Automatic Optic Disc/Cup Segmentation and Glaucoma Classification and Fovea Localization*

Xuesheng Bian[1,6], Yijie Huang[2,6], Xiaoxiao Li[3,6], Xiang Niu[4,6], Junyan Wu[5,6], and Yubin Xie[4,6]

[1] Xiamen University, Xiamen, Fujian, China
xsbian@stu.xmu.edu.cn
[2] Shanghai Jiao Tong University, Shanghai, China
huangyj_wuhan@sjtu.edu.cn
[3] Yale University, New Haven, US
xiaoxiao.li@yale.edu
[4] Tri-Institutional Training Program in Computational Biology and Medicine & Weill Cornell Medical College of Cornell University & Memorial Sloan Kettering Cancer Center, New York City, New York, US
{xin2001,yux2009}@med.cornell.edu
[5] Cleerly Inc. New York City, New York, US
mylotarg1989@gmail.com
[6] The authors are listed based on alphabetical order of last names.

Abstract. Glaucoma is the second leading cause of blindness in the world, according to the World Health Organization. Correctly diagnosis can prevent blindness. Current deep neural network empowered computer vision techniques show the potential of accurate classification and segmentation on the fundus camera images. We propose an ensemble frame work for automatic optic disc/cup segmentation, glaucoma classification and fovea localization. We achieved dice of 0.927 in disc and 0.860 in cup segmentation task, F1 score 0.875 in glaucoma classification task and average euclidean distance of 47.8 in fovea detection task using the REFUGE [15] off-site datasets.

Keywords: Glaucoma · Deep Learning · Optic Disc/Cup Segmentation · Glaucoma Classification · Fovea Detection.

1 Introduction

1.1 Background

Glacoma is a chronic ocular disease that leads to irreversible damage to the optic nerve. It is one of the leading causes of vision loss and blindness. According to a recent report [1], the number of people with the age range between 40 and 80 year-old with glaucoma worldwide in 2013 was estimated to be 64.3 million, increasing to 76.0 million in 2020 and 111.8 million in 2040. Early diagnosis and

* All the authors contribute equally to this study.

prompt treatment is essential that it can efficiently slow or even stop the progression of disease. Assessment of optic nerve head (ONH), whose morphology is changed during glacoma, is a convenient and widely used method for clinical examination[2]. However, ONH assessment requires experienced ophthalmologist and it is time-consuming. An automatic, and high accuracy assessment is preferred for glaucoma diagnosis.

Current research found that morphological features such as cup to disc ratio (CDR) is an important indicator of disease status. A CDR between 0.3 to 0.4 suggest a normal state. A larger CDR may indicate glaucoma. The determination of CDR usually requires ophthalmologists' manual annotation of optic disk and cup, which can be influenced by subjectivity. Also, CDR itself is not accurate enough given the fact that a large CDR can also be introduced by other disease. An unbiased method that integrates other features from ONH is in need.

There is an emerging of expect-level artificial intelligence in medical area over the past several years, largely driven by recent dramatic acceleration in deep learning. Thanks to the increasing availability of large amounts of data and advanced computational resources, deep learning approaches have gained popularity. For example, convolutional neural networks (CNNs) have outperformed the state-of-the-art in many computer vision applications such as ImageNet [11] and COCO [7] . Similarly, CNNs are successfully applied to medical image analysis problems, including the detection or segmentation of structures in CT/MRI images [19] and microscopy pathology images [20]. Deep learning based segmentation methods have gotten great achievement on medical imaging. And the deep convolutional neural network empowered classification and segmentation largely assist the clinical decision making.

1.2 The Retinal Fundus Glaucoma Challenge

REFUGE Challenge consists of THREE Tasks:

- Classification of clinical Glaucoma
- Segmentation of Optic Disc and Cup
- Localization of Fovea (macular center)

As stated in the REFUGE Challenge website [15]:

Task 1. The reference standard for glaucoma presence obtained from the health records, which is not based on fundus image ONLY, but also take OCT, Visual Field, and other facts into consideration. For training data, glaucoma and non-glaucoma labels (a.k.a. the reference standard) are reflected on the image folder names.

Task 2. Manual pixel-wise annotations of the optic disc and cup were obtained by SEVEN (was 3 as proposed) independent GLAUCOMA SPECIALISTS from Zhongshan Ophthalmic Center, Sun Yat-sen University, China. The reference standard for the segmentation task was created from the seven annotations, which were merged into single annotation by another SENIOR GLAUCOMA SPECIALIST. It is stored as a BMP image with the same size as the corresponding fundus image with the following labels:

128: Optic Disc (Grey color) 0: Optic Cup (Black color)

The numbers in front of the structures indicate the pixel-wise labels. All other pixels are labeled as 255(White Color).

Task 3. Manual pixel-wise annotations of the fovea (macular center) were obtained by 7 independent GLAUCOMA SPECIALISTS. The reference standard for localization task was created by using the average of selected annotations from the 7 annotations, for each individual images by another independent GLAUCOMA SPECIALIST.

Training and Off-site and On-site Test datasets A total of 1200 color fundus photographs are available. The dataset is split 1:1:1 into 3 subsets equally for training, offline validation and onsite test, stratified to have equal glaucoma presence percentage. Training set with a total of 400 color fundus image will be provided together with the corresponding glaucoma status and the unified manual pixel-wise annotations (a.k.a. ground truth). Testing consists of 800 color fundus images and is further split into 400 off-site validation set images and 400 on-site test set images.

Evaluation Framework This challenge evaluates the performance of the algorithms for: (1) glaucoma classification, and (2) optic disc/cup segmentation. Thus there will be two main leaderboards. The average score across the two leaderboards will determine the final ranking of the challenge. In case of a tie, the classification leaderboard score has the preference.

Classification results will be compared to the clinical grading of glaucoma. Receiver operating curve will be created across all the test set images and an area under the curve (AUC) will be calculated. Each team receives a rank (1=best) based on the obtained AUC value. This ranking forms the classification leaderboard.

Submitted segmentation results will be compared to the reference standard. the disc and cup Dice indices (DI), and the cup-to-disc ratio (CDR) will be calculated as segmentation evaluation measures. Each team receives a rank (1=best) for each evaluation measure based on the mean value of the measure over the set of test images. The segmentation score is then determined by adding the three individual ranks (2xDI and 1xCDR). The team with the lowest score will be ranked #1 on the segmentation leaderboard.

In Task 3 (fovea localization), the evaluation criterion is the Average Euclidean Distance between the estimations and ground truth, which is the lower the better.

2 Related Work

The tate-of-the-art deep learning driven methods for semantic segmentation [8] have achieved inspiring success in image segmentation in terms of accuracy, speed and reproducibility. Recent research [14] shows the polar transform helps the disc/cup segmentation.

For the classification task, while prior works mainly focus on global image screening. Cup-to-disc ratio(CDR) serves as an important role in diagnose. The

calculation requires accurate sub-organ segmentation. Recent research [9] highlights the benefit of region of interest (RoI) cropping, model ensemble and geometry transformation.

3 Methods

3.1 Optic Disc/Cup Segmentation

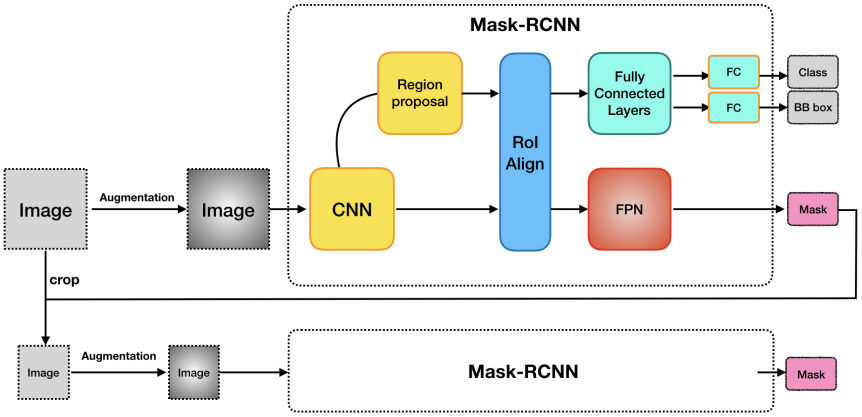

Fig. 1. Two-step segmentation with two-stage detector

Two-stage detector Current state-of-the-art object detectors are based on a two-stage detector, which uses proposal-driven mechanism. The first stage generates a set of candidate object regions and the second stage classifies each candidate region as one of the foreground classes or as background using a convolutional neural network. For optic disc/cup segmentation, we used state-of-the-art two-stage detector: Mask-RCNN [4]. Mask-RCNN is built on Faster R-CNN [6]. It uses CNN layers to extract image features, and it uses a CNN region proposal network to create region of interests (RoIs), which were then warped into fixed dimension. It is then feed into fully connected layers to make classification and boundary box prediction. In Mask-RCNN, After the ROIs proposal, Feature Pyramid Networks were added to build the mask for each instance. We applied Mask-RCNN for both optic disc and cup segmentation.

Given the fact that the size of optic disc and cup is relatively small when compared to the whole image, we propose a two-step segmentation. Firstly, we segment the optic disk. Then centered at the center of the segmented optic disk, we cropped a 512*512 optic disk image from the original image. Then we do segmentation for optic cup.

One-stage detector One stage detectors are applied over a regular, dense sampling of object locations, scales, and aspect ratios, which is simpler and faster [5]. For optic disc segmentation task, we propose a new network based on Unet [8] architecture, whose encoder part introduces dense block with multi-scale kernel filters for extracting richer semantic information. Dilate convolutional layer is added following every denseblock to obtain large receptive field.The detail is shown in Fig. 1. For optic cup segmentation task, we suppose that the optic

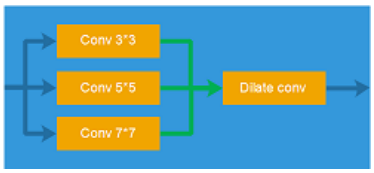

Fig. 2. Dense Block

disc area provides precise location information for optic cup's position. Thus, we regard the result of optic disc segmentation as attention area for current task. A new binary channel is added to the input layer to supply more prior information.The proposed network architecture is shown in Fig. 2.

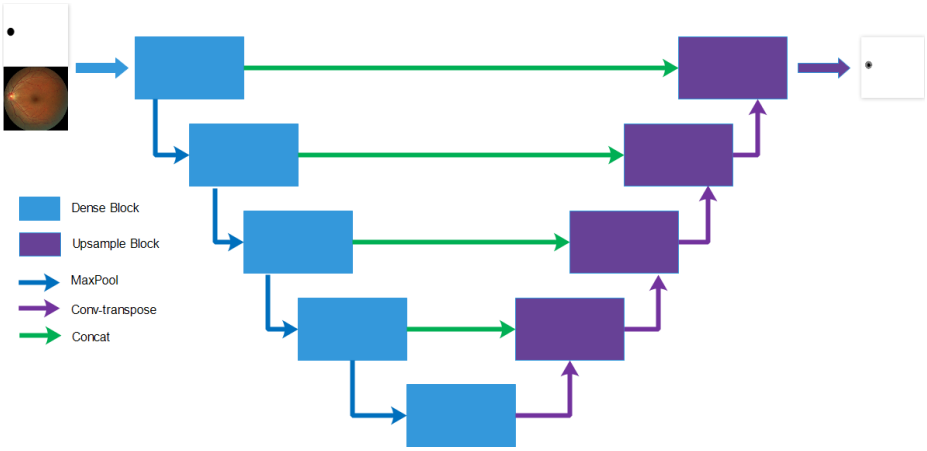

Fig. 3. Segmentation Network

We trained these two model with a same loss function which consists of two items .

$$Loss = Loss = \alpha \times Dice\_loss + \beta \times Cross\_entropy\_loss \tag{1}$$

$$Dice\_loss = 1 - \frac{2 \times \sum(p \times y)}{\sum p + \sum y} \tag{2}$$

$$Cross\_entropy\_loss = -(y \log p + (1-y) \log(1-p)) \tag{3}$$

where α and β is the weight of the two items. We set $\alpha = 1$ and $\beta = 0$ at the beginning of training to accelerate convergence, and decrease α while increase β in the iteration to get high accuracy.

3.2 Glaucoma Classification

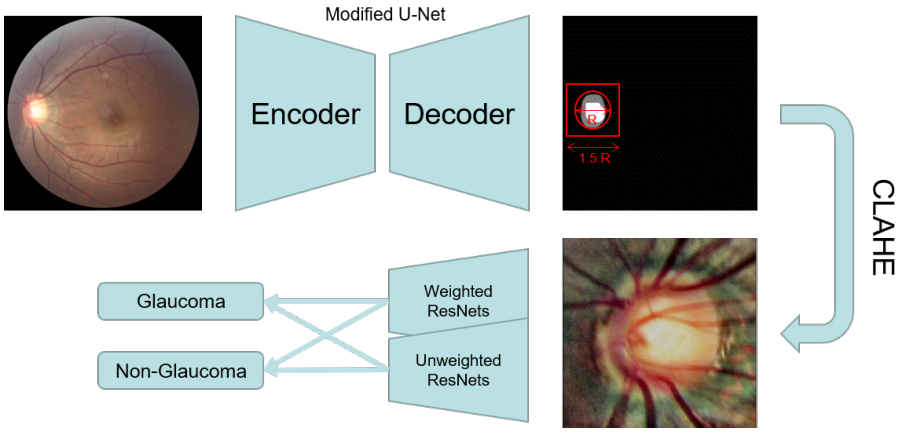

Fig. 4. Classification Workflow

Glaucoma displays its main clinical symptoms in the optic disc region. Based on this mechanism, with optic disc segmentation predictions, we first crop the regions of interest and resize the patches to size 448×448 to exclude irrelevant background contexts. We further employ CLAHE [10] for contrast enhancement and mean color normalization to images acquired under different imaging conditions.

For the glaucoma assessment task, we propose to use ImageNet [11] pre-trained ResNet-18 [12] trained on the training set with a four-fold cross validation strategy. To harness the imbalanced class distribution, we train an extra set of models using a weighted cross-entropy loss function.

$$Loss = -(w_p \times y \log p + w_n \times (1-y) \log(1-p)) \tag{4}$$

where weights $w_p = 9$ and $w_n = 1$. In the inference stage, both the weighted model set and the unweighted set, eight models in total, are averaged to acquire more robust final predictions.

3.3 Glaucoma Classification with semi-supervised GAN

Semi-supervised GAN [21] is a generative model based semi-supervised learning method. For Glaucoma and Non-Glaucoma two classes classification problem,

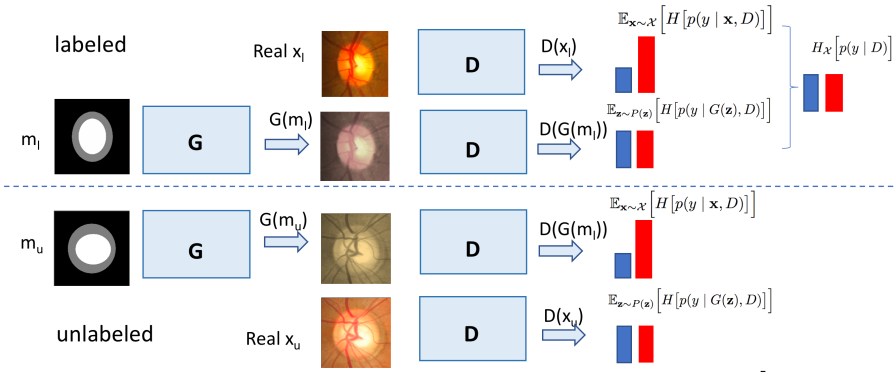

Fig. 5. Semisupevised Classification Workflow

by learning the latent feature of each class, we could use large amount unlabeled data to improve classification performance. GAN has been approved its latent feature representation ability by trading-off a generative model and discriminative model. Given segmented mask, the generator G generates fake image; and a discriminator D here is more like a classifier, giving each class a prediction score [22]. For the labeled image, we resampled the images and keep the sample size at each class. For the Generator, the loss function is defined as:

$$L_G = \min_G -H_G[p(y|D)] + \mathbb{E}_{\mathbf{m}\sim\mathbf{M}}[H[p(y|G(\mathbf{m}), D)] \tag{5}$$

, and for the Discriminator, the loss function is defined as:

$$L_D = \max_D H_{\chi^L}[p(\mathbf{y}|D)] - \mathbb{E}_{\mathbf{x}\sim\chi}[H[p(y|\mathbf{x}, D)]] + \mathbb{E}_{\mathbf{m}\sim\mathbf{M}}[H[p(y|G(\mathbf{m}), D)] + \lambda\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\chi^L}[CE[\mathbf{y}, p(y|\mathbf{x}, D)]]$$

where

$$H[p(y|\mathbf{x}, D)] = \sum_{k=1}^{K} p(y = k|\mathbf{x}, D) log p(y = k|\mathbf{x}^i, D) \tag{6}$$

and

$$CE[\mathbf{y}, p(y|\mathbf{x}, D)] = -\sum_{i=1}^{K} y_i log p(y = y_i|\mathbf{x}, D) \tag{7}$$

3.4 Fovea Localization

For fovea detection, we propose a 2D convolution regression network that takes in the input of the histogram equilization (CLAHE) enhanced [10] image along with the output of the adaptive watershed segmentation [17] to produce a density map D(x) that has high probability values at fovea centers. Our detection network has three layers, where each layer consists of a $N \times 3 \times 3$ convolution,

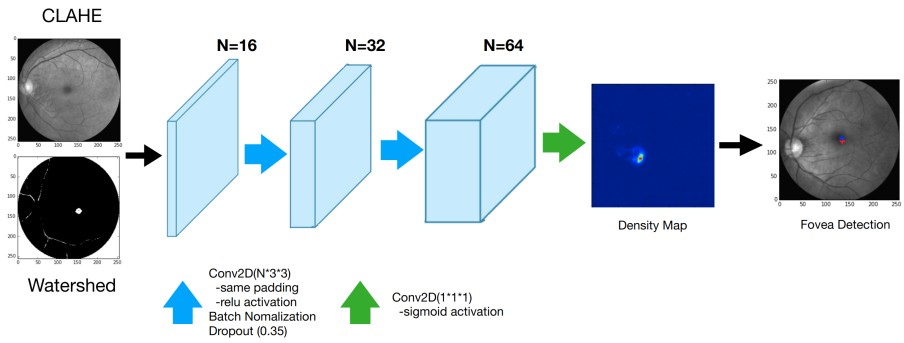

Fig. 6. Detection Network

where N is the number of convolution kernel, followed by a ReLU activation, batch normalization and a dropout [16] at the rate of 0.35. In the last layer, we use a $1 \times 1 \times 1$ convolution similar to the U-Net structure [8] to reduce the number of output channels to a regressed density map D(x). The input and output of our detection network are $100 \times 100 \times 8$ fovea region patches which are cropped based on the image and watershed masking. A graphical representation of the detection network architecture is shown in Fig. 5.

4 Experiments

4.1 Optic Disc/Cup Segmentation

4.1.1 Two-stage detector

Data splitting We split the training set of 400 images, including 40 glaucoma images and 360 non-glaucoma images provided by REFUGE to two parts: a training set consisting 32 glaucoma images and 288 non-glaucoma images, and a fixed validation set of 8 glaucoma images and 72 non-glaucoma images.

Augmentation The overcome the limit of small training set, we applied a series augmentation to our training data. 0-5 of the following transformation was randomly applied to the original image on the fly:

- Flip on left and right direction with 50% probability
- Flip on up and down direction with 50% probability
- One of 90, 180, 270 degree rotation
- Gaussian blur with random sigma within 0-20 pixel with 20% probability
- Scale the image randomly within 0.7 and 1.5 independently in x and y directions with 20% probability
- Multiply the image intensity randomly within 0.7 and 1.5 with 20% probability

Model We used ResNet-50 [12] that is pre-trained with COCO [7] as backbone for our CNN feature extractor. We first do whole image training and prediction on optic disk and cup. Then we used the cropped disk image to train and predict optic disk and cup again. For the cropped image, we then did polar transformation. Then training and prediction on polarized image were performed.

We keep 20% trainining data for integration test.

One-stage detector Intuitively, optic cup never exceed the scope of the optic disc, therefore, we segment optic cup first. All images for training are downsampled to decrease the consumption of memory and eliminate redundant information or noise , which is proviede by REFUGE. To cope with the big difference between training data and validation data in color contrast, we apply histogram equalization and gamma transform randomly in all training data. We sampled 300 images randomly for training and the rest for validation form the dataset provided by REFUGE. Every image was augmented 10 times with random parametres during training. Several different scales were tested form 128×128 to 1024×1024 in our experiments and each of them is acceptable. The smallest scale was choosed for saving time. Adam [13] optimizer was picked to train our model and the learning rate starts form 0.001 and multiplys by 0.95 every epoch.

In the validation stage, images in validation dataset are downsampled to fix size first(128×128 in our experiment),and upsample the result generated from our model with splineinterpolation.

4.2 Glaucoma Classification

We split the training set of 400 images, including 40 glaucoma images and 360 non-glaucoma images provided by REFUGE to two parts: a training and validation set consisting 32 glaucoma images and 288 non-glaucoma images, and a fixed test set of 8 glaucoma images and 72 non-glaucoma images. To acquire each model, we validate the ResNet-18 [] on non-overlapping 8 glaucoma images and 72 non-glaucoma images and train it on the rest of the training and validation set. In the training stage, Adam optimizer [13] is employed and learning rate starts from 0.0001 and decays by 0.9 per 30 epochs. The model of the lowest cross-entropy loss is saved for the inference stage.

4.3 Fovea Localization

We split the training set of 400 images, including 40 glaucoma images and 360 non-glaucoma images provided by REFUGE to two parts: a training set consisting 32 glaucoma images and 288 non-glaucoma images, and a fixed validation set of 8 glaucoma images and 72 non-glaucoma images. The performance of our detection model is evaluated based on euclidean distance with the ground truth fovea center in both validation and test set. We noticed that the size of training/validation and test set are different, so there is a scaling effect of our

detection error. The transformation from resized 256×256 image to original 1634×1634 leads to an enhancement of euclidean distance error by 8.3 times on the validation set and 6.4 times on the test set. During training, the biases are initialized with zero, and the learning rate is initialized to a = 0.001 and decreased by a factor of 10 every 10 epochs using Adam optimizer [13]. We saw good convergence of validation error starting around 20 epoch. The best model was selected with the lowest validation euclidean distance. We also explored the detection frame work using different neural network parameters and canonical methods of Watershed.

5 Results

5.1 Optic Disc/Cup Segmentation

Two-stage detector On self-held-out validation set, our best model (two-step, no polarization, shown in bold in the table) get 0.91 Iou Score on the optic disc segmentation task and 0.81 Iou Score on optic cup segmentation task. On official validation set, our best model get 0.860 on dice optic cup, 0.927 on dice optic disk, 0.06 on MAE CDR. We also tested one-step (direct prediction on cup without cropping) and two-step with polarization. Results shown in the following table. They are not as good as other methods.

One-stage detector On self-held-out validation set, our model get 0.95 Iou Score on the optic disc segmentation task and 0.81 Iou Score on optic cup segmentation task. On official validation set, our model get 0.756 on dice optic cup, 0.851 on dice optic disk, 0.09 on MAE CDR.

Table 1. Segmentation Results

	Our Validation		Official Validation		
Method	IoU Disc	IoU Cup	Dice Disc	Dice Cup	MAE CDR
Two-stage one-step	0.91	0.75	NA	NA	NA
Two-stage two-step polar	0.91	0.77	NA	NA	NA
One-stage model	**0.95**	**0.81**	0.851	0.756	0.09
Two-stage two-step	0.91	**0.81**	**0.927**	**0.860**	**0.06**

5.2 Glaucoma Classification

The weighted model set gets 0.85 F1 Score on the test set while the unweighted model set 0.82 F1 Score. By averaging 8 models, the F1 Score increases to 0.875.

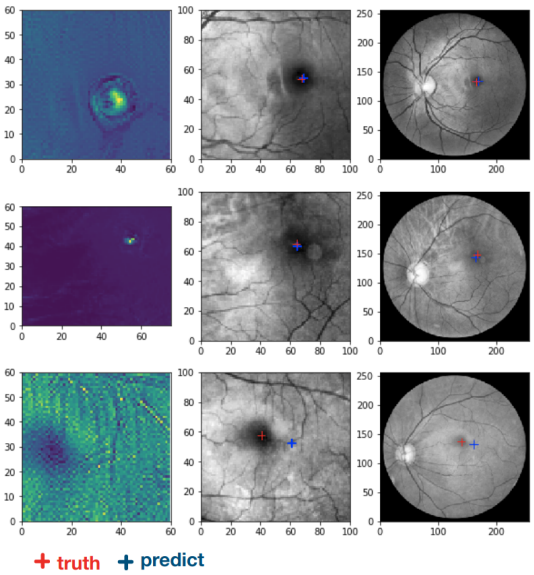

Fig. 7. Detection Network

5.3 Fovea Localization

Our detection model gets average euclidean distance of 47.8 on the off-site validation set images in size of 1634×1634. Representative examples were plotted in the Fig. 6. For the failure sample, our model fails to detect the fovea region because of the false activation in the blood vessels around fovea. This is probably due to image calibration variations that lead to abnormal pixel intensity. By increasing sample size and rare samples during training, our model will be able to learn and handle potential marginal cases such as false activation in blood vessels.

6 Conclusion

6.1 Optic Disc/Cup Segmentation

Here we applied one- and two-stage detector to address the segmentation task. Both of them have high performance, while Mask-RCNN based two-stage detector shows better performance on the official held-out.

6.2 Glaucoma Classification

To tackle this challenging task, we employ optic cup localization for irrelevant background context exclusion, weighted loss function for class distribution imbalance and model ensemble for better robustness.

6.3 Fovea Localization

We introduced an end-to-end 2D convolutional detection network for the problem of fovea detection in fundus images. We evaluate the proposed model and show that it outperforms the Watershed benchmark method and other parameter settings.

6.4 Overall

Here we present an automatic pipeline for glaucoma disease assessment. We applied one- and two-stage detector to address the segmentation task, designed an ensemble CNN model to do glaucoma classification, and developed a novel method for the fovea localization task. This pipeline has decent performance on the held-out validation set, and helpfully will help the diagnose of glaucoma.

7 Contribution

Yubin Xie performed the two-stage detector (Mask-RCNNand cropping task, and Xuesheng Bian performed the one-stage detector (U-Net) for optic disc/cup segmentation. Yijie Huang and Junyan Wu applied the weighted and unweighted ResNet for glaucoma classification. Xiaoxiao Li proposed the semi-supervised learning based classification algorithm. Xiang Niu developed the density-based detection network for the fovea localization task.

References

1. Y. C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C. Y. Cheng, Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis, Ophthal- mology, vol. 121, no. 11, pp. 20812090, 2014.
2. J. Jonas, W. Budde, and S. Panda-Jonas, Ophthalmoscopic evaluation of the optic nerve head, Survey of Ophthalmology, vol. 43, no. 4, pp. 293320, 1999.
3. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich fea- ture hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014. 1, 2, 5
4. Mask R-CNN. Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick. IEEE International Conference on Computer Vision (ICCV), 2017.
5. Focal Loss for Dense Object Detection. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollr. IEEE International Conference on Computer Vision (ICCV), 2017.
6. Faster R-CNN: Towards real-time object detection with region proposal networks S Ren, K He, R Girshick, J Sun Advances in neural information processing systems, 91-99
7. Tsung-Yi Lin, Michael Maire,et al. C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. 2014, ECCV DOI:10.1007/978-3-319-10602-1_48
8. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[M]// Medical Image Computing and Computer-Assisted Intervention MICCAI 2015. Springer International Publishing, 2015:234-241.

9. Fu H, Cheng J, Xu Y, et al. Disc-aware Ensemble Network for Glaucoma Screening from Fundus Image[J]. IEEE Transactions on Medical Imaging, 2018, PP(99):1-1.
10. Reza A M. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement[J]. Journal of Vlsi Signal Processing Systems for Signal Image & Video Technology, 2004, 38(1):35-44.
11. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3):211-252.
12. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2015:770-778.
13. Kingma D, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.
14. Fu, Huazhu, Cheng, Jun, et al. 2018, arXiv preprint arXiv:1801.00926
15. https://refuge.grand-challenge.org
16. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proc. Intl. Conf. Mach. Learning. pp. 1050 1059 (2016)
17. Nandy, K., Chellappa, R., Kumar, A., Lockett, S.J.: Segmentation of nuclei from 3D microscopy images of tissue via graphcut optimization. IEEE Trans. Sel. Topics Signal Process. 10(1), 140150 (Feb 2016)
18. Yang, Samuel J., et al. "Assessing microscope image focus quality with deep learning." BMC bioinformatics 19.1 (2018): 77.
19. Zhu, Wentao, et al. "Deeplung: 3d deep convolutional nets for automated pulmonary nodule detection and classification." arXiv preprint arXiv:1709.05538 (2017).
20. Anthimopoulos, Marios, et al. "Semantic Segmentation of Pathological Lung Tissue with Dilated Fully Convolutional Networks." arXiv preprint arXiv:1803.06167 (2018).
21. Salimans, Tim, et al. "Improved techniques for training gans." Advances in Neural Information Processing Systems. 2016.
22. Springenberg, Jost Tobias. "Unsupervised and semi-supervised learning with categorical generative adversarial networks." arXiv preprint arXiv:1511.06390 (2015).