions, and tissue influence on the retina and fundus in DR lesions. Discarding deep supervised training, replacing the ResNet50 backbone network, and integrating CAR The fusion of AFMA into feature extraction solved the problem of low segmentation accuracy caused by the complex structure and significant differences of fundus lesions. We validated the effectiveness of our model using IDRID and DDR datasets.

The rest of the paper is organized as follows: Chapter 2 reviews related work on DR lesion segmentation. Chapter 3 provides a detailed description of our improved structural model. Chapter 4 compares our model with other deep learning-based models, highlighting its superior performance. Finally, we conclude our work and provide suggestions for future research.

II. Related Work

DR lesion segmentation is generally achieved by analyzing color fundus images. The lesion segmentation methods can

Manuscript received Apr 1, 2024; revised Aug 4, 2024.

This work was supported by the Special Fund for Scientific Research Construction of University of Science and Technology Liaoning.

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).

be roughly divided into two categories: traditional methods-based and deep learning-based. The following will introduce these two types of methods separately.

A. Traditional methods

Early methods for DR lesion segmentation were primarily based on digital image processing. Wynne Hsu et al. [3] employed a clustering-based approach, first preprocessing retinal images with median filtering. They then divided the images into several blocks, identifying the maximum value in each block as the lesion center and the minimum value as the background center. Subsequently, segmentation was achieved through dynamic clustering iterations until convergence. Thomas Walter et al. [4] utilized morphology-based algorithms for retinal lesion segmentation. They leveraged the higher pixel intensity in lesion areas to detect lesions and then used morphology algorithms to reconstruct lesion contours. After eliminating interference from vascular structures, areas containing exudates were determined using local contrast. Alan et al. [5] employed region-growing algorithms for retinal image segmentation. They initially eliminated image intensity variations caused by retinal illumination changes through median filtering and normalized contrast. Then, they utilized closing operations to generate images and obtained edges through thresholding. Subsequently, they performed region growth along edge gradients and merged similar regions using the watershed algorithm.

However, the performance of the above methods is often hindered by limitations in the brightness and contrast of fundus images, resulting in poor robustness and difficulty in achieving ideal segmentation results, which do not meet the requirements of clinical screening.

B. Deep learning methods

Deep learning-based methods with promising results have recently been applied to DR lesion segmentation. In 2017, Tan et al. [6] 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. [7] 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. [8] introduced a small object segmentation network, Lseg, capable of simultaneously segmenting four types of lesions: microaneurysms, soft exudates, hard exudates, and hemorrhages. The backbone network was VGG-16, with the fully connected layers and the fifth pooling layer removed and an additional lateral extraction layer added. The output was obtained through a multi-scale weighted fusion of the lateral extraction layers. Subsequently, Yan et al. [9] 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 consisted of three components: GlobalNet, LocalNet, and Fusion module. GlobalNet received downsampled features of the image as input and generated coarse segmentation maps of the same size as the input. LocalNet processed cropped image patches as input and generated segmentation maps at the original resolution. The Fusion module was used to crop feature maps from GlobalNet and concatenate them into LocalNet to simultaneously capture global and local information. 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. This achieved a significant improvement in the segmentation accuracy of small-scale DR lesions.

This study aims to devise a method for segmenting DR lesions, aiming to overcome the limitations posed by existing approaches, including the small size of DR lesions and the significant variations between lesions across different samples. To address these challenges, we introduce an enhanced UNet++ architecture for the automated segmentation of multiple lesions in DR images.

III. Methods

Unet++ network [10] is a widely adopted and efficient 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 optimization of the algorithm and model structure is needed.

A. Image Preprocessing

Before training the network, we perform necessary preprocessing on the original images for lesion enhancement and data augmentation. Firstly, sizeable black background areas at the boundaries of fundus images do not contain any eye-related information and can waste hardware resources. We use the Canny edge detection algorithm and apply thresholding to detect the edges of the fundus images. Then, we adaptively crop the fundus images to remove areas with zero pixels. Secondly, since the sizes of images in different datasets vary, we resize all images to 1024×1024. Finally, due to the limited number of image samples, we apply data augmentation to the original images: (1) Scaling according to random scaling factors. (2) Random flipping of images based on random probabilities. (3) Color transformations on images based on color space conversion parameters.

B. Framework Design

To address the challenges above, we have devised a new medical image segmentation model based on the UNet++ framework. We utilize ResNet50 [11] as the backbone network to enhance the model's feature extraction capabilities. Additionally, the Conditional Convolution Attention Residual Module (CAR) dynamically covers, expresses, and utilizes relationships between samples, mitigating the significant

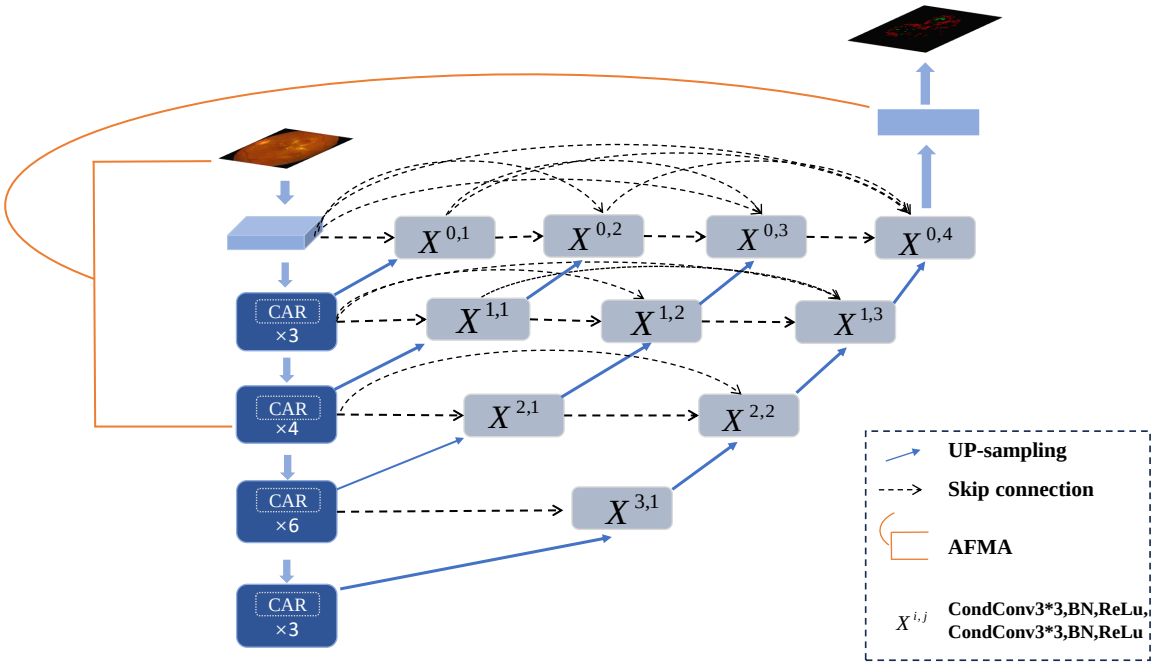

Fig. 1: Network structure

Fig. 2: Image Cropping

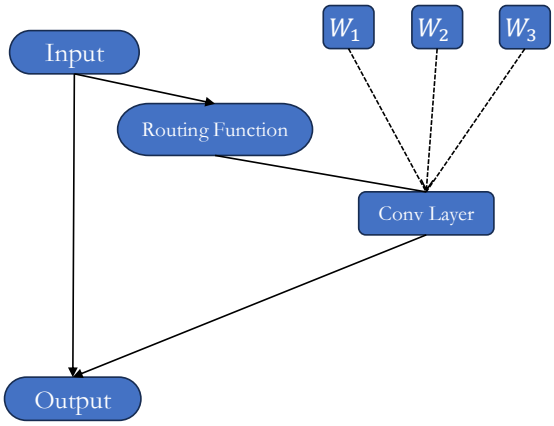

Fig. 3: CondConv Moudle

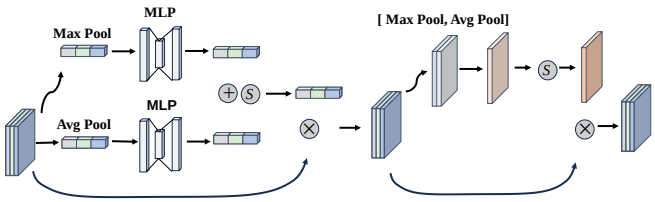

Fig. 4: CBAM Moudle

variability among different samples in DR fundus images, thereby further improving the segmentation accuracy of the model. To compensate for the loss of information in the feature extraction process for small-scale lesions in DR fundus images, we introduce Across Feature Mapping Attention (AFMA) [12]. AFMA represents the similarity of objects within the same category by computing the relationship matrix between intermediate feature blocks and original image blocks, compensating for the information loss associated with small-scale lesions by acting on the output module. Considering the large proportion of small target samples in DR samples, we have omitted deep supervision in UNet++. The overall architecture of the model is illustrated in Figure 1.

C. Encoder Design

1) Backbone Network: In order to improve the model's ability to fit different types or stages of lesions in DR fundus images, more complex calculations are needed to extract features. The Unet++ model only uses simple convolution and pooling operations in the feature extraction, resulting in weak feature capabilities. In theory, the deeper the network, the stronger the fitting ability. However, in practice, when the network depth reaches a certain level, the problem of network degradation will occur. ResNet effectively solves the problem of model degradation in deep neural networks by introducing residual structures, which allow specific layers in the neural network to skip the next feature extraction layer and connect to other layers, thereby weakening strong connections between layers. As the depth of the network increases, each residual block can learn additional feature changes, enhancing the network's expression and modeling capabilities. Considering hardware resources and feature extraction capabilities, the model chooses ResNet50 as the backbone network for feature extraction.

2) CAR Module: Considering the considerable variations in lesions among different samples of DR fundus images, we propose employing Conditional Convolution (CondConv)

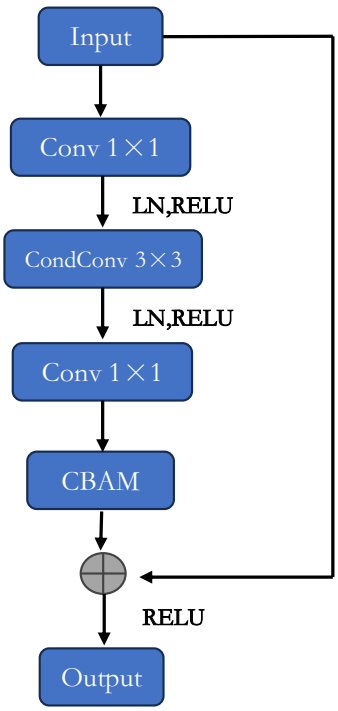

Fig. 5: CAR Moudle

[13] instead of traditional convolution modules for feature extraction. In traditional convolution, the same convolutional kernels are applied to all input samples, with fixed parameters determined solely by backpropagation during model training. However, in CondConv modules, the convolutional kernel parameters are determined by the input samples. The core idea is to equip each convolutional layer with multiple sets of weights and to weigh the convolutional kernels, thereby generating unique convolutional kernel parameters for each sample, which is particularly crucial for DR image segmentation, as the parameters of ordinary convolutions are fixed and cannot adapt well to the features of lesions with significant scale differences. This limitation results in suboptimal performance when dealing with DR images containing lesions of varying scales. The computation formula for CondConv is as follows:

$$Output(x) = \sigma((\alpha_1 \cdot W_1 + \cdots + \alpha_n \cdot W_n) * x) \quad (1)$$

Here α_i represents a scalar weight derived from the input feature x, W_i stands for the convolution parameters, n indicates the dimension of α_i, and α_i can be obtained through a routing function $r(x)$. The routing function is defined as:

$$r(x) = \text{Si}gmoid(GlobalAveragePool(x)R) \quad (2)$$

in the equation, R denotes a learnable routing weight matrix multiplied by the features after average pooling to map them to n scalar weights.

The Convolutional Block Attention Module (CBAM) [14] is a hybrid attention mechanism, as illustrated in Figure 3, consisting of a Channel Attention Module (CAM) and a Spatial Attention Module (SAM). CAM models relationships between different channels of the feature map to better capture dependencies among features. Conversely, SAM models relationships between different spatial positions of the feature map to better capture the spatial distribution of features.

$$CAM(x) = SigmoidMLP(AvgPool(x)) + MLP(MaxPool(x))) \quad (3)$$

As shown in Equation 3, CAM modifies the spatial dimension of the feature map x using both average pooling and max pooling layers while keeping the number of channels unchanged. 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.

$$SAM(x) = sigmioid(f^{n\times n}([AvgPool(x); \quad MaxPool(F)])) \quad (4)$$

SAM transforms the input features into feature maps with a single channel each through max pooling and average pooling layers. Subsequently, the two feature maps are concatenated and passed through a convolutional layer to create a one-channel feature map. Finally, an activation function is applied to generate the spatial attention feature map, as depicted in Equation 4.

Through conditional convolution, the network can dynamically generate convolutional kernel weights tailored to different samples, thereby better understanding and utilizing relationships among samples, thus improving model performance. Considering the complex structure and diverse lesion types in DR images, the CBAM module helps the network better understand the relationships among different channels and spatial positions in the images, thereby more accurately capturing and segmenting lesion areas, enhancing perception of lesion features, and improving segmentation accuracy of lesion areas. We propose a new module, CARMoudle, which combines the hybrid attention module with conditional convolution and residual structures, enhancing the model's ability to learn feature dependencies and feature extraction. This integration is more suitable for extracting features of multiple lesions in DR fundus images and replacing the residual modules in ResNet50. Our research results demonstrate that the CAR module improves the segmentation accuracy of lesions in DR images. The module architecture is shown in Figure 5.

D. AFMA

DR fundus images contain many different types of small-scale lesions, and statistics on the IDRID dataset with a resolution of 4288 * 4288 show that 50% of lesions are less than 269 pixels. At the same time, the shape and location of lesions, such as microaneurysms and bleeding points, can also vary depending on the patient's condition and stage, which poses challenges for small target segmentation of DR lesions. Traditional models often use convolution and pooling operations to capture high-level semantic features in images, which reduces the resolution of image features and causes information loss of some small objects (small targets) in the image, making it difficult for the model to recover the information of small targets from these low-resolution feature maps.

The AFMA module is designed to address information loss resulting from feature propagation by exploiting relationships among objects of similar sizes within the same category. It computes a relationship matrix between intermediate feature

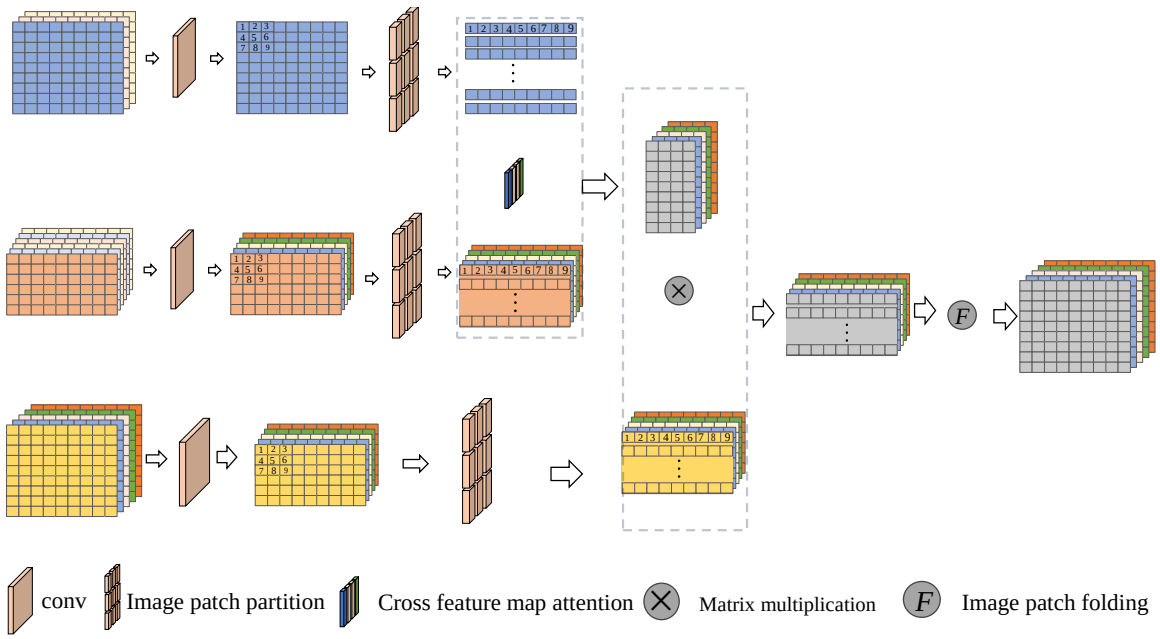

Fig. 6: AFMA Moudle

blocks and the input image to depict the similarity of objects within the same category. Subsequently, it utilizes these relationships to bolster the segmentation of small targets.Its structure is depicted in Figure six. It consists of two stages. In the first stage, within the encoder part, convolution is applied separately to the input image $img \in \mathbb{R}^{H\times W\times C}$ and a certain stage feature map $F_i \in \mathbb{R}^{H_i\times W_i\times C_i}$.Their channel numbers are transformed to 1 and the number of object classes N_c , resulting in $R_{img} \in \mathbb{R}^{H\times W\times 1}$ and $R_i \in \mathbb{R}^{H_i\times W_i\times N_c}$ respectively. R_{img} is segmented into fixed-size dd patches and reshaped into a two-dimensional patch $P_{img} \in \mathbb{R}^{\frac{HW}{d^2}\times d^2}$. The same segmentation and reshaping operations are performed for each channel of R_i, and the results are concatenated along the channel dimension to obtain $P_i \in \mathbb{R}^{\frac{H_iW_i}{d^2}\times d^2\times N_c}$. Finally, the dot product is computed between each channel category P_i^k of P_i and P_{img} to determine the relationship between each patch block in the image and the feature map related to the kth category. The resulting relationship matrices are concatenated along the channel dimension to obtain the final relationship map $A_i \in \mathbb{R}^{\frac{HW}{d^2}\times\frac{H_iW_i}{d^2}\times N_c}$. In the second stage, average pooling is applied to adjust the predicted map $M_{mask} \in \mathbb{R}^{H\times W\times N_c}$ from the decoder output to the size of R_i, denoted as $R_{mask} \in \mathbb{R}^{H_i\times W_i\times N_c}$. Similar segmentation, reshaping, and concatenation operations as those applied to R_i are applied to R_{mask} to obtain $P_{mask}\mathbb{R}^{\frac{H_iW_i}{d^2}\times d^2\times N_c}$. Then, the dot product is computed between R_{mask} and A_i, and the results from each channel are concatenated. The resulting $M_i \in \mathbb{R}^{\frac{HW}{d^2}\times d^2\times N_c}$ is unfolded into the same size as M_{mask} to obtain $O_i \in \mathbb{R}^{H\times W\times N_c}$. Finally,$O_i$ is added to M_{mask} to obtain the final output feature map $Pre \in \mathbb{R}^{H\times W\times N_c}$.

We choose to apply the AFMA module to the input image and the backbone feature network's third layer output feature map to obtain the relationship feature map. The obtained relationship feature map is modulated in the decoder output feature map to obtain the final segmentation map. Our experimental results indicate that The AFMA module significantly improves the accuracy of DR lesion segmentation.

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 abandons the deep supervision structure and uses the last layer of upsampled feature maps as output. The results of ablation experiments indicate that abandoning the deep supervision structure improves the segmentation accuracy of the model in DR.

2) Loss Function: The overall loss of the model is composed of two weighted losses. First, the standard segmentation loss aims to minimize the difference between the predicted values and basic facts of each pixel in the retinal image. Second, utilizing AFMA's auxiliary branch loss to supervise the generated relationship feature maps further improves their quality.

We use a weighted combination of cross-entropy loss and dice loss functions for the standard segmentation loss. Let P and G represent the predicted and ground truth maps, respectively. Let $< c, i, j >$ denote a pixel's channel, horizontal, and vertical coordinates. Let C represent the number of classes, where $1 \leq$ c $\leq$ C, and H, W represent the height and width of the output image. Then, the dice loss function

and the cross-entropy loss function can be represented by equations 5 and 6, respectively.

$$L_{\mathrm{Dice}}(P,G) = C - \sum_{c=1}^{C} \frac{2\sum_{i,j=1}^{H,W} p<c,i,j> G<c,i,j>}{\sum_{i,j=1}^{H,W} p^2<c,i,j> + \sum_{i,j=1}^{H,W} G^2<c,i,j>} \quad (5)$$

$$L_{CE}(P,G) = \frac{-1}{C \cdot H \cdot W} \sum_{c=1}^{C} \sum_{i,j=1}^{H,W} G<c,i,j> \log(p<c,i,j>) \quad (6)$$

the standard segmentation loss can be expressed as:

$$L_{seg} = \lambda_1 L_{CE} + \lambda_2 L_{Dice} \quad (7)$$

For the AFMA auxiliary branch, mean squared error loss is used.

$$L_{afma} = \frac{1}{C \cdot l_1 \cdot l_2} \sum_{c=1}^{C} \sum_{i=1}^{l_1} \sum_{j=1}^{l_2} [A_i<c,i,j> - A_{gt}<c,i,j>]^2 \quad (8)$$

In the equation,$A_i$represents the relationship feature map obtained by the AFMA module in the first stage, where l_1 and l_2 denote the height and width of A_i, respectively. A_{gt} is computed based on the label G, initially compressed to the size of $R_i \in \mathbb{R}^{H_i \times W_i \times N_c}$ using an average pooling layer, denoted as $R_{gt} \in \mathbb{R}^{H_i \times W_i \times N_c}$. Subsequently, the same segmentation, reshaping, and dot product operations are performed on R_{gt} and G for each corresponding channel using the AFMA module. Finally, the results from each channel are concatenated to obtain the feature $A_{gt} \in \mathbb{R}^{\frac{HW}{d^2} \times \frac{H_i W_i}{d^2} \times N_c}$. The specific formula is as follows:

$$R_{\mathrm{gt}}^k = \psi(\mathrm{G^k}, \frac{H}{H_i}, \frac{W}{W_i}, \frac{H}{H_i}, \frac{W}{W_i}) \quad (9)$$

$$A_{gt}^k = \phi(G^k, d) \otimes \phi(R_{gt}^k, d)^{-1} \quad (10)$$

$$A_{gt} = A_{gt}^1 || A_{gt}^2 \cdots || A_{gt}^C \quad (11)$$

here,$\Psi\left(input, K_h, K_w, S_h, S_w\right)$ represents the average pooling operation, where (K_h, K_w) are the kernel sizes and (S_h, S_w) are the stride sizes. $\phi(input, d)$ epresents the segmentation and reshaping operation, and $||$ denotes the concatenation operation. Thus, the overall loss can be represented as:

$$L = \alpha L_{seg} + \beta L_{afma} \quad (12)$$

IV. Experiments and Analysis

A. Dataset

We evaluated the segmentation performance of our model architecture on two retinal image datasets: IDRID and DDR. Here is some general information about the datasets:

1) IDRID: The images were captured using the Kova VX-10 50° FOV alpha retinal camera. It is used for challenges consisting of three subtasks: segmentation, grading, and localization. The segmentation subset of IDRID consists of 81 retinal images from India with a resolution of 4288 × 2848 pixels. Each retinal image has pixel-level labels for hemorrhages (HE), soft exudates (SE), microaneurysms (MA), and exudates (EX). In the experiments, 54 images were used for training and 27 for testing.In this experiment, we allocated 54 training set images as 46 training images and 8 validation images.

2) DDR: The CFP images in the dataset were captured using various retinal cameras with a 45° FOV. It contains 757 color retinal images of Chinese individuals, each corresponding to four lesions with manufactured labels. The image resolutions range from 1380 × 1382 to 2736 × 1824 pixels. In the experiments, 383 images were used for training, 149 for validation, and 225 for testing.

B. Experimental environment

This work experimented on a server with an NVIDIA GeForce RTX V100 (32GB) GPU. The comparative experiments on the DDR dataset were carried out on a server equipped with two NVIDIA TITAN XP (12GB) GPUs. The comparative experiments on the IDRID dataset were carried out on a workstation equipped with a single NVIDIA GeForce RTX 3090 (24GB) GPU.

We utilized PyTorch as our deep learning framework and trained and tested the models on each dataset separately using uninitialized networks. Prior to entering the model, images underwent preprocessing. We employed the Adam training strategy for rapid convergence training, with a batch size of 2 and a maximum iteration count of 650. The initial learning rate, momentum, and weight decay were set to 1e-4, 0.9, and 0.1, respectively. For the loss function hyperparameters, λ_1 and λ_2 were set to 1, while α and β were set to 5 and 1, respectively. We employed a strategy to adjust the learning rate adaptively based on the batch size.

C. Evaluation Criteria

To evaluate the segmentation performance of our method on the IDRID and DDR datasets, we adopted the widely accepted evaluation metrics in the field of semantic segmentation, namely the Dice Similarity Coefficient (Dice) and Intersection over Union (IoU). The definitions of these

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (13)$$

$$IOU = \frac{TP}{TP + FN + FP} \quad (14)$$

Where TP (True Positives) represents the number of correctly predicted positive cases, FP (False Positives) represents the number of incorrectly predicted positive cases, and FN (False Negatives) represents the number of incorrectly predicted negative cases. In some test images, specific lesions may not be present. Therefore, each pixel is treated as a case during evaluation, and the test set is considered a large pixel set. Evaluation metrics are then computed for all pixels accordingly.

D. Ablation Experiments

To explore the contributions of backbone networks, CAR modules, AFMA modules, and deep supervised learning to DR lesion segmentation performance, we tested them on the IDRID dataset. Compare the performance improvement of different modules using the deep supervised UNet++ as the benchmark model.

The results in Table 1 indicate that UNet++ without deep supervision has better accuracy in lesion segmentation than UNet++ with deep supervision. In addition, each added

TABLE I: Ablation study results on IDRID

Model	Encoder	DS	CAR	AFMA	mIou(%)	mDice (%)
Baseline	UNet++	√	-	-	38.98	55.58
(a)	UNet++	-	-	-	40.37	56.25
(b)	ResNet50	-	-	-	40.53	55.82
(c)	ResNet50	-	√	-	42.05	57.94
(d)	ResNet50	-	√	√	47.58	63.29

module significantly improves the segmentation ability of the model, especially the AFMA module, which has the most apparent ability to substitute for the model's segmentation effect.

E. Comparative Experiments

To validate the effectiveness of our proposed method, we compared our model with other mainstream semantic segmentation models on the DDR and IDRID datasets. The results are presented in Table 2 for the IDRID dataset and DDR dataset. The best results are highlighted in bold. We implemented the UNet, Deeplabv3+, TransUnet, and UNeXt results, with Deeplabv3+ and TransUnet utilizing the ResNet50 backbone. Other results are sourced from Paper [2].

The comparison results on the DDR dataset demonstrate that our model exhibits superior segmentation performance. Specifically, our model achieves the best performance on the MA lesion, with Dice and IOU metrics being 2.04% and 1.4% higher than the second-best segmentation result, respectively. Moreover, the model outperforms the overall average segmentation metrics for EX, HE, SE, and MA lesions, reaching 46.07% and 30.46% in Dice and IOU metrics, respectively.

The comparative analysis of the IDRID dataset reveals that our model exhibits the best performance on EX and SE lesions. Within EX lesions, Dice and IoU are 1.39% and 1.82% higher than the next-best structure. Similarly, in SE lesions, Dice and IoU are 2.24% and 2.42% higher. The segmentation performance for other lesions still surpasses most models, with the overall average segmentation metrics reaching state-of-the-art levels.

Overall, the proposed model in this paper demonstrates balanced segmentation capabilities in both DDR and IDRID datasets, showcasing competitive performance. Figure 7 depicts the segmentation results of partial fundus images, where (a) represents the ground truth labels, (b) shows the segmentation results of the UNet model, (c) shows the segmentation results of the Deeplabv3+ model, (d) shows the segmentation results of the TransUNet model, (e) shows the segmentation results of the UNeXt model and (f) shows the segmentation results of our model. In the segmentation images, red indicates EX lesions, green indicates HE lesions, yellow indicates SE lesions, and blue indicates MA lesions.

V. Conclusion

DR is one of the significant causes of blindness, making accurate detection and segmentation of DR lesions crucial. Deep learning-based DR lesion segmentation has made good progress in recent years. Still, it faces challenges such as significant differences in lesion shape and scale between samples and many small lesions. We have improved the Unet++ architecture and introduced ResNet50 as the backbone network to enhance feature extraction. Then, we used a new residual structure called CAR to perceive lesions better and address differences between DR lesions. Finally, we introduced the AFMA module to compensate for the loss of DR small lesions during the feature extraction process. In addition, we abandoned the deep regulatory structure and used a weighted mixed loss function. Although our model has improved the segmentation accuracy of DR lesions, our model still encounters missed and false detections of DR lesions due to the limitations of manually annotated high-quality DR lesion data, and the segmentation accuracy of small lesions such as MA still needs to be improved. In future research, we will focus on medical image generation based on deep learning to alleviate the impact of data on model segmentation while further optimizing our model to improve the segmentation accuracy of small lesions and exploring methods for creating lightweight network models without affecting segmentation accuracy.

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] W. Hsu, P. Pallawala, M. L. Lee, and K.-G. A. Eong, "The role of domain knowledge in the detection of retinal hard exudates," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 2. IEEE, 2001, pp. II–II.

[4] T. Walter, J.-C. Klein, P. Massin, and A. Erginay, "A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina," *IEEE transactions on medical imaging*, vol. 21, no. 10, pp. 1236–1243, 2002.

[5] A. D. Fleming, S. Philip, K. A. Goatman, G. J. Williams, J. A. Olson, and P. F. Sharp, "Automated detection of exudates for diabetic retinopathy screening," *Physics in medicine & biology*, vol. 52, no. 24, p. 7385, 2007.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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 [15]**	45.3	27.99	32.8	24.81	32.73	29.29	16.27	19.62	14.17	19.83
	TransUnet [16]	56.64	47.82	40.46	24.54	42.37	39.52	31.43	25.36	13.99	27.57
	UNeXt [17]	50.73	28.92	31.43	14.94	31.5	33.99	16.91	18.65	8.07	19.41
	Swin-base [18]	59.79	**50.53**	46.77	23.31	45.1	42.64	**33.82**	30.62	13.19	30.07
	M2MRF-C [2]	**60.62**	45.16	**47.78**	28.04	45.4	**43.49**	29.17	**31.39**	16.31	30.09
	Ours	57.43	49.9	46.87	**30.08**	**46.07**	40.29	33.25	30.61	**17.71**	**30.46**
IDRID	**UNet [15]**	77.31	49.04	54.85	35.48	54.17	63.01	32.49	37.80	21.56	38.72
	Deeplabv3+ [19]	70.18	51.35	59.47	37.94	54.74	54.07	34.55	42.32	23.44	38.60
	TransUnet [16]	78.81	**64.88**	63.01	**44.68**	62.84	65.03	**48.02**	46.0	**28.77**	46.95
	UNeXt [17]	73.90	52.79	46.75	27.79	50.30	58.60	35.87	30.50	16.14	35.28
	Ours	**80.20**	63.70	**65.25**	44.01	**63.29**	**66.95**	46.74	**48.42**	28.21	**47.58**

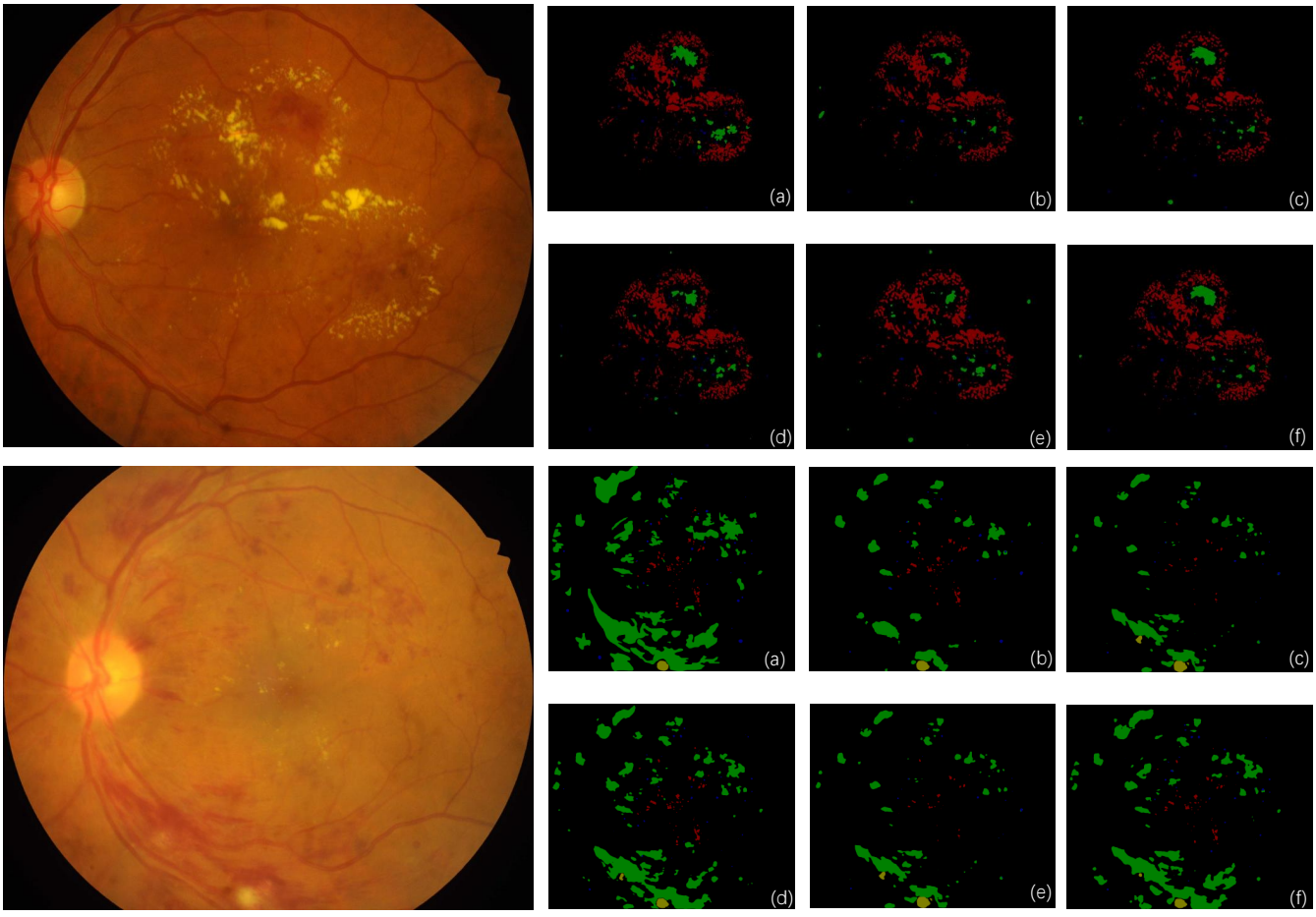

Fig. 7: Segmentation Images

[12] 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.

[13] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference," *Advances in neural information processing systems*, vol. 32, 2019.

[14] 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.

[15] 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, pro-*

ceedings, part III 18. Springer, 2015, pp. 234–241.

[16] 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.

[17] 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.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[19] 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.