Automatic segmentation of retinal edema in optical coherence tomography based on deep neural networks

Rongrong Lu, Lidan Wang, Shukai Duan*
School of Electronics and Information Engineering, Southwest University, Chongqing 400715, China
duansk@swu.edu.cn

Abstract. A multi-task network framework based on convolutional neural networks improvements is proposed for automatic segmentation and classification in three different lesion areas of retinal OCT images. We propose a dilated convolutions module to capture fine-grained semantic information and introduce dense skip connections to integrate features of multiple scales and different semantic levels. Furthermore, we use a joint loss function that balances the relative size of the data labels and the improved generalization of the network. The validity and reliability of our proposed method are trained and verified on OCT datasets. Extensive experimental results demonstrate that the average Dice coefficient of the three lesion areas is 2.61% higher than the most commonly used U-net model in the field of medical image segmentation and reaches 0.81501.

1. Introduction
Optical coherence tomography (OCT) is a non-invasive, non-radioactive, and secure imaging technology, it has already become established as a standard diagnostic modality for retinal edema. Accurate segmentation of retinal edema lesion areas in OCT images is necessary for quantification in clinical practice. In recent years, Long J et al [1] presented fully convolutional networks (FCN) for semantic segmentation, which was the first network trained end-to-end and combined with skip architecture. Ronneberger O et al. [2] presented U-net which could be trained end-to-end from a small quantity of images and surpassed the advanced best method. Hu J et al. [3] proposed the automatic segmentation of SRF, and PED using deep neural networks and stochastic atrous spatial pyramid pooling (sASPP). Lu D. et al [4] segmented RIC, SRF, and PED in OCT images based on the FCN together with a graph-cut algorithm. Zhang R et al. [5] proposed a boundary multi-scale OCT segmentation network (BMM-Net) to segment the retinal lesion areas.

The segmentation of retinal edema in OCT images is still a challenging task due to some unique features of the task. First, retinal lesion areas have various sizes, shapes, and locations within different OCT images, which hinder the automatic segmentation. Second, the scattering of light in retinal OCT imaging technology is likely to cause a large amount of speckle-noise resulting in low-intensity contrast of different lesions region. Third, the size of the lesion area is exceptionally unbalanced, as shown in Fig.1. (four slices, which represent the appearance of normal and abnormal tissues). As can be seen from the slices, that the area of the REA is relatively more substantial than that of the PED. In this case, it is essential and more challenging to retain detailed spatial information of small targets.
In this paper, we present an improved convolutional network model for the segmentation and detection of REA, SRF, and PED retinal edema lesions in OCT images.

(1) A series of nested, dense skip pathways together with an encoder-decoder model based on U-Net is used to efficiently segment the SRF, PED, and REA lesions in OCT images. The segmentation accuracy of our method is higher than several extensively used benchmark models. At the same time, a classification module added to our model between down-sampling and up-sampling. (2) We take advantage of dilated convolution, which could extend receptive fields of convolution filters without losing the resolution of multi-scale feature maps, using shortcut connection to combine dilated convolution with our model. The experimental results demonstrate that method could perform higher segmentation accuracy, even in the OCT image where the boundary of the edema area is very blurry and complicated. (3) Propose a joint loss function and it consists of three loss functions. One is the balanced cross entropy, which could evaluate the difference in the probability distribution of different categories, and the other is the weighted cross-entropy which could mitigate the imbalance of sample categories. The most important one is exponential logarithmic loss which balance the difficulty of labels segmentation by weighting. Besides, it could balance the labels by their relative sizes.

2. RELATED WORKS

Recently, a novel biomedical semantic image segmentation architecture named U-Net++, has been proposed by Zhou et al. [6]. It uses a series of nested and dense skip-connections. The resolution features from low to high, and the different scale features from small to large are all aggregated together. As a result, their architecture significantly reduced the semantic gap between the feature maps from contracting path and symmetric expanding path. The U-Net++ performs the advantages of capturing fine-grained details, hence generating higher segmentation accuracy rate than U-Net. Therefore, it is promising to exploit the potential of dense skip connections for semantic segmentation on OCT images.

Besides, in this paper, the primary improve in U-net is the middle of architecture, we add dilated convolution layers [7] between the encoder sub-network and decoder sub-network. The dilated convolutions module that could be added to existing convolutional neural network at any resolution. And it is capable of fusing multi-scale contextual information without losing any feature map resolution. The input and output of the dilated module have the same form, thus the module could be plugged into our architectures.

3. Method and Results and Discussions

3.1. Network Architecture

Propose the multi-task OCT segmentation and classification model. The model architecture is illustrated in Fig.2. As seen, it consists of the encoder (left side), the decoder (right side), and the dense skip-connections (center part). The network first extracts image features through down-sampling, which consists of a 3x3 convolution, each followed by a rectified linear unit (ReLU) and a 2x2 max pooling
operation with stride 2. Then feature maps are fed into the dilated module, which could multiply increase the receptive field of feature points and retain the detailed information. The nested, dense skip connections, which could combine the high-resolution features from the down-sampling with the up-sampling output to obtain rich detailed information, in the center part are adopted as the backbone to learning multiscale and different semantic levels of visual features representations. Every step in the decoder consists of an up-sampling of the feature map followed by a 4x4 transposed convolution, which could expand the size of the feature map, a concatenation with the correspondingly feature map from the encoder, and one 3x3 convolution, each followed by a ReLU and batch normalization.

3.2. Dilated Convolution
The dilated convolution module systematically aggregate multi-scale contextual information without losing resolution or analyzing rescaled images. It consists of the five times repeated application of 3×3 convolutions, each followed by a max-pooling operation. We then add the dilated module between the encoder and the decoder. It consists of four dilated convolution layers with the dilation rate of 1, 2, 4, 8 in the center feature map. The receptive fields of their corresponding convolution kernel are 3, 7, 15, 31 respectively. We find dilated convolution that could effectively retain sufficient the low-level features information and accurately locate lesion areas by enlarging receptive field size of kernels. Moreover, it does not introduce any additional parameters without losing resolution.

3.3. Joint Loss Function
Dice loss related to Dice coefficient is the most widely used loss function in semantic image segmentation. Dice coefficient is used to evaluate the similarity of two samples. However, the Dice coefficient is not conducive to measuring the segmentation performance of small targets, because the misclassification of very few pixel categories could lead to a significant decrease in the coefficient. Given the diversity of the shape of the lesion area and highly unbalanced object sizes in the OCT image, exponential logarithmic loss ($L_{EL,Dice}$) [8] is introduced in the paper, which balances the labels not only by their relative sizes but also by their segmentation difficulties. Furthermore, $L_{EL,Dice}$ is essentially more concerned with the structure of small objects that are less accurately segmented. The specific formula is as follows:
$$L_{EL\_Dice} = E\left[ w_i \left( -\ln (\text{Dice}_i) \right)^{\gamma_{iwv}} \right] \quad \text{and} \quad w_i = \left( \frac{\sum_k f_k}{f_i} \right)^{0.5}$$ \hspace{1cm} (1)$$

$$\text{Dice}(r, p) = \frac{2\|r \cap p\|}{\|r\| + \|p\|} = \frac{2\sum_{k=1}^{N} r_i(k) \cdot p_i(k)}{\sum_{k=1}^{N} (r_i(k) + p_i(k))}$$ \hspace{1cm} (2)$$

Where $\kappa$ represents the pixel position and $i$ represent the label. $\{1\}$ is the ground-truth label at $\kappa$. $E[\cdot]$ is the mean value concerning $i$ and $\kappa$ in $L_{EL\_Dice}$. Respectively, $r_i(k)$ is the K ronecker delta which is 1 when $i = 1$ and 0 otherwise. $p_i(k)$ is the softmax probability which acts as the portion of pixel $\kappa$ owned by label $i$ when computing Dice. $w_i$ with $f_k$ the frequency of label $k$, is the label weight for reducing the influences of more frequently seen labels. $\gamma_{iwv}$ controls the nonlinearities of the loss functions. Through several comparison experiments, we could get better results with $\gamma_{iwv} = 0.3$.

we introduce the weighted cross-entropy ($L_{WCE}$) \cite{9} which could measure the difference of probability distribution of different classes and the balanced cross entropy ($L_{BCE}$) \cite{5} which could be weak the effect of class imbalance problems. $L_{WCE}$ and $L_{BCE}$ are usually applied in semantic segmentation tasks. The specific formula of $L_{WCE}$ is as follows: the dice coefficient loss could be defined as:

$$L_{WCE}(y, p) = -\frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{L} \left( y_{i\mu}(\kappa) \log (p_i(\kappa)) \right)$$ \hspace{1cm} (3)$$

$y_{i\mu}(\kappa)$ represents the ground-truth label of the OCT image, $p_i(\kappa)$ represents the predicted label, $N$ represents the number of all pixels in the feature map.

$L_{BCE}$ is a frequently-used automatically balancing strategy function and it could be defined as:

$$L_{BCE}(y, p) = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{L} \left( I \{y_i(\kappa) = 1\} \log p_i(\kappa) \right)$$ \hspace{1cm} (4)$$

$L$ represents the number of category tags, in the paper, it is 3(SRF, REA, and PED). $y_i(\kappa)$ and $p_i(\kappa)$ are $1 \times 3$ matrix(if SPF exists in the OCT image, then the number in the (if REA exists in the OCT image, then the number in the corresponding position in the matrix is 1, otherwise 0). $y_i(\kappa)$ and $p_i(\kappa)$ correspond to the truth value and output value of the index $i$ , $I\{\}$ is a binary indicator. $I \{y_i(\kappa) = 1\}$ is 1 only when $y_i(\kappa) = 1$; otherwise, it is 0.

In this paper, the multi-task learning joint loss function could be defined as:

$$L_{Loss} = \lambda_1 L_{EL\_Dice} + \lambda_2 L_{WCE} + \lambda_3 L_{BCE}$$ \hspace{1cm} (5)$$

Among them, $\lambda_1$, $\lambda_2$, and $\lambda_3$ are three hyper parameters that refer to the weight between losses. We referred to the results of comparative experiments in multiple papers, and selected the value with the best empirical. $\lambda_1$, $\lambda_2$, and $\lambda_3$ are all set to 0.25.

### 3.4. Dataset

The dataset is sourced from the first medical image detection competition for retinal edema in China. The dataset has a total of 100 OCT cubes, and it consists of a training set containing 70 cubes and a test set containing 30 cubes. Each cube contains 128 slices with the resolution of 1024*512. Considering that the label of the test set is not public, we only perform comparison experiments on the validation set.
The dataset is devised as a Four-value segmentation problem, in which SRF, PED, and REA are labeled as different foregrounds and the rest is labeled as background.

3.5. Implementation Details

**Parameter Setting:** The input images and their corresponding segmentation maps are used to train the network with the stochastic gradient descent implementation of PyTorch. All these experiments are carried out on a server with Ubuntu OS and performed on 3 NVIDIA GeForce GTX GPUs with 12GB of memory. All experiments are optimized using Adam optimizer with a learning rate of 0.001 until the verification loss converges. Our network trains a total of 100 epochs, and every 30 epochs drop by one order. The momentum is set to 0.9, and the minimum batch size is 8.

**Evaluation Metrics:** To evaluate the performance of our classifier method, Area under Curve (AUC) metrics are applied. The AUC metrics could steadily assess the performance of the classification module when there is an imbalance of sample categories in the dataset. The Dice coefficient is used to estimate the performance of the baseline semantic segmentation models and our proposed. It is calculated based on two areas \(P\) and \(T\) as:

\[
\text{Dice}(P,T) = \frac{2|P \cap T|}{|P| + |T|}
\]

Where \(P\) is predicted area and \(T\) is ground-truth. \(|P \cap T|\) represents the same elements between two areas \(P\) and \(T\).

3.6. results

Fig.3 Comparison the segmentation results between U-Net++ and our models. (a)~(d) are the segmentation results of four different images and different methods. (e) denote the original images, (f) denote the ground truths, (g) denote the segmentation results of U-Net++ model, (h) denote the segmentation results of our model. In these figures, different colors indicate different types of lesion: SRF(green), PED(red), REA(blue).
Some typical segmentation results are showed in Fig.3. Our network could perform more precise segmentation on large and small lesion areas. Table 1 quantitatively compares the performance of our proposed network architecture with U-Net, U-Net++ models [5] through Dice coefficients. The same classification module is added into all models and calculate the classification AUC scores of three different lesion areas. The experimental results prove that our proposed model is 2.61% higher in the average Dice coefficient than U-Net which is a standard performance baseline model in the medical field. At the same time, our proposed model achieves a significant performance gain over advanced model U-Net++.

### Table 1: Comparison with state-of-the-art models

|         | Dice          | AUC          |
|---------|---------------|--------------|
|         | Mean          | SRF          | PED          | REA          | Mean          | SRF          | PED          | REA          |
| U-Net   | 0.7889        | 0.8378       | 0.7653       | 0.7637       | 0.9860        | 0.9990       | 0.9838       | 0.9752       |
| U-Net++ | 0.7978        | 0.8336       | 0.7814       | 0.7784       | 0.9761        | 0.9987       | 0.9472       | 0.9825       |
| Our model | 0.8150        | 0.8905       | 0.7808       | 0.7735       | **0.9888**    | 0.9992       | 0.9780       | 0.9892       |

### 4. Conclusion

In this paper, proposed an improved encoder-decoder framework to automatically segment SPF, PED, and REA lesions in retinal OCT images and identify three different lesions. Firstly, to decrease the semantic gap between the contextual semantic features and the precise positioning features, we introduce the nested, dense skip connections, which also could capture multi-scale feature maps from different semantic levels. Secondly, a dilated module with 4 dilated convolution layers is adopted to combine the encoder and the decoder. The dilated module not only expand the receptive field of the convolution kernel but also capture fine-grained characteristic information. Finally, to handle the highly unbalanced dataset of SRF, PED and REA lesions classification and the imbalanced size of three lesions problem, we combined the weighted binary cross-entropy loss, the weighted cross-entropy loss, and exponential logarithmic loss effectively. The effectiveness of our architecture is elaborately investigated through a series of comparative experiments. We achieve better performance than previous start-and-art methods.

### Acknowledgments

This work supported by the National Key R&D Program of China (Grant No. 2018YFB1306600); the National Natural Science Foundation of China (Grant Nos.62076207, 62076208, U20A20227).

### References

[1] Shelhamer Evan, Long Jonathan, Darrell Trevor. Fully Convolutional Networks for Semantic Segmentation.[J]. IEEE transactions on pattern analysis and machine intelligence,2017,39(4).

[2] 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.

[3] Hu J, Chen Y, Yi Z. Automated segmentation of macular edema in OCT using deep neural network[J]. Medical image analysis, 2019, 55:216.

[4] Lu D, Heisler M, Lee S, et al. Deep-Learning based Multiclass Retinal Fluid Segmentation and Detection in Optical Coherence Tomography Images using a Fully Convolutional Neural Network[J]. Medical Image Analysis, 2019.

[5] Ruru Zhang, Jiawen He, Shenda Shi, Haihong Ou, and Meina Song. 2020. BMM-Net: automatic segmentation of edema in optical coherence tomography based on boundary detection and multi-scale network. Association for Computing Machinery, New York, NY, USA, 51–59.

[6] Zhou Z, Siddiquee M M R, Tajbakhsh N, et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation[C]// 2018.

[7] Yu F, Koltun V, Funkhouser T. Dilated Residual Networks[J]. 2017.

[8] Wong, Ken CL, et al. "3D segmentation with exponential logarithmic loss for highly unbalanced object sizes." International Conference on Medical Image Computing and Computer-Assisted
Intervention. Springer, Cham, 2018.

[9] Sudre, Carole H, et al. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations." Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, 2017. 240-248.