Diabetic Retinopathy Image Segmentation Method Based on Fusion DenseNet and U-Net Network

Yaxuan Zhang, Yang Xu, Jinlong Zhai

Abstract—**Diabetic Retinopathy (DR) is a common and significant complication in patients with diabetes, and severely affecting their quality of life. Image segmentation plays a crucial role in the early diagnosis and treatment of DR. However, traditional methods are limited in terms of segmentation accuracy and generalization capability. This paper proposes a novel image segmentation method for diabetic retinopathy based on the integration of DenseNet and U-Net networks. Firstly, DenseNet is utilized to replace the encoder part of U-Net. The dense connection mechanism enhances the efficiency of feature propagation and improves the feature extraction capability of the encoder. Secondly, we introduce Omni-Dimensional Dynamic Convolution (ODConv) to replace traditional convolutions. ODConv is used to handle diverse input features effectively. The model's adaptability and segmentation accuracy for various samples is improved. Finally, we integrate the Convolutional Block Attention Module (CBAM) into the decoder. By leveraging channel and spatial attention, the ability to capture key features is enhanced, and the segmentation accuracy and boundary recognition capabilities are improved. To validate the effectiveness and feasibility of the model. This paper conducts experiments on the DDR dataset and the IDRID dataset. The experimental results show that the proposed method improved the mDice score by 0.53% compared to traditional methods. It also increased the mIoU score by 1.12% over traditional methods. This indicates that the proposed method has better performance in segmenting diabetic retinopathy images. It also demonstrates better generalization ability.**

Index Terms—**Diabetic retinopathy, DenseNet, U-Net, ODconv Dynamic convolution, CBAM attention mechanism**

I. Introduction

With the advancement of national economies, improvements in living standards, and changes in dietary patterns, the incidence of diabetes has risen significantly in recent years. Diabetic Retinopathy (DR) is a leading cause of vision impairment and blindness among diabetic patients, severely impacting their quality of life. The primary cause of DR is microvascular damage to the retina due to prolonged high blood sugar, leading to retinal hemorrhages, edema, and neovascularization [1]. If diabetic retinopathy is not treated in time, it can lead to severe vision loss or even blindness. Therefore, early and accurate diagnosis and effective treatment of diabetic retinopathy are very important [2]. In the diagnosis of diabetic retinopathy, image segmentation technology can help doctors quickly and accurately identify and quantify lesion areas, providing important references for treatment plans [3].

However, DR image segmentation faces significant challenges due to factors such as complex textures, uneven brightness and contrast, as well as subtle differences between lesion areas and normal tissues [4]. Thus, how to observe diabetic retinopathy in DR Images while saving doctors' time and effort is still an urgent need for computer-aided diagnosis in clinical practice.

Traditional methods for DR image segmentation usually rely on manually designed features and rules, leading to unstable segmentation results and poor generalization ability. These methods often fail to achieve the desired accuracy and robustness when dealing with complex DR images. In recent years, the rapid development of deep learning, particularly Convolutional Neural Networks (CNNs), has led to widespread applications in medical image segmentation, significantly improving DR image segmentation [5]. Deep learning-based methods make significant progress in the segmentation of diabetic retinopathy images.

Among these methods, U-Net [6] and its variations are widely used for image segmentation tasks. U-Net, with its encoder-decoder architecture, automatically learns features through end-to-end learning, overcoming many limitations of traditional methods and offering stronger image processing and segmentation capabilities [7]. However, despite the excellent performance of the U-Net network in medical image segmentation, it still has some shortcomings. For example, the bottleneck between the encoder and decoder in the U-Net network can limit its ability to handle the relationship between global and local features. The skip connections in U-Net, while preserving information, can also transfer a large amount of redundant information and noise to the decoder, leading to excessive information transfer and high memory consumption when processing large-scale images. Additionally, U-Net is not accurate enough in

Manuscript received July 10, 2024; revised December 17, 2024.

This work was supported by the National Natural Science Foundation of China(61775169), the Education Department of Liaoning Province (LJKZ0310) the Excellent Young Talents Program of Liaoning University of Science and Technology (2021YQ04).

Yaxuan Zhang is a postgraduate student of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (phone:86-13358900923; e-mail: 3238486193@qq.com).

Yang Xu is a Professor at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (corresponding author to provide phone: 86-13889785726; e-mail: xuyang_1981@aliyun.com).

Jinlong Zhai is a postgraduate student of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (phone:86-15184189808, e-mail: 927341159@qq.com).

handling image edges and details, which may result in blurred or incomplete segmentation results. To address these limitations, DenseNet [8] has been proposed, which improves network efficiency and feature representation through dense connections, feature reuse, and reduced parameters, leading to enhanced segmentation performance. Furthermore, the CBAM attention mechanism [9] combines channel attention and spatial attention. It can effectively regulate attention in both channel and spatial dimensions, and helping the network better capture important information and improving segmentation results. Based on this, this paper proposes a U-Net network (DenseCUNet) that integrates DenseNet and CBAM attention mechanisms for DR image segmentation.

The DenseCUNet model integrates the feature extraction capabilities of DenseNet with the image segmentation characteristics of U-Net, thereby enhancing the perception of target boundaries and details through the incorporation of the CBAM module. During training, we further improved the clarity of lesion area boundaries by modifying the loss function. This paper verifies the effectiveness of the DenseCUNet model on the DDR and IDRID datasets [10]. The experimental results show that the DenseCUNet model performs significantly better than other similar models. This result demonstrates the effectiveness and practicality of our model [11].

In summary, the main contributions and innovations of this paper are as follows:

1) A novel diabetic retinopathy image segmentation network model, termed DenseCUNet, is proposed. This model enhances prediction accuracy while simultaneously reducing prediction time.
2) The ODconv dynamic convolution network [12] is used to replace the traditional convolutional neural network, enabling it to dynamically select suitable convolution kernels according to the input content, further addressing the issue of differences among DR fundus images from different samples.
3) The CBAM attention mechanism (channel attention mechanism and spatial attention mechanism) is added to the original model to focus on important information and suppress irrelevant information.
4) The feasibility and effectiveness of the DenseCUNet model are verified through extensive experiments.

The following content will proceed as follows: The second part will provide an overview of other frameworks similar to the one proposed in this paper. The third part will explain the proposed model, including its structure and parameters. The fourth part will design experiments and conduct experiments on two real datasets to compare with other semantic segmentation models to verify the practical effectiveness and feasibility of the proposed method. Finally, the fifth part will summarize the work and look forward to future research directions and goals, especially in the field of diabetic retinopathy image segmentation.

II. Materials And Methods

A. Deep Learning-Based Diabetic Retinopathy Image Segmentation Methods

Deep learning-based methods for diabetic retinopathy image segmentation typically rely on data-driven supervised learning techniques. These methods can automatically learn features and predict vessel locations to some extent. In 2006, Hinton and his team [16] introduced the concept of deep learning and proposed several deep learning models. Compared to traditional algorithms, deep learning models possess the ability to autonomously learn and extract features, eliminating the subjectivity and instability associated with manual operations, while improving the robustness and accuracy of vessel image segmentation. In 2015, Wang et al. [17] proposed a method for retinal layer segmentation, which integrates image preprocessing, convolutional neural network (CNN)-based feature extraction, and random forest ensemble classification. That same year, Ronneberger et al. proposed the U-Net model, a fully convolutional network (FCN) specifically designed for image segmentation tasks.

U-Net consists of an encoder and a decoder. Its innovation lies in the introduction of skip connections, which link high-resolution feature maps from the encoder to low-resolution feature maps in the decoder. This allows U-Net to integrate information across multiple scales [18].

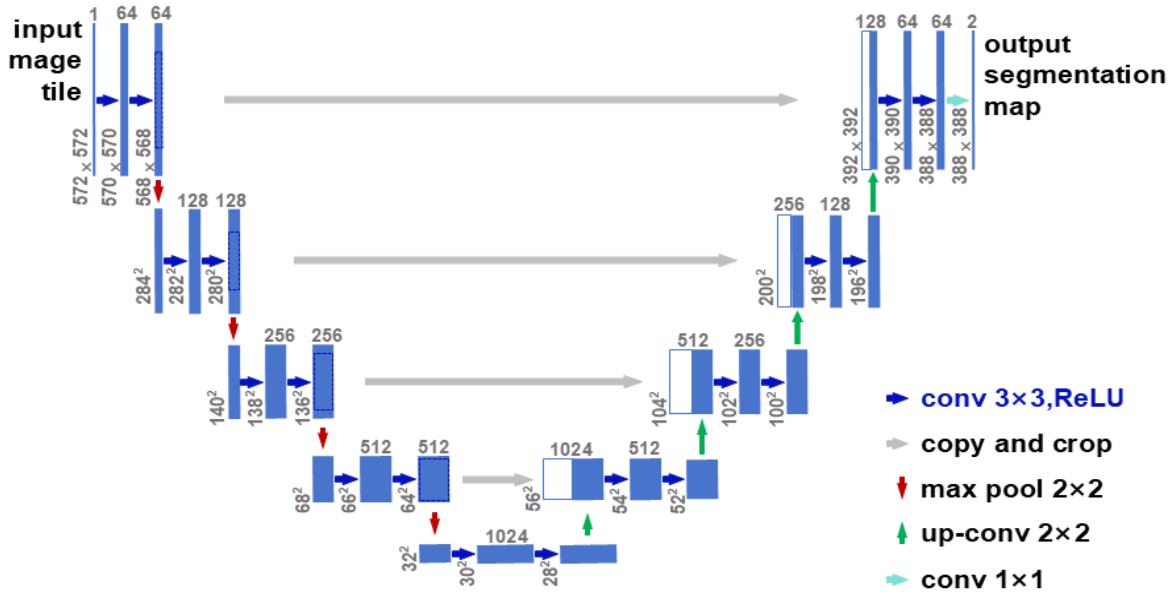

Fig. 1. U-Net Structure

The encoder progressively reduces the size of feature maps while extracting high-level features from the image. The encoder is composed of multiple convolutional and pooling layers. The convolutional layers capture local features of the image, while the pooling layers perform downsampling to reduce the size of the feature maps. The decoder maps these high-level features back to the original image size, aiming to recover as much detail as possible. U-Net employs skip connections to link feature maps from the encoder to corresponding feature maps in the decoder, which helps integrate information and recover fine details. The decoder typically consists of transposed convolutional layers (also known as deconvolution layers) and additional convolutional layers. The transposed convolutional layers are used to upsample the feature maps, while the skip connections merge the upsampled feature maps with their corresponding encoder maps, enhancing the integration of information and improving detail recovery. As shown in Figure 1, the architecture is symmetrical, with both the encoding and decoding paths forming a U-shaped structure, hence the name U-Net.

This structural design effectively integrates information across different scales, addressing the issues faced by traditional CNNs, such as the need for a large volume of labeled data for training and the high cost associated with annotating medical image data. By establishing feature fusion channels at different scales between the encoder and decoder, U-Net better captures both global and local features of the image. As such, it is particularly well-suited for medical image segmentation tasks with limited data annotation.

Despite its straightforward architecture and widespread use in medical image segmentation, U-Net may face challenges related to information loss when processing complex images. This issue stems from the traditional U-Net structure's insufficient mechanisms within the decoder to recover detailed information, particularly in images with intricate structures or textures. As a result, the accuracy of segmentation results may be compromised in such cases.

III. The Proposed Methods

In order to improve the performance of U-Net, address the issue of gradient vanishing, and enhance segmentation accuracy, this study focuses on two primary optimizations: optimizing the backbone network and adjusting the loss function.

Despite U-Net's excellent performance in image segmentation tasks, it tends to overfit when data is insufficient or classes are imbalanced (referenced from U-Net applications and challenges). And U-Net's downsampling and upsampling processes may lead to resolution loss and information loss. DenseNet effectively utilizes features through dense connections, enhancing

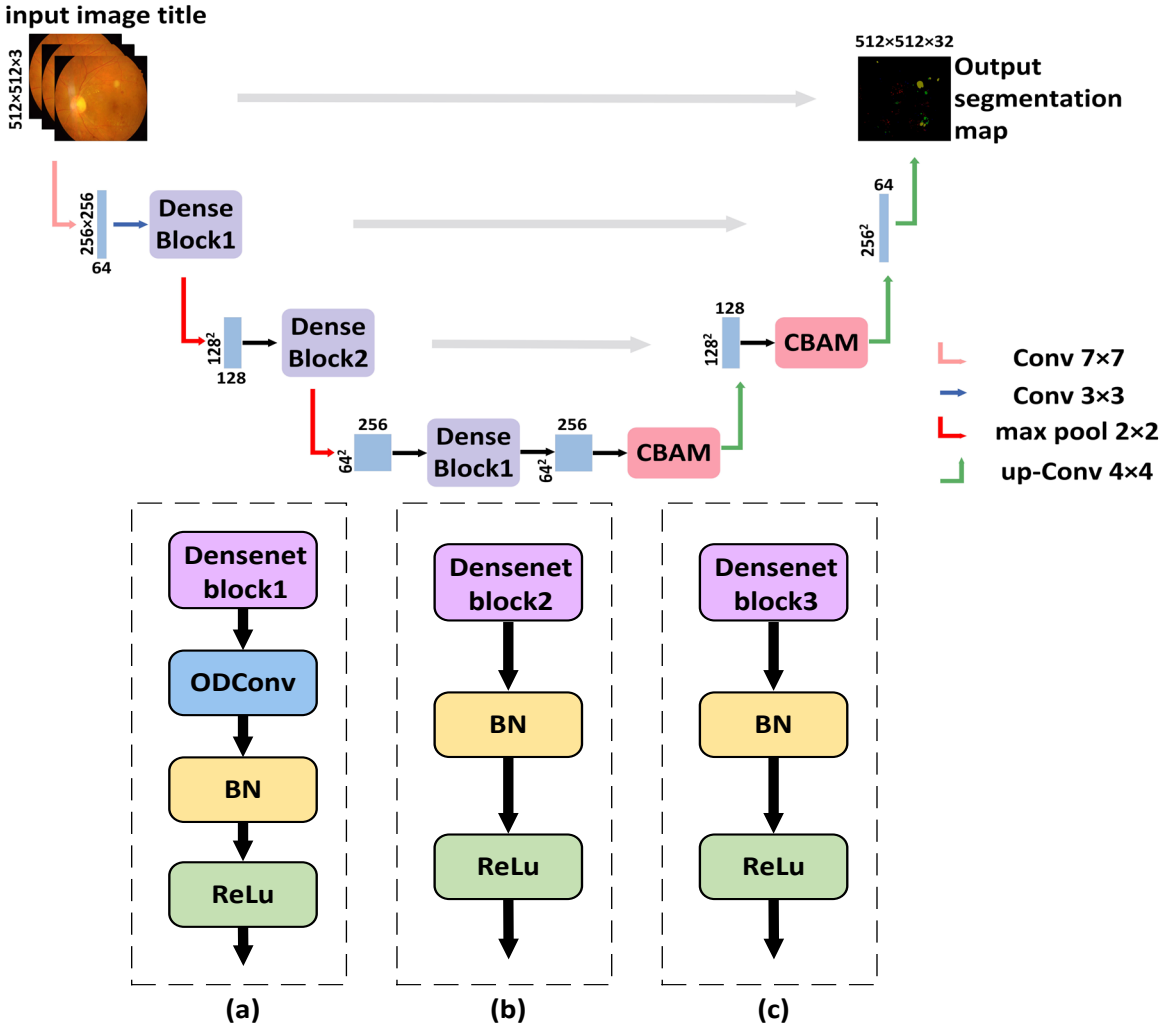

Fig. 2. DenseCUNet Structure

information flow and capturing subtle features in images. Therefore, it performs well on complex images and large datasets. However, DenseNet has issues with too many parameters and high computational costs. It increases training time and resource requirements. Therefore, this paper introduces DenseNet dense connections into the U-Net encoder to balance data volume and training time.

Part one is optimization the backbone network. The backbone network as the foundation of computer vision tasks, and its main function is to extract and output the essential features. The proposed model integrates parts of DenseNet and U-Net architectures. Each layer in DenseNet's dense blocks connects to all previous layers, ensuring better information flow and smoother gradient propagation. Additionally, using Omni-Dimensional Dynamic Convolution (ODConv) instead of traditional convolution dynamically selects convolution kernels based on input content, addressing the variability in DR fundus image samples and enhancing segmentation and detection accuracy. To improve focus on identifying lesion areas and suppress irrelevant information, this paper introduces an enhanced channel and spatial attention mechanism (CBAM) module after the dense blocks. Channel attention weights feature channels via global pooling, while spatial attention applies attention at different feature positions, improving long-term dependency capture and prediction accuracy.

Part two is adjustment the loss function. The loss function is adjusted to a hybrid loss function combining Dice loss, Weighted Binary Cross-Entropy, Structural Similarity (SSIM), and Shape-aware Loss. This adjustment effectively enhances the boundary clarity of lesion areas.

The improved DenseCUNet structure is shown in Figure 2, where (a), (b), and (c) are dense encoding modules Dense block n (n=1, 2, 3). ODConv represents dynamic convolution layers, BN represents batch normalization layers, and ReLU represents activation layers.

A. Optimizing the Backbone Network

1) *DenseNet Network:* DenseNet is a deep learning neural network structure initially. It is used for image classification tasks, but later it is also successfully applied to medical image segmentation tasks. In previous studies, when dealing with complex situations such as blurred lesion boundaries and uneven illumination. Gao Huang et al. proposed a network architecture called DenseNet in 2017. It utilizes a densely connected design that enables full utilization of feature reuse, thereby enhancing the model's feature extraction capabilities.

As shown in Figure 3, DenseNet121 achieves feature reuse through dense connections. In traditional convolutional neural networks, each layer only connects to the next layer. In DenseNet121, each layer connects to all subsequent layers. This dense connection structure allows each layer to receive feature information from all previous layers, enhancing feature propagation and utilization, as well as the representation of complex images, thereby improving network performance and accuracy. DenseNet121's dense connections allow features to be reused multiple times. It alleviates gradient vanishing and making the network easier to train. The dense connections also promote parameter sharing, reducing the network's parameter count and overfitting risk, particularly when data is insufficient, thus improving generalization ability.

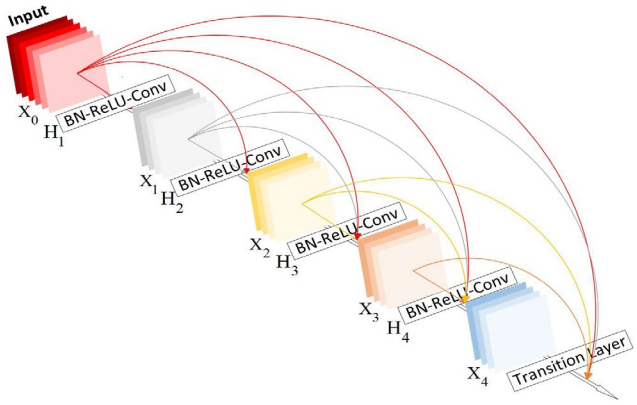

Fig. 3. DenseNet121 Structure

The advantage of the DenseNet network is its excellent performance in handling features at different scales. It can captures both local details and global structures in images. Therefore, this paper chooses to integrate the DenseNet network with the U-Net network as our backbone network for feature extraction.

2) ODConv Module: ODConv (OMNI-DIMENSIONAL DYNAMIC CONVOLUTION) was introduced by Chao Li et al. in 2022. This dynamic convolution algorithm adjusts convolution kernel parameters based on the content of input features. This dynamic tuning mechanism enables the network to generate a unique convolution kernel for each input sample. Therefore, it can extract more important features in the segmentation task with higher precision. The ODConv module first uses a lightweight neural network, often called the "kernel generator". It can be used to analyze the global information of the input feature maps, and generate specific kernel parameters based on this information. These dynamically generated kernel parameters are then used for convolution operations to extract richer and more targeted feature representations. The dynamic convolution formula can be expressed as:

$$y = \sum_{K=1}^{K} \alpha_k W_k \times x \tag{1}$$

Where y is the feature map obtained through the dynamic convolution operation, α_k represents the weights assigned to each convolution kernel by the kernel generator, W_k represents the specific convolution kernels generated for each input sample by the "kernel generator" neural network, K represents the number of convolution kernels, and x represents the original feature map entering the ODConv module.

As shown in Figure 4, W_i is the convolution kernel, α_{si} is the convolution parameters of each filter in the spatial position, α_{ci} represents assigning different scalars to the c_{in} channels of each convolution filter W_i^m, α_{fi} is to assign different scalars to the convolution filter, α_{wi} assigns scalars to the entire convolution kernel.

Where (a) means the different attention values are assigned to convolution parameters in spatial position, (b) means that different attention values are assigned to convolution filters in different input channels, (c) means that different attention values are assigned to convolution filters in different output channels, and (d) means that different values are assigned to n global convolution kernels.

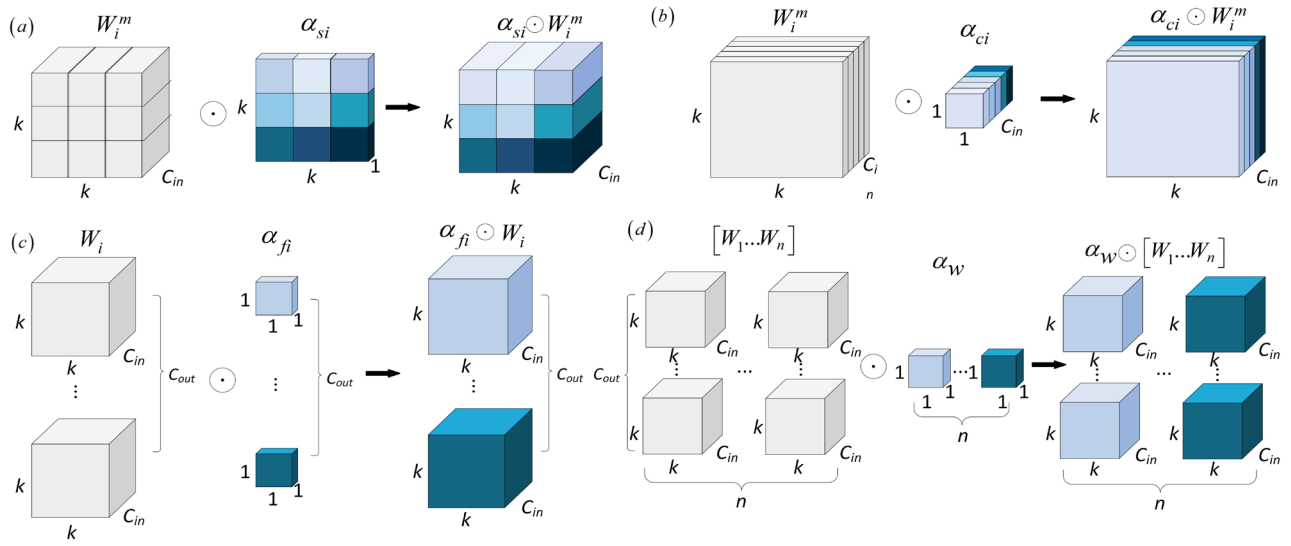

Fig. 4. ODConv Diagram

In the DenseCUNet model, the ODConv module is integrated into each dense block of DenseNet. The input feature map for each dense block is initially processed by the ODConv module, which performs feature extraction using a dynamically adjusted convolution kernel. Subsequently, the extracted features are fused with the original features to produce the output of the dense block. This design not only capitalizes on DenseNet's strengths in feature reuse and information flow but also significantly enhances the model's capability to discern subtle differences between lesion areas and normal tissue in diabetic retinopathy images through ODConv's dynamic feature extraction mechanism.

On one hand, the introduction of the ODConv module enhances the model's adaptability to the diverse range of input images, thereby enabling it to more effectively process DR images captured under varying conditions and from different devices. On the other hand, the dynamic nature of ODConv allows the model to adjust its behavior adaptively in response to task requirements and data characteristics. Consequently, the ODConv module achieves significant performance improvements without imposing an excessive computational burden.

3) *CBAM Attention Mechanism:* The Convolutional Block Attention Module (CBAM), introduced by Jeonghee Choo et al. in 2018, has garnered considerable attention within the deep learning community. Distinct from prior attention mechanism models, CBAM sequentially integrates channel attention (CAM) and spatial attention (SAM) modules. This sequential integration enables the model to concentrate more effectively on pertinent regions of interest while simultaneously suppressing irrelevant information.

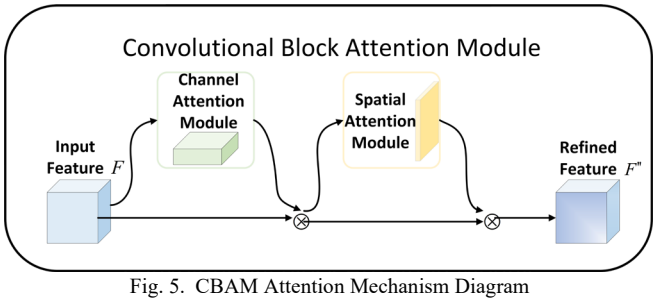

Fig. 5. CBAM Attention Mechanism Diagram

The CBAM attention mechanism is employed to enhance data prediction, particularly by emphasizing the significance of both channel and spatial dimensions in information processing. The integration of CBAM markedly improves the model's capacity to capture long-term dependencies, thereby increasing the accuracy of predictions. The computation process of CBAM sequentially generates a one-dimensional channel attention map and a two-dimensional spatial attention map from the input feature map, as illustrated in Figure 5:

CBAM first receives a given intermediate feature map as input. Then it is sequentially derives to obtain attention $A_c \in \mathbb{R}^{L\times1\times1}$ and spatial attention maps $A_s \in \mathbb{R}^{1\times W\times C}$. The calculations are as follows:

$$F' = A_c(F) \otimes F \tag{2}$$

$$F'' = T_c(F') \otimes F' \tag{3}$$

In this process, $\otimes$ denotes element-wise multiplication, and the necessary attention values are replicated as required: channel attention values are extended across spatial dimensions and vice versa. Furthermore, represents the final refinement calculation. The detailed procedure for each computational step is illustrated in Figure 6.

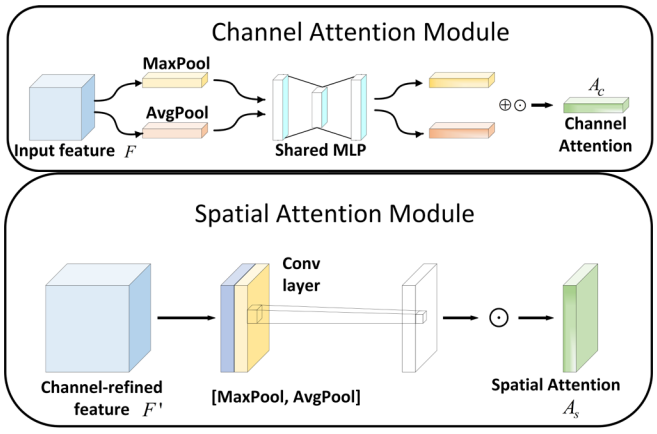

Fig. 6. Channel and Spatial Attention Mechanism Diagram

First, two distinct spatial context descriptors are generated by applying average pooling and max pooling operations to the feature map: one descriptor is derived from the features obtained through average pooling, while the other is based on

those acquired via max pooling. These descriptors are then fed into a shared network for the generation of the channel attention map. The shared network consists of a multi-layer perceptron (MLP) with a single hidden layer. To minimize the parameter count, the spatial dimensions of the input feature map are compressed using pooling operations. The calculation of channel attention is performed as follows:

$$\begin{aligned} A_c(F) &= \lambda(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \lambda(K_1(K_0(F_{avg}^c)) + K_1(K_0(F_{\max}^c))) \end{aligned} \tag{4}$$

Where λ represents the function *sigmoid* , F_{avg}^c represents the feature after average pooling, and $F_{\max}^c$ represents the feature after maximum pooling, $K_0 \in \mathbb{R}^{L/\alpha \times L}$, $K_1 \in \mathbb{R}^{L \times L/\alpha}$. Note that the *MLP* weights K_0 and K_1 are shared for both inputs.

We perform two types of pooling operations on the aggregated feature map, two 2D maps are generated: one representing the average pooled features in the channel $F_{\text{avg}}^s \in \mathbb{R}^{1 \times W \times C}$ and the other representing the max pooled features in the channel $F_{\max}^{\text{s}} \in \mathbb{R}^{1 \times W \times C}$. Based on these, a 2D spatial attention map is produced. The spatial attention calculation process is as follows:

$$\begin{aligned} A_{\text{s}}(F) &= \lambda(p^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \lambda(p^{7\times7}([F_{avg}^s; F_{\max}^s])) \end{aligned} \tag{5}$$

Where $A_{\text{s}}(F) \in R^{W \times C}$, λ represents a function of *sigmoid* and $p^{7\times7}$ represents a convolution operation with filter size 7×7.

The implementation of this method significantly improves the effectiveness of the application of attention mechanisms in deep learning models. This method enhances the ability of the model to identify and utilize key features through more precise information screening.

B. Optimizing the Loss Function

Currently, the majority of loss functions employed in medical image segmentation are based on the Dice loss function. However, it has been observed that the Dice loss function is sensitive to class imbalance, prone to boundary blurring, neglects structural information within images, and lacks shape constraints during the training process. To enhance both the accuracy and robustness of image segmentation, we propose a transition from traditional cross-entropy loss to a mixed loss function. This new approach integrates the Dice Loss function with Weighted Binary Cross-Entropy loss, Structural Similarity Index (SSIM), and Shape-aware loss.

1) *Dice Coefficient Loss:* The Dice loss function is extensively utilized to assess the similarity between the predicted segmentation map and the ground truth labels. It is particularly effective in addressing issues related to class imbalance.

In this paper, A and B are used to represent the prediction graph and label, respectively, and $\langle r, x, y \rangle$ is used to represent the channel, horizontal coordinate, and vertical coordinate of the pixel, respectively. The number of categories is represented by R , $1 \le r \le R$. H and W are represented as the height and width of the output image, respectively. Then the Dice loss function can be expressed by equation 6:

$$\begin{aligned} &L_{Dice}(A, B) \\ &= R - \sum_{r=1}^{R} \frac{2\sum_{x,y=1}^{H,W} A < r, x, y > B < r, x, y >}{\sum_{x,y=1}^{H,W} A^2 < r, i, j > + \sum_{i,j=1}^{H,W} B^2 < r, i, j >} \end{aligned} \tag{6}$$

2) *Weighted Binary Cross-Entropy Loss:* The weighted cross-entropy loss function builds upon the standard cross-entropy loss function by incorporating class-specific weights. This approach addresses issues arising from the uneven distribution of classes within the dataset. By assigning different weights to samples from various categories, this study enhances the ability to adjust the loss function's weight, thereby enabling the model to focus more on underrepresented classes. Consequently, this leads to an overall improvement in segmentation performance.

In this paper, y_i is used to represent the real label of the i category, and P_i is used to represent the probability of the i category predicted by the model, then the weighted cross entropy loss function can be expressed by equation 7:

$$L_{wce} = -\sum y_i \times \log(P_i) \tag{7}$$

3) *Structural Similarity Loss (SSIM Loss) :* To thoroughly address the structural information inherent in images, this paper incorporates Structural Similarity (SSIM) as an additional loss function. SSIM operates by maximizing the structural similarity between the predicted outcomes and the actual image. By evaluating the structural similarity between two images, SSIM aids in preserving both the structural integrity and spatial consistency of segmentation results.

In this paper, x and y are used to represent the prediction graph and label respectively, N represents the number of pixels, $K \ll 1$ is a constant, L represents the dynamic range of gray level, and $\overline{x_i}$ and $\overline{y_i}$ represent the average pixel value of position i respectively. Then the structural similarity loss function can be expressed by equation 8:

$$L_{SSIM} = \frac{\left[2\overline{x_i}\,\overline{y_i} + (K_1L)^2\right]\left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x_i})(y_i - \overline{y_i}) + (K_2L)^2\right]}{\left[\overline{x_i}^2 + \overline{y_i}^2 + (K_1L)^2\right]\left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x_i})^2 (y_i - \overline{y_i})^2 + (K_2L)^2\right]} \tag{8}$$

4) *Shape-aware Loss :* To enhance the accuracy of shape details in segmentation results, we incorporate a Shape-aware Loss function. This loss function specifically targets the optimization of segmentation contours and edge clarity. By aligning these contours more closely with the actual shape characteristics of the target, Shape-aware Loss significantly improves the model's ability to accurately recognize lesion boundaries.

In this paper, x and y are used to represent prediction graphs and labels respectively, N is used to represent the number of feature layers. $F_i(x)$ and $F_i(y)$ are used to represent the feature representation of layer i respectively. Then the shape perception loss function can be represented by equation 9:

$$L_{Shape} = \frac{1}{N}\sum_{i=1}^{N}\left(F_i(x) - F_i(y)\right)^2 \quad (9)$$

Based on the above discussion, our loss function effectively improves the performance and quality of DR image segmentation by integrating the advantages of multiple loss functions through proper weighting and adjustment. Therefore, the hybrid loss function in this paper can be expressed as Equation 10:

$$L = L_{Dice} + L_{wce} + L_{SSIM} + L_{Shape} \quad (10)$$

IV. Experiment

This experiment evaluates the segmentation performance of the enhanced DenseCUNet model on diabetic retinopathy (DR) fundus images. First, we compare the enhanced model with the original U-Net to assess the improvements made in DenseCUNet. Then, we conduct a comparison between DenseCUNet and other well-known deep learning-based semantic segmentation models. Finally, we present a comprehensive set of experimental results.

A. Datasets

In this experiment, two publicly available fundus image datasets, namely IDRID and DDR, have been selected to evaluate the performance of the proposed model. The detailed annotation information for both datasets is presented in Table I.

TABLE I
DSTASET STATISTICS

Dataset	Training set	Validation set	Test set	total
IDRID	40	14	27	81
DDR	383	149	225	757

The Indian Diabetic Retinopathy Image Dataset (IDRID) was acquired using a Kowa VX-10 alpha fundus camera, featuring a 50° field of view (FOV). This dataset contains fundus images specifically collected for the Indian population [13]. It is designed for various challenges, organized into three sub-tasks: segmentation, grading, and localization. The IDRID segmentation data subset consists of 81 images, each with a resolution of 4288 × 4288 pixels, annotated with pixel-level labels for Hard Exudates (HE), Soft Exudates (SE), Microaneurysms (MA), and Exudates (EX). Of these, 54 images are used for training, while 27 images are allocated for testing.

The DDR dataset, collected in China, contains 757 color fundus images obtained from various fundus cameras with a 45° FOV. The dataset provides extensive data with international standard annotations. Image resolutions range from 1380×1382 to 2736×1824 pixels, and lesion labels are provided for each image. Among these images, 383 are used for training, 149 for validation, and 225 for testing.

From the data presented in Table I, it is evident that the DDR dataset offers more annotated images than the IDRID dataset. However, the IDRID dataset has more uniform characteristics, with all images having the same resolution (4288 × 4288 pixels), while the DDR dataset features images with varying resolutions. Nevertheless, the DDR dataset provides a wider range of lesion severities, offering better diversity for training.

B. Experimental Platform

The experiment was conducted on a server with the following specifications: Intel Core i7-11700 CPU, NVIDIA GeForce GTX 3070 GPU, and 8 GB of RAM.

The deep learning framework used is TensorFlow. The model was trained and evaluated on the two datasets without any pre-initialization. Prior to input into the model, all images underwent preprocessing. The Adam optimizer was employed to facilitate rapid convergence during model training. A vector-based adaptive adjustment strategy was implemented based on the batch size, with a maximum iteration limit set at 100. The initial learning rate was set to 1e-3, dynamically adjusted via LearningRateScheduler. The momentum parameters for the Adam optimizer were configured at 0.9 and 0.999 via beta_1 and beta_2, respectively. A weight decay parameter of 1e-6 was applied to prevent overfitting.

C. Evaluation Metrics

To assess the segmentation performance of the model proposed in this paper on the IDRID and DDR datasets, we employ the most commonly utilized evaluation metrics in the domain of semantic segmentation: Dice coefficient and Intersection over Union (IoU). The definitions of these metrics are as follows:

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (11)$$

$$IoU = \frac{TP}{TP + FN + FP} \quad (12)$$

Where TP, FP, and FN represent the counts of true positives, false positives, and false negatives, respectively. It is important to note that some test images may not contain specific lesions. For evaluation, we consider each pixel as an individual case and treat the entire test set as a comprehensive dataset of pixels. The evaluation metrics were computed for all pixels in the dataset..

TABLE II
CO
MPARISON OF ABLATION EXPERIMENTAL DATA

Methods					Dice					IoU%				
UNet[5]	Densenet	ODConv	CBAM	Loss	EX	HE	SE	MA	mDice	EX	HE	SE	MA	mIoU
√	×	×	×	×	76.19	59.23	62.19	40.51	59.23	65.15	50.37	47.25	28.71	47.87
√	√	×	×	×	78.75	65.81	66.06	47.62	64.56	67.05	50.36	50.88	29.11	49.35
√	√	×	×	×	78.97	65.35	65.97	48.27	64.64	67.13	**51.15**	51.07	**32.77**	**50.53**
√	√	√	√	√	**80.19**	**65.99**	**67.66**	**48.88**	**65.68**	**67.69**	50.55	**51.27**	32.41	50.48

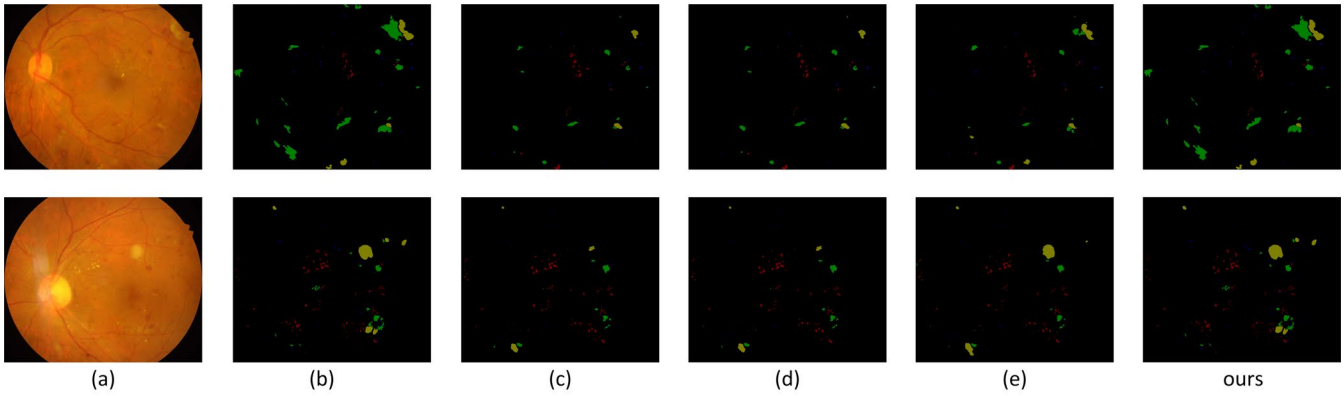

Fig. 7. Comparison of ablation results

D.Ablation Study

To validate the effectiveness of the DenseCUNet model and the optimized loss function, we conduct an ablation study. Table 2 presents the results comparing the contributions of various components, including the DenseNet backbone, ODConv dynamic convolution, and the CBAM module. And then shows the comparison between the experimental results and the segmentation renderings of other models. In order to investigate the individual contributions of the DenseNet backbone, ODConv module, and CBAM module to the segmentation performance of DR lesions, we evaluat key metrics—Dice and IoU—across four lesion types: microaneurysm (MA), hemorrhage (HE), hard exudate (EX), and soft exudate (SE). Each component is incrementally integrated into the model to assess its impact on segmentation accuracy. We conduct experiments by progressively adding each component to the baseline model. The results indicate that each module—DenseNet, ODConv, and CBAM—contribut significantly to the model's overall segmentation ability, improving performance to varying extents. The detailed comparison of segmentation performance for each component is summarized in Table II. This table highlights how the inclusion of each module improved both Dice and IoU scores for all lesion types :

The results presented in Table II indicate that the integration of DenseNet into the U-Net architecture significantly enhanced segmentation accuracy across all lesion types, particularly in addressing complex structures such as hard exudates (EX). The incorporation of ODConv further improved the model's adaptability to diverse lesion shapes and sizes by dynamically adjusting convolutional kernels. Additionally, the CBAM module provided notable advancements in boundary detection, especially for small and faint lesions, by refining attention mechanisms on both spatial and channel dimensions. Collectively, these enhancements resulted in an increase of 1.12% and 0.98% in mean Dice and IoU scores, respectively, when compared to the baseline U-Net model.

As shown in Figure 7, (a) is the input image, (b) is the ground truth image, (c) is the result of U-Net, (d) is the result of DenseNet integration in the backbone network, (e) is the result of adding the ODConv module in the integrated network, and ours are the result of adding the CBAM attention mechanism in the network. From the results in (b), it can be clearly seen that the segmentation performance after integrating DenseNet with U-Net is not very good. This is because the skip connection of DenseNet network introduces too many parameters, so that the detail information is not well recovered. However, in terms of the effect comparison of (c), (d), (e), and ours, DenseCUNet has effectively solved this problem. In the images of segmented lesion areas, we utilize blue to indicate MA lesions, green for HE lesions, red for EX lesions, and yellow for SE lesions. This method of color coding can assist medical professionals and researchers in distinguishing between different types of lesions more easily, thus aiding in further analysis and diagnosis. Through the labeling and categorization of these images, a more precise understanding of the patient's condition can be obtained, providing valuable references for personalized treatment plans. Additionally, employing color coding in medical imaging is advantageous for data visualization and communication as it allows for quick comprehension of image content and effective communication among professionals. In summary, utilizing specific colors to represent various types of lesions in segmented lesion area images is a common yet effective practice that equips doctors, researchers, and patients with essential information to intuitively comprehend the patient's condition and make informed treatment decisions clearly.

TABLE III
PERFORMANCE COMPARISON OF METHODS ON IDRID DATA SET

Methods	Dice					IoU%				
	EX	HE	SE	MA	mDice	EX	HE	SE	MA	mIoU
UNet++[23]	79.12	50.37	57.98	41.47	57.23	65.46	33.67	40.83	26.16	41.53
Deeplabv3+[24]	76.25	57.36	65.57	39.62	58.17	61.14	36.68	46.55	24.71	42.27
UNeXt[25]	73.67	51.53	58.96	30.41	53.64	58.32	34.71	41.81	17.93	38.19
TransUnet[26]	80.04	62.81	**68.82**	45.82	64.37	65.67	48.17	**54.09**	27.44	48.84
Swin-base[27]	79.71	65.19	68.19	48.79	64.53	67.26	48.36	48.54	30.86	48.76
M2MRF-D[28]	79.97	64.88	67.50	48.26	65.15	66.62	48.04	50.98	31.81	49.36
Ours	**80.19**	**65.99**	67.66	**48.88**	**65.68**	**67.69**	**50.55**	51.27	**32.41**	**50.48**

TABLE IV
PERFORMANCE COMPARISON OF METHODS ON DDR DATA SET

Methods	Dice					IoU%				
	EX	HE	SE	MA	mDice	EX	HE	SE	MA	mIoU
UNet++[23]	47.31	29.54	37.72	25.19	34.94	29.64	16.62	21.37	14.83	21.62
Deeplabv3+[24]	58.59	37.97	41.83	25.40	40.95	41.44	23.44	26.46	14.55	26.47
UNeXt[25]	50.69	28.93	31.42	14.90	31.48	33.98	16.93	18.62	8.09	19.40
TransUnet[26]	56.62	47.81	40.45	24.56	42.36	39.50	31.42	25.35	14.01	27.57
Swin-base[27]	59.79	**50.53**	46.77	23.31	45.10	42.64	**33.82**	30.62	13.19	30.07
M2MRF-D[28]	**61.15**	45.29	**48.02**	27.81	45.57	**44.04**	29.28	**31.60**	16.15	30.27
Ours	59.43	49.25	46.46	**28.98**	**46.03**	41.81	31.81	30.66	**17.36**	**30.41**

E. Comparative Experiments

To assess the effectiveness of the method proposed in this paper, we conducte a comparative analysis between our model and several mainstream semantic segmentation models using the DDR dataset and IDRID dataset. The results of the comparison for the IDRID dataset are presented in Table III, while those for the DDR dataset are shown in Table IV. The best-performing results within each table are highlighted in bold:

In contrast, based on the results from the IDRID dataset, the model struggles to fully learn the features of SE lesions due to their limited sample size and relatively low representation within the dataset. This limitation has led to insufficient emphasis being placed on this category by the model. Consequently, although our model does not achieve optimal performance for SE lesions, it still outperforms the second-best network—the M2MRF network—demonstrating improvements in Dice and IoU scores of 0.16% and 0.29%, respectively.

Similarly, the results obtained from the DDR dataset further substantiate the superior segmentation performance of outcomes for MA lesions, demonstrating improvements of 1.17% and 1.21% in Dice and IoU scores, respectively, when compared to the second-best network—the M2MRF network. This enhancement underscores the model's capability to effectively capture small and subtle characteristics of lesions.

Moreover, despite the inherent challenges posed by exudate (EX), hemorrhage (HE), and soft exudate (SE) lesions—characterized by their limited sample sizes and uneven distribution within the DDR dataset—our model consistently maintains superior segmentation performance across all lesion categories. In particular, it exhibits robust generalization capabilities, achieving an average Dice score of 46.03% and an IoU score of 30.41% for EX, HE, SE, and MA lesions combined. These findings highlight the model's adaptability and precision, especially in addressing more complex and underrepresented types of lesions.

To demonstrate the superiority of DenseCUNet, Figure 8 provides a typical example for visual comparison with alternative semantic segmentation models. In this example, microaneurysms (MA) are represented in blue, exudates (EX) in red, hemorrhages (HE) in green, and soft exudates (SE) in yellow. It is clear that DenseCUNet excels in detecting lesion areas, capturing complex details, and producing clear boundaries. Comparative analysis shows that our model performs significantly better than other models in identifying and segmenting minor lesions.

V. SUMMARY AND PROSPECT

In this paper, we propose an enhanced U-Net architecture for the segmentation of diabetic retinopathy (DR) lesions. DR is one of the most prevalent complications for diabetic patients, severely affecting their quality of life. Early and accurate diagnosis of DR lesions is crucial for preventing vision loss and providing effective treatment. Despite significant advancements in deep learning-based DR lesion segmentation, challenges remain, such as considerable variation in lesion shape and size across samples, as well as the presence of numerous small lesions that are difficult to segment accurately.

To overcome these challenges, we have developed an improved model that incorporates several advanced techniques: the ODConv dynamic convolution replaces traditional convolutions, DenseNet is integrated as the backbone network for enhanced feature extraction, and a novel residual structure is introduced to improve information flow and gradient propagation. Additionally, the Convolutional Block Attention Module (CBAM) is employed to reduce information loss, especially in small lesions, by applying attention mechanisms to both channel and spatial dimensions.

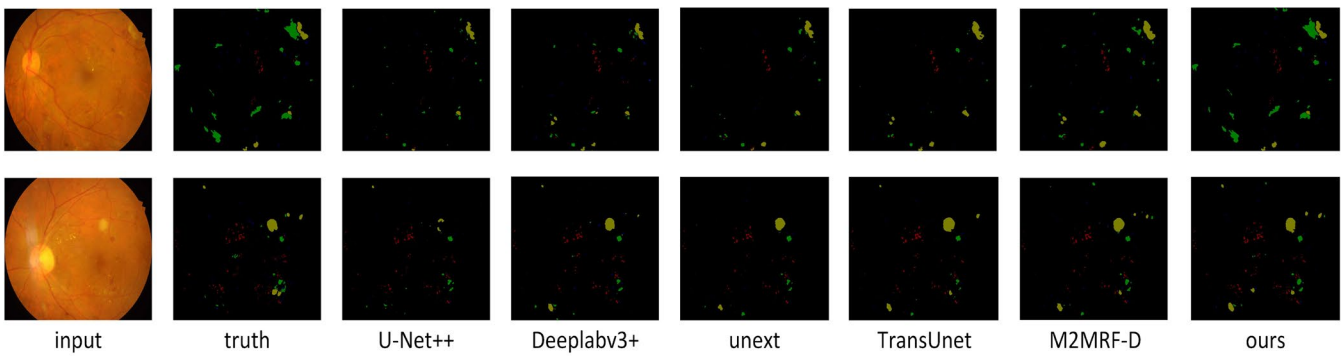

Fig. 8. Comparison of experimental results

Our experimental results demonstrate that the proposed DenseCUNet model significantly improves the segmentation accuracy of DR lesions compared to existing methods. Specifically, the model effectively handles complex lesions, small lesions, and images with varying resolutions. The combination of DenseNet, ODConv, and CBAM provides a robust solution for segmentation tasks, showing a marked improvement in both Dice coefficient and Intersection over Union (IoU) scores across multiple datasets, including IDRID and DDR.

While our method significantly advances the state of DR lesion segmentation, there are several avenues for future research. Firstly, further validation studies are essential to confirm the robustness and generalizability of the DenseCUNet model across a broader range of datasets and clinical environments. It will be important to test the model on other fundus image datasets that include diverse populations, varying levels of DR severity, and images captured with different imaging devices.

Additionally, addressing the issue of class imbalance and improving segmentation performance on underrepresented lesion categories, such as Soft Exudates (SE), will be crucial for refining the model's ability to handle diverse lesion types. Techniques such as class-weighted loss functions, data augmentation, or semi-supervised learning could be explored to improve performance on such cases.

Another interesting research direction is extending the use of DenseCUNet to other medical imaging tasks. The methodologies implemented in this paper could potentially be adapted to segment lesions in other types of medical images, such as lung nodules in chest X-rays, brain tumors in MRI scans, or skin lesions in dermoscopic images. The flexibility of the DenseCUNet architecture, especially with its attention mechanisms and dynamic convolutions, makes it a promising candidate for a range of medical image segmentation challenges.

Furthermore, integrating multi-modal imaging data, such as combining optical coherence tomography (OCT) images with fundus photography, could enhance segmentation accuracy and provide a more comprehensive diagnosis of DR and other retinal diseases. Exploring the use of multi-task learning frameworks, where multiple tasks (such as segmentation, classification, and lesion detection) are performed simultaneously, could also further improve the model's performance and generalization across different clinical settings.

Finally, real-time segmentation for clinical application remains a critical challenge. Optimization techniques, such as model quantization, pruning, or hardware acceleration (e.g., using FPGAs or TPUs), could be explored to improve inference speed without compromising accuracy. This would make the model more suitable for practical use in medical environments, enabling faster diagnosis and treatment planning for diabetic retinopathy patients.

REFERENCE

[1] J. W. Y. Yau et al. "Global Prevalence and Major Risk Factors of Diabetic Retinopathy," Diabetes Care, vol. 35, no. 3, pp. 556–564, Mar. 2012.

[2] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[3] Zengqiang Yan, Xin Yang, Kwang-Ting Cheng. (2024, June). A Three-Stage Deep Learning Model for Accurate Retinal Vessel Segmentation IEEE journal of biomedical and health informatics. (Online). 23(4). pp. 1427 – 1436. Available: https://ieeexplore.ieee.org/abstract/document/8476171.

[4] M. Niemeijer, B. Van Ginneken, J. Staal, M. S. A. Suttorp-Schulten, and M. D. Abramoff. "Automatic detection of red lesions in digital color fundus photographs," IEEE Trans. Med. Imaging, vol. 24, no. 5, pp. 584–592, May 2005.

[5] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." arXiv, May 18, 2015. Accessed: May 29, 2024. (Online). Available: http://arxiv.org/abs/1505.04597.

[6] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, vol. 9901, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., in Lecture Notes in Computer Science, vol. 9901, Cham: Springer International Publishing, 2016, pp. 424–432.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely Connected Convolutional Networks," 2017 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR, pp. 2261–2269, Jul. 2017.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. "CBAM: Convolutional Block Attention Module." arXiv, Jul. 18, 2018. (Online). Accessed: May 29, 2024. Available: http://arxiv.org/abs/1807.06521.

[9] P. Porwal et al. "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research," Data, vol. 3, no. 3, p. 25, Jul. 2018.

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

[11] C. Li, A. Zhou, and A. Yao. "Omni-Dimensional Dynamic Convolution." arXiv, Sep. 16, 2022. (Online). Accessed: May 29, 2024. Available: http://arxiv.org/abs/2209.07947.

[12] Q. Li, J. You, and D. Zhang. "Vessel segmentation and width estimation in retinal images using multiscale production of matched filter responses," Expert Syst. Appl., vol. 39, no. 9, pp. 7600–7610, Jul. 2012.

[13] M. Zardadi, N. Mehrshad, and S. M. Razavi. "Unsupervised Segmentation of Retinal Blood Vessels Using the Human Visual System Line Detection Model," vol. 4, no. 2, 2016, pp. 125–133.

[14] S. A. A. Shah, T. B. Tang, I. Faye, and A. Laude. "Blood vessel segmentation in color fundus images based on regional and Hessian features," Graefes Arch. Clin. Exp. Ophthalmol., vol. 255, no. 8, pp. 1525–1533, Aug. 2017.

[15] C. A. Lupascu, D. Tegolo, and E. Trucco. “FABC: Retinal Vessel Segmentation Using AdaBoost,” IEEE Trans. Inf. Technol. Biomed., vol. 14, no. 5, pp. 1267–1274, Sep. 2010.

[16] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks," Science, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[17] S. Wang, Y. Yin, G. Cao, B. Wei, Y. Zheng, and G. Yang. "Hierarchical retinal blood vessel segmentation based on feature and ensemble learning," Neurocomputing, vol. 149, pp. 708–717, 2015.

[18] Y. Li, X. Zhang, and L. Liu. "LiU-Net: Ischemic Stroke Lesion Segmentation Based on Improved KiU-Net," vol. 32, no. 2, pp.369-378, 2024.

[19] F. Milletari, N. Navab, and S.-A. Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation." arXiv, Jun. 15, 2016. (Online). Accessed: May 30, 2024. Available: http://arxiv.org/abs/1606.04797.

[20] S. Xie and Z. Tu. "Holistically-nested edge detection," in Proceedings of the IEEE international conference on computer vision, (Online). 2015, pp. 1395–1403. Accessed: Jul. 02, 2024. Available: http://openaccess.thecvf.com/content_iccv_2015/html/Xie_Holistically-Nested_Edge_Detection_ICCV_2015_paper.html.

[21] Z. Wang, E. P. Simoncelli, and A. C. Bovik. "Multiscale structural similarity for image quality assessment," in The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Ieee, (Online). 2003, pp. 1398–1402. Accessed: Jul. 02, 2024. Available: https://ieeexplore.ieee.org/abstract/document/1292216/.

[22] J. Liu, C. Desrosiers, and Y. Zhou. "Semi-supervised Medical Image Segmentation Using Cross-Model Pseudo-Supervision with Shape Awareness and Local Context Constraints," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, vol. 13438, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., in Lecture Notes in Computer Science, vol. 13438, Cham: Springer Nature Switzerland, 2022, pp. 140–150.

[23] Z. Zhou, M. R. Siddiquee, N. Tajbakhsh, and J. Liang. "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support 4th

Int. Workshop DLMIA 2018 8th Int. Workshop ML-CDS 2018 Held Conjunction MICCAI 2018 Granada Spain S, 2018.
[24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." arXiv, Aug. 22, 2018. Accessed: Jun. 27, 2024. (Online). Available: http://arxiv.org/abs/1802.02611.
[25] J. M. J. Valanarasu and V. M. Patel. "UNeXt: MLP-based Rapid Medical Image Segmentation Network." arXiv, Mar. 09, 2022. (Online). Accessed: Jun. 27, 2024. Available: http://arxiv.org/abs/2203.04967.
[26] J. Chen et al. "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation." arXiv, Feb. 08, 2021. (Online). Accessed: May 29, 2024. Available: http://arxiv.org/abs/2102.04306.
[27] Z. Liu et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002.
[28] Q. Liu, H. Liu, W. Ke, and Y. Liang. "Automated lesion segmentation in fundus images with many-to-many reassembly of features," Pattern Recognit., vol. 136, p. 109–191, Apr. 2023.