Appendix: Weakly-supervised Metric Learning with Cross-Module Communications for the Classification of Anterior Chamber Angle Images

Jingqi Huang,[1] Yue Ning,[2] Dong Nie,[3*] Linan Guan,[1] Xiping Jia [1]
[1]*Guangdong Polytechnic Normal University*, [2]*Stevens Institute of Technology*
[3]*University of North Carolina at Chapel Hill*

1. Dataset Labeling

In ACA dataset, each image is labeled by a senior ophthalmologist. A two-stage check is performed to ensure the quality of labeling. In the first stage, 5 undergraduates with medical and non-medical backgrounds are trained to perform the check. The check quality is controlled based on the following standards: (1) the image should not contain severe resolution reductions or significant artifacts; (2) the ACA structure should be complete; (3) the image's illumination should be acceptable (i.e., not too dark or too bright); (4) the image should be focused on the four structures. In the second stage, there are 3 examiners to perform the check. One is a board-certified ophthalmologist with more than 10 years' experience and the other two are postgraduate ophthalmology trainees who have passed a pretraining test. The two postgraduate ophthalmology trainees label each image separately according to the Scheie angle depth system. Then, the ophthalmologist makes the final decision on each image.

2. Dataset Details

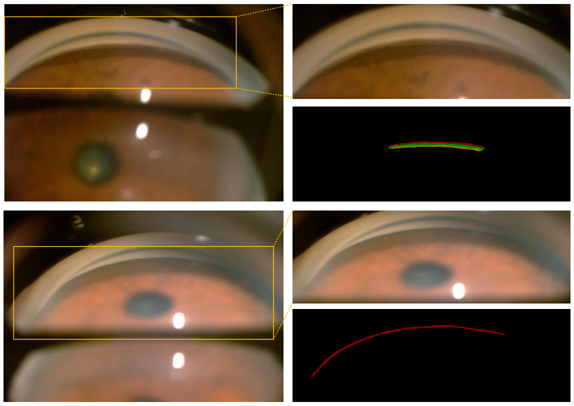

Figure 1. Example of cropped ACA images by YoLo detector. The top row is N2 and the bottom row is N4.

All datasets are randomly divided into training, validation and testing sets. Training/validation sets are used to develop the methods, while testing sets are used for final evaluation. Note that the testing sets are not used during method development in any way.

2.1. ACA dataset

ACA dataset are first randomly divided into a training/validation set, and a hold-out testing set. As shown in Figure 1, to eliminate the influence of background noise, YoLo detector is applied to automatically crop the target region of SL, TM, SS and CBB. The image size ranges from 215×765 to 1272×3264 after croping. To balance the classification performance and computational cost, we resize all the images to 700×2100 using bilinear interpolation.

As described in the main paper, we partitioned the dataset into a training set (80%) and a testing set (20%) based on a random seed of 72. We have 802 images for training and 197 for testing. Besides, the **GCNet** framework is designed for multi-task image classification with image-level annotations and partial pixel-level annotations. We have 999 image-level labels and 100 pixel-level labels. In the training part, we shuffle the 802 images and select 642 for training and 160 for validating using random seed 42. During training, we have 642 images with image-level labels and 83 images with pixel-level labels among a total of 642 training samples; We have 160 images with image-level labels and 17 images with pixel-level labels among a total 160 validation samples. In the testing process, we have 197 images with whole image-level labels but no pixel-level labels.

2.2. REFUGE dataset

REFUGE dataset has 1200 images with 10% of glaucoma (positive) images and 90% for non-glaucoma (negative) images. We integrate the training set and validation set. In the training process, we shuffle the 800 images and select 640 for training and 160 for validation using random seed 42. Although all the images in the REFUGE dataset have whole image-level labels and whole pixel-level labels,

[*] Now at Alibaba Inc.

we randomly select 100 images with pixel-level labels from the integrated 640 images, and among these 100 images, 83 of them are for training and 17 for validation.

2.3. SIGF dataset

SIGF dataset is randomly divided into training, validation and testing sets. It consists of 3,671 images, 71.82% for training, 9.15% for validation and 19.03% for testing. Positive samples account for 4.16% of the entire training set and the rest are negative. Images are labeled to positive glaucoma according to the retinal nerve fibre layer defect, rim loss and optic disc hemorrhage [2]. The main basis for doctors to judge glaucoma are these three feature. Thus, we randomly select 58 of the positive sample and 103 negative sample from the training set to label some pixel-level annotations by trained volunteers. Then, the ophthalmologist makes the final check on each image. Pixel-level annotations consist of three components: optic cup, optic disc and background. The retinal nerve fibre layer defect, rim loss and optic disc hemorrhage near the optic cup and optic disc may be noticed by the network through pixel-level annotations.

3. Training Details

When training the **GCNet** framework on two different datasets, some settings are the same and some are different. On the one hand, we use the same SGD optimizer with momentum set to 0.9 and a weight decay of 0.0005. We initialize the backbone network weights by the ResNet50 weights trained on the ImageNet dataset. We set the initial learning rate 1e-3 and mini-batch size of 4. Then, we decay the learning rate with proportional decline. Data augmentation is adopted to expand the training dataset by pepper noise and horizontal flipping. After each epoch, we save the current best-performing model weights by validating the model on the validation set. On the other hand, our experiment is optimized by a total loss composed of three losses: classification loss L_{cla}, segmentation loss L_{seg} and embedding loss L_{em}.

- On the ACA dataset. We use standard cross-entropy loss as L_{cla}. Besides, according to the original size of two images, the ACA dataset are resized into 128×256 resolution to train our model. We use random pepper noise and horizontal flipping as data augmentation. We set $\alpha = \beta = \gamma = 1.0$, $\lambda = \rho = 1.0$, and $\omega = 0.01$ using the ACA validation set.

- On the REFUGE dataset. We use binary cross-entropy loss as L_{cla}. The REFUGE dataset are resized into 256×256 with the full use of the computer memory. We use random pepper noise, vertical flipping and horizontal flipping as data augmentation. We set $\alpha = \beta = \gamma = 1.0$, $\lambda = \rho = 1.0$, and $\omega = 0.01$ using the REFUGE validation set.

- On the SIGF dataset. Because of the extreme imbalanced class distributions of SIGF between positive and negative sample, we use focal loss as L_{cla}. The SIGF dataset are resized into 224×224. We use random pepper noise, vertical flipping and horizontal flipping as data augmentation. We set $\alpha = 1.0$, $\beta = \gamma = 0.1$,$\lambda = \rho = 1.0$ and set $\omega = 0.01$ using the SIGF validation set.

4. Baselines

All tested baselines use the following settings unless otherwise stated.

In the training process, we use the SGD optimizer with learning rate of 1e-3 with proportional decay. Then, we use the same MLP (as the main classifier) as Equation 5 to complete the evaluation of five levels on the ACA dataset. Note that dropout layers are used in MLP to alleviate overfitting.

In this paper, we compare GCNet with other state-of-the art methods on two datasets: ACA dataset and REFUGE dataset. Six baselines used in the experiments can be divided into two categories: traditional methods and weakly-supervised based methods, which are described as follows.

4.1. Traditional methods

In this paper, we use VGG, GoogLeNet, and ResNet-50 as traditional deep learning methods.

- VGG [5]. We initialize the VGG weights trained on the ImageNet dataset and update the model weights on the output layer only.

- GoogLeNet [7]. In our experiment, we use inception v3 without auxiliary classifiers. We freeze the first 27 layers and train the model on the rest layers and the main classifier.

- ResNet-50 [1]. We initialize the VGG weights trained on the ImageNet dataset. Then we freeze the first convolutional layer and layers 1 and 2 and train on the rest layers.

4.2. Weakly-supervised based methods

Consistency regularization and pseudo-labeling are two common strategies in weakly-supervised based methods.

- FixMatch [6] uses both consistency regularization and pseudo-labeling to optimize its framework. In our experiments, weak augmentation is a standard random pepper noise and horizontal flipping. For strong augmentation, we experiment with "RandAugment" as mentioned in the original paper. We use cross-entropy loss for pseudo pixel-level labels and dice loss for real pixel-level labels. We set $\tau = 0.7$ which is a scalar hyperparameter denoting the threshold of retaining a pseudo-label.

- CCT [3] uses consistency regularization to obtain similar output distribution between the main decoder predictions and those of the auxiliary decoders. In our experiments, we add a classifier to CCT after encoder for ACA evaluation. We use the cross-entropy loss for pixel-level labeled data and mean squared error loss for pixel-level unlabeled data to measure distance.

- UPS [4] is an uncertainty-aware pseudo-label selection framework which aims to improve the performance of classification. The contributions of UPS include negative pseudo label selection and confidence-based pseudo labels selection. We use UPS for segmentation in our experiments. To generate confidence-based pseudo labels, dropout layers are moved to decoder from encoder. As described in UPS, τ_p and τ_n are the confidence thresholds for positive and negative pseudo labels, κ_p and κ_n are uncertainty thresholds. In our experiments, we set $\tau_p = 0.75, \tau_n = 0.05, \kappa_p = 0.05, \kappa_n = 0.005$. We use Equation 9 to calculate loss between dense prediction and positive pseudo label while for negative pseudo labels, we define:

$$L^{u'}_{\text{dice}} = \frac{1}{N_s} \sum_{j=1}^{N_s} \left(1 - \frac{2 \sum_{i=1}^{N_j} m_i \hat{y}_i^s (1 - \tilde{y}_i^s)}{\sum_{i=1}^{N_j} m_i \hat{y}_i^s + \sum_{i=1}^{N_j} m_i \tilde{y}_i^s} \right). \tag{1}$$

The total loss of this segmentation module is:

$$L_{seg} = L^{u}_{\text{dice}} + L^{u'}_{\text{dice}}. \tag{2}$$

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[2] Liu Li, Xiaofei Wang, Mai Xu, Hanruo Liu, and Ximeng Chen. Deepgf: Glaucoma forecast using the sequential fundus images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 626–635. Springer, 2020. 2

[3] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 3

[4] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2020. 3

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[6] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 2