GLER-Unet: An ensemble network for hard exudates segmentation

Siyu Liu[1], *Dan* Wang[1,*], and *Xiaoxi* Wang[2]

[1]Faculty of Information Technology, Beijing University of Technology, Beijing, China
[2]State Grid Management College, Beijing, China

Abstract. The detection of hard exudation in diabetic retinopathy is a hot topic in medical image segmentation. Aiming at the irregular shape and different size of lesion area in Hard Exudates segmentation task and the common few-shot learning challenge in medical image segmentation task, a Global-Local Ensemble Robust U-Net is proposed. The network consists of a Global Contour Extraction network for extracting long-range semantics and hard exudates contour which use complete image for training, a Local Refined Feature Segmentation network for extracting local refined segmentation rules which use patch image for training, and a Feature Revise network for fusing the features extracted by the first two networks and generating binary masks. The proposed method obtains DICE, TPR and PPV of 0.8741, 0.8752, 0.8730 and 0.8960, 0.8964, 0.8956 respectively on E-Ophtha and IDRiD. At the same time, the proposed methods shows strong robustness in cross dataset testing, better than other baseline models.

Keywords: Diabetic retinopathy, Segmentation, Ensemble, Deep learning.

1 Introduction

Diabetic retinopathy (DR) is a complication of diabetes that causes retinal blood vessels to swell and ooze fluid and blood, and can even lead to vision loss in advanced stages[1]. Hard exudates is a common lesion in DR. Hard exudates appears as bright yellow spots on the retina, caused by plasma leakage, with sharp edges that can be found on the surface of the retina. As an important symptom for identifying DR, automatic segmentation of hard exudation has practical significance to improve the efficiency of DR discrimination and reduce the rate of artificial misdiagnosis.

Xue et al.[2] proposed an improved Mask R-CNN network to extract microaneurysms and exudates. Guo et al.[3] proposed a L-SEG network for simultaneous segmentation of exudates, haemorrhages and microaneurysms. In order to improve computing efficiency, many researchers[4-6] also use segmented patches as network input for training. However, this training method makes the network reduce the ability to extract long-distance semantics, resulting in a certain decline in training accuracy.

* Corresponding author: wangdan@bjut.edu.cn

The irregular shape of the hard exudates lesion area makes it difficult to extract the features. In addition, the wide existence of few-shot learning in the hard exudates segmentation datasets makes the network less robust and difficult to be applied in real diagnosis. To solve these problems, this paper proposed a global-local ensemble network for hard exudates segmentation.

2 Method

The structure of GLER-Unet is shown in figure 1. It consist of three part: Global Contour Extraction (GCE) network, Local Refined Feature Segmentation (LRFS) network and Feature Revise (FR) network.

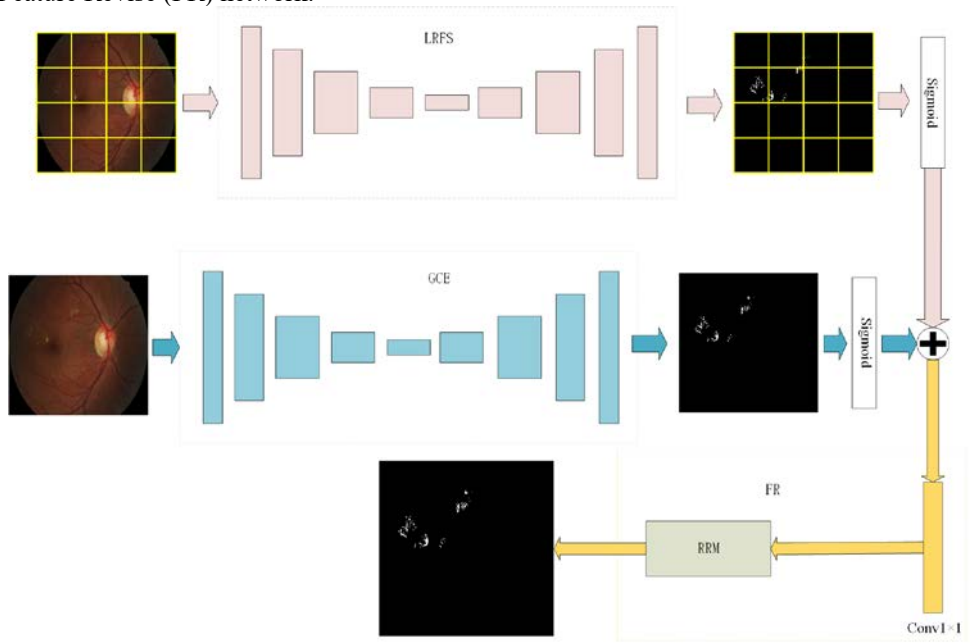

Fig. 1. Structure of GLER-Unet.

2.1 Network architecture

2.1.1 Global contour extraction network

Using complete fundus images for training, GCE is used to extract long distance semantics and complete contours of hard exudates to limit the area of extraction area. The different shape and size of hard exudates requires multi-scale and arbitrary shape feature extraction ability. At the same time, GCE also need to pay more attention to the features extracted from the shallow network such as the boundary.

The GCE network as shown in Figure 2 inherits the architecture of U-Net [7]. For encoder, GCE replaces the normal convolutions with deformable convolutions [8] and inception blocks[9]. Deformable convolution offers GCE the ability of transforming the receptive field adaptively into the shape-which is more obvious in shallow feature-of the

target lesion. Inception block enhanced the ability of multi-scale feature extraction. For decoder, GCE use normal convolutions and deformable convolutions to restore the features.

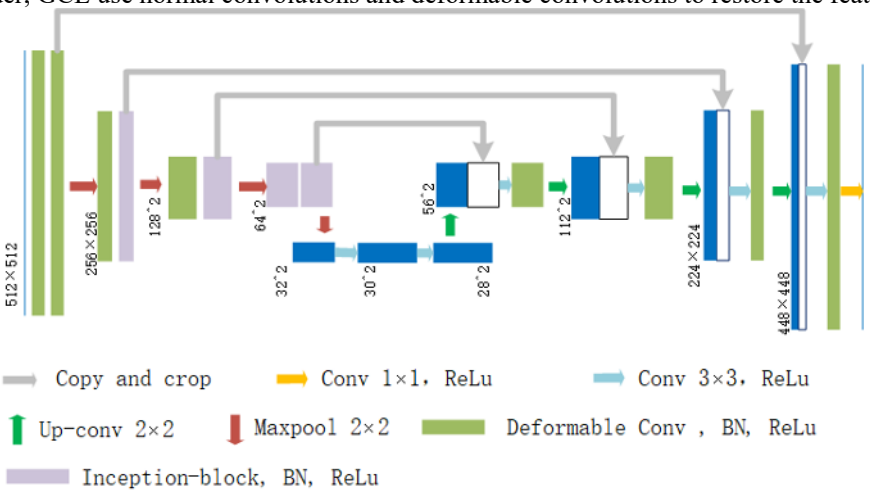

Fig. 2. Structure of GCE.

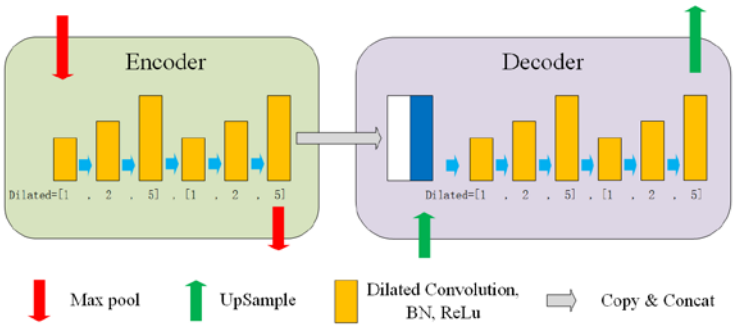

Fig. 3. Encoder & decoder of LRFS.

2.1.2 Local refined feature segmentation

Using patches cropped from origin images for training, LRFS is used to extract local fine segmentation rules. LRFS use U-Net++ as backbone, and multiple dilated convolutions [10] of different sizes are stacked in sawtooth shape to form the dilated conv layer to replace the origin conv layer. Dilated conv layer can extract multi-scale features without increase the amount of computing. The structure of encoders and decoders in LRFS is shown in figure 3.

2.1.3 Feature revise

Used to fuse features extracted from GCE network and LRFS network，and mapped to a binary mask. Firstly, the features extracted by GCE and LRFS are weighted and fused by Sigmoid, and then mapped into a non-binary mask with original resolution through a full-connection layer, then the non-binary mask is transformed into the final binary mask by a Residual Rebuilt Module (RRM). The RRM as shown in figure 4 is a simple network based on residual blocks. Each residual block is composed of a conv layer, a batch normalization layer and a PReLu activation layer. RRM offers FR the ability to adjust network layers adaptively and realize mask map mapping.

2.2 Training

2.2.1 Training strategy

The training process of GLER-UNET network can be divided into two stages – feature extraction network training and FR network training. Feature extraction network training includes the training of GCE and LRFS, which is trained independently without interfering with each other. When the network state is optimal, freeze the parameters remove the classifier, then the FR network training is carried out. Therefore, GCE and LRFS only play a role of inference in the training of FR network.

2.2.2 Loss function

Semantic segmentation task itself is a binary classification task, and Binary Cross Entropy (BCE) Loss is a common loss function in semantic segmentation, which can be applied to both GCE and LRFS networks. Focal Loss can also be used to solve the problem of extremely uneven distribution of front and background samples in hard exudates images. Thus, the weighted combination of Focal Loss and BCE is used as the loss function of the network. The formula is as follows:

$$L = \lambda L_{BCE} + (1-\lambda) L_{FL} \tag{1}$$

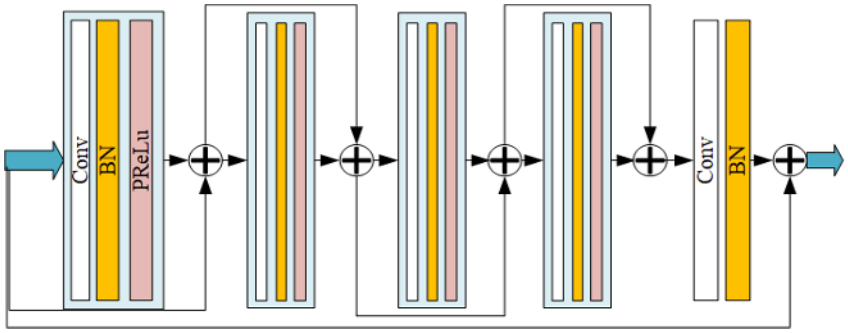

Fig. 4. Structure of RRM **Fig.5.** Segmentation diagrams of Sub-networks

3 Experimental results

3.1 Implementation Details

The Adam optimizer was used in the experiment, and the initial learning rate was 0.002. The dynamic learning rate adjustment strategy was used to downregulate every 30 epochs by 10 times, and a total of 120 epochs were trained. The model is implemented with pytorch v1.9 framework and is based on Nvidia Tesla T4 GPU training.

IDRiD and E-Ophtha dataset which contains 80 and 47 hard exudates images were used as training sets. For GCE networks, use the complete image directly as input. For LRFS, the Patch sizes of 384*384 and 128*128 were used respectively to perform random clipping on IDRiD and E-Ophtha, and the datasets were expanded to eight times of the original by means of data enhancement. Finally, 48,600 Patch images and 49,500 Patch images were obtained. The datasets was divided into training sets and test sets in a 7:3 ratio.

3.2 Evaluation

The model structure of GLER-UNET is relatively complex, and the role of each sub-network is different. Therefore, Dice, TPR, PPV, OR and UR are used to evaluate the model from all aspects. The formulas is as follows:

$$\text{Dice} = \frac{2|X \bigcap Y|}{|X| + |Y|}, \text{TPR} = \frac{TP}{TP + FN}, \text{PPV} = \frac{TP}{TP + FP}, OR = \frac{FP}{FP + FN}, UR = \frac{FN}{FP + FN} \quad (2)$$

X is the ground truth and Y is the prediction. True Positive (TP) describes the number of correctly predicted pathological pixels, False Negative (FN) describes the number of wrongly predicted background pixels, False Positive (FP) describes the number of wrongly predicted pathological pixels.

Dice coefficient describes the similarity of two samples. The True Positive Rate (TPR) describes the proportion of identified lesions to True lesions. Positive Predictive Value (PPV) describes the proportion of lesions identified as true lesions. Over Segmentation Rate (OR) and Under Segmentation Rate (UR) are respectively used to describe the ratio of pixels outside the actual prediction result and the ratio of pixels that the actual prediction structure lacks in the Ground Truth.

3.3 Ablation experiment

We conducted ablation experiments on each part of the integrated model to verify the effectiveness of the sub-networks, and the results are shown in the table. The results is shown in Table 1.

Table 1. Ablation experiment of GLER-Unet.

Dataset	Model	Dice/%	TPR/%	PPV/%	OR/%	UR/%
E-Ophtha	GCE	84.98	84.49	85.48	12.55	13.56
	LRFS	86.34	91.23	81.95	16.73	**7.30**
	GCE+LRFS	**87.42**	**87.87**	86.98	11.63	10.72
	GCE+LRFS+FR	87.41	87.52	**87.30**	**11.29**	11.07
IDRiD	GCE	86.21	86.94	85.49	12.86	11.38
	LRFS	88.97	**93.43**	84.92	14.23	**5.64**
	GCE+LRFS	89.59	89.94	89.24	9.78	9.08
	GCE+LRFS+FR	**89.60**	89.64	**89.56**	**9.46**	9.38

It can be found that GCE tends to be under-segmented, its Dice coefficient is lower than that of LRFS, but PPV index is higher, which indicates that GCE network is more inclined to improve the accuracy of pixel classification in the segmentation region. On the contrary, LRFS tends to be over-segmented, its Dice coefficient is higher and TPR index is generally higher, which indicates that LRFS's segmentation strategy is to cover the real lesion region as much as possible, while the relative PPV index is lower. The combination of GCE and LRFS network further improves the network accuracy, while the OR and UR rate are more balanced and the value is relatively low, indicating that the ensemble network absorbed the advantages of GCE and LRFS network and get balanced. The main function of FR network is to fuse the feature, but numerically, FR network can also further balance the OR and UR rate.

Figure 5 shows the segmentation diagrams of each component. There are a lot of noises in the segmentation diagram of LRFS network, which classifies irrelevant areas into lesion areas, but this also makes the LRFS network successfully cover most lesion areas. In terms of the shape and number of lesion areas, the segmentation result of GCE network is closer to the real label with less noise.

In the ensemble network of GCE and LRFS, the segmentation region of GCE effectively defines the final segmentation region, and the ensemble network classifier filters part of the noise points in the segmentation region of LRFS. Overall, the shape of the final lesion region is close to the segmentation result of GCE, but locally, it covers as much real lesion as possible, effectively improving the segmentation accuracy.

3.4 Robust test

In order to verify the robustness of GLER-UNET, this paper uses Unet, Unet++ and attention-UNET as baseline models to conduct robustness tests. The network trained with E-OPHTHA dataset was applied to IDRiD dataset and the network trained with IDRiD dataset was applied to E-OPHTHA dataset to observe the change of segmentation accuracy, so as to analyze the network robustness. The results is shown in table 2.

Table 2. Robust test.

Training set	Model	Test Dice/%	
		E-Ophtha	IDRiD
E-Ophtha	Unet	82.21	64.47
	Unet++	84.76	64.72
	Attention-Unet	85.09	65.94
	GLER-Unet	87.41	**73.21**
IDRiD	Unet	70.59	83.94
	Unet++	73.24	84.26
	Attention-Unet	72.79	86.42
	GLER-Unet	**78.78**	89.60

GLER-Unet is superior to the baseline experiment in all data sets, and at least 5.54% higher than the benchmark experiment in robustness test, which proves that GLER-UNET has excellent segmentation effect and robustness.

4 Conclusion

This paper proposes an ensemble network called GLER-UNET with global-local structure and training strategy, which obtained high accuracy and strong robustness in hard exudates segmentation. This has practical significance in the real diagnosis of hard exudation. In the future, the improvement of hard exudation segmentation task results from more data sets and image pre-processing methods, and we will conduct further research on this.

References

1. Chakrabarti R, Harper C A, Keeffe J E. Diabetic retinopathy management guidelines [J]. Expert Review of Ophthalmology, 2012, 7(5):417-439.
2. Guo S, Li T, Kang H, et al. L-Seg: An end-to-end unified framework for multi-lesion segmentation of fundus images [J]. Neurocomputing, 2019, 349: 52-63.
3. Xue J, Yan S, Qu J, et al. Deep membrane systems for multitask segmentation in diabetic retinopathy [J]. Knowledge-Based Systems, 2019, 183: 104887.
4. Wang Z, Yin Y, Shi J, et al. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2017: 267-275.
5. Lam C, Yu C, Huang L, et al. Retinal lesion detection with deep learning using image patches [J]. Investigative ophthalmology & visual science, 2018, 59(1): 590-596.

6. Chudzik P, Majumdar S, Caliva F, et al. Exudate segmentation using fully convolutional neural networks and inception modules[C]//Medical Imaging 2018: Image Processing. International Society for Optics and Photonics, 2018, 10574: 1057430.
7. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.
8. Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 764-773.
9. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.
10. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions [J]. arXiv preprint arXiv:1511.07122, 2015.