



CARNet: Cascade attentive RefineNet for multi-lesion segmentation of diabetic retinopathy images

Yanfei Guo¹ · Yanjun Peng^{1,2}

Received: 5 June 2021 / Accepted: 17 December 2021 / Published online: 4 January 2022
© The Author(s) 2021

Abstract

Diabetic retinopathy is the leading cause of blindness in working population. Lesion segmentation from fundus images helps ophthalmologists accurately diagnose and grade of diabetic retinopathy. However, the task of lesion segmentation is full of challenges due to the complex structure, the various sizes and the interclass similarity with other fundus tissues. To address the issue, this paper proposes a cascade attentive RefineNet (CARNet) for automatic and accurate multi-lesion segmentation of diabetic retinopathy. It can make full use of the fine local details and coarse global information from the fundus image. CARNet is composed of global image encoder, local image encoder and attention refinement decoder. We take the whole image and the patch image as the dual input, and feed them to ResNet50 and ResNet101, respectively, for downsampling to extract lesion features. The high-level refinement decoder uses dual attention mechanism to integrate the same-level features in the two encoders with the output of the low-level attention refinement module for multiscale information fusion, which focus the model on the lesion area to generate accurate predictions. We evaluated the segmentation performance of the proposed CARNet on the IDRiD, E-optha and DDR data sets. Extensive comparison experiments and ablation studies on various data sets demonstrate the proposed framework outperforms the state-of-the-art approaches and has better accuracy and robustness. It not only overcomes the interference of similar tissues and noises to achieve accurate multi-lesion segmentation, but also preserves the contour details and shape features of small lesions without overloading GPU memory usage.

Keywords Fundus image · Diabetic retinopathy · Multi-lesion segmentation · RefineNet · Attention fusion

Introduction

Diabetic retinopathy (DR) is one of the major complications of diabetes and has become a leading cause of blindness. It has been reported that approximately 600 million people will suffer from diabetes, and one-third of them will have DR by 2040 worldwide [1,2]. DR is caused by diabetic microvascular disease, which is divided into three stages: blood vessel rupture and haemorrhage, release of growth factors and blood vessel obstruction. Common pathological features of

DR include hard exudate (EX), soft exudate (SE), microaneurysms (MA) and haemorrhage (HE) [3,4], as shown in Fig. 1. The International Clinical Diabetic Retinopathy Disease Severity Scale classifies DR into five stages, including normal, mild, moderate, severe and proliferative, based on lesion symptoms [5]. There were no obvious symptoms in the early stages of DR, but the severity gradually increased over time. At present, ophthalmologists manually observe lesions from fundus images for DR screening in real clinical applications. However, this method is not only labour intensive and time-consuming but also susceptible to the subjective factors of experts, which has difficulty ensuring detection accuracy. Therefore, it is particularly crucial to create an automatic lesion segmentation method for DR diagnosis.

Although great progress has been made in the task of lesion segmentation, it is still full of challenges. The main reasons are as follows. (1) The proportion of lesions (e.g., MA) in the high-resolution fundus image is so small that is easily confused as noises. (2) The structure of the lesion is complex, and various kinds of lesions have differences

✉ Yanjun Peng
pengyanjun@163.com

Yanfei Guo
guo_1992_fei@163.com

¹ College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, Shandong, China

² Shandong Province Key Laboratory of Wisdom Mining Information Technology, Shandong University of Science and Technology, Qingdao 266590, Shandong, China

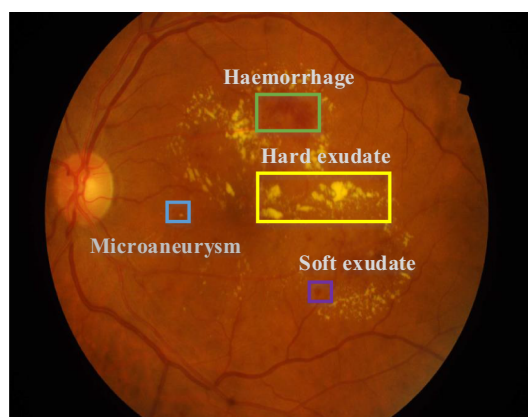


Fig. 1 Different types of DR lesions in fundus images

in shape, size, colour and brightness. (3) The colour, contour and texture of tissues on the retina (e.g., blood vessels and optic disc) are similar to those of lesions, which are prone to false-positive results. (4) The appearance of the retinal images varies due to differences in the camera system and light source intensities. (5) The tissue pigmentation in fundus images of patients of different races varies greatly, which increases lesion segmentation difficulty [6]. Specifically, compared with the fundus images of white people and yellow people, the retinal pigmented epithelium of black people contains more melanocytes, which forms the most outer layer of the retina. Compared with the light-colored retina, the darker retina obscures some lesions and vascular changes.

Recently, convolutional neural networks (CNNs) have become widespread in many fields of real life [7–11], numerous deep learning-based methods have been presented for lesion segmentation of DR. The existing methods [12–17] for lesion segmentation of DR are categorized into encoder–decoder structures and non-encoder–decoder structures. On the one hand, due to the high resolution of fundus images and GPU memory limitations, works [12–15] first cropped the original image into patches or resized and input them into U-Net and its variants for lesion segmentation. However, patch images lack global information, and the deconvolution operation cannot preserve the detailed information of small lesions, such as MA, which makes generating accurate predictions difficult. On the other hand, considering that different types of lesions are inconsistent in size and scale, algorithms [16,17] fed the full image to VGG or ResNet to extract the contextual information and used a 1×1 convolution operation to fuse multiscale feature maps. However, these methods cannot make full use of multilevel features, especially low-level features, such as edges. The main reason is that the strategy uses fixed fusion weights for multilevel information while ignoring the content differences between them. Therefore, efficiently fusing multiscale context fea-

tures without excessively consuming computing resources is crucial for accurate multi-lesion segmentation of DR images.

To address the above issues, this paper proposes a cascade attentive RefineNet (CARNet) for multiple lesion segmentation. CARNet adopts a dual-input encoder–decoder structure and trains in an end-to-end manner. The input is the whole image and the patch image, and the output is the segmentation result of four lesions. The proposed model includes the whole image encoder, patch image encoder and attention refinement decoder. First, the whole images and the patch images are sent to ResNet50 and ResNet101, respectively, for downsampling to extract features. Second, the deep features of the fourth residual block in ResNet50 and ResNet101 are simultaneously fed to the bottom-level attention refinement module (ARM) to fuse the global and local features. Third, the feature maps from the same level in the two encoders and the output of the previous ARM are fed to the current ARM to fuse multiscale lesion features. Finally, the fused feature maps of the top-level ARM are sent to the dense sigmoid layer to obtain the final lesion segmentation results. We evaluate the proposed CARNet on three public data sets, i.e., IDRiD [18], E-Ophtha [19] and DDR [20]. Experimental results show that our method has good robustness and accuracy. It not only overcomes the interference of similar tissue and noise but also reserves the fine details of the lesion area without overloading GPU memory usage. To the best of our knowledge, this is the first study to apply attentive RefineNet for multi-lesion segmentation from DR images.

The main contributions are summarized as follows:

- 1) To make full use of the coarse global information and fine local features in fundus images, this paper takes the full images and patches as dual inputs, and proposes a cascade attentive RefineNet (CARNet) to achieve multi-lesion segmentation of DR.
- 2) This paper proposes an attention refinement module (ARM) to fully integrate the multilevel context features of different lesions. It consists of a residual convolution unit (RCU), attention fusion block (AFB) and chained residual pooling (CRP). ARM uses fine-grained local features as auxiliary information to promote the reorganization of coarse-grained global information, which focuses the network on the lesion area.
- 3) We carry out comprehensive experiments on three fundus image data sets of IDRiD, E-Ophtha and DDR. Ablation studies show that the whole image and patch image as the dual input and attention fusion block contribute to improving the performance of lesion segmentation. Comparison experiments demonstrate that the proposed framework outperforms the state-of-the-art approaches and has better robustness and accuracy. It can effectively avoid the interference of other similar tissues and unde-

sirable noise for automatic and accurate multiscale lesion segmentation from retinal images.

The rest of this paper is organized as follows. “Related work” introduces related works on lesion segmentation. “Proposed method” describes the proposed approach in detail. “Experiments and results” presents the experimental data sets, implementation details and results analysis. “Discussion” discusses the advantages and limitations of the proposed framework. “Conclusion draws conclusions and suggests future work.

Related work

Recently, the task of lesion segmentation from DR images has attracted considerable attention from numerous researchers, and they have proposed several methods for the problem. The existing methods can be divided into traditional methods and deep learning methods based on the features used in lesion segmentation.

Traditional methods

Traditional methods for lesion segmentation are grouped into four categories: region growing methods [21], thresholding algorithms [22], mathematical morphology approaches [23] and machine learning-based methods [24].

Wu et al. [21] first preprocessed the original fundus image to make the MAs clearer, then used the region growing method to locate the MA candidate areas, and extracted dimensional features to feed to AdaBoost, Bayesian net and k-nearest neighbour (KNN) classifiers to segment MA from fundus images. Long et al. [22] combined fuzzy C-means clustering with a dynamic threshold to determine the candidate HE regions, extracted the texture features from fundus images, and finally fed them into the support vector machine (SVM) classifier for automatic HE detection. Colomer et al. [23] extracted granulometric profiles and local binary patterns (LBP) to calculate the morphological and texture features of the fundus images and then used Gaussian processing, random forest and SVM for EX, MA and HE segmentation on the DIARETDB1 [25] and E-Ophtha_EX [19] data sets. Amin et al. [24] first applied the Gabor filter for image enhancement, then extracted geometric and statistical features from the candidate lesion area, and finally adopted Bayesian net, KNN and SVM for EX detection on the DIARETDB1 and E-Ophtha EX data sets.

However, the performances of the above methods are easily limited by the brightness and contrast of the fundus images. Therefore, the robustness is poor and inefficient for meeting the need for clinical screening.

Deep learning methods

In recent years, deep learning algorithms [12–17,26–29] have shown outstanding performance and outperformed traditional methods in lesion segmentation. The existing methods can be classified into two types: encoder–decoder and non-encoder–decoder structure.

Zhou et al. [26] proposed a semi-supervised collaborative learning model based on attention mechanism to realize lesion segmentation and DR classification. They combined U-Net with the Xception module to generate segmentation masks for four kinds of lesions. Foo et al. [13] presented MTUnet for lesion segmentation and DR classification on the IDRiD and SiDRP14–15 [30] data sets. They replaced the encoder with a VGG16 network and passed the skip connection to the decoder before each max-pooling operation. Furthermore, transposed convolution instead of the usual upsampling operation is used in the decoder. Yang et al. [14] proposed a two-stage network for lesion detection (e.g., EX, HE and MA) and DR grading. The original fundus images are divided into patches overlapping grid and input into the local network to generate weighted lesion maps, and then the weighted lesion maps are fed to the global network to classify the DR severity. Sambyal et al. [15] developed an improved U-Net framework for MA and HE segmentation. The model adopts ResNet34 as the encoder and applies the upsampling operation of periodic shuffling convolution to realize the rapid convergence of the network. However, periodic shuffling with subpixel convolution is suitable for the super-resolution task. In summary, the deconvolution operation in the encoder–decoder framework cannot recover the low-level features lost after downsampling, which makes generating accurate high-resolution segmentation results difficult.

Guo et al. [16] modified VGG16 and proposed L-Seg, using a multiscale feature fusion approach and a multichannel bin loss function to address the issues of small lesions that are difficult to accurately segment and class imbalances, respectively. However, they adopted fixed weights for the fusion of multiple side output features. Mo et al. [17] proposed a fully convolutional residual network for EX segmentation. However, this model is only suitable for single lesion segmentation. Xue et al. [27] proposed a hybrid structure-based dynamic membrane system to segment HE and MA in the IDRiD, E-Ophtha and Messidor [31] data sets. Li et al. [20] established a new retinal image data set named DDR and applied holistically nested edge detection (HED) [32] and DeepLabv3+ [33] for EX, HE, MA, and SE segmentation. Tan et al. [28] adopted a ten-layer fully CNN to realize automatic and simultaneous segmentation of HE, MA and EX on the CLEOPATRA data set. However, this method only used pixels instead of the lesion area to evaluate the segmentation performance. Zhou et al. [34] introduced a new data

set named FGADR for DR segmentation and grading and used FCN, DeepLabv3+, U-Net and its variations for lesion segmentation.

The goal of this study was to create a lesion segmentation method from DR images that overcomes the limitations of traditional algorithms and deep learning approaches. Therefore, a cascade attentive RefineNet is proposed in this paper to realize automatic multi-lesion segmentation of DR images.

Proposed method

Problem formulation

Let $T = \{(X, Y)\}$ denote the training set, where $X = \{x_i, i = 1, \dots, m\}$ denotes the original fundus image, $Y = \{y_i, i = 1, \dots, m, y_i \in (0, 1)\}$ denotes the lesion label, and m denotes the number of images in the training set. Each original image corresponds to a maximum of four labels due to different types of lesions in fundus images. Before inputting the segmentation network, we crop each training image and the lesion mask (x_i, y_i) into n patches $p = \{(x_i^k, y_i^k), k = 1, \dots, n\}$. In this paper, we take the whole images and the patch images as dual inputs and train the segmentation model by minimizing the difference between the predictions and ground-truth masks. Our goal is defined as:

$$\min_{\eta} = \sum_{i=1}^m \sum_{k=1}^n L(C_{\eta}(x_i, x_i^k), (y_i, y_i^k)), \quad (1)$$

where C_{η} and η represent the lesion segmentation model and parameters, respectively, and L denotes the loss function.

Preprocessing

We perform necessary preprocessing on the original image for lesion enhancement and data augmentation before training the proposed network.

First, all images are uniformly resized to 1024×1024 due to inconsistent image size in different data sets, and the inner circle of retina is appropriately cropped and filled into squares to remove zero-pixel areas. Second, the original fundus images are cropped into patches with a resolution of 1024×1024 . Third, due to the differences in illumination and contrast between fundus images, we use contrast-limited adaptive histogram equalization (CLAHE) to enhance the contrast between the lesion area and the background, as shown in Fig. 2. Finally, data augmentation operations are performed on the original image because of the limited number of image samples. (1) Scaling randomly by a factor of $[0.8, 1.2]$ in a step of 0.1. (2) Rotating randomly within the range of $[0, 360^\circ]$ in a step of 60° . (3) Translating the orig-

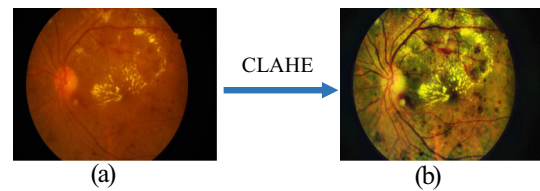


Fig. 2 Comparison of fundus image before and after preprocessing. **a** Original DR image. **b** DR image after CLAHE operation with channel separation

inal image from -50 pixels to 100 pixels in 30 -pixel steps vertically and horizontally.

Framework overview

The overall architecture of the proposed CARNet is shown in Fig. 3. The model adopts the encoder–decoder structure and consists of a global image encoder, a patch image encoder and an attention refinement decoder. The two encoders send the whole images and patches to ResNet50 and ResNet101, respectively, for downsampling to obtain feature maps of different resolutions. The decoder consists of five attention refinement modules (ARMs). ARM combines the output feature of the same level in the two encoders and low-level ARM for upsampling, and then passes them to the high-level ARM to fuse multiscale lesion feature. The ARM is composed of residual convolution unit (RCU), attention fusion block (AFB) and chain residual pooling (CRP). AFB uses local features as auxiliary information to promote the reorganization of global information, which focus the model on the lesion area. In this way, the integration of both fine-grained local information and coarse global information is beneficial to generate accurate lesion segmentation results.

Deep residual network

The whole image encoder $\text{Enc}_{\text{whole}}$ and patch image encoder $\text{Enc}_{\text{patch}}$ use ResNet50 and ResNet101, respectively, to extract multiscale contextual features from full images and patch images. The encoding process of $\text{Enc}_{\text{whole}}$ and $\text{Enc}_{\text{patch}}$ can be expressed as

$$\begin{aligned} F_w^1, F_w^2, F_w^3, F_w^4 &= \text{Enc}_{\text{whole}}(x_i), \\ F_p^1, F_p^2, F_p^3, F_p^4 &= \text{Enc}_{\text{patch}}(x_i^k), \end{aligned} \quad (2)$$

where F_w and F_p represent the encoding features of the whole image and the patch image in different hidden layers, respectively.

Both ResNet50 and ResNet101 [35] are composed of four residual blocks, and the resolution of the output image after each residual block is reduced to $1/4$ of the original image. The structure of residual block is shown in Fig. 4, which

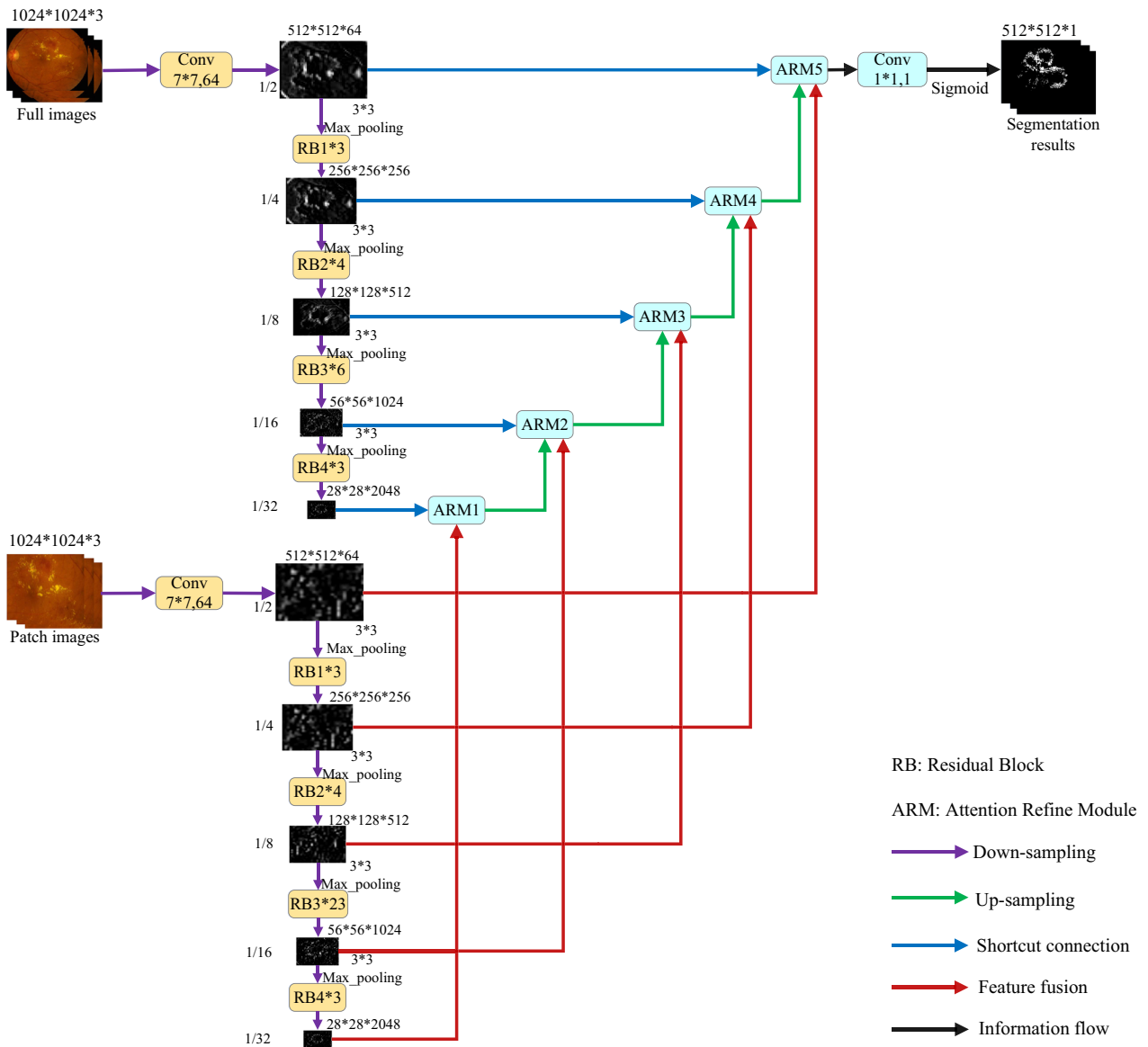


Fig. 3 Overall architecture of proposed CARNet framework

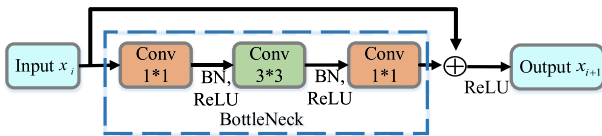


Fig. 4 Structure of residual block, both 1*1 and 3*3 indicate the kernel size of the convolution layers

includes three convolutional layers, two batch normalization (BN) layers and three rectified linear unit (ReLU) layers. ResNet introduces a residual block through skip connection to improve the information flow, and it consists of multiple shallow networks to speed up network convergence.

Attention refinement module

The decoding process of the attention refinement module is formulated as

$$\begin{aligned}
 S_j &= \text{Dec}_{\text{ARM}}(F_w^j, F_p^j), j = 1, \\
 S_j &= \text{Dec}_{\text{ARM}}(F_w^j, F_p^j, S_{j-1}), j = 2, 3, 4, 5,
 \end{aligned}
 \tag{3}$$

where S_j indicates the predictions of the j th ARM, and S_5 denotes the final segmentation results.

Figure 5 shows the framework of the proposed ARM, which includes RCU, AFB and CRP. The arrows in different colours represent input features of various types. First, the whole image features, the patch image features and the

Fig. 5 Overall framework of attention refinement module

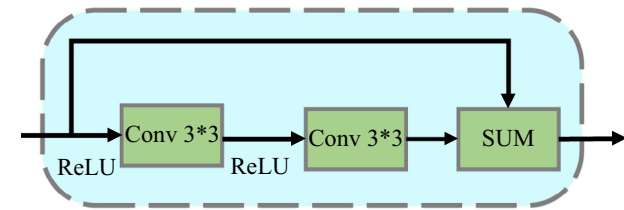
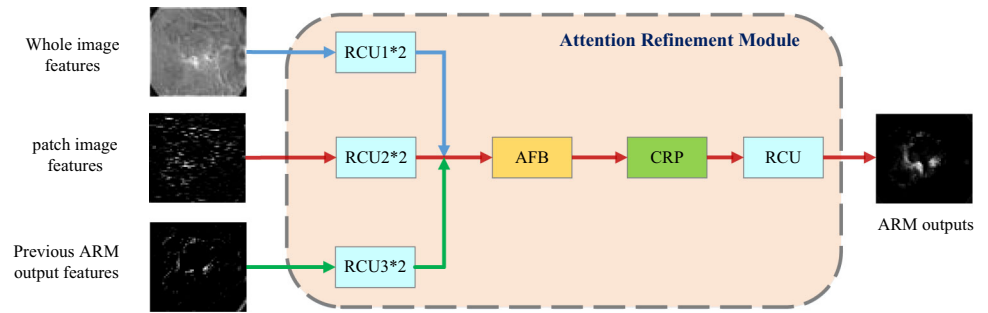


Fig. 6 Illustration of residual convolution unit

output features of the previous ARM are input into the RCU to extract multiscale features. Second, the feature maps of three RCUs are integrated into the AFB to locate the lesion area. Third, the fused attention features are fed to the CRP to capture contextual features in a larger image area. Finally, the outputs of the pooling operation are fed to the RCU to obtain the final predictions. Formally, the input of j th ARM is the features of full images x_i and patches x_i^k , as well as the output S_{j-1} of the previous ARM, then the predictions of ARM are calculated as follows:

$$M_{\text{fusion}} = \text{fusion}(x_i, x_i^k, S_{j-1}), \tag{4}$$

$$W_{\text{attention}} = W(\gamma_1 M_{\text{fusion}}^c + \gamma_2 M_{\text{fusion}}^s), \tag{5}$$

$$S_j = W_{\text{attention}} * M_{\text{fusion}} + S_{j-1}, \tag{6}$$

where M_{fusion} , M_{fusion}^c and M_{fusion}^s represent the fusion feature of the three inputs, channel attention feature and spatial attention feature, respectively. $W_{\text{attention}}$ and S_j denote the weight of the ARM and the output of j th ARM, respectively. γ_1 and γ_2 represent factors used to balance the weight of the channel attention feature and the spatial attention feature, respectively.

Residual convolution unit

The detailed structure of the RCU is illustrated in Fig. 6. It includes two ReLU and two convolution layers with a kernel size of 3×3 . Moreover, it uses addition to fuse the feature maps before and after the convolution. Compared with the original ResNet [35], the RCU removes the BN layer.

Attention fusion block

Different from the previous method using a fixed fusion weight [16,17], the proposed CARNet adopts a dual attention mechanism to adaptively learn different types of input features for information fusion. The architecture of the AFB is shown in Fig. 7, which consists of channel attention and spatial attention. This module can not only enhance specific semantic information to improve the dependence between channels, but also model rich contextual feature representation in space.

Channel attention: The fusion feature from the three types of input in the ARM is defined as $M_{\text{fusion}} \in R^{W \times H \times C}$, where W , H and C represent the width, height, and channel number of the feature map, respectively. First, the fusion features are reshaped to $M_{\text{fusion}}^{c1} \in R^{(W \times H) \times C}$ and $M_{\text{fusion}}^{c2} \in R^{C \times (W \times H)}$, respectively. Then we use matrix multiplication and softmax operation to obtain the channel attention map $t^c \in R^{C \times C}$. Finally, the final channel attention feature is computed by $M_{\text{fusion}}^c = M_{\text{fusion}} + \text{Reshape}(t^c M_{\text{fusion}}^{c3})$.

Spatial attention: We perform three convolution operations on the fusion feature map to obtain three branches. First, the channel dimension of feature maps from the first two branches is reduced to C/f to reduce the computational complexity. Second, we multiply the reshaped feature $M_{\text{fusion}}^{s1} \in R^{(W \times H) \times (C/f)}$ and $M_{\text{fusion}}^{s2} \in R^{(C/f) \times (W \times H)}$. Third, we use a softmax function to obtain the spatial attention map. Finally, the spatial attention feature is calculated by $M_{\text{fusion}}^s = M_{\text{fusion}} + \text{Reshape}(t^s M_{\text{fusion}}^{s3})$. In the experiment, we set f as 8, 16, 32 and 64 to adapt to the following feature maps of different scales $28 \times 28 \times 2048$, $56 \times 56 \times 1024$, $128 \times 128 \times 512$ and $256 \times 256 \times 256$. The final output of the ARM is the sum of weighted pixel-by-pixel addition of M_{fusion}^c and M_{fusion}^s , as well as the output S_{j-1} of the previous ARM.

To explore the relationship between different types of input features, we compare three fusion strategies of attention, addition and concatenation, and show the experimental results in the ablation study.

Addition fusion: It directly adds the corresponding channels of input features with different types. The computational complexity of this strategy is low, but the relationship

Fig. 7 Structure of attention fusion block. The pink arrows represent the convolution fusion block. The orange arrows represent the convolution operations. The orange arrows denote the reshape operations

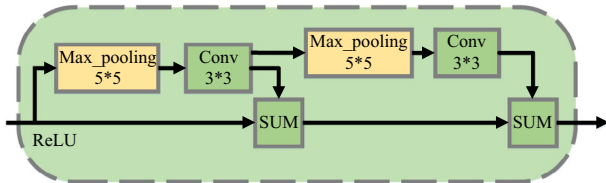
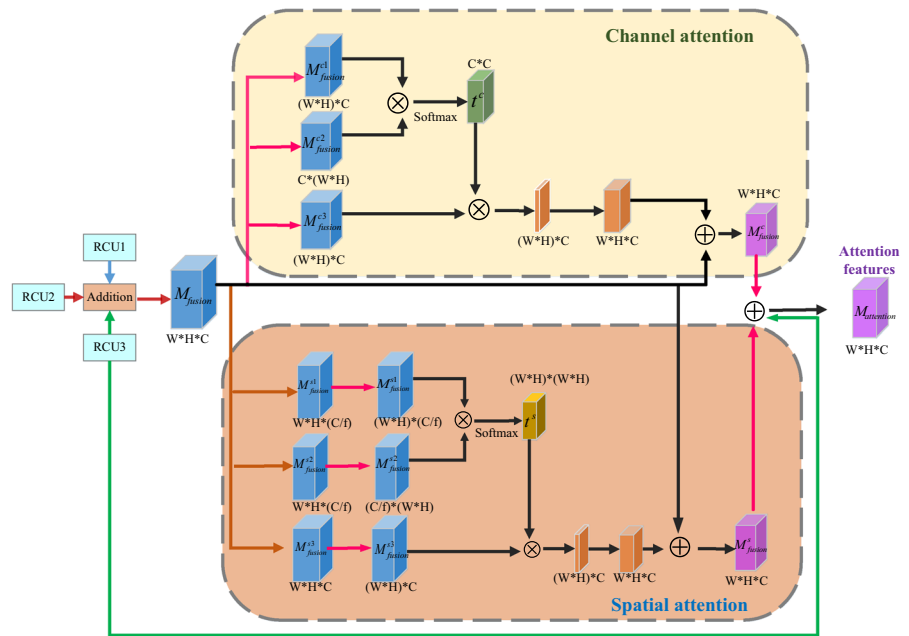


Fig. 8 Details of chained residual pooling

between features in different channels will be destroyed during the addition process, which causes information loss.

Concatenation fusion: It fuses different input features to increase the number of feature channels in terms of channel dimensions. In the concatenation operation, the contribution of each feature channel is the same.

Attention fusion: This method can automatically learn the attention weights between different types of input features, focusing on learning the features of the lesion area. We use fine-grained local features as auxiliary information to promote the reorganization of coarse global features, which makes the ARM focus on the lesion area.

Chained residual pooling

The specific structure of CRP is shown in Fig. 8. It is a chain composed of multiple pooling operations, which is similar to CRP of RefineNet [36]. Each pooling module contains a convolutional layer and a max-pooling layer. The next CRP takes the output of the previous CRP as input, so the current CRP can reuse the result of the previous pooling operation. Therefore, context features are captured from a larger background area without using a larger pooling window.

Loss function

The loss function is computed by the difference between the prediction and the label of each pixel in the retinal image. However, due to the serious imbalance between the lesion pixels and the normal pixels in the retinal image, the predictions tend to be more prone to healthy pixels if not considered the sample imbalance problem. Therefore, we use the weighted sum of the class-balanced binary cross-entropy loss and the Dice loss as the overall loss function of the proposed CARNet to solve the potential overfitting problem. Let $Y = \{y_i, i = 1, \dots, m, y_i \in (0, 1)\}$ denote the lesion label and $S = \{s_i, i = 1, \dots, m, s_i \in (0, 1)\}$ denote the prediction of the network, then the loss function is defined as

$$L(Y, S) = \lambda_1 L_{BCE}(Y, S) + \lambda_2 L_{Dice}(Y, S) \\ = \lambda_1 \cdot [-\beta y_i \log s_i - (1 - \beta)(1 - y_i) \log(1 - s_i)] \\ + \lambda_2 (1 - \frac{2 \langle y_i, s_i \rangle}{\|y_i\|_1 + \|s_i\|_1}), \tag{7}$$

where L_{BCE} and L_{Dice} represent the balanced cross-entropy loss and Dice loss function, respectively. λ_1 and λ_2 represent the factors used to balance L_{BCE} and L_{Dice} . β indicates that the weighting factor used to balance the difference between the lesion and normal pixel. $\langle y_i, s_i \rangle$ denotes the matrix dot product of the ground truth and the prediction results of each channel, and $\|\cdot\|_1$ denotes a norm, i.e., the sum of the absolute values of the matrix elements.

Table 1 Overview of experimental data set

Data set	Total Size	HE	MA	SE	EX	Healthy	Country
IDRiD	81	80	81	40	81	0	India
E-ophtha_EX	82	–	–	–	47	35	France
E-ophtha_MA	381	0	148	–	–	233	France
DDR	757	601	570	239	486	–	China

Experiments and results

Data sets

We evaluate the effectiveness of our CARNet on three fundus image data sets of IDRiD [18], E-Ophtha [19] and DDR [20]. The general information of the data sets is shown in Table 1.

- 1) IDRiD: This data set includes 81 fundus images from India with a resolution of 4288×4288 . Each fundus image has corresponding pixel-level labels of HE, SE, MA and EX. In the experiment, 54 images were used for training, and 27 images were used for testing.
- 2) E-ophtha: This data set includes 463 retinal images of French, and the resolution of images ranges from 1440×960 to 2544×1696 . It contains two subsets of E-ophtha_EX and E-ophtha_MA. E-ophtha_EX contains 82 original fundus images and corresponding EX labels, including 47 EX images and 35 healthy images. E-ophtha_MA contains 381 fundus images and corresponding ground truth mask, which includes 148 MA images and 233 healthy images.
- 3) DDR: This data set contains 757 colour fundus photographs of Chinese individuals. Each image corresponds to the manufactured labels of four kinds of lesions. The resolution of images ranges from 1380×1382 to 2736×1824 . In the experiment, 383 images were used for training, 149 images were used for validation, and 225 images were used for testing.

Implementation details

All experiments were performed on an Ubuntu 18.04 system with an NVIDIA GeForce RTX 2080Ti graphics card with 11 GB of RAM. The framework was implemented based on the PyTorch platform. We used ResNet pretrained on ImageNet [37] as the backbone of the encoder. We applied the stochastic gradient descent (SGD) method to train the model for rapid convergence. Moreover, we used the multilearning strategy to update the learning rate. The batch size was set to 4, and the initial learning rate, momentum and weight decay were set to 10^{-4} , 0.9 and $2e^{-3}$, respectively. The learning rate was the

initial learning rate multiplied by $(1 - \frac{\text{iter}}{\text{max_iter}})^{\text{power}}$, where the power was 0.9, and the maximum number of iterations was 300. The balance parameters λ_1 , λ_2 and β in the loss function were set to 0.6, 0.4 and 0.7, respectively. The balance factors γ_1 and γ_2 in the weight of the attention fusion module are set to 0.3 and 0.7, respectively.

Evaluation metrics

We adopted sensitivity (Sen), specificity (Spe), accuracy (Acc) and Dice to evaluate the segmentation performance of various networks. The evaluation metrics are calculated as follows:

$$\text{Sen} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (8)$$

$$\text{Spe} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad (9)$$

$$\text{Acc} = \frac{1}{2}(\text{sen} + \text{spe}), \quad (10)$$

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

Specifically, Sen denotes the proportion of real lesion pixels classified as lesion pixels. Spe shows the percentage of actual healthy pixels correctly classified by the network. Acc represents the ratio of the number of correct predictions to the total number of predicted pixels. Dice refers to the similarity extent between the prediction of the algorithm and the ground truth. In Eqs. (8)–(11), TP is true positive, which indicates the number of lesion pixels that are correctly classified. FP is false-positive, which denotes the number of normal pixels misclassified as lesion pixels. TN is true negative, which shows the number of normal pixels that are correctly predicted by the network. FN is false negative, which represents the number of lesion pixels misclassified as healthy pixels.

Moreover, we adopted the receiver operating characteristic (ROC) curve to visualize the performances of different approaches for lesion segmentation. The AUC value represents the area under the ROC curve. The higher the AUC value is, the better the model works.

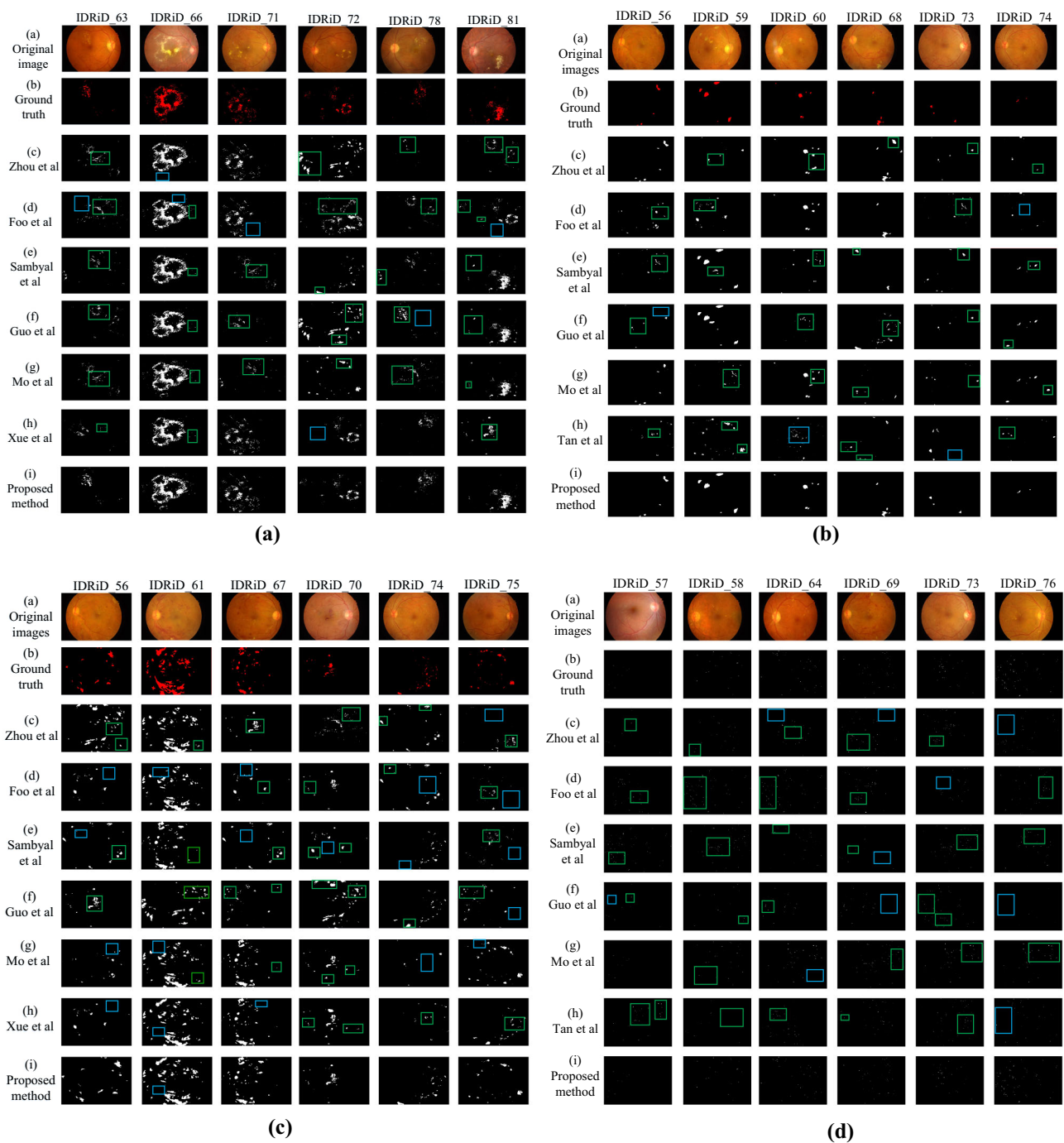


Fig. 9 Multi-lesion segmentation performance on the IDRiD data set. The green and blue box represent false positives and false negatives, respectively. **a** EX results. **b** EX results. **c** HE results. **d** MA results

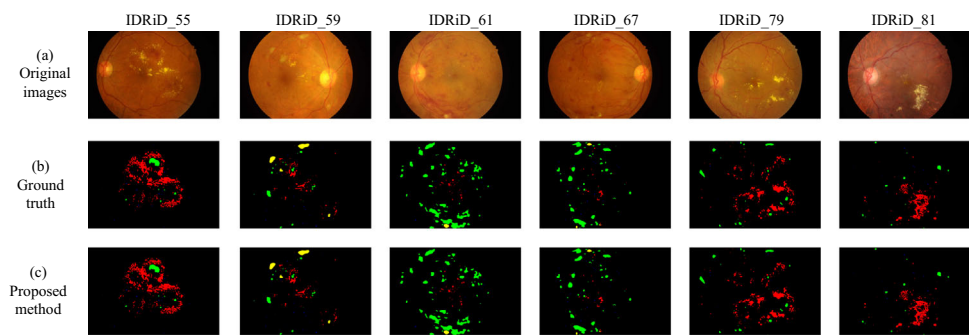
Comparisons with state-of-the-art methods

Performance on IDRiD

We compared the proposed approach to other state-of-the-art methods in Fig. 9, including Zhou et al. [26], Foo et al. [13], Sambyal et al. [15], Guo et al. [16], Mo et al. [17] and Tan

et al. [28]. As shown in Fig. 9, there were lesions of inconsistent size and other noises in the fundus images, which increased the difficulty of simultaneous segmentation of the four types of lesions. Zhou et al. [26] segmented large EX and SE regions accurately but segmented small and inconspicuous MA and HE areas poorly. The main reason is that this method uses dilated convolution to extract features for

Fig. 10 Segmentation results of four lesions on IDRiD data set



each channel, which introduces coarse feature subsampling results in losing the detailed information of small lesions. The approach of Foo et al. [13] had difficulty preserving the fine boundary details of the HE region, which leads to the coarse contour of segmentation results. Moreover, this method easily confuses MA and small HE. The main reason is that the scale and shape between some MA and HE regions are similar, which makes accurately distinguishing the two kinds of lesions difficult. Although Sambyal et al. [15] segmented EX and SE accurately, it is easy to misclassify some large and bright cotton wool spots as HE and ignore small HE. Guo et al. [16] could not accurately detect tiny EXs in high-contrast fundus images, because they cannot distinguish EXs from other bright lesions, such as drusen and artefacts. Mo et al. [17] sometimes misclassified cotton wool spots as SE, leading to false-positive results in SE segmentation. The main reason is that this model is only suitable for segmentation of a single lesion and does not consider the variety in scale between different lesions. Xue et al. [27] easily confused EX and SE, which indicates that the method has poor performance in the segmentation of similar lesions. Tan et al. [28] distinguished the obvious EX and SE, but it was easier to ignore the tiny MA. The main reason is that this method performs lesion segmentation based on pixel features while not considering the context information of the whole image.

Compared with other methods, the proposed CARNet can not only segment large-scale EX and SE regions but also retain details of small-scale MA and HE. It has good robustness, which can effectively overcome the interference of reflective vessels, artefacts and drusen noise. The main reason is that we use the whole images and the patch images as the dual input of the network. First, ResNet is used for down-sampling to extract rich features, and then the ARM is used for up-sampling to fuse multilevel contextual information. The ARM combines spatial attention and channel attention, which can not only enhance specific semantic information to improve the dependence between different channels but also model rich contextual feature representations in space, making the network focus more on lesions areas. Although multiscale lesions respond at different levels of the network, the ARM can adaptively learn different types of input features

and perform effective fusion to generate more accurate lesion segmentation results. In summary, the proposed method can make full use of coarse global features and fine local features to realize accurate segmentation of multiple lesions.

Furthermore, Fig. 10 displays the four types of lesion segmentation results of the proposed method for better visualization. We use four colours to denote the different kinds of lesions. Red indicates EX, yellow indicates SE, blue indicates MA, and green indicates HE. As shown in Fig. 10, the proposed approach can overcome the interference of similar tissues (e.g., intertwined blood vessels and optic discs) and other lesion noises (e.g., cotton wool spots and drusen), and segment lesions of various scales and shapes from complex fundus images accurately.

Table 2 quantitatively compares the four lesions segmentation performance of our method and seven other state-of-the-art methods on the IDRiD data set. It is observed from Table 2 that the Acc and Dice of our method in segmenting EX, SE, MA and HE reached 0.9823/0.9752/0.9257/0.9537 and 0.9782/0.9642/0.9183/0.9478, respectively. Compared with Foo et al. [13], the Acc and Dice of EX, SE, MA, and HE are increased by 8.06%/6.28%/12.71%/13.23% and 8.89%/12.15%/16.68%/18.86%, respectively. Experimental results on the IDRiD data set demonstrate the superiority of CARNet over other competing approaches on multi-lesion segmentation.

In addition, Table 3 shows the quantitative comparison results between the proposed approach and the top five teams in the IDRiD competition on average-precision (AP) scores. AP represents the area under the precision–recall curve. In general, the higher the AP value is, the better the segmentation model. Table 3 shows that the proposed CARNet ranks third, first, first, and third in the segmentation of EX, SE, MA, and HE, respectively. Although the top five teams cannot segment the four lesions simultaneously, they are better at segmenting a single lesion. Compared with the other five state-of-the-art methods, the proposed method uses an end-to-end architecture to simultaneously segment the four lesion areas in the fundus image and performs better in the segmentation of small lesions.

Table 2 Quantitative comparison results of different methods on IDRiD

Methods	EX				SE				MA				HE			
	Sen	Spe	Acc	Dice	Sen	Spe	Acc	Dice	Sen	Spe	Acc	Dice	Sen	Spe	Acc	Dice
Zhou et al. [26]	0.9438	0.9475	0.9673	0.9563	0.9426	0.9582	0.9467	0.9413	0.7852	0.7348	0.7933	0.7508	0.8452	0.7894	0.8261	0.7593
Foo et al. [13]	0.8573	0.8562	0.9017	0.8893	0.8697	0.8453	0.9124	0.8427	0.7314	0.7259	0.7986	0.7515	0.8169	0.8273	0.8214	0.7592
Sambyal et al. [15]	0.9335	0.9482	0.9574	0.9681	0.9452	0.9613	0.9487	0.9546	0.8458	0.9017	0.8945	0.8728	0.8126	0.7985	0.8263	0.8171
Guo et al. [16]	0.8165	0.7613	0.8564	0.8971	0.7869	0.8154	0.8352	0.8793	0.6172	0.5463	0.5249	0.4657	0.6374	0.6581	0.6128	0.5987
Mo et al. [17]	0.9630	0.9304	0.9408	0.9517	0.8697	0.9371	0.9384	0.9267	0.8152	0.7986	0.9126	0.9042	0.8563	0.8347	0.7698	0.9156
Xue et al. [27]	0.7791	0.9675	0.9731	0.9652	0.7723	0.9753	0.9723	0.9534	0.6723	0.6586	0.7544	0.8096	0.7214	0.7639	0.8067	0.9231
Tan et al. [28]	0.8758	0.9873	0.8632	0.8539	0.8147	0.8263	0.8759	0.9246	0.4606	0.9563	0.8542	0.7697	0.6257	0.9893	0.9452	0.9137
Proposed	0.9627	0.9852	0.9823	0.9782	0.9457	0.9685	0.9752	0.9642	0.8936	0.9328	0.9257	0.9183	0.9382	0.9521	0.9537	0.9478

The best results are shown in bold

Figure 11 shows the ROC curves of the five methods in EX, SE, MA and HE segmentation. The closer the ROC curve is to the upper-left boundary, the more accurate the training model is. It can be seen in Fig. 11a, b that the ROC curve of the proposed CARNet is the upper-left curve of the five models, and the curve of Guo et al. [16] is the lowest curve of the five curves. The data in the lower right corners of Fig. 11c, d show that the AUC value of the proposed method is the largest, Sambyal et al. [15] is second, and Guo et al. [16] is the smallest.

Performance on E-ophtha

To demonstrate the effectiveness of our method, we compared the proposed framework with recent state-of-art methods on the E-Ophtha data set in Fig. 12. As shown in Fig. 12, Zhou et al. [26] detected a large EX but ignored inconspicuous MA. The main reason is that compared with the size of the input image, MA is so small that it is easier to focus on other parts of the input image for the model. The EX and MA contours of the method proposed by Foo et al. [13] are very coarse. The main reason is that the deconvolution operation cannot restore the low-level visual features lost after downsampling, which makes it difficult to generate accurate high-resolution segmentation results. The method presented by Sambyal et al. [15] more easily misclassifies reflective vessels as EX, which results in false-positive results. The approach proposed by Guo et al. [16] is susceptible to interference from other lesion noises and thus cannot distinguish EX from artefacts and drusen. The main reason is that this method uses VGG to extract the contextual information and uses a 1 × 1 convolution operation to fuse multiscale feature maps. This strategy uses fixed fusion weights for multilevel features, but it ignores the content differences between them. Xue et al. [27] misclassified cotton wool spots as EX, leading to false-positive results in EX segmentation.

The experimental results in Fig. 12 demonstrate the superior performance of our method over the other five methods [13,15,16,26,27] on DR image lesion segmentation. It can overcome the adverse effects of optical artefacts from the anatomical structure and vascular reflections, as well as the light intensity and shape of the lesion area, and segment the multiscale lesions from the fundus image accurately. Moreover, it can better retain the contour details of EX and MA.

Table 4 shows the quantitative results of the two lesions segmentation on the E-ophtha data set with different methods. It is observed from Table 4 that our method is superior to other methods on most metrics, and Acc and Dice in segmenting EX and MA reach 0.9746/0.9489 and 0.9761/0.9513, respectively. Compared with Guo et al. [16], the Acc and Dice of EX and MA increased by 20.03%/25.07% and

Table 3 Results comparison with the top five teams in IDRiD competition

Methods	AP on EX	AP on SE	AP on MA	AP on HE
VRT (1st)	0.7127	0.6995	0.4951	0.6804
PATech (2nd)	0.8850	–	0.4740	0.6490
IFLYTEK-MIG (3rd)	0.8741	0.6588	0.5017	0.5588
SOONER (4th)	0.7390	0.5395	0.4003	0.5395
SAIHST (5th)	0.8582	–	–	–
CARNet	0.8675	0.7125	0.5148	0.6389

The best results are shown in bold

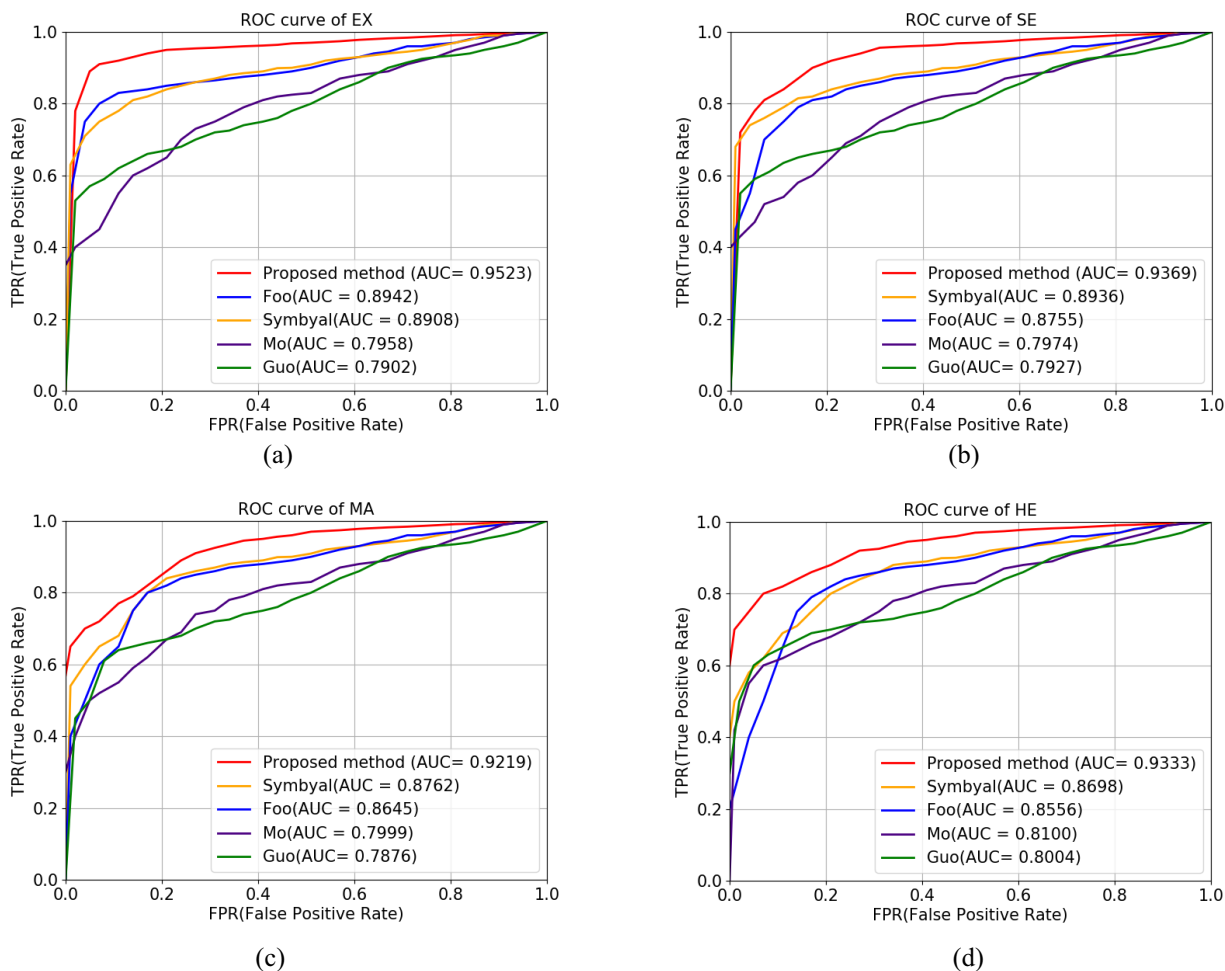


Fig. 11 ROC curves of the four lesions segmentation on the IDRiD data set

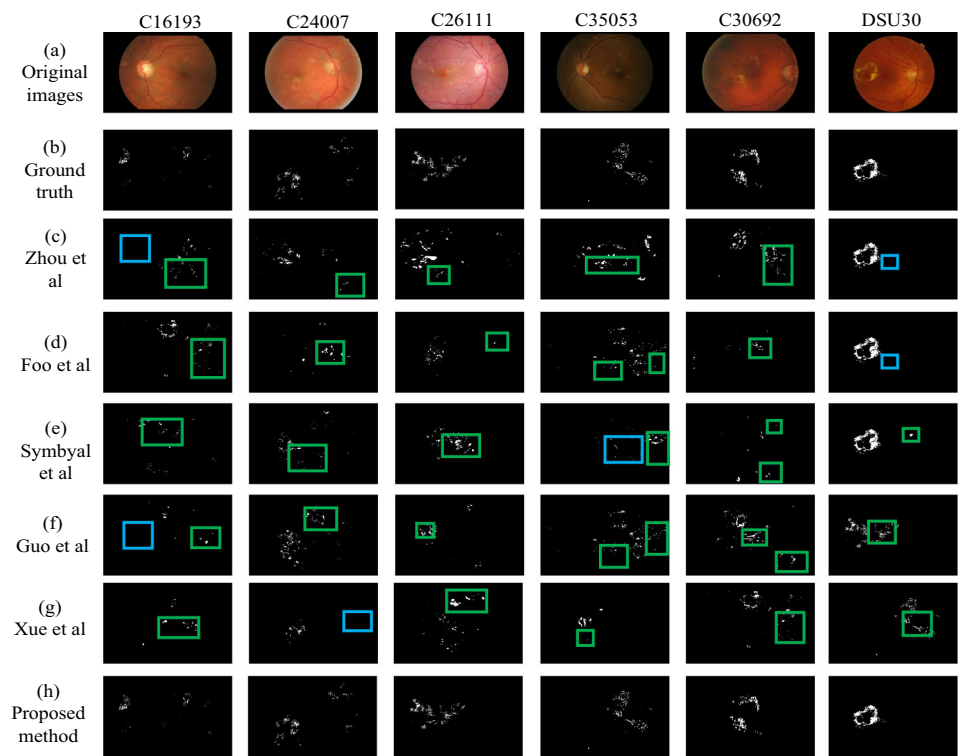
26.02%/29.66%, respectively. The above results demonstrate the superiority of the proposed model over the-state-of-art methods for both the segmentation of yellow and red lesions.

Figure 13 displays the ROC curves of different networks in segmenting EX and MA to evaluate the experimental performance. We can observe that the AUC value of the proposed CARNet is higher than that of the other four approaches in terms of the performance of EX segmentation or MA segmentation.

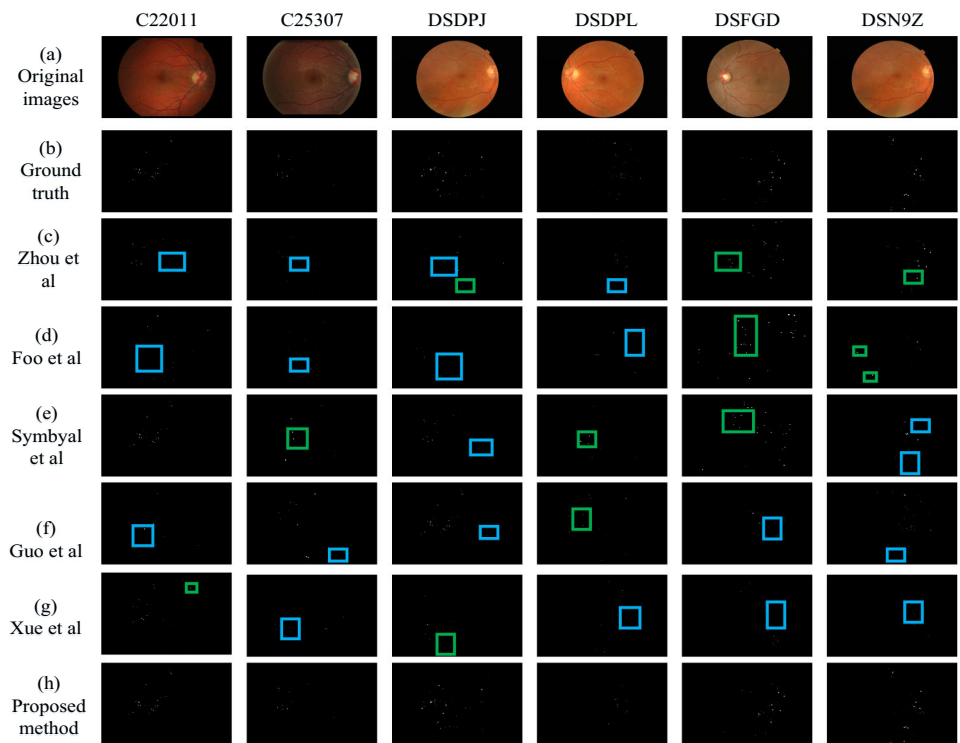
Performance on DDR

We compared the proposed CARNet with other existing methods on the DDR data set to validate the effectiveness of our framework. As shown in Fig. 14, the result of the proposed CARNet is closer to the ground truth. Although there are similar tissues and other lesion noises, our method can overcome undesirable interference and segment multiscale lesions from fundus images accurately. Zhou et al. [26] can

Fig. 12 Qualitative comparison results of the two lesions segmentation on the E-optha data set. The green and blue box represent false positives and false negatives, respectively. **a** EX results. **b** MA results



(a)

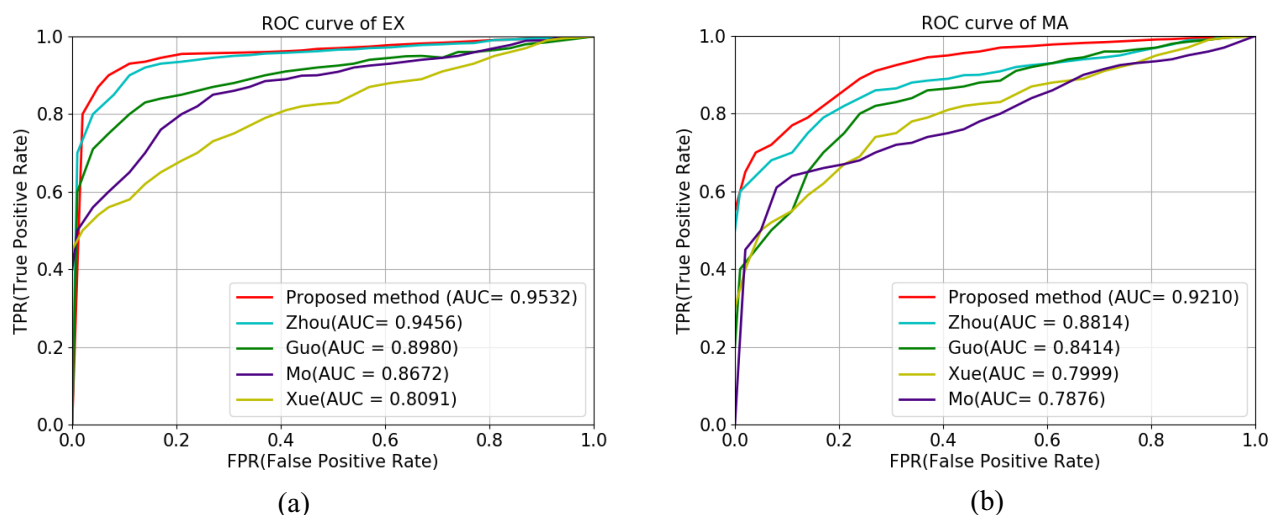


(b)

Table 4 Quantitative results of different methods on E-ophtha

Methods	EX				MA			
	Sen	Spe	Acc	Dice	Sen	Spe	Acc	Dice
Zhou et al. [26]	0.9542	0.9387	0.9546	0.9481	0.9016	0.8957	0.9214	0.9025
Foo et al. [13]	0.8427	0.8257	0.8692	0.8893	0.7984	0.8256	0.7862	0.7593
Sambyal et al. [15]	0.9451	0.9673	0.9652	0.9739	0.9317	0.9463	0.9459	0.9295
Guo et al. [16]	0.8064	0.6813	0.7743	0.7159	0.6517	0.7258	0.6982	0.6547
Mo et al. [17]	0.9227	0.9186	0.9053	0.9247	0.8546	0.8163	0.7985	0.8628
Xue et al. [27]	0.7723	0.9738	0.9614	0.9564	0.6725	0.9518	0.9473	0.9426
Tan et al. [28]	0.8759	0.9867	0.8541	0.8497	0.4606	0.9782	0.8563	0.9148
Xu et al. [12]	0.8755	0.7998	0.9217	0.9138	0.5933	0.5276	0.8867	0.7189
Proposed	0.9652	0.9845	0.9746	0.9761	0.9352	0.9582	0.9489	0.9513

The best results are shown in bold

**Fig. 13** ROC curve on the E-ophtha data set

detect large EX, SE and HE but cannot reverse the detailed contours of MA. The method proposed by Foo et al. [13] sometimes misclassifies small HE as MA due to the similarity of the two lesions in colour and scale. The method of Sambyal et al. [15] tends to ignore HE with low contrast, resulting in false negative results. The approach of Guo et al. [16] misclassifies other bright lesions, such as artefacts and drusen as EX, which results in false-positive results. The main reason is that the multichannel loss function can only address the issue of misclassification to a certain extent but cannot overcome the interference of other noises on lesion segmentation completely. Tan et al. [28] can detect the obvious EX and HE but ignores the tiny MA. The main reason is that this method only performs lesion segmentation based on pixel features; however, it does consider context information of whole images.

Table 5 displays the quantitative results for segmentation over four lesions on different metrics. It is observed from Table 5 that the proposed approach achieves the better scores

on most evaluation metrics, and Sen and Spe of segmenting EX, SE, MA and HE reach 0.9664/0.9548/0.9361/0.9327 and 0.9782/0.9642/0.9651/0.9678, respectively. Compared with Xue et al. [27], the proposed model improves Sen and Spe of the four lesions by 20.41%/19.23%/25.78%/25.43% and 5.44%/0.24%/0.04%/2.5%, respectively.

In addition, Fig. 15 shows the ROC curves of various models for segmenting four types of lesions. The data in the lower right corner of Fig. 15 show that the AUC value of the proposed CARNet is the largest, the model presented by Zhou et al. [26] is second, and the approach proposed by Guo et al. [16] is the smallest. The above results show the superior performance of our method over other competitors on multi-lesion segmentation of the DDR data set.

Although the fundus images of experimental data sets are from three different races, we found from Tables 2, 3, 4 and 5 that the segmentation performances of MA and HE is worse than those of EX and SE. There are two main reasons. First, MAs are so small and inconspicuous that even ophthalmol-

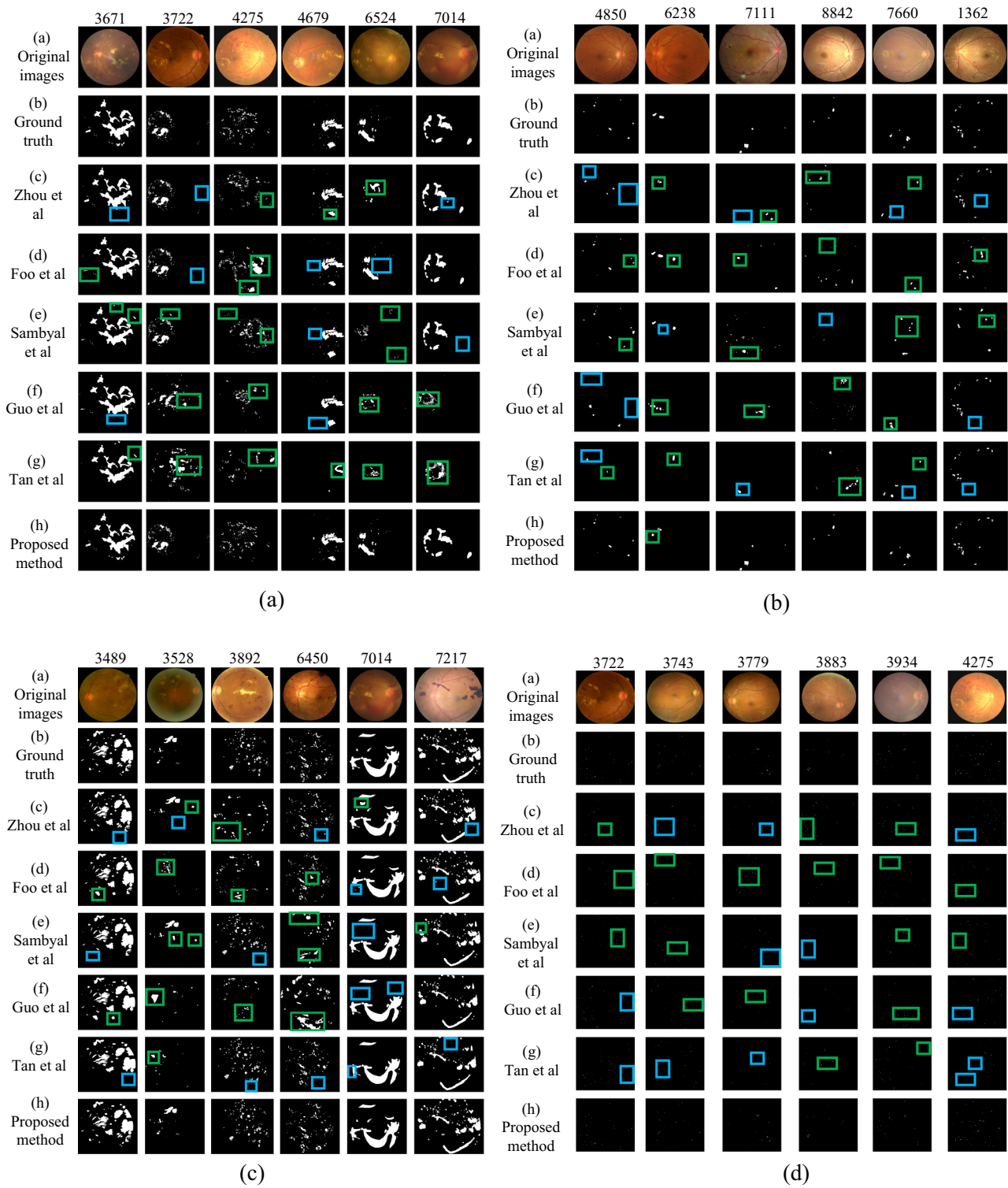


Fig. 14 Qualitative comparison results of the four lesions segmentation on the DDR data set. The green and blue box represent false positives and false negatives, respectively. **a** EX results. **b** SE results. **c** HE results. **d** MA results

Table 5 Performance comparisons of lesion segmentation on the DDR data set

Methods	EX			SE			MA			HE		
	Sen	Spe	Acc	Dice	Sen	Spe	Acc	Dice	Sen	Spe	Acc	Dice
	Zhou et al. [26]	0.9562	0.9348	0.9743	0.9461	0.9451	0.8673	0.9162	0.9137	0.9216	0.8938	0.9231
Foo et al. [13]	0.8427	0.8516	0.8468	0.8457	0.8752	0.8413	0.8259	0.8354	0.7894	0.8357	0.8147	0.7923
Sambyal et al. [15]	0.9432	0.9547	0.9658	0.9652	0.9328	0.9457	0.9482	0.9257	0.9025	0.9158	0.9316	0.9284
Guo et al. [16]	0.8764	0.7931	0.7815	0.8749	0.7951	0.8542	0.7982	0.8756	0.5647	0.7481	0.7769	0.6787
Mo et al. [17]	0.9217	0.9259	0.9357	0.9421	0.8645	0.8268	0.8085	0.8537	0.8457	0.8749	0.8064	0.8523
Xue et al. [27]	0.7623	0.9238	0.9518	0.9664	0.7625	0.9438	0.9253	0.9438	0.6783	0.9647	0.9436	0.9247
Tan et al. [28]	0.8452	0.9763	0.8782	0.8493	0.7612	0.8728	0.8632	0.9147	0.4657	0.9658	0.8653	0.9126
Xu et al. [12]	0.8452	0.7956	0.9237	0.8927	0.5934	0.6259	0.7876	0.8812	0.5933	0.5312	0.8657	0.7188
Proposed	0.9664	0.9782	0.9738	0.9687	0.9548	0.9462	0.9573	0.9465	0.9361	0.9651	0.9457	0.9413

The best results are shown in bold

ologists cannot detect them with the eyes. Compared with red lesions (MA and HE), yellow lesions (EX and SE) have relatively higher contrast and larger area, which reduces the difficulty of accurate segmentation to a certain extent. Second, MAs and small HEs are very similar in shape and scale, which makes accurately distinguish the two kinds of lesions difficult.

Ablation study

To verify the effectiveness of each component in the proposed CARNet and its contribution to the experimental results, we carry out comprehensive ablation studies. Since the DDR data set contains four abundant kinds of lesion labels and the number of fundus images is much greater than that of the other two data sets, we choose to perform the following ablation experiments on the DDR data set.

Ablation study of dual inputs

To prove the effectiveness of using the whole image and the patch image as the dual input of the network, we compared the performance of the segmentation network with the single input and dual input. The quantitative results are shown in Table 6. A single input means that only full fundus images are used as the input to the model. Dual input refers to the input of the model using multi-size whole images, as well as the whole images and the patch images. For a single-input model, we compared U-Net, RefineNet50, and combined ResNet50 with U-Net frameworks. For the dual-input model, we used ResNet50 and ResNet101 as the whole image encoder and patch image encoder, respectively, and adopted the ARM as the decoder for comparison. Acc and Dice of Table 6 both represent the average scores of the four lesion segmentation results on the DDR data set.

As shown in Table 6, compared with the framework with a single input, the segmentation performances of the dual-input network are better than those of the single-input network. Specifically, compared with the baseline, Model 1 and Model 2, the Acc and Dice of the proposed CARNet increased by 1.96%/1.21%/0.54% and 2.58%/1.31%/0.4%, respectively. This shows that dual input provides more context information of fundus images, which is beneficial to the framework to extract richer multiscale features to realize lesion segmentation. In addition, compared with the multi-size whole image as input, the whole image and the patch image as dual input improved Acc and Dice by 0.35% and 0.17%, respectively. The above results demonstrate that using the whole image and the patch image as the dual input of the network can fully exploit the global coarse features and local detailed features of the fundus image, which helps the model achieve accurate segmentation of lesions.

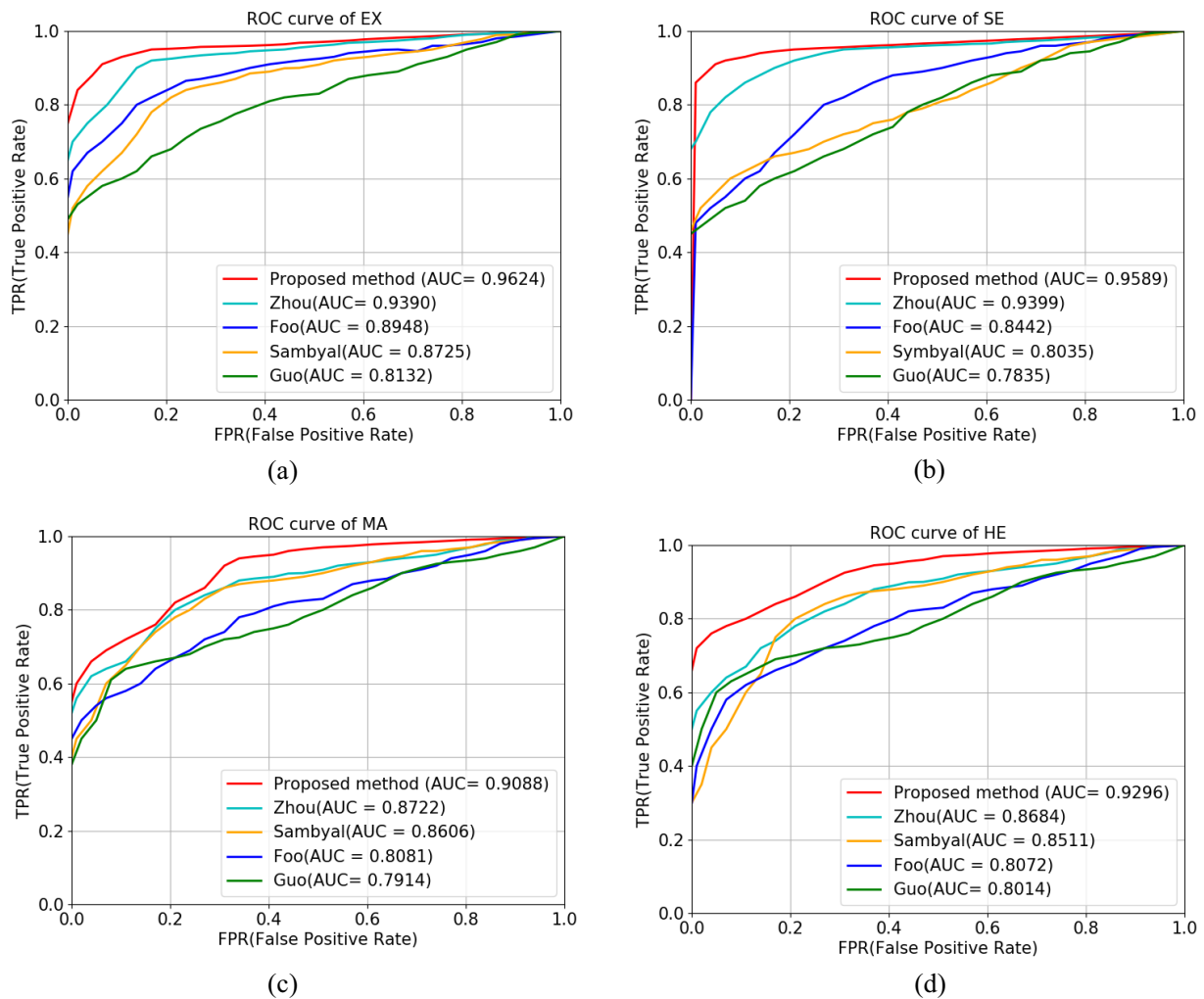


Fig. 15 ROC curve on the DDR data set

Table 6 Verify the effect of different inputs on the segmentation results

Methods	Encoder	Decoder	Input	Acc	Dice
Baseline	U-Net	U-Net	Whole image	0.9382	0.9265
Model1	ResNet50	U-Net	Whole image	0.9457	0.9392
Model2	ResNet50	RefineNet	Whole image	0.9524	0.9483
Model3	ResNet50+ResNet101	ARM	Multi-size whole image	0.9543	0.9506
CARNet	ResNet50+ResNet101	ARM	Whole image + patch image	0.9578	0.9523

The best results are shown in bold

Figure 16 visually compares the HE segmentation results of models with different inputs on the DDR data set. As shown in Fig. 16, compared with the single-input network, Model 3 and our method can better overcome the interference of other noises, fully use the multiscale context features of dual inputs, leading to the segmentation results clearer and more accurate. Moreover, compared with Model 3, the pro-

posed model uses fine-grained local information as auxiliary information to promote the reorganization of coarse-grained global features, which makes the ARM pay more attention to the lesion area. Therefore, the segmentation results can better preserve the contour details and shape features of HE areas.

Fig. 16 Visual segmentation results of networks with different inputs

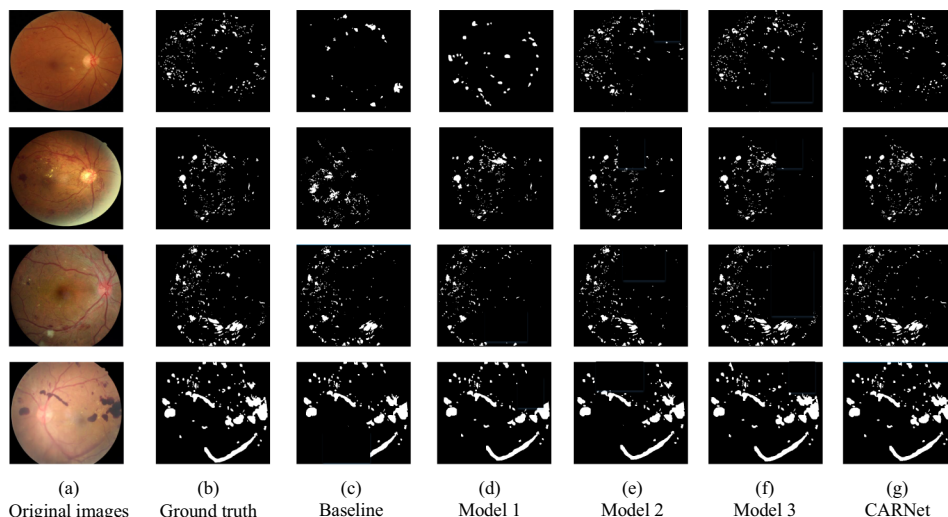


Table 7 Verity the effectiveness of the attention fusion in the ARM

Dual-input encoder	Patch image encoder	Feature fusion way	Acc	Dice
Whole image encoder	ResNet101	Concatenation	0.9485	0.9471
		Addition	0.9462	0.9483
		Attention	0.9578	0.9523
ResNet101	ResNet101	Concatenation	0.9512	0.9513
		Addition	0.9487	0.9492
		Attention	0.9561	0.9517

The best results are shown in bold

Ablation study of attention fusion

We compared the segmentation performance of the three feature fusion methods in the ARM on two dual-input encoders to explore the effectiveness of the AFB. Table 7 shows the quantitative comparison results of different feature fusion methods. Acc and Dice both represent the average values of the four lesion segmentation results on the DDR data set.

It can be seen from Table 7 that the segmentation performance of the attention fusion on any dual-input encoder is significantly better than that of the other two feature fusion methods. This shows that the AFB can adaptively learn different types of input features and perform efficient information fusion. The quantitative results demonstrate that AFB plays an irreplaceable role in the ARM, and it has an important contribution to accurately segmenting the lesion. In addition, a good feature fusion method can compensate for the limitation of network depth and improve the robustness of the framework. It is observed from the experimental results of addition and concatenation that because the depth of ResNet50 is lower than that of ResNet101, the feature extraction ability of ResNet50 as the whole image encoder is worse than that of ResNet101. However, the results of the attention fusion are the opposite, and the performance of the ResNet50 encoder

is better than that of ResNet101. This shows that the attention fusion method helps the network converge quickly and generate better segmentation results.

Computation time

To analyse the time complexity of the proposed approach, we compared the training time and test time of various networks on the DDR data set, as shown in Table 8. To ensure a fair comparison, we applied the same strategies (e.g., training epoch, learning rate and optimizer) to train all competing models on an NVIDIA GeForce Titan GPU in the experiment. It can be seen in Table 8 that our method takes 7 h to train 300 epochs on a GPU, and it takes 0.2 s to test a fundus image with a resolution of 1024×1024 . Compared with other models, the training time of our method is shorter than most methods, and the test time is the shortest.

The main reason is that the proposed CARNet adopted the AFB to fuse the global image feature, the patch image feature, and the output feature of the previous ARM to speed up the network convergence, thereby shortening the training time. Zhou et al. [26] combined U-Net and an Xception module to achieve lesion segmentation and used dilated convolution to calculate lesion features for each channel. However,

Table 8 Comparison of computation time of different methods

Methods	Training time (h)	Test time (s)
Zhou et al. [26]	9.2	1.45
Foo et al. [13]	7.8	1.21
Sambyal et al. [15]	8.5	1.53
Xu et al. [12]	8.2	0.69
Mo et al. [17]	7.4	1.1
Guo et al. [16]	6.7	0.65
Ours	7	0.2

The best results are shown in bold

dilated convolution introduces more calculation parameters and increases the training time. Sambyal et al. [15] used ResNet34 for downsampling to extract fundus image features and applied periodic shuffling with convolution for upsampling to generate lesion segmentation results. However, the subpixel convolution increases the number of calculations. Therefore, the training time is much longer than that of the proposed method. Guo et al. [16] improved VGG16 for lesion segmentation. Since VGG adopts a smaller convolution kernel and a shallower network, it can not only ensure the receptive field but also reduce the parameters of the convolution layer; therefore, the training time is less than that of our method. However, it comes at the cost of reduced segmentation performance and increased testing time.

Discussion

DR is a leading cause of blindness worldwide. Therefore, accurate lesion segmentation plays a vital role in DR detection and grading. Although the existing methods have achieved promising performance in lesion segmentation, they are still full of challenges. The difficulty is mainly manifested in the various sizes and scales, complex structure, and similarity with other fundus tissues in appearance. To address this issue, this paper proposes a cascade attentive RefineNet for automatic and accurate multi-lesion segmentation of DR images. The proposed framework adopts whole images and patch images as the dual input and applies ResNet50 and ResNet101 to extract global features and local features. Then, we applied the ARM to fuse the multiscale context features from the two encoders and previous ARM to generate high-resolution segmentation results. Dual input can provide the coarse features and fine features of the fundus images, and the ARM can effectively integrate the global features, local features, and the output features of the low-level ARM to focus more on lesion regions and promote network convergence. Extensive comparison experiments show the superiority of the proposed CARNet against other competing approaches on various fundus image data sets of IDRiD, E-Ophtha and

DDR. It can eliminate the adverse effects of similar retinal tissues and other lesion noise and segment different kinds of lesions from complex fundus images accurately.

Currently, two mainstream ideas exist for solving the task of multi-lesion segmentation. On the one hand, works [12–15,26] resized or cropped the original fundus images into patches and fed them into the U-Net based network for lesion segmentation. However, they only use a single input and the deconvolution operation cannot better retain the contour detail information of the lesion, resulting in the coarse segmentation result. On the other hand, approaches [16,17] used non-encoder–decoder models (e.g., VGG, ResNet) to extract lesion features and used 1×1 convolution to perform multiscale feature fusion to obtain predictions. However, these methods use a fixed fusion weight and ignore the content difference between multilevel feature maps, which makes exploiting middle layer features difficult. The proposed approach can effectively overcome the limitations of the current methods. It adopted the whole images and the patch images as the dual input, and applied the dual attention mechanism to adaptively learn the coarse global features and fine detail information for multilevel feature fusion.

To verify the effectiveness of our backbone, we compared different models on the IDRiD data set in terms of parameters, segmentation performance and test time in Table 9. Acc and Dice both represent the average values of the four lesion segmentation results. Table 9 shows that the calculation parameters of U-Net [38] and SegNet [39] are far less than those of RefineNet [36], but the accuracy and Dice score are lower than those of RefineNet, and the test time is longer than RefineNet. RefineNet [36] uses more residual connections to form short-range connections within the network, and it forms long-range connections with ResNet to effectively transfer the gradient to the whole framework. It helps to improve the flow of information and speed up network convergence. Compared with DeepLabv3+ [33] and VGG16 [40], RefineNet [36] has fewer calculation parameters, and the test time is shorter while ensuring better segmentation performance.

Although the proposed method achieves satisfactory segmentation results on the IDRiD, DDR and E-Ophtha data sets, it still has certain limitations. First, the proposed approach sometimes misclassified the noise as MA in the fundus image with very low contrast, resulting in false-positive results. In fact, it is difficult for ophthalmologists to identify such small and inconspicuous lesions with eyes in clinical practice. Therefore, it will be the focus of the next step of research. Furthermore, we experiment on the fundus images of Caucasian and Mongolian and mixed races for lesion segmentation but do not consider the fundus images of the black people. In general, tissue pigmentation in retinal images of black people increases the difficulty of lesion segmentation. Next, we will perform lesion segmentation experiments on

Table 9 Comparison of various network backbones

Network	Parameters	Acc	Dice	Test time (s)
U-Net [38]	7.76 M	0.8954	0.8796	0.48
SegNet [39]	14.2 M	0.8876	0.9013	0.51
DeepLab v3+ [33]	42.1 M	0.9127	0.8695	1.23
VGG16 [40]	36.5 M	0.8467	0.8512	0.62
ResNet50 [35]	25 M	0.9374	0.9257	0.35
ResNet+RefineNet [36]	22 M	0.9592	0.9521	0.2

The best results are shown in bold

the fundus images of black people to prove the effectiveness of our approach.

Conclusion

This paper proposes a cascade attentive RefineNet to realize automatic and accurate multi-lesion segmentation of DR images. The framework consists of three parts: the whole image encoder, the patch image encoder and the attention refinement decoder. First, the dual input of the whole image and the patch image are input into ResNet50 and ResNet101, respectively, to extract the features of multi-size fundus images. Then, the features of the same level in the two encoders and the output of the previous attention refinement module are fed to the current attention refinement module to fuse multiscale context features of lesions and obtain the final segmentation results. We evaluate the segmentation performance on fundus images from mixed races, Caucasian and Mongolian race. Comprehensive experimental results demonstrate the superiority and effectiveness of our CARNet over state-of-the-art networks. It can overcome the interference of similar fundus tissue and noise and fully use coarse-grained global features and fine-grained local details to achieve precise segmentation of multiscale lesions.

However, the darker retina obscures some lesion changes, because there are more melanocytes in the retinal pigmented epithelium of black people. Therefore, in the future, we will study lesion segmentation in retinal images of black people to improve the generalization of the proposed approach. Moreover, we plan to grade the severity of DR based on the lesion segmentation results in the next step.

Acknowledgements The authors gratefully acknowledge the financial supports by the National Natural Science Foundation of China (Grant No. 61976126), Shandong Nature Science Foundation of China (Grant No. ZR2019MF003, ZR2020MH291, ZR2020MF132).

Author Contributions YG: conceptualization, methodology, software, writing the original draft. YP: data curation, supervision.

Funding This work was supported in part by the National Natural Science Foundation of China (Grant No. 61976126), Shandong Nature

Science Foundation of China (No. ZR2019MF003, ZR2020MF132, ZR2020MH291).

Availability of data and materials Data related to the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare there are no conflicts of interest regarding the publication of this paper.

Code availability The codes used during the study are available from the corresponding author by request.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Ting DSW, Cheung GCM, Wong TY (2016) Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clin Exp Ophthalmol* 44(4):260–277
2. Yau JWY, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, Chen S-J, Dekker JM, Fletcher A, Grauslund J et al (2012) Global prevalence and major risk factors of diabetic retinopathy. *Diabetes care* 35(3):556–564
3. Salamat N, Missen SMM, Rashid A (2019) Diabetic retinopathy techniques in retinal images: a review. *Artif Intell Med* 97:168–188
4. Stolte S, Fang R (2020) A survey on medical image analysis in diabetic retinopathy. *Med Image Anal* 64:101742
5. Varun G, Lily P, Marc C, Martin CS, Derek W, Arunachalam AN, Subhashini V, Kasumi W, Tom M, Jorge C et al (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316(22):2402–2410
6. Asiri N, Hussain M, Al-Adel F, Alzaidi N (2019) Deep learning based computer-aided diagnosis systems for diabetic retinopathy: a survey. *Artif Intell Med* 99:101701

7. Guo J, Yi P, Wang R, Ye Q, Zhao C (2014) Feature selection for least squares projection twin support vector machine. *Neurocomputing* 144:174–183
8. Yu Y, Gao Y, Wang H, Wang R (2018) Joint user knowledge and matrix factorization for recommender systems. *World Wide Web* 21(4):1141–1163
9. Shamsolmoali P, Zareapoor M, Zhou H, Wang R, Yang J (2020) Road segmentation for remote sensing images using adversarial spatial pyramid networks. *IEEE Trans Geosci Remote Sens* 59(6):4673–4688
10. Zheng H, Wang R, Ji W, Zong M, Wong WK, Lai Z, Lv H (2020) Discriminative deep multi-task learning for facial expression recognition. *Inf Sci* 533:60–71
11. Cheng G, Cheng J, Luo M, He L, Tian Y, Wang R (2020) Effective and efficient multitask learning for brain tumor segmentation. *J Real Time Image Process* 17(6):1951–1960
12. Xu Y, Zhou Z, Li X, Zhang N, Zhang M, Wei P (2021) Ffu-net: feature fusion u-net for lesion segmentation of diabetic retinopathy. *BioMed Res Int* 2021:12
13. Foo A, Hsu W, Lee ML, Lim G, Wong TY (2020). Multi-task learning for diabetic retinopathy grading and lesion segmentation. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 13267–13272
14. Yang Y, Li T, Li W, Wu H, Fan W, Zhang W (2017) Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 533–540
15. Sambyal N, Saini P, Syal R, Gupta V (2020) Modified u-net architecture for semantic segmentation of diabetic retinopathy images. *BioCybern Biomed Eng* 40(3):1094–1109
16. Guo S, Li T, Kang H, Li N, Zhang Y, Wang K (2019) L-seg: an end-to-end unified framework for multi-lesion segmentation of fundus images. *Neurocomputing* 349:52–63
17. Mo J, Zhang L, Feng Y (2018) Exudate-based diabetic macular edema recognition in retinal images using cascaded deep residual networks. *Neurocomputing* 290:161–171
18. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabudhe V, Meriaudeau F (2018) Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data* 3(3):25
19. Etienne D, Guy C, Xiwei Z, Guillaume T, Klein J-C, Fernand M, Beatriz M, Gwénoél Q, Mathieu L, Ronan D et al (2013) Teleophtha: machine learning and image processing methods for teleophthalmology. *Irbm* 34(2):196–203
20. Li T, Gao Y, Wang K, Guo S, Liu H, Kang H (2019) Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf Sci* 501:511–522
21. Bo W, Zhu W, Shi F, Zhu S, Chen X (2017) Automatic detection of microaneurysms in retinal fundus images. *Comput Med Imaging Graph* 55:106–112
22. Long S, Huang X, Chen Z, Pardhan S, Zheng D (2019) Automatic detection of hard exudates in color retinal images using dynamic threshold and svm classification: algorithm development and evaluation. *BioMed Res Int* 2019:13
23. Colomer A, Igual J, Naranjo V (2020) Detection of early signs of diabetic retinopathy based on textural and morphological information in fundus images. *Sensors* 20(4):1005
24. Amin J, Sharif M, Yasmin M, Ali H, Fernandes SL (2017) A method for the detection and classification of diabetic retinopathy using structural predictors of bright lesions. *J Comput Sci* 19:153–164
25. Kälviäinen RVJPH, Uusitalo H (2007) Diaretdb1 diabetic retinopathy database and evaluation protocol. *Med Image Underst Anal* 2007:61
26. Zhou Y, He X, Huang L, Liu L, Zhu F, Cui S, Shao L (2019) Collaborative learning of semi-supervised segmentation and classification for medical images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2079–2088
27. Xue J, Yan S, Jianhua Q, Qi F, Qiu C, Zhang H, Chen M, Liu T, Li D, Liu X (2019) Deep membrane systems for multitask segmentation in diabetic retinopathy. *Knowl Based Syst* 183:104887
28. Tan JH, Fujita H, Sivaprasad S, Bhandary SV, Rao AK, Chua KC, Acharya UR (2017) Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network. *Inf Sci* 420:66–76
29. Sun J, Peng Y, Guo Y, Li D (2021) Segmentation of the multimodal brain tumor image used the multi-pathway architecture method based on 3d fcn. *Neurocomputing* 423:34–45
30. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, Hamzah H, Garcia-Franco R, Yeo IYS, Lee SY et al (2017) Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama* 318(22):2211–2223
31. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, Gain P, Ordonez R, Massin P, Erginay A et al (2014) Feedback on a publicly distributed image database: the messidor database. *Image Anal Stereol* 33(3):231–234
32. Xie S, Tu Z (2015) Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision, pp 1395–1403
33. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818
34. Zhou Y, Wang B, Huang L, Cui S, Shao L (2020) A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. *IEEE Trans Med Imaging* 40(3):818–828
35. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
36. Lin G, Liu F, Milan A, Shen C, Reid I (2019) Refinenet: multi-path refinement networks for dense prediction. *IEEE Trans Pattern Anal Mach Intell* 42(5):1228–1242
37. Deng J, Dong W, Socher R (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255
38. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 234–241
39. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 39(12):2481–2495
40. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.