RMCA U-net: Hard Exudates Segmentation for Retinal Fundus Images

Yinghua Fu[a,*], Ge Zhang[a], Xin Lu[a], Honghan Wu[b], Dawei Zhang[a,c,*]

[a]*School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China*
[b]*Institute of Health Informatics, University College London, London, United Kingdom*
[c]*Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai, China.*

Abstract

Hard exudate plays an important role in grading diabetic retinopathy (DR) as a critical indicator. Therefore, the accurate segmentation of hard exudates is of clinical importance. However, the percentage of hard exudates in the whole fundus image is relatively small, and their shapes are often irregular and the contrasts are usually not high enough. Hence, they are prone to misclassifications e.g., misclassified as part of the optic disc structure or cotton wool spots, which results in the low segmentation accuracy and efficiency. This paper proposes a novel neural network RMCA U-net to accurately segmentation hard exudate in fundus images. The network features a U-shape framework combined with a residual structure to obtain the subtle features of hard exudate. A multi-scale feature fusion (MSFF) module and an improved channel attention (CA) module are designed and involved to effectively segmentation sparse small lesions. The proposed method in this paper has been trained and evaluated on three data sets: IDRID, Kaggle and one local data set. Experiments are shown and indicate that RMCA U-net of this paper is superior to the other convolutional neural networks. The method in this paper is increased by 6% higher in PR-MAP than U-net on the IDRID dataset, increased by 10% in Recall than U-net

[*]Corresponding author

Email address: fuyh@usst.edu.cn(Yinghua Fu), usstzg@163.com(Ge Zhang), 515654139@qq.com(Xin Lu), honghan.wu@ucl.ac.uk(Honghan Wu), dwzhang@usst.edu.cn (Dawei Zhang)

on the Kaggle dataset and increased by 20% in F1-score than U-net on the local dataset.

Keywords: RMCA U-net, Channel Attention, Multi-Scale Feature Fusion Module, Residual module, Hard exudate

1. Introduction

Diabetic retinopathy (DR) is one of the usual complications of diabetes and the main cause of blindness among adults. Some research estimates that about 93 million people worldwide suffer from DR (Reichel & Salz, 2015; Wild et al., 2004). The size, quantity and location of exudates are used to grade the severity of DR (Guo et al., 2020c) in clinical practice. Exudates are classified into two categories: hard exudates and soft exudates as Fig.1 shows. Hard exudates appear as bright yellow crystalline granules having sharper definition which is more related to diabetic retinopathy, whereas soft exudates exist as whitish gray in color having fuzzy boundaries which are the expression of hypertensive retinopathy. Hard exudate also is an important index for grading macular edema. If hard exudate occurs in a range of one diameter of the optic disc (OD) from fovea the center of the macula, the patients are regarded to be of symptom of macular edema (Bresnick et al., 2000).

Automatically segmenting hard exudates is necessary for the computer aided diagnosis (CAD) system to grade DR and DME. In the existing literature, the segmentation methods can be divided into three categories: the unsupervised method, the coarse-to-fine supervised approach and an end-to-end way.

Unsupervised methods mainly utilize the brightness and morphological features to segment hard exudates (Walter et al., 2002; Rajput & Patil, 2014; Kaur & Mittal, 2018). Walter et al. (Walter et al., 2002) took morphological operations and watershed algorithms to remove blood vessels and OD at first and then combined local window variance with morphological methods to segment hard exudates. Rajput et al. (Rajput & Patil, 2014) enhanced the fundus image by adaptive equalization and removed blood vessels and OD, then segmented hard

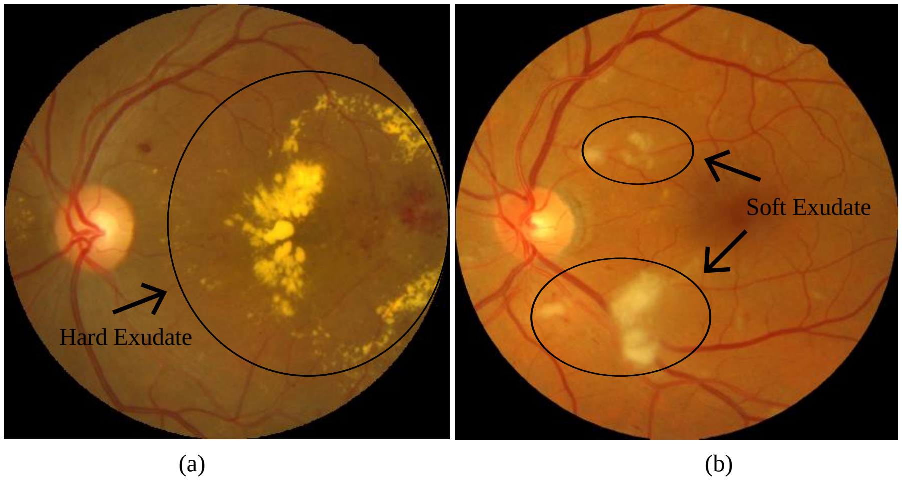

Figure 1: Two categories of exudate: (a) hard exudate, (b) soft exudate

exudates through K-means clustering and threshold. Similarly, after removing blood vessels and OD by a multiscale filter and Hough transform, J. Kaur and D. Mittal. (Kaur & Mittal, 2018) took a dynamic threshold method to segment hard exudates.

The advantage of unsupervised methods is that they do not require manual labels of ophthalmologists. However, this kind of methods might end up with a large amount of false negatives and false positives when the contrast between background and hard exudates is not high enough. Most of approaches need to remove OD and blood vessels before obtaining the segmentation of hard exudate. The high-performing preprocessing is the prerequisite to get accurate segmentation.

Coarse to fine supervised segmentation is data-driven and requires the expert's experience and labelling. It can be divided into two stages: (1) the coarse detection stage to obtain the candidate regions and (2) the fine segmentation stage to segment hard exudates in candidate regions. Wang et al. (Wang et al., 2020a) adopted the mathematical morphology to extract the candidate hard exudate regions and then trained a deep convolution network to learn the features of hard exudate. Finally, a random forest was adopted to identify actual hard exudates. Liu et al. (Liu et al., 2017) removed OD and blood vessels by morphological operations first and then divided and classified the image patches into two categories by random forest. Finally, the hard exudates were segmented by local variance and contrast. Khojasteh et al. (Khojasteh et al., 2019) chose ResNet-50 to select the patches with hard exudates and finally used support vector machines (SVM) to segment hard exudates.

Compared with the unsupervised methods, the coarse to fine methods can usually achieve better accuracies, but it requires to select patches with hard exudates. How to select the patches is a problem and often not straightforward.

The end-to-end method generally adopts deep learning architectures to segment hard exudates (Mo et al., 2018; Guo et al., 2019a; Tan et al., 2017; Liu et al., 2021). Mo et al. (Mo et al., 2018) used Resnet-50 as the encoder and built a full convolutional residual network (FCRN) in the decoder by fusing different stage feature layers to obtain the segmentation of hard exudates. Guo et al. (Guo et al., 2019a) proposed a multi-scale feature fusion (MSFF) method to segment hard exudates, but only Vgg-16 with straight cylinder structure was used as backbone to extract the feature, which limits the ability to extract features. J H Tan et al. (Tan et al., 2017) designed a 10-layer CNN architecture to segment four types of lesions: hard exudate, soft exudate, hemorrhage, and microaneurysm. However, this method divides the image into patches of 48×48, which means that the fundus image needs to be cut into lots of patches for prediction, which is obviously time-consuming. Liu et al. (Liu et al., 2021) proposed a dual-branching network and a new dual-sampling modified (DSM) loss function to solve large scale range and class imbalance in hard exudates respectively.

Compared with unsupervised and coarse-to-fine supervised methods, the end-to-end deep learning method significantly improves the segmentation performance, and most of deep learning methods have lower requirements on preprocessing. However, all these end-to-end models are only validated on fundus images with 45-degree or 50-degree field of view (FOV) which have relatively clean fundus background. The fundus images with 200-degree FOV have more complex background and involve eyelids, eyelashes and equipment frame in most cases. Hence, segmenting hard exudates is full of challenge for ultra-widefield fundus images.

Based on the aforementioned improvemnets and challenge, a new deep learning architecture is developed in this paper, an U-shaped structure of the encoder-decoder is designed, and a residual module is involved in the encoder part to obtain the subtle characteristics of hard exudates. Inspired by MSFF proposed in (Guo et al., 2020b), different scale feature maps for fusion are utilised to improve the segmentation accuracy in this paper. In addition, a new attention mechanism is added to the decoder module to enhance features of hard exudates (Jie et al., 2017; Wang et al., 2020b; Gu et al., 2021).

The contributions of this paper are summerized as follows. First, the residual module is involved in the encoder part to learn the subtle characteristics of hard exudates, and the new channel attention is added into the decoder module to further focus on the hard exudate. Second, the adjustment fusion of the multi-scale modules is used to learn different layers of feature instead of only a single layer in skip connection, which makes the segmentation more robust. Finally, the proposed method is validated on an ultra-widefield fundus images in addition to two public datasets IDRID as well as Kaggle, which proves that the proposed segmentation method can obtain a highly generalizable performance in the task of hard exudate segmentation.

The organization of this paper is arranged as follows. The relative works are discussed in the next section. The proposed segmentation method RMCA U-net is introduced including the architecture as well as loss function in section 3. The extensive experiments are described and the discussion is presented in section 4. The conclusion is made and the future work is given in the last section.

2. Related Works

In the last several years, deep learning is widely used in the community of biological and medical image analysis. End-to-end full convolution neural network (FCN) (Long et al., 2015) was developed with the convolution layer instead of full connection layer and de-convolution in up sampling firstly, but the segmentation is not accurate due to its high multiple up-sampling. U-net (Ronneberger et al., 2015) fusing different encoder levels of features with decoder through jumping connection was proposed by Ronneberger et al. in 2015, which makes up for the problem of poor segmentation precision caused by the loss of spatial information in the decoding process. As jumping connection strengthens the feature conduction and feature reuse, which effectively alleviates the gradient disappearance and improves the ability to extract features, U-net is still a popular medical image segmentation and its variants is widely used in segmenting organs (Guo et al., 2019a).

U-net still has some challenges in dealing with complex biomedical images such as the distractions of background in low quality images and poor accuracy of segmentation. The variants of U-net are mainly divided into two categories: improving the framework of U-net (Li et al., 2018; Zhou et al., 2020; Guo et al., 2020b) and adding attention modules (Oktay et al., 2018; Gu et al., 2021; Li et al., 2022).

H-DenseUNet (Li et al., 2018) mainly takes Dense-net block to construct 2D U-net and 3D U-net to extract features and then introduces hybrid feature fusion (HFF) to optimize and segment liver and lesions. U-net++ (Zhou et al., 2020) adopts a series of nested and dense skip connections which reduces the semantic gap between encoder and decoder and obtains good segmentation performance. DR-net (Guo et al., 2020b) also takes the strategy to improve the jump connection through multi-scale aggregation integrating different scales of features to extract global information, but this may introduce irrelevant interference features. CE-Net (Gu et al., 2019) adds dense atrous convolution (DAC) block and the residual multi-kernel pooling (RMP) block to the top of the encoder to capture wider and deeper semantic features. U2-net (Qin et al., 2020) keeps the basic framework of U-net but introduces an U-net to construct a nested architecture, which can capture more context information from different scales and achieves good segmentation results, but the number of parameters of this architecture is far larger than U-net. U-net GAN (Schnfeld et al., 2020) takes U-net as a discriminator in the GAN network to guide the generator to generate more realistic false images. DC U-net (Lou et al., 2020) designed dual-channel block as the encoder and decoder of the model, which helps capture detailed features.

These variants often obtain the better segmentation by improving the structure of encoder, decoder, jump connection and loss function, but sometimes the parameters of the architecture is over-cumbersome in order to a slight of improved performance.

Attention U-net (Oktay et al., 2018) introduces the attention gate (AG) module in the skip connection which suppresses the irrelevant areas in the encoder layers and highlights the significant features. CA-Net (Gu et al., 2021) adds an improved AG module to the skip connection and integrates an improved squeeze and excitation (SE) module and a proportional attention module (LA) in the decoder. CPFNet(Feng et al., 2020) adds scale-aware pyramid fusion module (SAPF) at the top of the encoder to obtain multi-scale contexts and adds global pyramid guidance module (GPG) in the jump connection to fuse the global context information flows. Dual encoder-based dynamic-channel graph convolutional network with edge enhancement (DE-DCGCN-EE) (Li et al., 2022) makes use of the dual-path encoder composed of edge detection and CNN to extract the feature of edge and adds a graph convolution network at the top of the encoder to solve the insufficient utilization of channel information, at the end it obtains the accurate segmentation of fundus blood vessels. CAR U-net (Guo et al., 2020a) adds the modified efficient channel attention (MECA) module to the jump connection and also accomplishes the segmentation of blood vessels in the fundus image. The AG module and non-local block was proposed in MsT-GANet (Wang et al., 2021), and position coding was introduced to obtain the global feature and identify the drusen in fundus image. TransUNet (Chen et al., 2021) uses the hybrid CNN-Transformer architecture to fuse the high-resolution CNN feature with the global context information from Transformer to achieve the accurate segmentation of multiple organs and hearts.

This kind of architectures is often superior to the former one in segmentation performance by limiting the attention on the targets. Hence, in order to achieve accurate segmentation of hard exudates, residual multi-scale feature fusion channel attention (RMCA) U-net is proposed in this paper to improve U-net from both the framework and attention.

3. Method

The proposed architecture RMCA (Residual Multi-scale feature fusion Channel Attention) U-net is introduced in this section and illustrated in Fig.2. It makes use of the encoder-decoder architecture as the basic framework which is divided into seven stages.The main components involve the encoder part, the multi-scale feature fusion module and the decoder part where the channel attention is involved.

The multi-scale feature fusion (MSFF) is used to replace the skip connection which guides the model to obtain the global salient features and suppress the uncorrelated local features. The improved channel attention is inserted into the top of the encoder path and the decoder path to extract more useful feature channels.

3.1. Encoder Module

The down sampling block of U-net is a plain feed-forward model indicating that the input is completely determined by the output of the only preceding layer. However, the gradients may disappear or explode as the network architecture deepens or gets complex. As shown in the Fig.2, the encoder path is improved and includes four blocks, and each consists of convolution module, drop block, pooling layer as well as residual module. Drop block can effectively

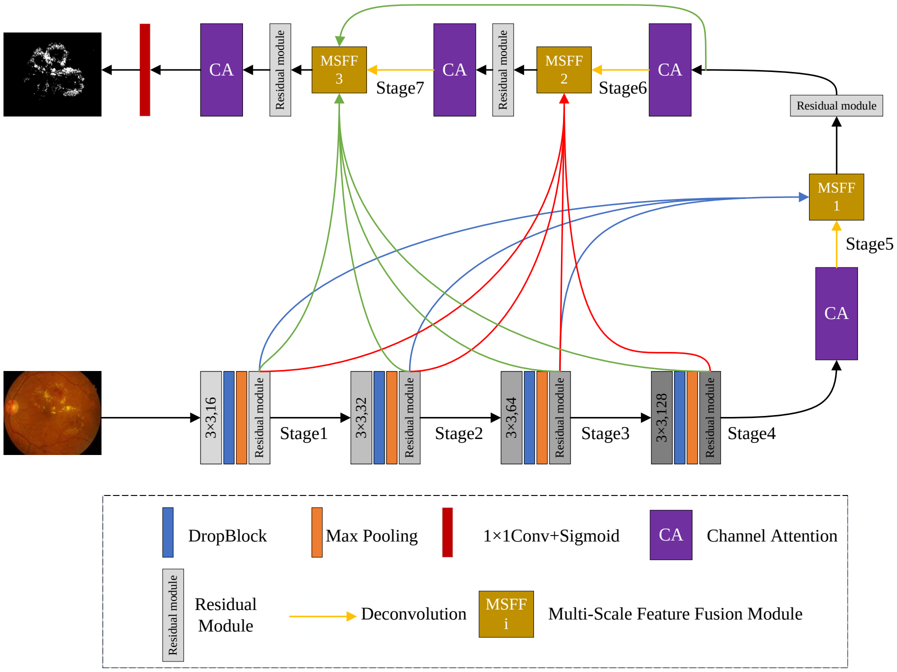

Figure 2: The framework of RMCA U-net

alleviate the overfitting by random dropping a continuous area rather than an independent unit (Ghiasi et al., 2018) as dropout does. In fact, drop block is a structured dropout.

Residual module as shown in the Fig.3. The residual module is made up of two residual blocks. Each residual block has two layers including normalization block, Relu activation function, 3×3 convolution and drop block. After the residual module is added, the output is not only related to the previous network, but also retains the network information of the previous layer. This structure not only effectively alleviates gradient disappearance and gradient explosion (Guo et al., 2020b) but also is very important to extract hard exudate features with different sizes and complex shapes.

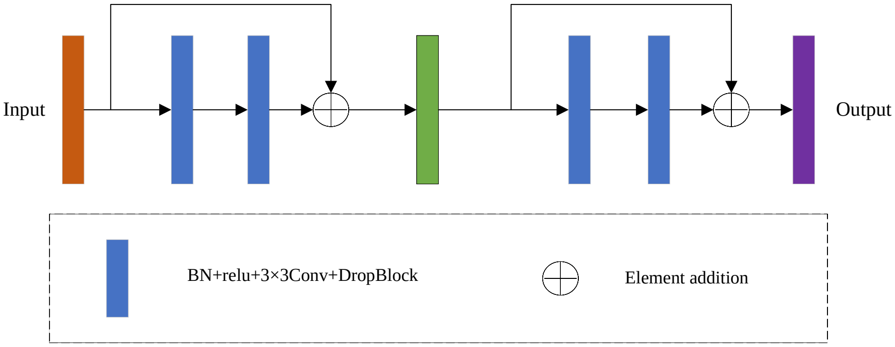

Figure 3: The residual model

3.2. Multi-scale feature fusion (MSFF)

U-net and its variants play an important role in medical image segmentation. However the simple skip connection ignores the global features and may introduce the distraction of local irrelevant features (Wang et al., 2021). This paper uses MSFF module to obtain the feature information from different receptive fields.

The structure of MSFF1 is shown in Fig.4. Features from stage 1, stage 2 and stage 3 are input into the drop block and 1×1 convolution, and the number of channels for each stage is adjusted uniformly by the 1×1 convolution to 64. The feature layers from stage 1 and stage 2 are then adjusted to the same size of feature layer as from stage 3 and stage 5 by performing 4×4 pooling and 2×2 pooling respectively. Concatenating all the features together from four stages, one can get the feature layer with 256 channels. The fusion features are fed into 3×3 convolution and drop block, then the number of channels is reduced to 64.

The structure of MSFF2 is illustrated in Fig.5. Features from stage 1 to stage 4 are convolved with 1×1 kernel and then pass by drop block where the channel number is 32. The feature layers from stage 1, stage 3 and stage 4 are then adjusted to the same size of feature layer as from stage 2 and stage 6 by performing 2×2 pooling, 2×2 de-convolution and 4×4 de-convolution respec-

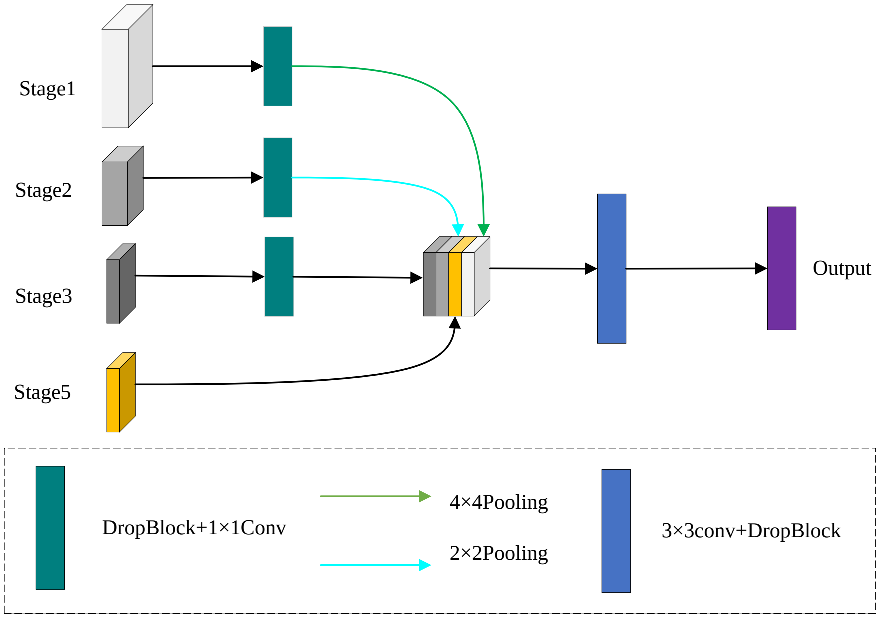

Figure 4: Multi-scale feature fusion module 1

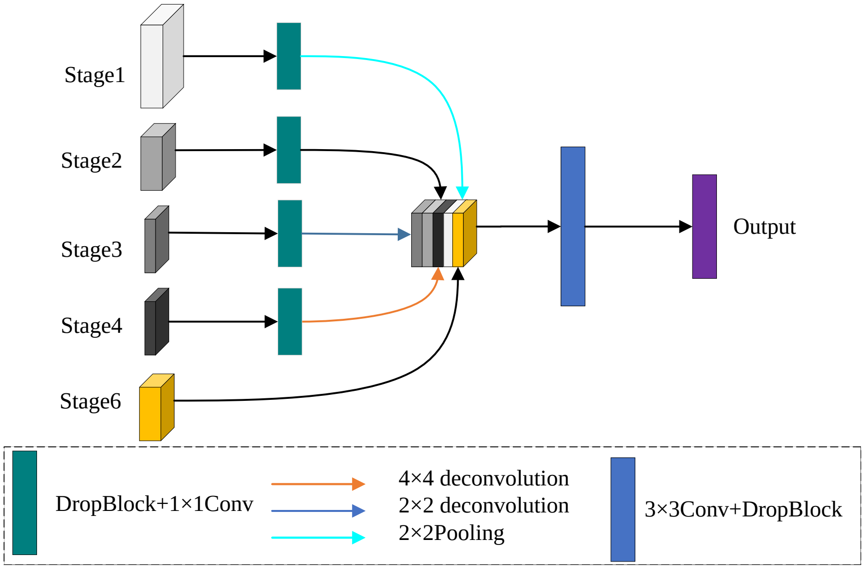

Figure 5: Multi-scale feature fusion module 2

tively. Concatenating all the features together from five stages, one can get the feature layer with 160 channels. The fusion feature layer are feeded into the 3×3 convolution and drop block, then the channels are reduced to 32.

The structure of MSFF3 is presented in Fig.6. The features from stage 1, stage 2, stage 3, stage 4 and Residual Module 1 are convolved with 1×1 kernel and then pass by drop block, then the number of channels is turned to 16. The feature layers from stage 2, stage 3, stage 4 and Residual Module 1 are then adjusted to the same size of feature layer as from stage 1 and stage 7 by de-convoluting with 2×2, 4×4, 8×8 and 4×4 respectively. Concatenating all the features from six stages together, one can obtain the feature layer with 96 channels. The fusion feature layer is de-convoluted with 2×2 to keep the size as the same as the input of the architecture and the number of channel is turned to 16, then and fed into convolutional layers, consisting of 3×3 convolution and drop block.

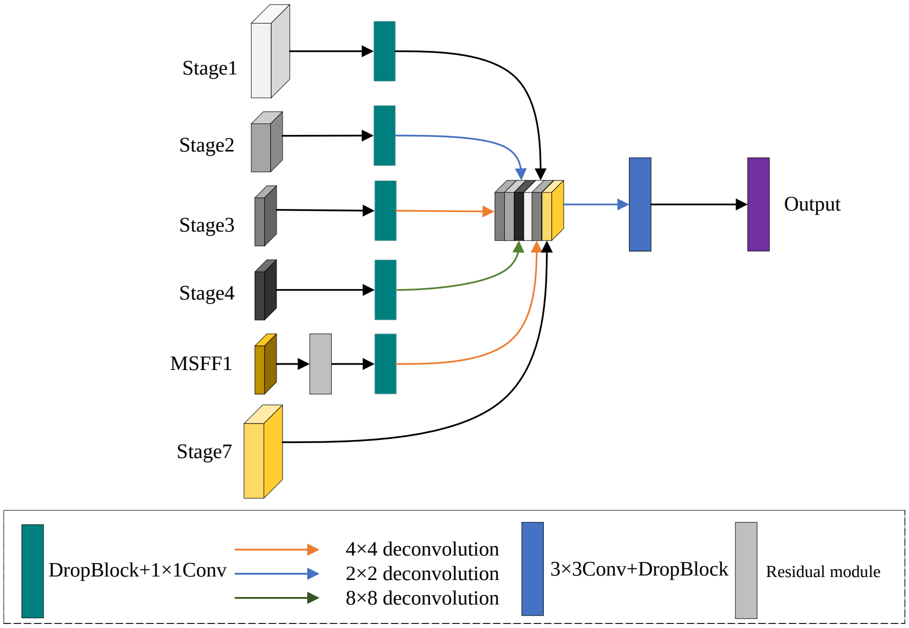

Figure 6: Multi-scale feature fusion module 3

3.3. Channel Attention and Decoder Module

In the decoder part, we introduce the channel attention module which is inserted into the top of the encoder path and the decoder path as Fig.7 shows. The output at the different stages of the encoder mostly contains some lower-level information, and the corresponding channel of the decoder contains more semantic information. To make better use of the obtained feature from MSFF, the channel attention is introduced in the proposed architecture. The core operation of SE-Net (Jie et al., 2017) is squeeze and excitation which are used to extract the global spatial information from the squeeze module by the global average pooling. Inspired by CA-Net (Gu et al., 2021) and ECA-Net (Wang et al., 2020b), we parallel a max pooling on the basis of SE-Net to store more spatial information, and the full connection layer in SE-Net is replaced by one-dimensional convolution which is used to determine the interaction between channels and reduce the computation without the loss of the accuracy.

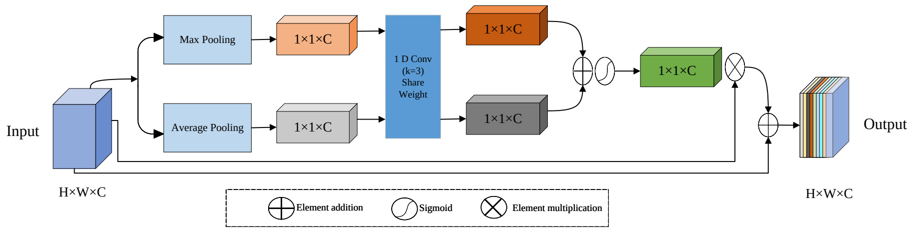

Figure 7: Channel Attention Module

In Fig.7, H, W, C denote the height, width and the number of channels related to the input respectively. After paralleling the max pooling and average pooling, two groups of 1×1×C channel attention coefficients are obtained, and then pass through a one-dimensional convolution layer with shared weights where the convolution kernel size is 3, and the correlation between the two groups of attention coefficients is calculated. Two vectors of the attention coefficient are added together and normalized by the sigmoid function, and the normalised sum is then multiplied with the input to obtain the feature layer with the channel attention. Finally, the feature layer with the channel attention weight is integrated with the input by element-wise addition to obtain the output.

For decoder part, in order to effectively recover the high-level feature mapping extracted by MSFF and the channel attention module, this paper reuses the residual block in the decoder just as the encoder does, because the jumping connection of the residual block is also helpful to recover the spatial position information of small lesions, and de-convolution is used for up-sampling.

3.4. Loss Function

To optimize the proposed architecture of this paper, binary cross entropy loss is taken to segment hard exudates at pixel level, and the fromula is presented in (1):

$$L_{BCE} = -\sum_{h,w}(1-Y)log(1-X) + Ylog(X) \tag{1}$$

where X means the segmentation result and Y for the label groundtruth, and h and w are the coordinates of pixels.

4. Experiment and analysis

The experiments are performed on an NVIDIA GeForce RTX 2080 Ti GPU with 12G video memory. The proposed architecture in this paper is validated on two public datasets IDRID and Kaggle as well as one local dataset ultra-widefield fundus images.

4.1. Datasets

IDRID (India diabetic retinopathy image dataset) (Porwal et al., 2018) was publicated online at the 2018 international conference on biomedical retina image challenge sponsored by international symposium on biomedical imaging. There are 81 images involving hard exudates with size 4288×2848 which are divided into two parts: 54 for training and 27 for testing. The images are clipped, filled by zero operations andresized to 608×608. In order to compensate the insufficient training samples, data augmentation are adopted such as image contrast enhancement, horizontal flip, up-down flip and random rotation.

Kaggle (diabetic retinopathy detection competition) includes large set of high-resolution retina images taken under a variety of imaging conditions and occupies 82GB in the storage. The amount of the images is more than 10,000 in their diabetic retinopathy detection competition. Each subject involves a left and right field with with the subject id as well as either left or right. In this paper, 120 fundus images with the whole retinal region are selected. After clipping, scaling and zero filling, they are resized to 448×448 where 80 are used as training, 20 validating and 20 testing.

Ultra-widefield fundus image data is the local dataset from Xin Hua Hospital Affiliated Shanghai Jiao Tong University School of Medicine. They were captured under the 200 degrees field of view (FOV) and include 261 fundus images with the size of 3900×3072.The images are resized to 1300×1024 with 165 for training and 96 for testing. The proportion of hard exudate is smaller and the lesions are sparer in ultra-widefield fundus images than the images of the normal FOV. In this paper, we train the network on patches with 128×128 randomly selecting 800 patches from the region of interest for each image, and get a total of 132,000 patches. For each iteration, 20% of patches are randomly selected to validate the method. The ultra-widefield fundus images are corrected and synthesized by adding an extra channel into the original two ones red and green, and then image contrast enhancement is taken as the input of the proposed architecture.

4.2. Training parameter and evaluation

The experiments are done on Keras and optimized through adaptive moment estimation (Adam). The batch size is 2 on IDRID, 4 on Kaggle and 48 on ultra-widefield fundus images. The initial learning rate and the iterations epoch are set to 0.001 and 100 respectively. The number of channels in the first coding block of RMCA U-net is 16 and doubled after each downsampling till to 128. The size of drop block and the retention probability of each active unit are set to 7 and 0.9 separately.

The quantitative evaluation are presented in formula (2) to (6) denoting accuracy, recall, specificity, precision and F1 score separately. ROC (receiver operating characteristic) curve and PR (precision recall) curve are plotted respectively. The abscissa and ordinate of ROC given in (7) and (8) denote the false positive rate (FPR) and true positive rate (TPR) indicating the proportion of predicted positive but actually negative samples in all negative samples as well as the proportion of predicted positive and actually positive samples in all positive samples separately. IoU (intersection over union) and dice coefficient are presented in formula (9) and (10). The abscissa and ordinate of PR curve is recall and precision respectively.

Here true positive (TP) denotes the number of samples with the positive label predicated to be also positive. False negative (FN) denotes the number ones with the positive label predicated to be negative. False positive (FP) means the number ones with the negative label predicated to be positive and true negative (TN) with the negative label predicated to be also negative.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

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

$$IoU = \frac{TP}{TP + FN + FP} \tag{9}$$

$$Dice = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{10}$$

4.3. Experimental results

Table 1 gives the quantity analysis results on IDRID. The results of U-net (Ronneberger et al., 2015) and DR-net (Guo et al., 2020b) were given by our own implementations on the dataset, while HED-Net (Xie & Tu, 2017), FCRN (Mo et al., 2018) and DeepLab V3+ Chen et al. (2018) were from LWE-Net

Table 1: Comparative segmentation of exudates on IDRID

Method	Year	Accuracy	Recall	Specificity	Precision	F1	AUC	MAP	IoU	Dice
U-Net(Ronneberger et al., 2015)	2015	99.39%	64.19%	**99.87%**	**86.72%**	73.78%	0.9669	0.8291	58.45%	73.78%
HED(Xie & Tu, 2017)	2017	-	76.18%	-	74.14%	75.15%	-	-	-	-
FCRN(Mo et al., 2018)	2018	-	68.62%	-	60.18%	64.12%	-	-	-	-
Avula et al.(Benzamin & Chakraborty, 2018)	2018	96.60%	41.40%	98.30%	-	-	-	-	-	-
DeepLab v3+(Chen et al., 2018)	2018	-	70.12%	-	65.71%	67.84%	-	-	-	-
LWENet(Guo et al., 2019b)	2019	-	**78.03%**	-	78.26%	78.15%	-	-	-	-
Xue et al.(Xue et al., 2019)	2019	99.20%	77.90%	99.60%	-	-	-	-	-	-
DR-Net(Guo et al., 2020b)	2020	99.43%	72.76%	99.80%	83.30%	77.68%	0.9836	0.8714	63.50%	77.68%
Liu et al.(Liu et al., 2021)	2021	-	76.30%	-	77.39%	76.84%	-	-	-	-
Azat et al.(Garifullin et al., 2021)	2021	-	76.70%	99.70%	75.30%	-	**0.995**	0.842	-	-
Ours	2021	**99.47%**	77.41%	99.78%	82.01%	**79.65%**	0.9863	**0.8792**	66.18%	79.65%

(Guo et al., 2019b). The results of (Benzamin & Chakraborty, 2018; Guo et al., 2019b; Xue et al., 2019; Liu et al., 2021; Garifullin et al., 2021) are from their own papers. Our approach achieves the highest accuracy (99.47%) and F1 score (79.65%), which improves the segmentation performance obviously, while LWE-Net (Guo et al., 2019b) the highest recall (78.03%) and U-net (Ronneberger et al., 2015) the highest specificity (86.72%). Our approach achieves the second best ROC-AUC (0.0087 lower than (Garifullin et al., 2021)) and the best PR-MAP (0.0372 higher than the second (Garifullin et al., 2021)). Compared with U-net and DR-net, IoU and Dice of RMCA U-net increase by 7.73%, 2.68% and 5.87%, 1.98% respectively. ROC and PR on IDRID given by U-net, DR-net and RMCA U-net are illustrated in Fig.8. AUC and MAP of our architecture are 0.9863 and 0.8792 separately.

Table 2 presents the quantitative analysis on Kaggle. Similarly, results of U-net and DR-net are given by our experiments. Except for specificity, our architecture is better than U-net and DR-net on all other metrics. F1 is up to 80.99%, and ROC-AUC and PR-MAP are 0.9919 and 0.8936 respectively. IoU and Dice of RMCA U-net increased by 5.52%, 0.9% and 4.04%, 0.64% respectively against U-net and DR-net. ROC and PR of three models are shown in Fig.9.

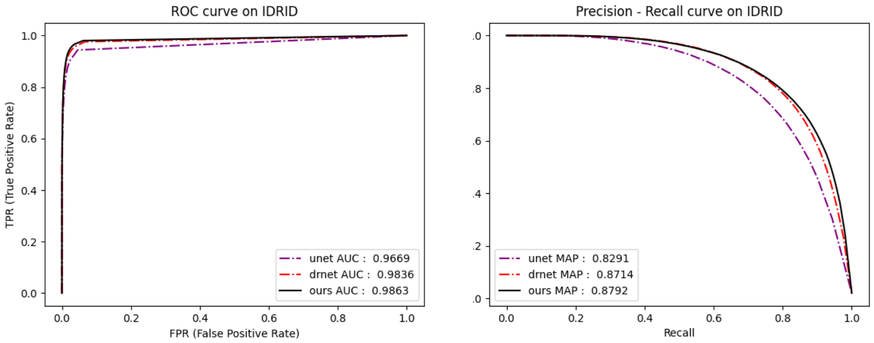

Figure 8: ROC and PR curves of U-net, DR-net and RMCA U-net on IDRID

Table 2: Comparative segmentation of exudates on Kaggle

Method	year	accuracy	Recall	specificity	Precision	F1	AUC	MAP	IoU	Dice
U-Net(Ronneberger et al., 2015)	2015	99.57%	71.69%	**99.85%**	83.05%	76.95%	0.9874	0.8457	62.54%	76.95%
DR-Net(Guo et al., 2020b)	2020	**99.63%**	77.21%	**99.85%**	**83.76%**	80.35%	0.9861	0.8843	67.16%	80.35%
Ours	2021	**99.63%**	**79.47%**	99.83%	82.56%	**80.99%**	**0.9919**	**0.8936**	68.06%	80.99%

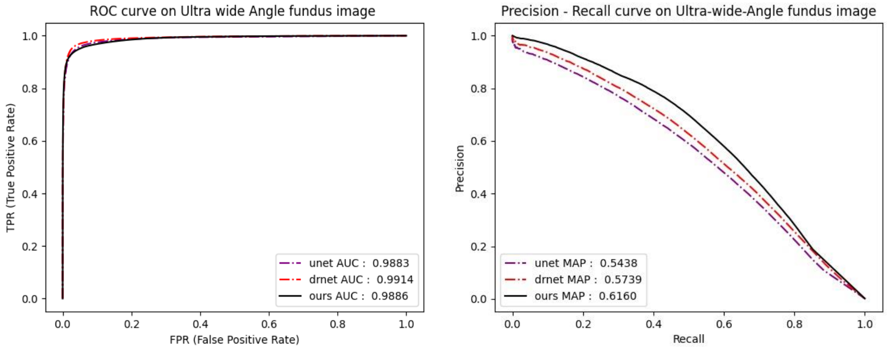

Figure 10: ROC and PR of U-net, DR-net and RMCA U-net on ultra-widefield fundus images

Table 3: Comparative segmentation of exudates on ultra-widefield fundus images

Method	year	accuracy	Recall	specificity	Precision	F1	AUC	MAP	IoU	Dice
U-Net(Ronneberger et al., 2015)	2015	99.88%	31.50%	99.98%	75.89%	44.52%	0.9883	0.5438	28.63%	44.52%
DR-Net(Guo et al., 2020b)	2020	99.88%	28.11%	**99.99%**	**81.67%**	41.83%	**0.9914**	0.5739	26.45%	41.83%
Ours	2021	**99.89%**	**41.96%**	99.98%	77.45%	**54.43%**	0.9886	**0.6160**	37.39%	54.43%

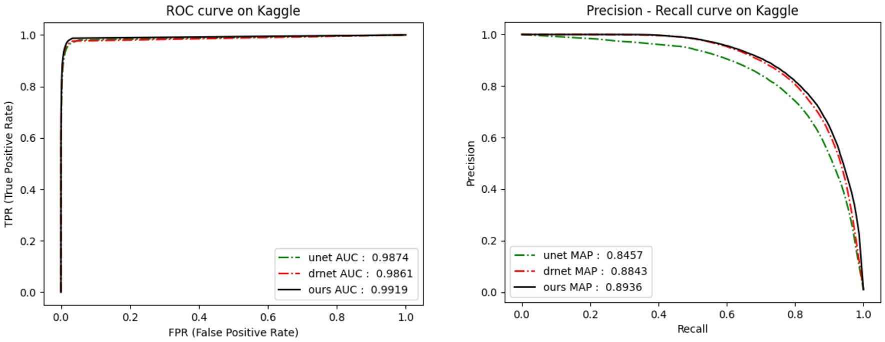

Figure 9: ROC and PR curves of U-net, DR-net and RMCA U-net on Kaggle

Table 3 shows the quantitative comparison on ultra-widefield fundus images. Precision, ROC and other metrics of RMCA U-net are better than U-net and DR-net except for specificity, where recall is 41.96% and MAP of PR curve is 0.6160. ROC and PR curves on ultra-widefield fundus images are shown in Fig.10 where our model has ROC-AUC and PR-AUC up to 0.9886 and 0.6160 respectively. IoU and Dice of RMCA U-net increased by 8.76%, 10.94% and 9.91%, 12.6% respectively.

The segmentation on IDRID, Kaggle and ultra-widefield fundus images are shown in Fig.11 to Fig.13 separately. For the first column in Fig.11 on IDRID, there are two kinds of similar lesions exudates and cotton wool spots in the fundus image, our method proposed in this paper effectively removes the distraction of cotton wool spots and obtains more accurate results than the other two approaches. For the fourth column, U-net brings obvious false negativity segmentation. Although DR-net can segment accurately the exudate, the brightness is obviously lower than our method, which means that RMCA U-net is more sensitive to the exudates and has more probability to detect exudates.

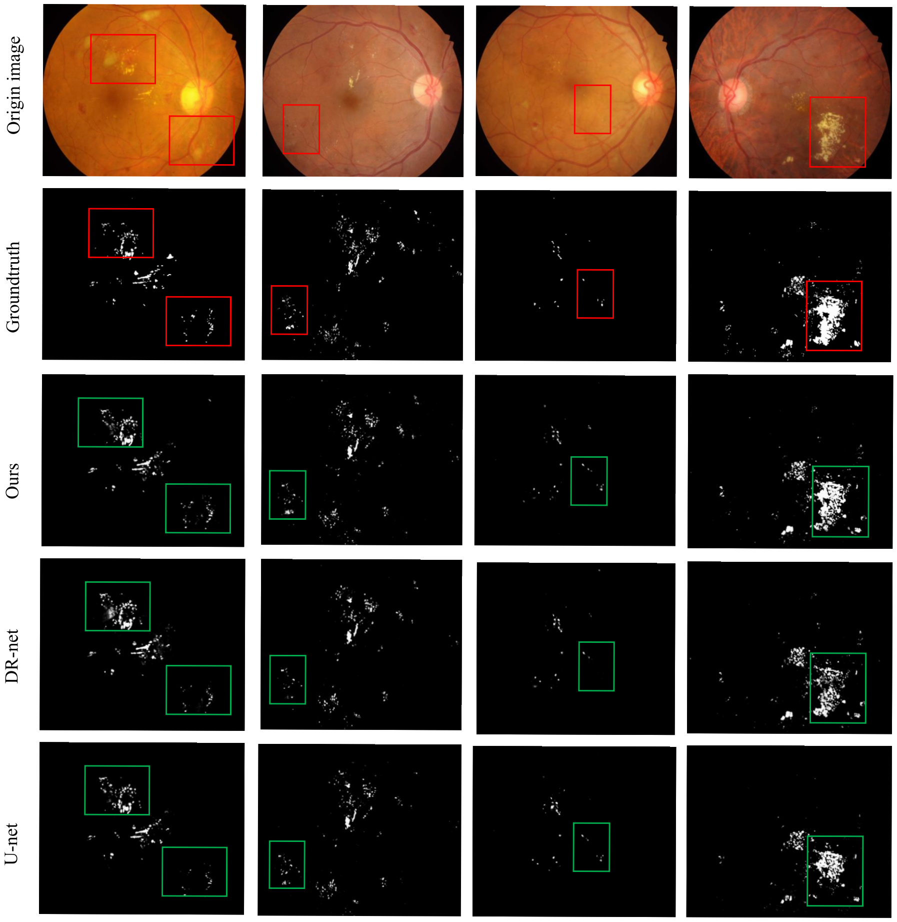

Figure 11: Segmentation of U-net, DR-net and our method on IDRID

For Fig.12 on Kaggle, there are large distractions from cotton wool spots and laser scars in the first two columns. The first column indicates U-net is heavily 370 distracted by the laser scars. The second column shows the proposed method in this paper is very robust to cotton wool spots and obtains most accurate segmentation regions among three architectures. The third and fourth column present the results of fundus images with different quality, and our method also obtains the best segmentations. When the contrast between hard exudate and background is not so obvious in the fourth column, DR-net and RMCA U-net give better segmentations than U-net. Since RMCA U-net involves the channel attention, it gets better results in terms of details than DR-net. There are false positive regions at some extend for U-net and DR-net, whereas RMCA U-net obtains more robust segmentation.

Fig.13 illustrates the segmentations on ultra-widefield fundus images. The low contrast of these fundus images and the sparsity of lesions make the segmentation of exudates full of challenges. The green boxes indicate the original position regions for the origin images, ground-truth and segmentation results. The red box is the enlarged version of the green one. The orange circles show a large number of false positive regions, and the yellow circle are the enlarged one. In the first column, the green box indicates the faint exudates and a little far from the center of the fundus image, and our method gives the relatively clear segmentation result while the other two methods are failed in the segmentation. The second and the fourth column involve many small distractions a slight similar with the exudates. In this case, our method effectively removes the distractions. In the third column, the region of interest only occupies a small part, and there are many distractions such as the eyelid and eyelash. Here, our method gives the segmentation with the highest intensity and the most integrated areas among the discussed techniques.

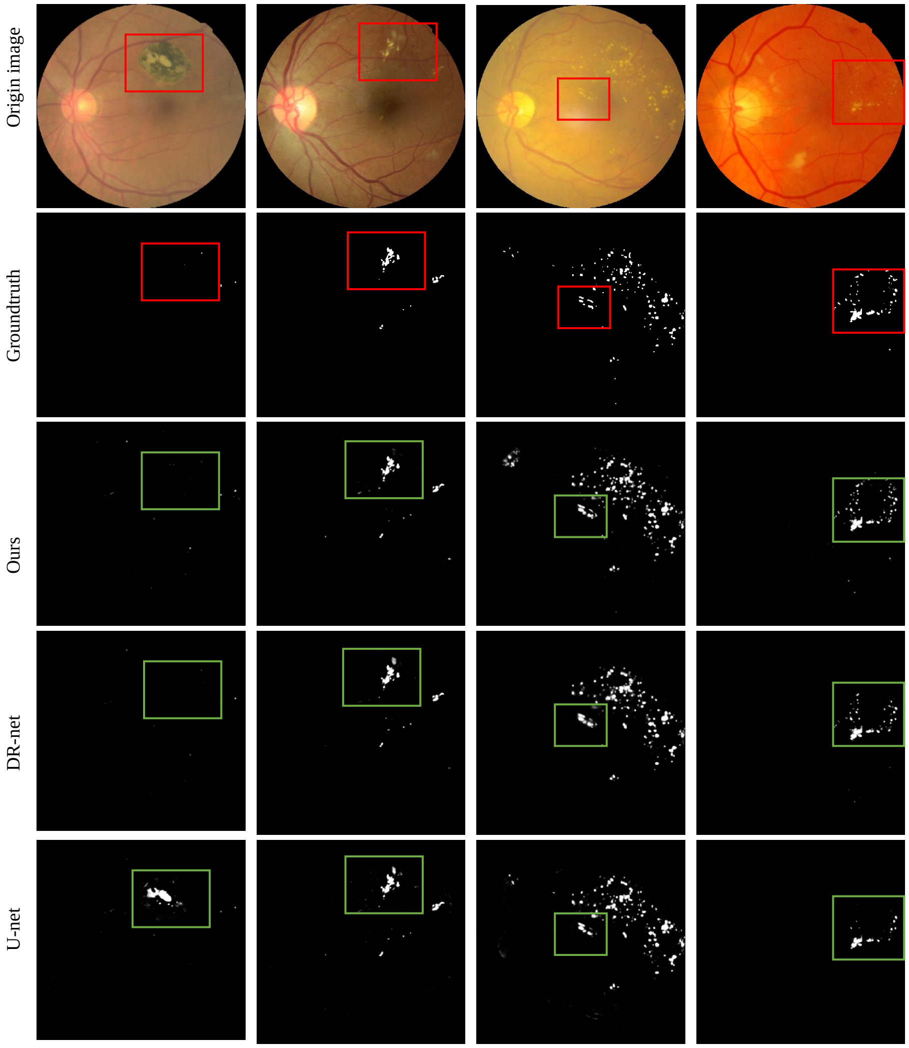

Figure 12: Segmentation of U-net, DR-net and our method on Kaggle

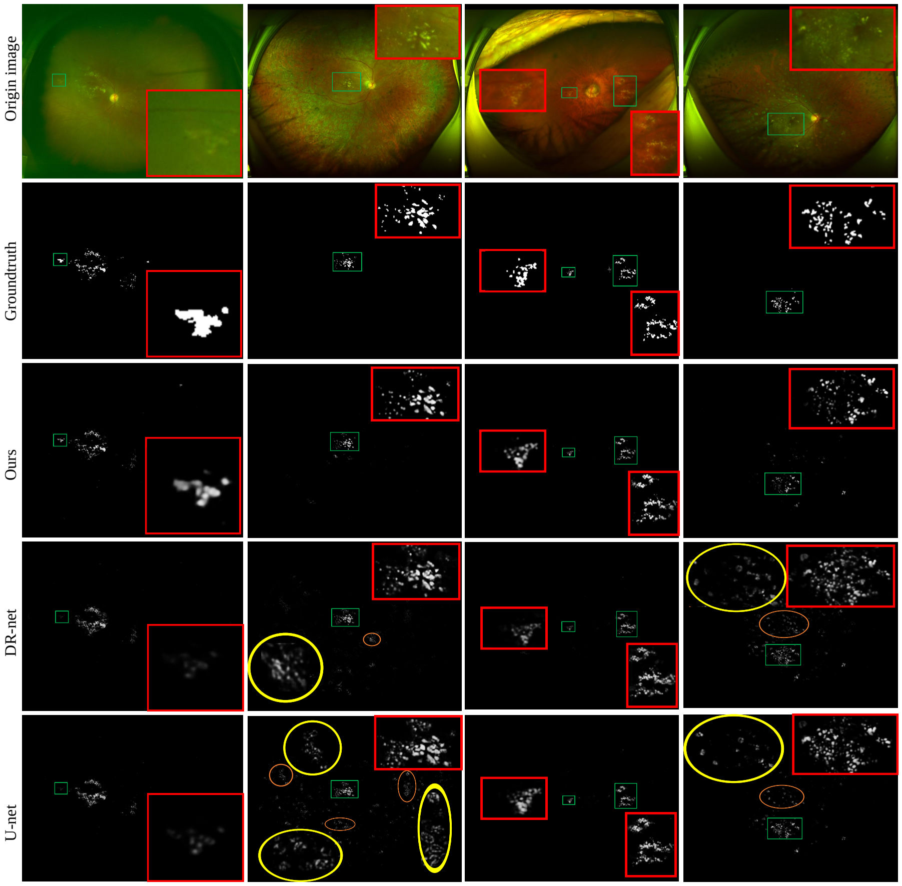

Figure 13: Segmentation of U-net, DR-net and our method on Ultra-widefield fundus images

5. Conclusion

The segmentation of hard exudates is significant in aiding to diagnose the diabetic retinopathy. This paper proposes a U-shaped encoder-decoder architecture with MSFF, channel attention and the residual module (RMCA U-net) to capture subtle characteristics of hard exudates. MSFF module is developed to learn diverse features under more subtle scales instead of just single layer of the architecture, the channel attention is introduced to achieve robust segmentation, and the residual module is to alleviate gradient disappearance and explosion as well as extract the features of hard exudates. The proposed method obtains high segmentation performances for the fundus images from both the normal FOV and the ultra-widefiled, which is validated on two public datasets and a local private dataset. RMCA U-net can not only effectively alleviate the disadvantage of class imbalance in training samples but also can learn the subtle features of hard exudates, which make it robust to the distractions of optic disc, cotton wool spots as well as laser scars.

Although RMCA U-net obtains satisfied results on the fundus images under the normal field of views, segmenting exudates on ultra-widefield fundus images still remains big challenges because of the poor image quality, low contrast and sparsity of hard exudates, which will be our future work in the next step.

References

Benzamin, A., & Chakraborty, C. (2018). Detection of Hard Exudates in Retinal Fundus Images Using Deep Learning. In *2018 Joint 7th International Conference on Informatics, Electronics Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision Pattern Recognition (icIVPR)* (pp. 465–469). doi:10.1109/ICIEV.2018.8641016.

Bresnick, G. H., Mukamel, D. B., Dickinson, J. C., & Cole, D. R. (2000). A screening approach to the surveillance of patients with diabetes for the presence of vision-threatening retinopathy. *Ophthalmology*, *107*, 19–24. doi:10.1016/S0161-6420(99)00010-X.

Chen, J., Lu, Y., Yu, Q., Luo, X., & Zhou, Y. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, . doi:10.48550/arXiv.2102.04306.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss

(Eds.), *Computer Vision – ECCV 2018* Lecture Notes in Computer Science (pp. 833–851). Cham: Springer International Publishing. doi:10.1007/978-3-030-01234-2_49.

Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., Xiang, D., Zhu, W., & Chen, X. (2020). CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, *39*, 3008–3018. doi:10.1109/TMI.2020.2983721.

Garifullin, A., Lensu, L., & Uusitalo, H. (2021). Deep Bayesian baseline for segmenting diabetic retinopathy lesions: Advances and challenges. *Computers in Biology and Medicine*, *136*, 104725. doi:10.1016/j.compbiomed.2021.104725.

Ghiasi, G., Lin, T.-Y., & Le, Q. V. (2018). DropBlock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc. volume 31.

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., & Zhang, S. (2021). CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation. *IEEE transactions on medical imaging*, *40*, 699–711. doi:10.1109/TMI.2020.3035253.

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., & Liu, J. (2019). CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, *38*, 2281–2292. doi:10.1109/TMI.2019.2903562.

Guo, C., Szemenyei, M., Hu, Y., Wang, W., Zhou, W., & Yi, Y. (2020a). Channel Attention Residual U-Net for Retinal Vessel Segmentation, . doi:10.48550/arXiv.2004.03702.

Guo, C., Szemenyei, M., Yi, Y., Xue, Y., Zhou, W., & Li, Y. (2020b). Dense

Residual Network for Retinal Vessel Segmentation. (pp. 1374–1378). doi:10.1109/ICASSP40776.2020.9054290.

Guo, S., Li, T., Kang, H., Li, N., Zhang, Y., & Wang, K. (2019a). L-Seg: An end-to-end unified framework for multi-lesion segmentation of fundus images. *Neurocomputing*, *349*, 52–63. doi:10.1016/j.neucom.2019.04.019.

Guo, S., Li, T., Wang, K., Zhang, C., & Kang, H. (2019b). A Lightweight Neural Network for Hard Exudate Segmentation of Fundus Image. In I. V. Tetko, V. Kůrková, P. Karpov, & F. Theis (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2019: Image Processing* Lecture Notes in Computer Science (pp. 189–199). Cham: Springer International Publishing. doi:10.1007/978-3-030-30508-6_16.

Guo, S., Wang, K., Kang, H., Liu, T., Gao, Y., & Li, T. (2020c). Bin loss for hard exudates segmentation in fundus images. *Neurocomputing*, *392*, 314–324. doi:10.1016/j.neucom.2018.10.103.

Jie, H., Li, S., Gang, S., & Albanie, S. (2017). Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PP*. doi:10.1109/TPAMI.2019.2913372.

Kaur, J., & Mittal, D. (2018). A generalized method for the segmentation of exudates from pathological retinal fundus images. *Biocybernetics and Biomedical Engineering*, *38*, 27–53. doi:10.1016/j.bbe.2017.10.003.

Khojasteh, P., Passos Júnior, L. A., Carvalho, T., Rezende, E., Aliahmad, B., Papa, J. P., & Kumar, D. K. (2019). Exudate detection in fundus images using deeply-learnable features. *Computers in Biology and Medicine*, *104*, 62–69. doi:10.1016/j.compbiomed.2018.10.031.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C. W., & Heng, P. A. (2018). H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes. *IEEE Transactions on Medical Imaging*, .

Li, Y., Zhang, Y., Cui, W., Lei, B., Kuang, X., & Zhang, T. (2022). Dual Encoder-Based Dynamic-Channel Graph Convolutional Network With Edge Enhancement for Retinal Vessel Segmentation. *IEEE Transactions on Medical Imaging*, *41*, 1975–1989. doi:10.1109/TMI.2022.3151666.

Liu, Q., Liu, H., Zhao, Y., & Liang, Y. (2021). Dual-Branch Network with Dual-Sampling Modulated Dice Loss for Hard Exudate Segmentation in Colour Fundus Images. *IEEE Journal of Biomedical and Health Informatics*, (pp. 1–1). doi:10.1109/JBHI.2021.3108169.

Liu, Q., Zou, B., Chen, J., Ke, W., Yue, K., Chen, Z., & Zhao, G. (2017). A location-to-segmentation strategy for automatic exudate segmentation in colour retinal fundus images. *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, *55*, 78–86. doi:10.1016/j.compmedimag.2016.09.001.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3431–3440). doi:10.1109/CVPR.2015.7298965.

Lou, A., Guan, S., & Loew, M. (2020). DC-UNet: Rethinking the U-Net Architecture with Dual Channel Efficient CNN for Medical Images Segmentation, . doi:10.48550/arXiv.2006.00414.

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. doi:10.1016/j.neucom.2018.02.035.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning Where to Look for the Pancreas. *arXiv:1804.03999 [cs]*, . arXiv:1804.03999.

Porwal, P., Pachade, S., Kamble, R., Kokare, M., & Meriaudeau, F. (2018). Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. *Data*, *3*, 25. doi:10.3390/data3030025.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, *106*, 107404. doi:10.1016/j.patcog.2020.107404.

Rajput, G. G., & Patil, P. N. (2014). Detection and Classification of Exudates Using K-Means Clustering in Color Retinal Images. In *Fifth International Conference on Signal & Image Processing* (pp. 126–130).

Reichel, E., & Salz, D. (2015). Diabetic retinopathy screening. In R. P. Singh (Ed.), *Managing Diabetic Eye Disease in Clinical Practice* (pp. 25–38). Cham: Springer International Publishing. doi:10.1007/978-3-319-08329-2_3.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* Lecture Notes in Computer Science (pp. 234–241). Cham: Springer International Publishing. doi:10.1007/978-3-319-24574-4_28.

Schnfeld, E., Schiele, B., & Khoreva, A. (2020). A U-Net Based Discriminator for Generative Adversarial Networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, .

Tan, J. H., Fujita, H., Sivaprasad, S., Bhandary, S. V., Rao, A. K., Chua, K. C., & Acharya, U. R. (2017). Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network. *Information ences*, (pp. 66–76). doi:10.1016/j.ins.2017.08.050.

Walter, T., Klein, J.-C., Massin, P., & Erginay, A. (2002). A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina. *IEEE Transactions on Medical Imaging*, *21*, 1236–1243. doi:10.1109/TMI.2002.806290.

Wang, H., Yuan, G., Zhao, X., Peng, L., Wang, Z., He, Y., Qu, C., & Peng, Z. (2020a). Hard exudate detection based on deep model learned information and multi-feature joint representation for diabetic retinopathy screening. *Computer Methods and Programs in Biomedicine*, *191*, 105398. doi:10.1016/j.cmpb.2020.105398.

Wang, M., Zhu, W., Shi, F., Su, J., Chen, H., Yu, K., Zhou, Y., Peng, Y., Chen, Z., & Chen, X. (2021). MsTGANet: Automatic Drusen Segmentation from Retinal OCT Images. *IEEE Transactions on Medical Imaging*, (pp. 1–1). doi:10.1109/TMI.2021.3112716.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020b). ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11531–11539). Seattle, WA, USA: IEEE. doi:10.1109/CVPR42600.2020.01155.

Wild, S., Roglic, G., Green, A., Sicree, R., & King, H. (2004). Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030. *Diabetes Care*, *27*, 1047–1053. doi:10.2337/diacare.27.5.1047.

Xie, S., & Tu, Z. (2017). Holistically-Nested Edge Detection. *International Journal of Computer Vision*, *125*, 3–18. doi:10.1007/s11263-017-1004-z.

Xue, J., Yan, S., Qu, J., Qi, F., Qiu, C., Zhang, H., Chen, M., Liu, T., Li, D., & Liu, X. (2019). Deep membrane systems for multitask segmentation in diabetic retinopathy. *Knowledge-Based Systems*, *183*, 104887. doi:10.1016/j.knosys.2019.104887.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2020). UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. arXiv:1912.05074.