Automatic Diagnosis of Glaucoma with Retinal Fundus Images

Menglu Zhang[1,2], Shuang Yu[1], and Kai Ma[1]

[1] Tencent HealthCare
shirlyyu, kylekma@tencent.com
[2] Computer Vision Institute, College of Computer Science and Software Engineering of Shenzhen University, Shenzhen, China
zhangmenglu2018@email.szu.edu.cn

Abstract. Glaucoma is an irreversible disease that gradually leads to the deterioration of optic nerve and vision loss. Manual diagnosis of glaucoma and annotation of cup/disc is both labor-intensive and time-consuming. In order to alleviate the workload of doctors, automatic glaucoma classification and cup/disc segmentation have been actively investigated in recent years. In this study, we have developed three efficient end-to-end deep learning based frameworks for the glaucoma classification, cup/disc segmentation and fovea detection tasks, respectively. The proposed models have been evaluated on the REFUGE2 challenge dataset and experiments indicate that our method achieves outstanding performance compared with others.

Keywords: classification · segmentation · detection.

1 Introduction

Glaucoma is a chronic neuro-degenerative disease and it is rated as the leading causes of irreversible blindness [11]. It is estimated that more than 110 million people will suffer from glaucoma world-widely by year 2040 [12]. However, manual screening of glaucoma is time consuming and labor intensive. Given the limited numbers of ophthalmologists or glaucoma specialists available, population wide screening of glaucoma is not realistic. Therefore, there is a strong clinical need for the automatic and accurate diagnosis for glaucoma, including both glaucoma classification and cup/disc segmentation.

Cup/disc segmentation and the estimation of cup-disc-ration (CDR) are among the most commonly adopted screening tests for glaucoma. Clinically, a CDR value larger than 0.6 is considered at risk for glaucoma, and the larger the CDR, the higher the risk. Recent years, along with the development of deep learning, many automatic methods have been developed for the task of cup/disc segmentation. For example, Fu *et al.* proposed a M-shaped network with multi-scale strategy for polar transformed images to produce the cup/disc segmentation maps [3]. Wang *et al.* adopted domain adaptation frameworks to increase the cross-datasets cup/disc segmentation performance [13].

Apart from cup/disc segmentation, automatic glaucoma classification with deep learning has also been actively investigated. Fu *et al.* proposed an ensemble network that combined feature extraction from multiple streams for the glaucoma screening task [4]. Li *et al.* developed an attention guided network for the localization of pathological region and achieved outstanding performance for glaucoma classification [8]. At the same time, a large-scale glaucoma classification database is also publicly released by Li *et al.* [8].

For the REFUGE2 challenge, in total of three tasks related to glaucoma are provided, including the classification of glaucoma, segmentation of cup/disc and the detection of fovea center. In order to tackle with the challenges, three independent deep learning frameworks are utilized in this report, which achieves excellent performance for the online validation set. More details about the framework and experimental results will be described in the following sections.

2 Method

2.1 Clinical Glaucoma Classification

Glaucoma classification aims to classify an input fundus image into glaucomatous or non-glaucomatous. In the classification task, Res2NeXt, which is the combination of ResNeXt [14] and Res2Net [5], is adopted as the backbone of the network. Since the cardinality parameter (the size of the set of transformations) is an essential factor and more effective than the dimensions of depth and width, we use ResNeXt with 32 cardinalities instead of ResNet to aggregate a set of transformations with the same topology. It has been reported that stronger multi-scale representation ability leads to consistent performance gains on a wide range of applications. We replace the bottleneck block in ResNeXt with Res2Net module to improve the feature representation on top of the original layer-wise feature aggregation.

In addition, in order to increase the quality and efficacy of the extracted feature maps, attention mechanism is utilized. The whole architecture is shown in Figure 1. The network is pretrained on ImageNet and then fine-tuned on REFUGE2 training set, and optimized with cross entropy loss. Since the pathological region of glaucoma focuses on the optic disc region, the model is trained on the 3 disc diameter cropped region surrounding the optic disc, i.e. 3DD region.

2.1.1 Convolutional Block with Attention Module (CBAM)

We apply CBAM on the feed-forward convolution outputs in block 2 to increase the network's attention to the relevant features and suppress the unnecessary features. The architecture sequentially infers attention maps along two separate dimensions, channel and spatial. Afterwards, the attention maps are multiplied to the input feature map for adaptive feature refinement.

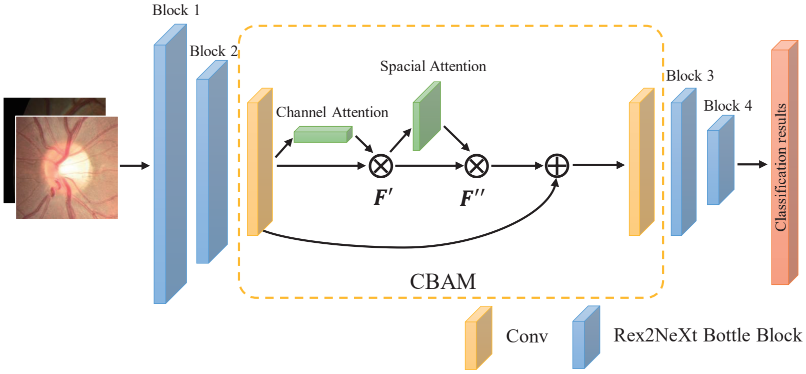

Fig. 1. The pipeline of the proposed classification method

2.2 Optic Cup and Disc Segmentation

In order to develop a more accurate model for the cup and disc segmentation task, we propose a two-stage framework based on a U-Net architecture [15]. To begin with, a Unet [10] with ResNet18 [7] as encoder is used to coarsely segment the OD region. The Region of Interest (ROI) is then cropped as the 2 disc diameter region surrounding the disc area (2DD). The cropped ROI is then fed to the proposed modified Unet model for more accurate segmentation.

The detailed architecture for the second-stage segmentation network is shown in Fig 2. The decoder produces a two-channel probability map for optic cup and disc, respectively. The original implementation of ResUNet [15], which uses Mean Square Error(MSE) as the loss function, can hardly produce satisfactory results. Therefore, we replaced the MSE loss function with Binary Cross Entropy(BCE) loss in this study.

On top of the ResNet as encoder, feature maps generated by the encoder are then passed through the Atrous Spatial Pyramidal Pooling (ASPP) module. The ASPP effectively enlarges the field-of-view of the filters and provides a broader context, by employing multiple parallel filters with different rates. Deep supervision block is also adopted at different levels of the decorder part to improve the performance. Finally, a 1×1 transposed convolution with sigmoid activation is used to produce the final segmentation map. A brief explanation of each individual module is provided in the following subsections.

2.2.1 Deep Supervision (DS)

Deep supervision has been widely adopted in the segmentation network and proved effective in enhancing the performance. It can be conveniently introduced to the network by adding extra side output layers at the decoder section as auxiliary supervision. Detailed architecture of the deep supervision block is shown in the framework at Fig 2.

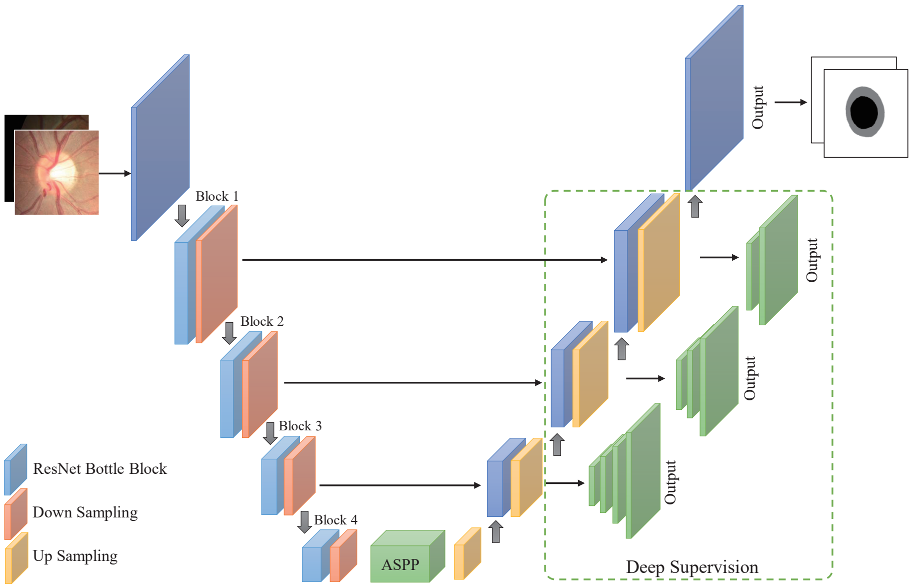

Fig. 2. The pipeline of the proposed segmentation method

2.2.2 Atrous Spatial Pyramidal Pooling (ASPP)

The concept of ASPP comes from spatial pyramidal pooling [6], which is effective at re-sampling features at multiple scales. In the proposed architecture, ASPP acts as a bridge between encoder and decoder, in which the contextual information is captured at various scales [1, 2]. Many parallel atrous convolutions[1] with different rates in the input feature map are fused.

2.2.3 Loss Function

Cross entropy loss is used for the optimization of the segmentation network, for the network final prediction and the auxiliary tasks of the deep supervision. Since the segmentation of optic cup is relatively more difficult than that of disc, in order to increase the importance of cup, the class weight between the cup channel and disc channel is empirically set as 1.5:1 in this study.

$$Loss = BCE(output, GT) + \frac{1}{3}\sum_{i=1}^{3} BCE(side_i, GT) \tag{1}$$

$$BCE(pred, target) = -\sum_{c=1}^{2} \mu_c \cdot target_c \cdot \log(pred_c) \tag{2}$$

2.3 Fovea Detection

Fovea is the central of human vision and thus the accurate localization of it is very important. In this study, we propose to transform the landmark localization of fovea into the framework of objection detection task. In order to realize the task transformation, we use the fovea location as the center of the object and use the disc radius as the width and height of the object. Then, the latest You-Only-Look-Once (YOLO) framework, YOLO5, is adopted for the detection of fovea. An framework of this method is shown in Fig 3.

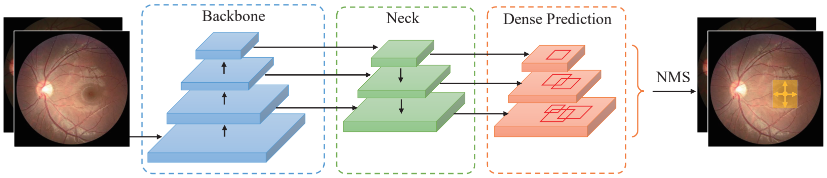

Fig. 3. The pipeline of the proposed detection method

YOLO framework specially separates the bounding boxes and associated class probabilities [9]. The confidence of the object in bounding boxes and the probabilities of the object class are predicted directly from full images. Non maximum suppression is utilized to filter out the redundant proposals, finally the optimal prediction with the highest probability on each image is adopted as the final detection result.

3 Experiments

The proposed approach is trained and evaluated on the REFUGE2 Challenge dataset to demonstrate its effectiveness.

3.1 Datasets

The REFUGE2 challenge database consists of 1200 retinal images stored in JPEG format, with 8 bits per color channel, acquired by ophthalmologists or technicians from patients sitting upright and using one of two devices: a Zeiss Visucam 500 fundus camera with a resolution of 2124 × 2056 pixels (400 images) and a Canon CR-2 device with a resolution of 1634 × 1634 pixels (800 images). The images are centered at the posterior pole, with both the macula and the OD visible, to allow the assessment of the glaucoma. The REFUGE2 online validation set is collected from a different device with the dimension of 1940 × 1940.

3.2 Implement Details

All experiments are performed on an NVIDIA Tesla P40 GPU with 24 GB of memory. Adam optimizer is chosen to optimize the entire framework in order to accelerate the training process. Different initial rates and training epochs are set for each tasks. The detailed experimental settings for different tasks are listed in Table 1.

Table 1. Detailed training settings for different tasks.

Settings	**Classification**	**Segmentation**	**Detection**
Optimizer	Adam	Adam	Adam
Learning rate	1e-2	1e-3	1e-3
Image Size	256	256	640
Batch Size	128	16	16

3.3 Pre-processing Strategy

Similar pre-processing operations are applied in different tasks. In glaucoma classification and optic cup/disc segmentation tasks, diagnosis mainly depends on an analysis of the OD, where the clinical interest is focused on. The acquired full image comprises of extra irrelevant details in large area, which results in both the increase of computation time and the reduction of accuracy. Thus, we use patches centered at the OD intead of full image for training to avoid irrelevant details. We cut out the targeted disc regions by the first-stage ResUNet18, and resize the cropped to 256 × 256 pixels for the second-stage training. In order to avoid overfitting, we employ scaling, rotation and flipping to improve the performance of the model.

For the fovea detection task, the full image is resized to the dimension of 640 × 640 and fed to YOLO5 for training. The labels for detection is set as x, y, w, h, where x, y is the provided fovea center, w, h is emprically set as the radius of the optic disc.

3.4 Post-processing Strategy

In order to reach better performance, we adopt Test-Time Augmentation(TTA) to yield transformed versions of the image for prediction and merge the result together. Different TTA methods are tried for classification, segmentation and detection tasks, respectively. Experiments show that even if simple augmentations (such as rotation or flipping) are applied, TTA can significantly improve prediction performance.

3.5 Performance Evaluation

To show the effectiveness of our methods, we evaluate different improvement techniques on REFUGE2 Datasets in classification, segmentation, and detection tasks, respectively. Fig 4 visualizes some examples of prediction results for glaucoma diagnosis.

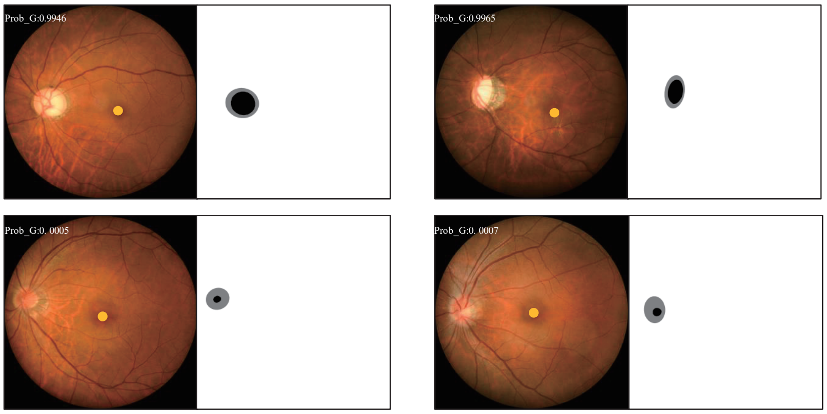

Fig. 4. The examples of prediction results. The classifcation probability of glaucoma and fovea localization are shown on the left image, while cup/disc segmentation masks on the right.

3.5.1 Glaucoma Classification

For glaucoma classification experiments, Area Under Curve(AUC) is used to assess the performance of the algorithm. We verify the effectiveness of CBAM attention module and TTA individually, as listed in Tabel 2. Note that different TTA methods have been tested, *i.e.*, rotation, flipping, noise addition, etc. And finally we choose rotation and resize for the classification task.

It is observed that Res2NeXt50 backbone achieves better performance than ResNet50. And the integration of CBAM module and TTA effectively boost the classification performance by 0.5% and 0.2%, respectively. Finally, the proposed framework achieves an AUC value of 97.04% for the glaucoma classfication on the online validation data.

Table 2. Performance Evaluation on Classification Task.

Method	CBAM	TTA	AUC
ResNet50			0.9585
Res2NeXt50			0.9632
Res2NeXt50	✓		0.9687
Res2NeXt50	✓	✓	**0.9704**

3.5.2 Optic Cup and Disc Segmentation

Three metrics are used to evaluate the performance of the optic cup and disc segmentation task, including the cup dice, disc dice and CDR Error. Ablation studies are performed to validate the efficacy of the deep supervision auxiliary tasks, ASPP modules and TTA methods, on top of the ResUnet architecture. As listed in Table 3, the integration of deep supervision and ASPP improves the cup dice by 0.6% and 0.2%, respectively. TTA further boosts the segmentation performance to achieve the final result of 86.89%, 96.36% and 0.039 for cup dice, disc dice and CDR error, respectively.

Table 3. Performance Evaluation on Segmentation Task.

Method	**DS**	**ASPP**	**TTA**	**Dice$_{\text{Cup}}$**	**Dice$_{\text{Disc}}$**	**CDR Error**
ResUNet				0.8481	0.9597	0.0433
	✓			0.8540	0.9618	0.0430
		✓		0.8504	0.9609	0.0434
	✓	✓		0.8625	0.9621	0.0409
	✓	✓	✓	**0.8689**	**0.9636**	**0.0393**

3.5.3 Fovea Detection

For the fovea detection task, the Average Euclidean Distance (AED) between the predicted fovea center and ground truth is adopted as the metric to evaluate the performance of the model. As listed in Table 4, the YOLOV5 framework achieves an AED error of 13.42, and the utilization of TTA method further reduces the AED error to 10.09.

Table 4. Performance Evaluation on Classification Task.

Method	**TTA**	**MED**
YOLOV5		13.42
	✓	10.09

4 Conclusion

In this study, we proposed three efficient end-to-end deep learning framework for the three critical tasks related to the screening of glaucoma, including the glaucoma classification, cup disc segmentation and fovea detection. Extensive experiments have validated the effectiveness of the proposed frameworks.

References

1. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. CoRR **abs/1606.00915** (2016), http://arxiv.org/abs/1606.00915

2. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017), http://arxiv.org/abs/1706.05587
3. Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Transactions on Medical Imaging **37**(7), 1597–1605 (2018)
4. Fu, H., Cheng, J., Xu, Y., Zhang, C., Wong, D.W.K., Liu, J., Cao, X.: Disc-aware ensemble network for glaucoma screening from fundus image. IEEE transactions on medical imaging **37**(11), 2493–2501 (2018)
5. Gao, S., Cheng, M., Zhao, K., Zhang, X., Yang, M., Torr, P.H.S.: Res2net: A new multi-scale backbone architecture. CoRR **abs/1904.01169** (2019), http://arxiv.org/abs/1904.01169
6. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. CoRR **abs/1406.4729** (2014), http://arxiv.org/abs/1406.4729
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
8. Li, L., Xu, M., Liu, H., Li, Y., Wang, X., Jiang, L., Wang, Z., Fan, X., Wang, N.: A large-scale database and a cnn model for attention-based glaucoma detection. IEEE transactions on medical imaging **39**(2), 413–424 (2019)
9. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. CoRR **abs/1506.02640** (2015), http://arxiv.org/abs/1506.02640
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR **abs/1505.04597** (2015), http://arxiv.org/abs/1505.04597
11. Tham, Y.C., Li, X., Wong, T.Y., Quigley, H., Aung, T., Cheng, C.y.: Global prevalence of glaucoma and projections of glaucoma burden through 2040 a systematic review and meta-analysis. Ophthalmology **121** (06 2014). https://doi.org/10.1016/j.ophtha.2014.05.013
12. Tham, Y.C., Li, X., Wong, T.Y., Quigley, H.A., Aung, T., Cheng, C.Y.: Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology **121**(11), 2081–2090 (2014)
13. Wang, S., Yu, L., Yang, X., Fu, C.W., Heng, P.A.: Patch-based output space adversarial learning for joint optic disc and cup segmentation. IEEE Transactions on Medical Imaging **38**(11), 2485–2495 (2019)
14. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. CoRR **abs/1611.05431** (2016), http://arxiv.org/abs/1611.05431
15. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. CoRR **abs/1711.10684** (2017), http://arxiv.org/abs/1711.10684