Region Focus Network for Joint Optic Disc and Cup Segmentation

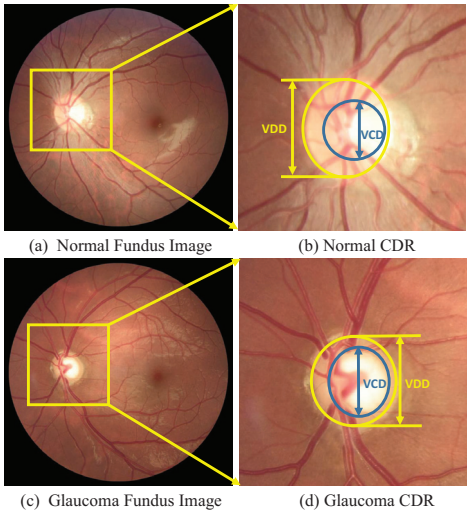
Ge Li,[1*] Changsheng Li,[2†] Chan Zeng,[1*] Peng Gao,[1] Guotong Xie,[1†]
[1]Ping An Technology (Shenzhen) Co. Ltd., Shenzhen, China
[2]School of Computer Science and Technology, Beijing Institute of Technology
{lige676, zengchan517, gaopeng712, xieguotong}@pingan.com.cn
changsheng_li_507@hotmail.com

Abstract

Glaucoma is one of the three leading causes of blindness in the world and is predicted to affect around 80 million people by 2020. The optic cup (OC) to optic disc (OD) ratio (CDR) in fundus images plays a pivotal role in the screening and diagnosis of glaucoma. Existing methods usually crop the optic disc region first, and subsequently perform segmentation in this region. However, these approaches come up with high complexities due to the separate operations. To remedy this issue, we propose a Region Focus Network (RF-Net) that innovatively integrates detection and multi-class segmentation into a unified architecture for end-to-end joint optic disc and cup segmentation with global optimization. The key idea of our method is designing a novel multi-class mask branch which generates a high-quality segmentation in the detected region for both disc and cup. To bridge the connection between the backbone and multi-class mask branch, a Fusion Feature Pooling (FFP) structure is presented to extract features from each level of the pyramid network and fuse them into a final feature representation for segmentation. Extensive experimental results on the REFUGE-2018 challenge dataset and the Drishti-GS dataset show that the proposed method achieves the best performance, compared with competitive approaches reported in the literature and the official leaderboard. Our code will be released soon.

1. Introduction

Glaucoma is a chronic disease that damages the optic nerves and leads to irreversible vision loss (Tham et al. 2014). Early screening and detection methods are essential to preserve vision. To detect glaucoma, the ratio of vertical cup diameter (VCD) to vertical disc diameter (VDD) is an important factor in clinical, which is called CDR for measurements. Normal CDR is 0.3 to 0.4 and the larger may indicate glaucoma or other diseases such as neuro-ophthalmic diseases (Jonas et al. 2000). Figure 1 gives an intuitive illustration. It is extremely time-consuming for acquiring those measurements manually, hence developing accurate algorithms to automatically segment optic disc (OD) and optic cup (OC) from fundus images is pretty meaningful for prompting the large-scale glaucoma screening.



(a) Normal Fundus Image (b) Normal CDR

(c) Glaucoma Fundus Image (d) Glaucoma CDR

Figure 1: Comparisons of normal CDR and glaucoma CDR in fundus images. Figure 1(a) is a normal fundus image and (c) is glaucoma. The Figure 1(b) and (d) are the partially enlarged images for (a) and (c) respectively, in which the region enclosed by the yellow circle is optic disc (OD) and the central bright zone enclosed by the blue circle is optic cup (OC). The VCD is vertical cup diameter, and the VDD is vertical disc diameter. The vertical cup to disc ratio (CDR) is calculated by the VCD to VDD. The OC region in (d) which has a higher CDR is relatively bigger than that in (b).

As illustrated in Figure 1(a) and (c), the optic disc occupies only about 10% of the total area of the fundus image. For the low area proportion and heterogeneous appearance, the segmentation of OD and OC is an extremely challenging task. Thus, it is an important pre-processing to locate the optic disc for segmentation, since correct optic disc location can not only reduce the computational complexity of subsequent optic disc segmentation, but also reduce the noise interference caused by non-optic disc areas in fundus images. Early works in OD and OC segmentation are

*Equal contribution.
†Corresponding author.

based on color, contrast or boundary feature in fundus images (Li et al. 2018a). Among these methods, the pixels or patches of fundus images are determined as background, disc, and cup regions, through a learned classifier with various visual features. While most of the methods based on hand-crafted features are easily affected by pathological regions and low contrast quality. In addition, they segment the OD and OC in two separate steps without considering the mutual relation between them. Recently, deep learning has triumphed over various computer vision tasks, such as image classification (Krizhevsky, Sutskever, and Hinton 2012), segmentation (Long, Shelhamer, and Darrell 2015; Chen et al. 2018), and object detection (Ren et al. 2017). A flurry of research has leveraged Convolution Neural Networks (CNNs) for fundus image segmentation and achieved encouraging results. For example, the state-of-the-art like M-Net (Fu et al. 2018), ET-Net (Zhang et al. 2019) and ResU-net (Shankaranarayana et al. 2017) localize the disc center firstly for explicitly suppressing irrelevant information and highlighting the Region of Interest (RoI), and generate the multi-class probability maps for optic disc and cup. However, these methods usually require much more complicated operations such as traditional method, detection network and pre-segmentation network to obtain cropped boxes first. They extremely rely on localization accuracy because the wrong cropped regions will lead to the wrong prediction.

To address these issues, we propose an end-to-end network architecture to simultaneously conduct the optic disc localization and the multi-class probability segmentation in a joint framework, by the spirit of Mask-RCNN (He et al. 2017). However, it is not a reasonable choice to directly apply Mask-RCNN to solve our problem. Because it detects the optic disc and cup separately, and predicts two label probability maps for foreground and background respectively in each detected region. And the detection of the optic cup is difficult due to the low contrast boundary between the optic cup and the optic disc region, which can lead to poor segmentation. Our experimental results also verify this point. Inspired by this, we propose a novel Region Focus Network (RF-Net) for automatic glaucoma screening, which can detect the optic disc in fundus images and learn a high-quality segmentation mask for disc and cup. The main contributions are summarized as follows:

- Our RF-Net is an end-to-end deep learning system, innovatively integrating detection and segmentation into a unified network without cropping the optic disc in advance, which is also beneficial for unified optimization. In contrast, existing methods (Fu et al. 2018; Zhang et al. 2019; Sevastopolsky et al. 2018) are more complex due to the separate operations.
- A novel multi-class mask branch is designed to generate a high-quality segmentation in the detected region for both disc and cup. Besides, we also present a Fusion Feature Pooling (FFP) structure to extract features from each level of feature pyramid network and fuse them on different levels, so as to further improve the performance.
- For joint OD and OC segmentation, we adopt the weighted focal loss and dice loss in our framework, which can remedy the issue of imbalance data during pixel-wise segmentation for fundus images.
- Furthermore, with much ablation study on the REFUGE[1] dataset and the Drishti-GS (Sivaswamy et al. 2014) dataset, we evaluate the effectiveness of the proposed method, and the results demonstrate that our method achieves better performance, compared with the state-of-the-art.

2. Related Work

Early works employ the handcrafted visual features, including the color features, gradient information, and superpixel-based classifier (Cheng et al. 2013; Li et al. 2018b). Before OD (or OC) segmentation, OD localization (Dehghani, Moghaddam, and Moin 2012; Li et al. 2014) is an inevitable task which can not only reduce the computational complexity, but also reduces the noise caused by non-optical disc areas in fundus images. However, most of the methods are easily affected by pathological regions and low contrast quality. Considering the blood vessel occlusion and the lower-contrast boundary, the OC segmentation is usually more challenging than OD. Therefore, OD and OC segmentation are usually studied independently. For the OD segmentation, Yin *et al.* (Yin et al. 2012) presented a method combined knowledge-based circular hough transform and a novel optimal channel selection for segmentation of the OD. Zheng *et al.* (Zheng et al. 2013) integrated the OD segmentation within a graph-cut framework and then used a Gaussian Mixture Model to decide a posterior probability of the pixel. Both of them are unsuitable for fundus images with low contrast. For the OC segmentation, Narasimhan *et al.* (Narasimhan and Vijayarekha 2011) predicted the OC region based on the K-means clustering technique and then used elliptical fitting to refine the boundaries. Based on the gradient information, Liu *et al.* (Liu et al. 2008) proposed a method using threshold level set to extract the optic cup. These methods are simple and easy to implement, but the robustness is not high.

As convolutional neural networks (CNNs) have recently achieved great success in medical image segmentation tasks (Milletari, Navab, and Ahmadi 2016; Cicek et al. 2016), several attempts have been made to realize simultaneous segmentation of joint optic disc and cup. Sharath *et al.* (Shankaranarayana et al. 2017) proposed a novel improved architecture named ResU-net building upon FCNs (Long, Shelhamer, and Darrell 2015) by using the concept of residual learning and used adversarial training to evaluate the segmentation result. Fu *et al.* (Fu et al. 2018) proposed a deep learning architecture named M-Net, which generates the multi-class probability maps for optic disc and cup after localizing the disc center and transfers the original fundus image into the polar coordinate system. The Stack-U-Net (Sevastopolsky et al. 2018) designed a special cascade network, which is based on the U-Net (Ronneberger, Fischer, and Brox 2015) architecture as building blocks and the idea of the iterative refinement for image segmentation on the example of OD and OC. The Psi-Net (Murugesan et al. 2019)

[1] https://refuge.grand-challenge.org/

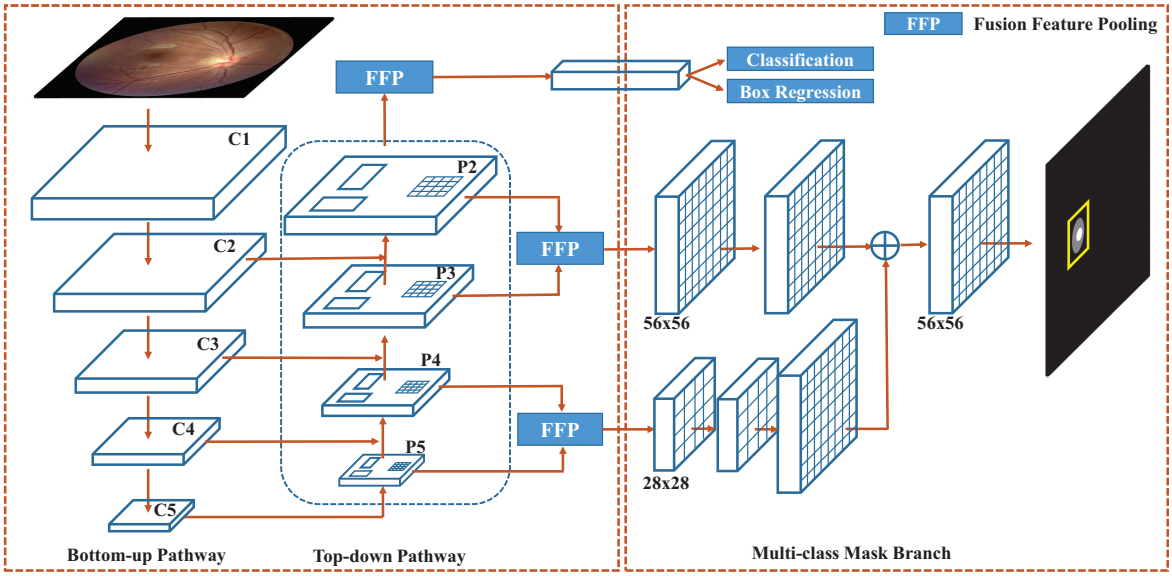


Figure 2: Illustration of RF-Net framework. The left part in the figure is the feature extraction network, which constructed by FPN with MobileNet-v1, and the right part is our proposed multi-class mask branch, which first gets high-level (28×28) and low-level features (56×56) of pyramid structure by Fusion Feature Pooling, then fuses the upsampling of high-level and low-level features to obtain the final representation.

proposed the use of parallel decoders which along with performing the mask predictions also performs contour prediction and distance map estimation for producing segmentation masks with smooth boundaries. Wang *et al.* (Wang et al. 2019) presented a patch-based output space adversarial learning framework to jointly and robustly segment the optic disc and cup from different fundus image datasets. Besides, a novel morphology-aware segmentation loss was proposed to guide the network to generate accurate and smooth segmentation. The method achieved an effective segmentation performance. Edge-aTtention guidance Network (ET-Net) was proposed by Zhang *et al.* (Zhang et al. 2019), which can generate edge-attention representations and sufficient edge information, and achieved an excellent result for optic disc and cup segmentation. The method is the same as the above, localizing the optic disc firstly by the traditional method, then segmentation. The segmentation result depends on the localization accuracy, and needs post-processing.

It is necessary to note that these methods crop images by area of their optic disc (cup) before performing segmentation. It makes the methods not applicable to new, unseen fundus images since it requires a bounding box of optic disc and cup to be available in advance. Compared to the previous methods, we integrate the 'crop-step' into a unified network, which can simultaneously conduct detection and multi-class segmentation in a joint framework for end-to-end optic cup and disc segmentation.

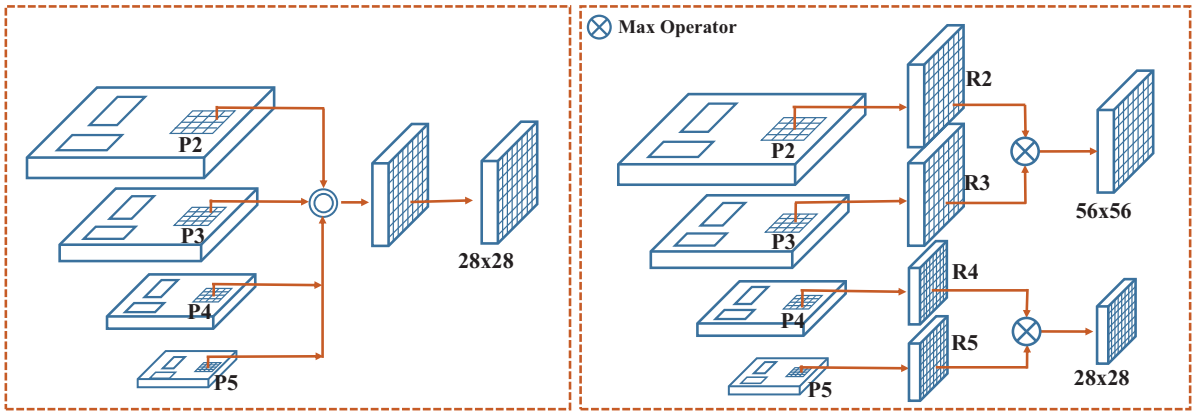
3. Method

Our task is to segment the optic disc and cup in the fundus image. As shown in Figure 1, the optic cup is contained in the optic disc, and both are only around one-tenth of the whole image, which brings a great challenge to segmentation. To this end, we propose an end-to-end method by building a detection network with a multi-class mask branch, which can detect only the optic disc region and simultaneously predict segmentation masks (both cup and disc) on each RoI. Next, we will introduce our architecture in detail.

3.1 Network Architecture

Figure 2 illustrates the architecture of our network which is composed of a backbone network, a detection branch and a multi-class mask branch. Here we use FPN (Lin et al. 2017) structure with MobileNet-v1 (Howard et al. 2017) as the backbone for feature extraction, which uses a top-down architecture with lateral connections to build an in-network feature pyramid from a single-scale input. The feature pyramid is constructed from the backbone network with levels from P_2 through P_5, where l is the pyramid level and P_l has $1/2^l$ resolution of the input image. After that, we propose a new feature fusion strategy, called Fusion Feature Pooling (FFP), to fuse feature grids from different levels. Connected after FFP, the detection branch consists of two stages. The first stage proposes candidate object bounding boxes, and the second stage extracts features using RoIAlign from each candidate box and performs classification and regression. The classification subnet predicts the probability of objects for each RoI. The regression subnet predicts the 4-dimensional class-agnostic offset for each RoI if the object exists.

More importantly, we design a multi-class mask branch that can predict multi-class probability maps for each RoI proposed by the classification subnet and the regression subnet. $\{P_2, P_3\}$ in the feature pyramid are used to get the



(a) Pyramid Feature Pooling (b) Fusion Feature Pooling

Figure 3: Comparison of Pyramid Feature Pooling and Fusion Feature Pooling.The (a) is Pyramid Feature Pooling, which only matches the RoIs from the previous network to the corresponding scale of feature pyramid and proposes feature pooling. The (b) is the proposed Fusion Feature Pooling, P_2, P_3 are the low-level features in pyramid structure and P_4, P_5 are high-level. P_2, P_3 are used to generate fixed-size (56×56) feature maps (R_2, R_3), and the max operator is adopted for the R_2, R_3 to get final features. P_4, P_5 are the same.

56×56 feature maps by FFP and $\{P_4, P_5\}$ are used to obtain 28×28 feature maps. Consequently, we fuse the features with two scales to obtain final segmentation results.

3.2 Fusion Feature Pooling

Considering image segmentation is a pixel-level classification task, whose result depends on the contribution of multi-scale feature maps, we propose a new feature pooling structure called FFP to obtain multi-scale feature information, as illustrated in Figure 3(b). The idea of FFP is to implement pooling layer in each level of feature pyramid network, namely, the region proposals generate a fixed-size (56×56) feature map (R_2 and R_3) by mapping to the P_2 and P_3 layer and another fixed-size (28×28) (R_4 and R_5) in P_4 and P_5 layer. Besides, we fuse the feature maps generated from the P_2, P_3 layer and P_4, P_5 layer separately after feature pooling, to properly align the extracted features with the input. Then, the two branches are input to the multi-class mask branch for segmentation task.

Our proposed FFP structure is different from the Pyramid Feature Pooling in Mask R-CNN, as shown in Figure 3(a). The Pyramid Feature Pooling only matches the RoIs from the previous network to the corresponding scale of feature pyramid and proposes feature pooling in the layer, which ignores the information fusion with different receptive fields. To verify the effectiveness of FFP, we also present a comparison with Pyramid Feature Pooling in Table 2, which indicates the FFP improves the accuracy of segmentation and the performance.

3.3 Multi-class Mask Branch

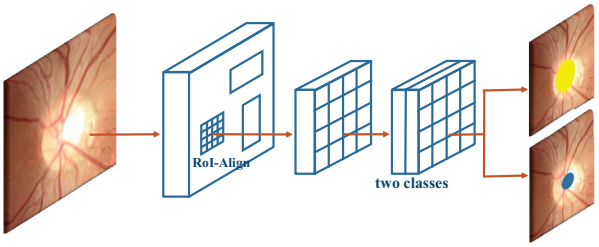
In general, the contrast boundary in the optic cup and disc region is subtle, resulting in that the optic cup cannot be identified correctly by instance-aware semantic segmentation. Motivated by this, we design a detection network for only optic disc detection which avoids the incorrect detection of the optic cup. Besides, we propose a novel multi-class mask branch to produce high-quality probability mask for optic disc and optic cup by pixel-level classification in the detected region.

Recall that in Section 3.2, we introduce FFP to generate feature maps with two scales. After that, they are fed into the proposed multi-class mask branch, as shown in Figure 4(b). Different from the binary mask branch only predicting two classes (foreground and background) in each bounding box, as depicted in Figure 4(a), our branch can predict multi-class object in the disc region. Also, our multi-class mask branch is distinct from traditional semantic segmentation which just generates the probability maps by several convolutions and one upsampling, so as to not make full use of the effective information in the multi-level features. Consequently, we fuse high-level features and low-level features with two scales to obtain finer details.

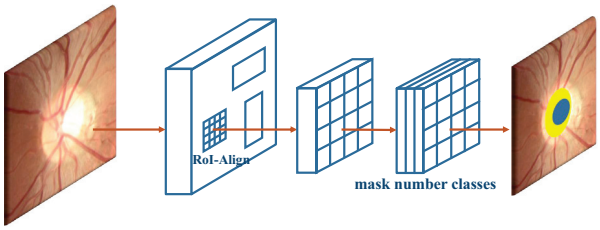
3.4 Loss Function

Our task is a multi-class segmentation problem, which can be seen as multiple binary classification. The overlap of optic disc and cup make the segmentation somewhat challenging in the regions. Moreover, for the glaucoma case, the boundary of the optic cup and optic disc is more indistinguishable. Thus, the multi-class method is more suitable for addressing these issues by treating OD and OC as two independent binary classification problems.

Our loss function in the multi-class mask branch section uses the weighted sum of focal loss and dice loss. Focal loss balances the proportion of positive and negative samples by dynamic weighting, which reduces the weight of a large number of simple negative samples in training. Dice loss was first proposed in U-Net and can reflect the similarity of two contour regions. Since the label of the mask will



(a) Instance–aware binary mask branch



(b) Our multi-class mask branch

Figure 4: Comparison of the binary mask branch in the instance-aware network and our multi-class mask branch. The top of the figure is the binary mask branch which can only perform two classes (foreground and background) segmentation in each RoI. It needs to detect two types of the bounding box, and segment each type into one class. The bottom is our proposed multi-class mask branch, different from the binary mask branch, it just detects one bounding box and segment three classes (optic disc, cup and background) in it.

change according to the RoIs during each iteration, the network is not accurate enough during the initial training, leading to the situation in Figure 5. This makes the proportion of the negative sample much larger than the positive sample, so we choose the focal loss to optimize the parameters of the network. Our loss function is as follows:

$$L_{mask}(p, p^*) = \lambda L_{FL}(p, p^*) + L_{DL}(p, p^*)$$

where

$$L_{FL}(p, p^*) = \sum_{k=1}^{K} [-p_k \alpha (1 - p_k^*)^\gamma \log p_k^* - (1 - p_k)(1 - \alpha)(p_k^*)^\gamma \log (1 - p_k^*)]$$

$$L_{DL}(p, p^*) = 1 - \sum_{k=1}^{K} \frac{2 p_k p_k^*}{(p_k)^2 + (p_k^*)^2}$$

where $p = \{0, 1\}$ is a binary ground truth label, and $p^* \in [0, 1]$ denote predicted probability. K is the number of classes. After detailed tuning parameters, the weighting factor α is set to 1 and the tunable parameter γ is set to 3. λ is a constant and is set to 0.3.

4. Experiments

4.1 Data and Evaluation Metric

In our experiments, we use two glaucoma screening datasets. The first one is the REFUGE dataset, provided by the

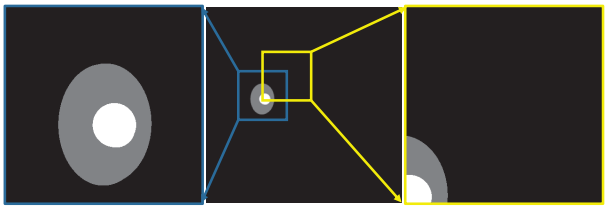


Figure 5: Diagram of sample imbalance. The images on the left and right are partially enlarged for the middle image. When the detection bounding box is precise, the positive and negative sample ratios are balanced. On the contrary, the background ratio is seriously more than the positive sample ratio.

REFUGE-2018 challenge, including both normal and glaucomatous cases and ground-truth of segmentation from multiple human experts. The dataset consists of 1200 color fundus photographs which are split 1:1:1 into 3 subsets equally for training, offline validation and onsite test. The testing set is not available currently, we validate the effectiveness of our proposed method on the validation set. The second one is the Drishti-GS dataset, provided by Medical Image Processing (MIP) group, IIIT Hyderabad. It contains 50 training images and 51 validation images.

As the same in most works (Fu et al. 2018; Zhang et al. 2019), the dice coefficients of the optic cup (Dice_{OC}) and disc (Dice_{OD}), mean intersection-over-union (mIoU), as well as FPS (Frames Per Second), are employed as the evaluation metrics.

4.2 Implementation Details

Training The backbone of the network we use is FPN structure with MobileNet-v1 and without using a pretrained model. The RoIs mentioned above is considered positive if it has IoU with a ground-truth bounding box of at least 0.5 and negative otherwise. The mask loss L_{mask} is defined only on positive RoIs. The mask target is the intersection between a RoI and its associated ground-truth mask.

We adopt image-centric training. Images are resized and padded with zeros to get a square image such that their scale is 512 pixels. Each mini-batch has 4 images per GPU and each image has N sampled RoIs, with a ratio of 1:3 of positives to negatives. N is set to 200. In our experiments, we train the model on 1 NVIDIA Tesla P100 GPU for 100 epochs and employ stochastic gradient descent (SGD) for optimizing the deep model. We use a gradually decreasing learning rate starting from 0.01 and a momentum of 0.9. We employ the piecewise constant learning rate policy where the learning rate is multiplied by 0.1 every 30 epoch. For data augmentation, we apply random flipping (horizontally, vertically) and rotation (90, 180, 270).

Inference At test time, the number of the proposals is 1000. We run the box prediction branch on these proposals, followed by non-maximum suppression. The mask branch is then applied to the highest scoring 10 detection boxes which speeds up inference and improves accuracy (due to the use of fewer, more accurate RoIs). The mask branch can pre-

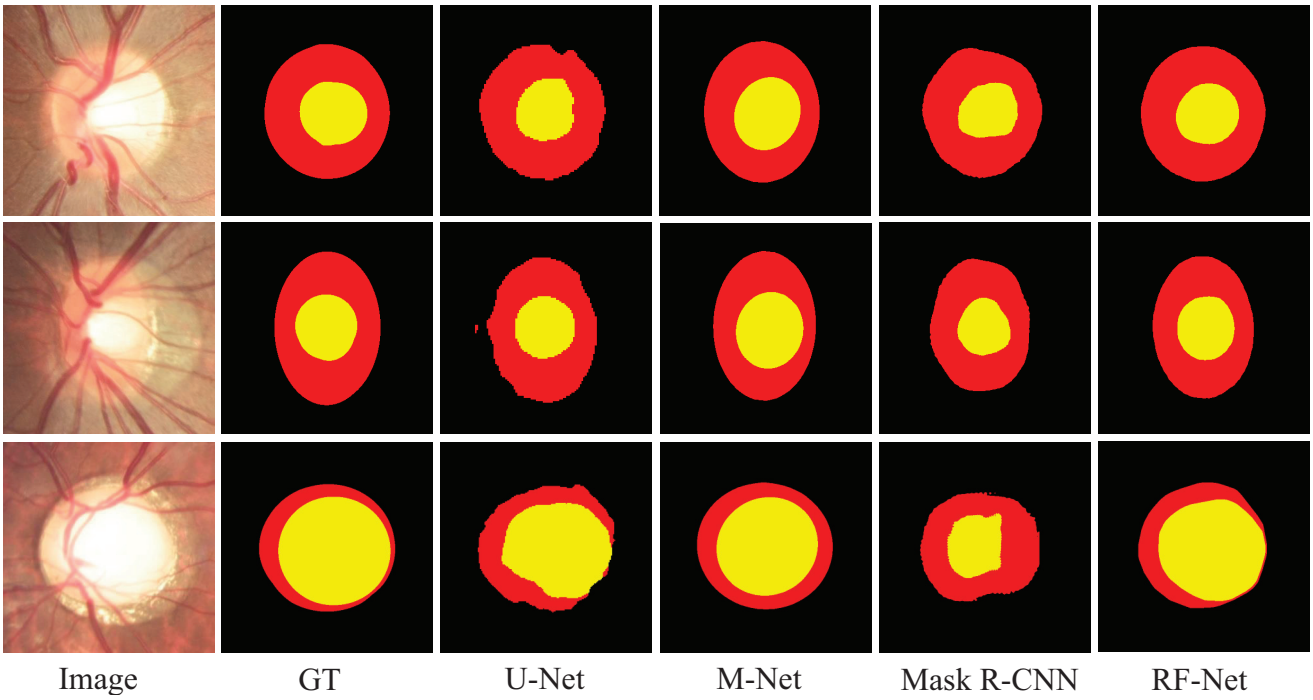


Figure 6: The visual examples of the optic disc and cup segmentation, where the yellow and red region denote the cup and disc segmentation, respectively. From the left to right: fundus image, ground truth (GT), U-Net, M-Net with the polar transformation, Mask R-CNN, and our RF-Net.

dict k masks per RoI, where k is the predicted class by the classification branch. The $m \times m \times k$ floating-number mask output is then resized to the RoI size, where each pixel value represents the probability for OD, OC, and background.

4.3 Comparison with State-of-the-art

Our RF-Net is compared to the state-of-the-art approaches (i.e., U-Net (Cicek et al. 2016), M-Net (Fu et al. 2018), ET-Net (Zhang et al. 2019), Mask R-CNN (He et al. 2017), and AIML[2]. The performance comparison is presented in Table 1. Our proposed method achieves scores of 89.79%, 96.06% and 87.08% in terms of Dice_{OC}, Dice_{OD} and mIoU on the REFUGE dataset, respectively. And it achieves scores of 94.60%, 97.75% and 88.13% on the Drishti-GS dataset. As shown in Table 1, compared to the second best scores achieved by AIML, it can be seen that our method has a marginal improvement of 0.64% for the optic cup and 0.23% for the optic disc. Compared with Mask R-CNN, our method has a huge improvement of 7.43% for the optic cup, 5.61% for the optic disc. On the Drishti-GS dataset, our RF-Net has slender improvement of 1.46% for the optic cup, 0.23% for the optic disc and 0.21% for the mIoU respectively, compared to the second best scores achieved by ET-Net. The experimental results show that Mask R-CNN is not suitable for the segmentation of overlapping bounding boxes which are the mostly same size. In addition, our network can reach a speed of 4 FPS, faster than other works.

[2]https://refuge.grand-challenge.org/Results-ValidationSet/

(a) mask shape of 28x28 **(b) mask shape of 56x56**

Figure 7: Comparison of the different shapes of mask groundtruth. The left part is 28×28 result and the right part is 56×56. It can be clearly seen the edge of the left is serrated, and the details on the right part remain more complete.

We also show a visual comparison of the OD and OC segmentation results of various methods in Figure 6. For better demonstration, the segmentation results are the partially enlarged images for the whole images. The first two rows are normal cases and the last row is glaucoma. For the U-Net method, the segmentation is coarse due to the 16 pixels stride at the final prediction layer, causing the loss of boundary details. The M-Net method obtains more accurate boundaries, but it easily changes the original shape of the optic cup with polar transformation. In terms of Mask R-CNN, the optic cup is not fine because of the challenging cup detection. By contrast, our RF-Net without post-processing can effectively and accurately segment OD and OC regions.

Method	REFUGE			Drishti-GS			
	Dice$_{OC}$(%)	**Dice$_{OD}$(%)**	**mIoU(%)**	**Dice$_{OC}$(%)**	**Dice$_{OD}$(%)**	**mIoU(%)**	**FPS**
U-Net	85.44	93.08	83.12	88.06	96.43	84.87	3.1
M-Net	86.48	93.59	84.02	88.60	96.58	85.88	2.0
ET-Net	89.12	95.29	86.70	93.14	97.52	87.92	3.8
Mask R-CNN	82.36	90.45	76.68	83.53	77.87	75.61	2.2
AIML[1]	89.15	95.83	-	-	-	-	-
Our RF-Net	**89.79**	**96.06**	**87.08**	**94.60**	**97.75**	**88.13**	**4.1**

Table 1: Optic disc/cup segmentation results on retinal fundus images. Comparison experiment of the state-of-the-art approaches (U-Net, M-Net, ET-Net, Mask R-CNN and AIML) on the REFUGE and the Drishti-GS datasets.

Our approach can generate better segmentation results, especially at the boundary regions.

4.4 Ablation Study

Feature Pooling Comparison The performance comparison between Pyramid Feature Pooling and Fusion Feature Pooling is shown in Table 2. When we use the Pyramid Feature Pooling, it achieves 87.83%, 93.96% and 85.76% in terms of Dice$_{OC}$, Dice$_{OD}$ and mIoU, respectively. When we adopt the Fusion Feature Pooling, it yields more favorable scores due to the ability to obtain multi-scale feature layer information with consideration of the size of proposals. Experimental results indicate Pyramid Feature Pooling using only one scale feature information is not sufficient for detailed segmentation. In contrast, the Fusion Feature Pooling using multi-scale feature map information can capture more details and bring great benefit to segmentation.

Feature Pooling	Dice$_{OC}$(%)	Dice$_{OD}$(%)	mIoU(%)
Pyramid	87.83	93.96	85.76
Fusion	**89.79**	**96.06**	**87.08**

Table 2: Optic disc/cup segmentation results from different feature poolings (Pyramid and Fusion Feature Pooling) on REFUGE dataset.

Loss Comparison To evaluate the contribution of the presented loss function, we conduct experiments with different settings on the REFUGE dataset. Our first attempt to train RF-Net using standard Binary Cross Entropy (BCE) loss with our learning rate policy mentioned above. BCE loss is a traditional two-class loss function, while Dice loss defines the similarity between two contour regions. The focal loss is a dynamically scaled cross entropy loss for dealing with class imbalance. As shown in Table 3, when we only use the dice loss, the result outperforms the other two loss functions. However, when we combine BCE loss with dice loss, it does not help in the task. When we append focal loss on dice loss with a weight of 0.3, it results in more favorable scores due to the ability to maintain the stability in the training process, in particular for the optic disc.

Mask Shape Comparison The performance comparison between different mask shape are demonstrated in Table 4, employing the mask shapes of 56×56 improves the performance remarkably. Compared with the mask shape of 28×28 result, it increases 2.25%, 3.03%, 1.96% in terms of Dice$_{OC}$, Dice$_{OD}$ and mIoU respectively. Particularly, in early experiments, we attempt to train with the mask shape of 28×28. However, as illustrated in Figure 7(a), too many times of downsampling make the edge information distortion seriously, which results in poor segmentation. Then we attempt to use the mask shape of 56×56. The results are visualized from Figure 7(b). The segmentation is refiner, demonstrating the superiority of the larger mask shape. The results show that the larger mask shape brings a more fantastic benefit to segmentation.

Loss Function	Dice$_{OC}$(%)	Dice$_{OD}$(%)	mIoU(%)
BCE Loss	88.54	93.84	84.26
Dice Loss	88.50	94.47	84.73
Focal Loss	86.48	93.25	82.26
BCE+Dice Loss	88.25	93.94	84.10
Focal+Dice Loss	**89.79**	**96.06**	**87.08**

Table 3: Optic disc/cup segmentation results from different loss function on the REFUGE dataset.

Mask Shape	Dice$_{OC}$(%)	Dice$_{OD}$(%)	mIoU(%)
28×28	87.54	93.03	85.12
56×56	**89.79**	**96.06**	**87.08**

Table 4: Optic disc/cup segmentation results from different mask shape on the REFUGE dataset.

5. Conclusion

In this paper, we propose a novel RF-Net, an end-to-end trainable architecture, which integrates detection and multi-class segmentation into a unified network for joint optic disc and cup segmentation with a global optimization. To achieve simultaneous segmentation of disc and cup, we design a multi-class mask branch that can predict multi-class probability maps. Moreover, we propose a new feature fusion strategy to utilize multi-scale information, so as to make the segmentation more subtle. In addition, we construct a loss function that can balance the positive and negative samples at the beginning of the training. In online testing, it costs only 0.244s on 1 NVIDIA Tesla P100 GPU to generate the final segmentation map for one fundus image, which is faster than other methods. Finally, we perform our method on the REFUGE and the Drishti-GS datasets, which achieves the state-of-the-art performance.

References

Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. 833–851.

Cheng, J.; Liu, J.; Xu, Y.; Yin, F.; Wong, D. W. K.; Tan, N. M.; Tao, D.; Cheng, C.; Aung, T.; and Wong, T. Y. 2013. Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. *IEEE Transactions on Medical Imaging* 32(6):1019–1032.

Cicek, O.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3d u-net: Learning dense volumetric segmentation from sparse annotation. 424–432.

Dehghani, A.; Moghaddam, H. A.; and Moin, M. 2012. Optic disc localization in retinal images using histogram matching. *Eurasip Journal on Image and Video Processing* 2012(1):19.

Fu, H.; Cheng, J.; Xu, Y.; Wong, D. W. K.; Liu, J.; and Cao, X. 2018. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Transactions on Medical Imaging* 37(7):1597–1605.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. B. 2017. Mask r-cnn. *international conference on computer vision* 2980–2988.

Howard, A.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv: Computer Vision and Pattern Recognition*.

Jonas, J. B.; Bergua, A.; Valckenberg, P. S.; Papastathopoulos, K. I.; and Budde, W. M. 2000. Ranking of optic disc variables for detection of glaucomatous optic nerve damage. *Investigative Ophthalmology & Visual Science* 41(7):1764–1773.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*.

Li, C.; Liu, Q.; Liu, J.; and Lu, H. 2014. Ordinal distance metric learning for image ranking. *IEEE transactions on neural networks and learning systems* 26(7):1551–1559.

Li, C.; Wang, X.; Dong, W.; Yan, J.; Liu, Q.; and Zha, H. 2018a. Joint active learning with feature selection via cur matrix decomposition. *IEEE transactions on pattern analysis and machine intelligence* 41(6):1382–1396.

Li, C.; Wei, F.; Dong, W.; Wang, X.; Liu, Q.; and Zhang, X. 2018b. Dynamic structure embedded online multiple-output regression for streaming data. *IEEE transactions on pattern analysis and machine intelligence* 41(2):323–336.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, J.; Wong, D. W. K.; Lim, J.; Jia, X.; Yin, F.; Li, H.; Xiong, W.; and Wong, T. Y. 2008. Optic cup and disk extraction from retinal fundus images for determination of cup-to-disc ratio. 1828–1832.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. *computer vision and pattern recognition* 3431–3440.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571. IEEE.

Murugesan, B.; Sarveswaran, K.; Shankaranarayana, S. M.; Ram, K.; and Sivaprakasam, M. 2019. Psi-net: Shape and boundary aware joint multi-task deep network for medical image segmentation. *arXiv: Computer Vision and Pattern Recognition*.

Narasimhan, K., and Vijayarekha, K. 2011. An efficient automated system for glaucoma detection using fundus image.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6):1137–1149.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. *medical image computing and computer assisted intervention* 234–241.

Sevastopolsky, A.; Drapak, S.; Kiselev, K.; Snyder, B. M.; and Georgievskaya, A. 2018. Stack-u-net: Refinement network for image segmentation on the example of optic disc and cup. *arXiv: Computer Vision and Pattern Recognition*.

Shankaranarayana, S. M.; Ram, K.; Mitra, K.; and Sivaprakasam, M. 2017. Joint optic disc and cup segmentation using fully convolutional and adversarial networks.

Sivaswamy, J.; Krishnadas, S. R.; Joshi, G. D.; Jain, M.; and Tabish, A. U. S. 2014. Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation. 53–56.

Tham, Y. C.; Li, X.; Wong, T. Y.; Quigley, H. A.; Aung, T.; and Cheng, C. 2014. Global prevalence of glaucoma and projections of glaucoma burden through 2040 a systematic review and meta-analysis. *Ophthalmology* 121(11):2081–2090.

Wang, S.; Yu, L.; Yang, X.; Fu, C.; and Heng, P. 2019. Patch-based output space adversarial learning for joint optic disc and cup segmentation. *IEEE Transactions on Medical Imaging* 1–1.

Yin, F.; Liu, J.; Wong, D. W. K.; Tan, N. M.; Cheung, C.; Baskaran, M.; Aung, T.; and Wong, T. Y. 2012. Automated segmentation of optic disc and optic cup in fundus images for glaucoma diagnosis. In *International Symposium on Computer-based Medical Systems*.

Zhang, Z.; Fu, H.; Dai, H.; Shen, J.; Pang, Y.; and Shao, L. 2019. Et-net: A generic edge-attention guidance network for medical image segmentation. *arXiv preprint arXiv:1907.10936*.

Zheng, Y.; Stambolian, D.; Brien, J. M. O.; and Gee, J. C. 2013. Optic disc and cup segmentation from color fundus photograph using graph cut with priors. 16:75–82.