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*

Abstract

As the basis for developing glaucoma treatment strategies, Anterior Chamber Angle (ACA) evaluation is usually dependent on experts' judgements. However, experienced ophthalmologists needed for these judgements are not widely available. Thus, computer-aided ACA evaluations become a pressing and efficient solution for this issue. In this paper, we propose a novel end-to-end framework ***GCNet*** *for automated Glaucoma Classification based on ACA images or other Glaucoma-related medical images. We first collect and label an ACA image dataset with some pixel-level annotations. Next, we introduce a segmentation module and an embedding module to enhance the performance of classifying ACA images. Within* ***GCNet*** *, we design a Cross-Module Aggregation Net (****CMANet****) which is a weakly-supervised metric learning network to capture contextual information exchanging across these modules. We conduct experiments on the ACA dataset and two public datasets REFUGE and SIGF. Our experimental results demonstrate that* ***GCNet*** *outperforms several state-of-the-art deep models in the tasks of glaucoma medical image classifications. The source code of* ***GCNet*** *can be found at* https://github.com/Jingqi-H/GCNet.

1. Introduction

Glaucoma is a leading cause of irreversible blindness in the world [44]. The main basis for determining clinical treatment protocols when glaucoma is diagnosed is using Anterior Chamber Angle (ACA) images and gonioscopy is widely regarded as the "Gold Standard" in ACA evaluation. There are five ACA levels that correspond to different glaucoma treatments. In order to determine these ACA levels, ophthalmologists examine four local structures in ACA images: Schwalbe Line (SL), Trabecular Meshwork (TM), Scleral Spur (SS), and Ciliary Body Band (CBB), with the help of microgonioscopy [8, 11] as shown in Figure 1. However, ACA evaluation needs expert ophthalmologists, while the availability of experienced ophthalmologists is severely sparse given the large number of glaucoma patients. Therefore, computer-aided systems are urgently needed for efficient ACA evaluation.

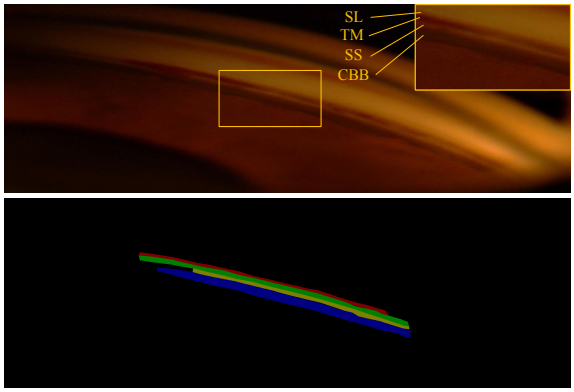


Figure 1. The main structures in ACA consist of Schwalbe line (SL), pigmentation of trabecular meshwork (TM), scleral spur (SS), and ciliary body band (CBB). The bottom figure shows the color annotations of these four structures.

Recently, Deep Neural Networks (DNNs) have become a default choice given its successful applications in Glaucoma-related medical image analysis. For instance, DNNs have been used to segment optic cup (OC) regions and optic disk (OD) regions [9], and to detect glaucomatous optic neuropathy [25]. Li *et al.* [21] apply a ResNet-18 model for automatic measurement of trabecular-iris angle (TIA). Peroni *et al.* [36] exploit a dense U-Net architecture to segment irido-corneal interface images. However, directly applying traditional DNN models for ACA level classification is suboptimal and the following challenges need to be overcome:

(1) In ACA images, the four spatial structures (SL, TM, SS, and CBB) are the main basis for ophthalmologists in their ACA level evaluations. However, these four structures concentrate in a small area of a full ACA picture as shown in Figure 1. This leads to great challenges for machine learn-

*Now at Alibaba Inc.

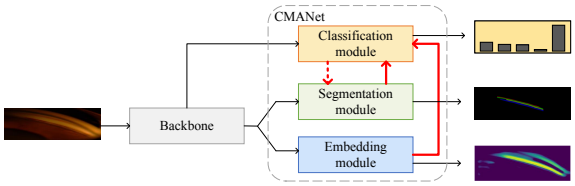


Figure 2. The illustration of **GCNet**. We introduce a weakly-supervised metric learning network for capturing pixel-level details of spatial features and a cross-module communication strategy (two type of pipes in red) for augmenting information across three sub-modules.

ing models to distinguish this important information from background noise.

(2) The appearance of SL, TM, SS, and CBB corresponds to different levels of glaucoma diseases. In ACA images, these four structures are adjacent with blurred boundaries between each other. Recognizing each feature and learning their hidden representations are crucial in ACA classification. However, pixel-level annotations of these structures are rare for supervised learning.

(3) With sparse pixel-level labels of the four structures, deep learning models often suffer from capturing inter-class similarities and intra-class variations of these structures. Fusing the knowledge of these four structures in the task of image classification is another obstacle.

To overcome the above challenges, we propose a deep neural network with weakly-supervised metric learning and interactive module communications on glaucoma image classification. To the best of our knowledge, this is the first work for classifying ACA images using deep learning techniques with three sub-modules. The main contributions of our paper include:

• We propose a weakly-supervised metric learning framework for glaucoma image classification. We augment glaucoma ACA images with pixel-level annotations for four structures (SL, TM, SS, and CBB) and utilize these annotations to enhance image classification.

• We propose a cross-module communication strategy to fuse features from multiple granularity levels as shown in Figure 2. The proposed network, **CMANet**, is able to capture semantic information in picture levels as well as pixel levels. This method can also be extended to other types of Glaucoma clinical images such as the REFUGE dataset [32] including both image-level and pixel-level annotations.

• We collect and label an ACA evaluation image dataset. The dataset includes 999 ACA images which are labeled by several senior ophthalmologists. In addition, 100 ACA images are labeled with SL, TM, SS, and CBB at pixel levels.

We conduct extensive experiments on three real-world datasets to demonstrate the effectiveness of the proposed framework in classification compared with several deep learning baselines.

2. Related Work

2.1. Glaucoma Detection

Early detection and intervention are essential to prevent the deterioration of glaucoma. With the overwhelming application of DNNs in medical image analysis in recent years, more and more DNN models have been designed for glaucoma detection [20, 48]. In these models, most are based on anterior segment optical coherence tomography (AS-OCT) [10,12,15,16] or fundus photographs [25,28,50].

Some recent work focuses on ACA structures. For instance, DNN models are designed to segment geometrical structures such as AS-OCT [9, 33, 35]. A fully automatic segmentation method is proposed to segment corneal boundary, iris region, and trabecular-iris contact [13]. Moreover, a model based on transfer learning and multi-level convolutional neural networks is designed to detect the angle-closure glaucoma [10].

2.2. Weakly-supervised Learning

Medical image labeling is expensive because of its need for senior doctors [48]. Thus, weakly-supervised learning [40] with little supervision has been actively studied in medical image analysis recently. As a branch of weakly-supervised learning, semi-supervised learning (SSL) has made remarkable achievements in representation learning in recent years [19, 27].

Consistency regularization and pseudo-labeling are two common strategies in SSL. The goal of consistency regularization is to obtain similar output distribution, which can be achieved by adding various degrees of augmentation [29], by embedding different perturbations [34], or by embedding different networks [18, 26]. The goal of pseudo-labeling is to assign reasonable labels to unlabeled samples for the purpose of training. Knowledge distillation is a technology to train a student model with pseudo-labeled data and labeled data [5, 51–53], of which, the pseudo-labels are predicted with a pre-trained teacher model.

Considering the erroneous high confidence predictions from poorly calibrated models, UPS framework [38] is proposed to alleviate noisy training with negative pseudo-labels. Based on MixMatch [3], UDA [49], and ReMixMatch [2], FixMatch [42] produces artificial labels using both consistency regularization and pseudo-labeling.

2.3. Aggregation

Many classification methods based on fully convolutional networks achieve remarkable classification performance in recent years. However, convolution as a local operation establishes pixel relationships in a local neighborhood. Long-range dependency modeling is necessary [22]. In order to model long-range information dependency [31], two categories of context aggregation have been used: (1)

Table 1. ACA classification system described by Shaffer [1]. Five levels from wide to narrow angle are based on visible structures.

Level	Clinical Feature	Percentage (%)
Narrow I	part of CBB visible	22.12
Narrow II	SS visible	20.42
Narrow III	posterior TM visible	12.91
Narrow IV	only SL visible or none	25.43
Wide	all structures visible	19.12

pairwise based [14, 46]: non-local modules utilize pairwise similarity to learn the global context for each location; (2) context fusion based: considering large computational capacities, a channel attention mechanism [4] is proposed to distinguish important features from minor features. A spatial attention mechanism [45] is proposed to find where to focus on features. A Convolutional Block Attention Module [47] is proposed to combines the advantages of channel attention and spatial attention by cascade connections. These methods enhance information representations, but such enhancement is self-correlation and is limited to one module only.

Panoramic segmentation requires both semantic segmentation and instance segmentation to perform well. Thus, a bidirectional aggregation network [6] is proposed to enable feature-level interaction between instance segmentation and semantic segmentation. However, different from panoramic segmentation, our framework has three different modules (classification, segmentation, embedding). Semantic and discriminative information cannot be fed back to the classification module from a unidirectional learning pipeline, since the classification module is before segmentation/embedding modules.

Semantic segmentation concentrates on capturing spatial details of local structures and the embedding module makes the information of the four structures more recognizable. Such information can be used in classification to understand both entire object features and local contexts. Based on these observations, the following two important problems need to be solved when applying DNN-based methods to ACA evaluation. (1) How do we find and label useful information for learning on ACA images based on domain knowledge and (2) How do we design a strategy to synthesize useful semantic information into classification?

3. Method

Our **GCNet** contains two major components: a backbone network and a cross-module aggregation network, as shown in Figure 2. The backbone network is composed of a deep residual network (ResNet) and a feature pyramid network (FPN). The former is used for feature extraction and the latter is used for resolution recovery. The cross-module aggregation network includes three sub-modules: a classification module, a segmentation module, and an embedding module. The communication among these three sub-modules is described in the section of Framework Overview.

3.1. Problem Formulation

To overcome the challenges in computer-aided ACA classification, we propose an end-to-end DNN framework named **GCNet** in this paper. The goal of **GCNet** is to map an ACA image into five levels as described in Table 1. Besides the image-level classification labels of the whole dataset, we also provide pixel-level labels (SL, TM, SS, and CBB) for some ACA images.

• Image-level labels: each image is associated with a label indicating the level of ACA: Narrow I, Narrow II, Narrow III, NarrowIV, or Wide.

• Pixel-level labels: for some images in the dataset, each image has pixel-level labels. Each pixel label denotes if this pixel belongs to SL, TM, SS, CBB, or background.

We design two auxiliary tasks with these weakly labeled samples: (1) dense prediction; (2) pixel embedding. The dense prediction is developed to distinguish the four structure areas SL, TM, SS, and CBB, while the pixel embedding is developed to map each pixel into a semantic vector so that pixels from the same structure are close to each other in terms of semantic distance. **GCNet** is able to learn effective spatial details and semantic embeddings for the four structures, while improving ACA classification performance.

3.2. Dataset

To establish an automatic ACA classification system, 1038 ACA images from 2015 to 2017 are collected from the hospital we are collaborating with. For each patient, two to five digital images are taken for each eye. Normally gonioscopy is only used when glaucoma has been diagnosed. So the size of our ACA dataset is limited. In our dataset, each image is labeled by a senior ophthalmologist with over 10 years of working experience.

We apply an examination on these images, and 39 images are removed due to quality issues. The remaining 999 images are labeled as one of the following: Narrow I (N1), Narrow II (N2), Narrow III (N3), Narrow IV (N4) and Wide (W) as shown in Table 1. The distribution of the labeled data is depicted in the last column of the table.

Furthermore, we can observe that the four types of structures named SL, TM, SS, and CBB are distributed near the boundary between cornea and iris as shown in Figure 1. Given that the background is relatively large compared to these structures and the majority of the useful area is exceedingly narrow, it will be a big challenge for ophthalmologist to manually annotate such small and concentrated structures in ACA images. We use a labeling tool, *Labelme*, to perform pixel-level masking with manual scribbles on images. Among the 999 labeled images, 100 of them are an-

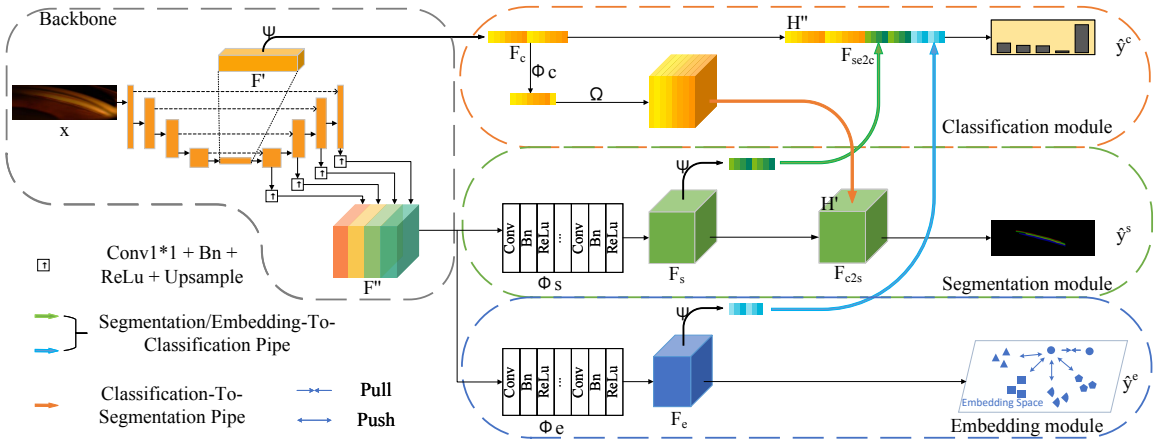


Figure 3. Our proposed architecture is composed of a shared backbone and cross-module aggregation net (**CMANet**). **CMANet** has three sub-modules: (i) a classification module; (ii) a segmentation module for dense prediction; (iii) an embedding module for pixel embedding. Meanwhile, **CMANet** has two key pipes for information exchanging among multi-granularity feature representations: Classification-To-Segmentation (c2s) and Segmentation/Embedding-To-Classification (se2c).

notated at pixel levels for these four structures. Our work is conducted according to the Declaration of Helsinki. Given the fully anonymity of ACA images, we are exempted by the medical ethics committee to inform the patients.

3.3. Framework overview

As shown in Figure 3, given an ACA image $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, **GCNet** is designed to output an image-level prediction $\hat{\mathbf{y}}^c \in \mathbb{R}^{N_c}$, a dense prediction $\hat{\mathbf{y}}^s \in \mathbb{R}^{N_s\times H\times W}$, and a pixel feature map $\hat{\mathbf{y}}^e \in \mathbb{R}^{q\times H\times W}$. In the training phase, we have image-level ground-truth label $\mathbf{y}^c \in \{0, 1, 2, 3, 4\}$ for each image where 0, 1, 2, 3, 4 denote N1, N2, N3, N4, W respectively and pixel-level labels $\mathbf{y}^s \in \mathbb{R}^{N_s\times H\times W}$ for the annotated images. N_c denotes the class number in classification, N_s denotes the class number in segmentation, and q denotes the length of embedding vector. In this work, $N_c = 5$ because we have 5 class levels (N1, N2, N3, N4, W) and $N_s = 5$ because we have four structures (SL, CBB, SS, TM) and background.

3.3.1 Backbone

The backbone of **GCNet** is composed of three parts: an encoder, a decoder, and a feature pyramid network (FPN) [24], as show in Figure 3. The encoder is a typical pre-trained convolutional neural network by freezing first two layers, and the decoder is an upsampling subnet like that of U-Net [39]. Given an input image x, the backbone outputs two latent feature maps F' and F''. While F' is the output of the encoder, and F'' is the output of the FPN.

3.3.2 Cross-Module Aggregation Network

In **GCNet**, the role of the *classification module* is to classify an ACA image into 5 classes as described in Table 1. We apply a global average pooling (GAP) [23] to map the latent feature map F' to a feature vector by $F_c = \mathrm{AvgPool}(\mathrm{F}')$.

In clinical practices, the basis of ACA classification for ophthalmologist is the previous introduced four structures. Thus, the spatial distribution of these features is critical in classifying ACA levels. Given this motivation, we design a segmentation module to induce the classification module to pay attention to these four structures. The segmentation module assigns each pixel with a class label. Given that multi-scale features have been fused in FPN by concatenation, rich spatial details and semantic information can be captured in the *segmentation module*:

$$F_s = \phi_s(\mathrm{F}'') = g\left(W_s^{(1)}\left(g(W_s^{(0)}\mathrm{F}'' + b_s^{(0)})\right) + b_s^{(1)}\right), \tag{1}$$

where g denotes the ReLu activation function, as it will for the rest of the paper. F'' is the output of the FPN model.

However, when incorporating the four structures using a segmentation module, we face a new challenge due to inter-class similarities and intra-class variations in the four structures:

• Inter-class similarity: the neighboring structures of SL, TM, SS, and CBB looks similar in shape, color, and texture. For example, the adjacent structures TM and SS share a similar fuscous color.

• Intra-class variation: same structures in different ACA images may have different appearances for different genes. For example, the color depth of TM is related to congenital pigment deposition. Some people are born with deeper color of TM due to more pigment deposition while others are born with lighter color of TM due to less pigment.

Thus, it is difficult for computers to distinguish the four structures in dense predictions. However, the shape and appearance of these four structures are the basis for the ophthalmologist to evaluate the ACA levels. A weak dense pre-

diction will result in poor classification performance. To overcome this challenge, we add an *embedding module* to the framework. We use the module to map each pixel into an embedding vector. Here five convolution layers are used to map F'' into embedding feature map: $\hat{\mathbf{y}}^e = \phi_e(\mathrm{F}'')$.

With three sub-modules, our weakly-supervised metric learning framework is formed. To resolve communication issues among ACA classification, dense prediction, and pixel embedding, we propose a cross-module communication strategy **CMANet**. Specifically, **CMANet** provides two communication mechanisms among the three modules: Classification-To-Segmentation (c2s) and Segmentation/Embedding-To-Classification (se2c).

(1) c2s. The segmentation model cannot be trained efficiently due to the small number of pixel-level annotated samples. Moreover, the contextual information of F_s may be lost after applying upsampling and convolutions for many times. We propose to supplement the contextual information of segmentation module by building a communication pipe from the ACA classification module to the segmentation module. We aggregate F_s and F_c as follows:

$$\mathrm{F}_{\text{c2s}} = \mathrm{F}_s \odot \Omega(\phi_c(\mathrm{F}_c)), \tag{2}$$

where ϕ_c is a 1×1 convolution layer and Ω is a repeat operation to align the feature space. $\odot$ denotes the aggregation operation which is element-wise multiplication. As a result, feature F_{c2s} aggregates the semantic information of F_s and F_c. Then we apply 3×3 and 1×1 convolution layers to obtain the dense prediction $\hat{\mathbf{y}}^s$:

$$\hat{\mathbf{y}}^s = W_s^{(3)}\left(g\big(W_s^{(2)}\mathrm{F}_{c2s} + b_s^{(2)}\big)\right) + b_s^{(3)}. \tag{3}$$

(2) se2c. The ACA classification model is inadequate in capturing hidden features of the four structures. The segmentation model can capture the spatial details for four structures while embedding module can learn the inter-class similarities and intra-class variations among SL, TM, SS, and CBB. Therefore, we propose to supplement the ACA classification module with hidden semantic information of these four structures by building two communicate pipes. As shown in Figure 3, one is from the segmentation module to the ACA classification module, and the other is from the embedding module to the ACA classification module. Global average pooling and concatenation are used to fuse F_s and F_e into F_c:

$$\mathrm{F}_{\text{se2c}} = \mathrm{F}_c \,||\, \text{AvgPool}\,(\mathrm{F}_s) \,||\, \text{AvgPool}\,(\phi_e\,(\mathrm{F}'')), \tag{4}$$

where $||$ denotes concatenation. Then we apply a multi-layer perceptron (MLP) to map aggregated latent features $\mathbf{F}_{\text{se2c}}$ to class distributions:

$$\hat{\mathbf{y}}^c = W^{(2)}\left(g\left(W^{(1)}\big(g(W^{(0)}\mathbf{F}_{\text{se2c}} + b^{(0)})\big) + b^{(1)}\right)\right) + b^{(2)}. \tag{5}$$

3.4. Loss Function

The **GCNet** framework is designed for multi-task image classification with image-level annotations and partial pixel-level annotations. Some images have pixel-level labels while some do not. If we have K ACA images as training samples, we have K image-level annotations and ηK ($0<\eta<0.5$) pixel-level annotations. The classification module is trained with fully annotated images, while the segmentation module and the embedding module are optimized with weakly labeled samples. Note that pseudo labels are generated based on the predicted probabilities of pixels ($\hat{\mathrm{y}}_i^s$) and are only used in the segmentation module. In **GCNet**, the loss is composed of three parts: the classification module uses Cross-Entropy loss, denoted as L_{cla}; the segmentation module uses Dice Loss, denoted as L_{seg}; and the embedding module uses discriminative loss, denoted as L_{em}. The total loss function L_{total} is defined as:

$$L_{\text{total}} = \alpha \cdot L_{\text{cla}} + \beta \cdot L_{\text{seg}} + \gamma \cdot L_{\text{em}}, \tag{6}$$

where α, β, γ are the weights of three loss items.

The segmentation task takes advantage of the spatial annotations of the four structures to assist the ACA classification. However, it is hard and expensive to obtain pixel-level labeled data in the medical image analysis domain because only experts can provide reliable annotations. We focus on weakly-supervised segmentation approaches because it is relatively easy to acquire a large amount of image-level labels. In addition, we adapt a discriminative loss based on distance metric learning [7, 30] to learn inter-class similarities and intra-class variations among the four structures.

3.4.1 Weakly-supervised Segmentation Loss

We train the segmentation model using a small number of samples with pixel-level annotations for SL, TM, SS, and CBB. In the segmentation module, each pixel is classified into five categories: SL, TM, SS, CBB, and background. In our framework, the segmentation loss is defined as:

$$L_{\text{seg}} = L_{\text{dice}}^{l} + L_{\text{dice}}^{u}, \tag{7}$$

where L_{dice}^{l} is a loss used in pixel-labeled data and L_{dice}^{u} is used in pixel-unlabeled data. Specifically, we denote

$$L_{\text{dice}}^{l} = \frac{1}{N_s}\sum_{j=1}^{N_s}\left(1 - \frac{2\sum_{i=1}^{N_j}\hat{\mathrm{y}}_i^s \mathrm{y}_i^s}{\sum_{i=1}^{N_j}\hat{\mathrm{y}}_i^s + \sum_{i=1}^{N_j}\mathrm{y}_i^s}\right), \tag{8}$$

where $\hat{\mathrm{y}}_i^s$ and y_i^s are the predicted probability and ground truth label for pixel i, respectively; N_j denotes the number of pixels in structure j.

Let $m_i = \mathbb{I}[\hat{\mathrm{y}}_i^s \geq \tau]$ be the selected pseudo-label for pixel i, where $\mathbb{I}$ is the indicator function and τ is a hyper-parameter which denotes a confidence threshold. For instance, if the probability score is sufficiently high ($\hat{\mathrm{y}}_i^s \geq \tau$) then the corresponding pseudo label is selected. We choose $\tau = 0.7$ empirically. For the unlabeled pixels, we define

$$L_{\text{dice}}^{u} = \frac{1}{N_s}\sum_{j=1}^{N_s}\left(1 - \frac{2\sum_{i=1}^{N_j} m_i \hat{\mathrm{y}}_i^s \tilde{\mathrm{y}}_i^s}{\sum_{i=1}^{N_j} m_i \hat{\mathrm{y}}_i^s + \sum_{i=1}^{N_j} m_i \tilde{\mathrm{y}}_i^s}\right), \tag{9}$$

where $\tilde{\mathrm{y}}_i^s$ indicates the pseudo label for pixel i.

3.4.2 Discriminative Loss

Inspired by the success of pixel embeddings in instance segmentation [7, 30], we use a discriminative loss function to guide the model to learn structural details in feature space. Discriminative losses enforce the model to map each pixel in an image to a q-dimensional vector such that embedding vectors of pixels with the same label (same structure) should be close to each other while embedding vectors of pixels with different labels should be far apart.

By doing so, different structures of an image and the same structure in different ACA images should be well recognized through this loss. The discriminative loss is described as a weighted sum of three parts:

$$L_{\text{em}} = \lambda \cdot L_{\text{var}} + \rho \cdot L_{\text{dist}} + \omega \cdot L_{\text{reg}}, \tag{10}$$

$$L_{\text{var}} = \frac{1}{N_s} \sum_{j=1}^{N_s} \frac{1}{N_j} \sum_{i=1}^{N_j} \left[\left\| \boldsymbol{\mu}_j - \hat{\mathbf{y}}_i^e \right\| - \boldsymbol{\delta}_v \right]_+^2, \tag{11}$$

$$L_{\text{dist}} = \frac{1}{N_s(N_s-1)} \sum_{j_A=1}^{N_s} \sum_{j_B=1}^{N_s} \left[2\boldsymbol{\delta}_{\text{d}} - \left\| \boldsymbol{\mu}_{j_A} - \boldsymbol{\mu}_{j_B} \right\| \right]_+^2, \tag{12}$$

$$L_{\text{reg}} = \frac{1}{N_s} \sum_{j=1}^{N_s} \left\| \boldsymbol{\mu}_j \right\|, \tag{13}$$

where $\hat{\mathbf{y}}_i^e$ is the embedding vector of a pixel i; $\boldsymbol{\mu}$s are class mean vectors and $\boldsymbol{\delta}_v, \boldsymbol{\delta}_d$ are the margins for the variance and distance loss. We set $\boldsymbol{\delta}_v = 0.0001$ and $\boldsymbol{\delta}_d = 1.0$ empirically. Equation 11 represents the variance term which applies a pull force on each embedding vector towards the mean vector of a structure. Equation 12 denotes a distance term that pushes the cluster centers away from each other.

4. Experimental Evaluation

4.1. Dataset Preprocessing

To eliminate the influence of background noise and to extract the features of four structures, critical region detection is applied to automatically crop the target region of SL, TM, SS and CBB. Here, we use the YoLo detector [37] to preprocess our original images. These segmented ACA images are then resized to $[700 \times 2, 100]$ for better quality.

In addition, we also investigate the REFUGE dataset [32] which contains 1,200 images with two classes (10% positive and 90% negative) and pixel-level annotations (background, optic disc, and optic cup). SIGF dataset [20] is used for a glaucoma forecasting task. This dataset contains sequential fundus images of a patient, which is different from glaucoma evaluation. Therefore, we convert SIGF to a standard classification task by ignoring sequential information. All fundus images are annotated with binary labels of glaucoma, i.e., positive (4%) or negative (96%), corresponding to 3,671 images in total.

Table 2. Inference time for an input tensor of size $2*3*128*256$.

Methods	VGG	GoogLeNet	ResNet	FixMach	CCT	UPS	GCNet
Time(ms)	39.27	56.19	47.51	87.18	50.79	86.75	56.12
Flops(G)	20.07	22.33	24.05	770.75	46.20	770.75	119.44
Params(M)	18.93	3.35	5.37	87.76	46.86	87.76	64.85

4.2. Training and Evaluation

We conduct experiments on our ACA dataset and the public REFUGE dataset [32]. Table 2 shows the inference time of all baseline models and **GCNet**. For the ACA dataset, we partition the dataset into a training set (80%) and a test set (20%) based on a random seed of 72. We shuffle the training set and select a subset as our validation set from training samples. The validation set accounts for one-fifth of the training set. For the REFUGE dataset, We use the standard train/val/test split with 800 images for training/validating and 400 images for testing. Note that we use all image-level annotations and 83 images with pixel-level annotations during training to carry out weakly-supervised metric learning framework.

We use standard metrics of Accuracy (ACC), Area Under Curve (AUC), and F1 to evaluate the classification accuracy for ACA and REFUGE. Given the extreme imbalanced class distributions of SIGF, we only use AUC to demonstrate the evaluation results. All methods are implemented in Ubuntu 18.04 with NVIDIA GeForce RTX 3080 graphics cards with 10 GB memory. Both pepper-noise and horizontal flipping are used as data augmentation in the ACA training images. On the basis of this augmentation, we also add vertical flipping when training on REFUGE and SIGF. In our experiments, we enumerate hyperparameter values with a tolerance of 0.1 in the interval $[0, 1]$. Then we set $\alpha = \beta = \gamma = 1.0$, $\lambda = \rho = 1.0$, and $\omega = 0.01$ empirically on the ACA dataset.

4.3. Results

4.3.1 Classification results

To evaluate the performance of **GCNet**, we first compare it with other state-of-the-art methods on two datasets: ACA dataset and REFUGE dataset. We improve the original UPS with additional segmentation modules to get the results in Table 3, which makes our comparison fairer. UPS is designed for classifications initially. With an additional segmentation module, UPS performs better.

As shown in Table 3, **GCNet** outperforms all baselines on all metrics for both datasets. More details of baseline models can be found in the Appendix. From the results, we have the following interesting observations. First, our method achieves the best Accuracy (ACC) (79.19%), AUC (94.32%), and F1 (78.53%) scores compared with state-of-the-art approaches on ACA classification. Second, weakly-supervised models (e.g., FixMatch [42], CCT [34],

Table 3. Classification results of Accuracy (ACC), Area Under Curve (AUC), and F1 on two test sets.

Methods	ACC(%)	AUC(%)	F1(%)
	ACA		
VGG	68.32±1.18	88.85±0.96	66.64±1.79
GoogLeNet	70.66±1.41	91.62±0.73	70.04±1.20
ResNet	74.82±0.52	91.64±0.41	73.03±1.07
FixMatch	73.81±1.89	92.96±0.32	73.02±1.66
CCT	72.69±1.08	92.37±0.23	70.92±1.41
UPS	76.75±1.55	93.74±0.68	75.70±1.56
GCNet	**79.19±0.72**	**94.32±0.42**	**78.53±0.29**
	REFUGE		
VGG	92.15±0.78	92.38±0.84	95.62±0.46
GoogLeNet	93.00±1.99	93.19±1.07	96.13±1.04
ResNet	95.25±0.57	94.00±1.92	97.37±0.32
FixMatch	93.65±1.75	93.85±0.82	96.43±1.03
CCT	90.00±0.42	92.35±0.60	94.51±0.31
UPS	93.20±1.79	94.60±0.72	96.22±1.07
GCNet	**96.30±0.43**	**97.20±0.326**	**97.97±0.23**

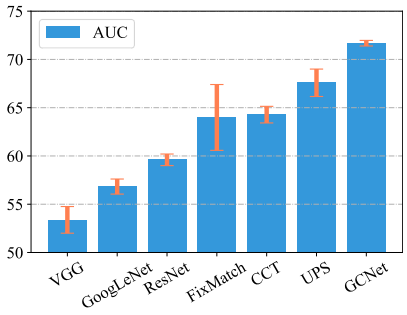


Figure 4. Classification results (AUC) on the SIGF test set.

and UPS [38]) are mostly superior to traditional methods (including VGG [41], GoogLeNet [43], and ResNet [17]) on AUC. For instance, the weakly-supervised model UPS achieves 1.93% higher ACC than the traditional method ResNet (without considering pixle-level information).

We also test the proposed method on the SIGF dataset. We add pixel-level annotations to 36% of the training set, leading the model to pay more attention to the information near the optic and disc. As shown in Figure 4, the results show that weakly-supervised models deal effectively with the challenging task of glaucoma evaluation under the guidance of pixel-level annotations. Due to the extreme imbalance of classes in SIGF, we only report AUC scores. We observe that **GCNet** obtains the highest AUC score on glaucoma classification.

To further illustrate the predictive power of the proposed model, we analyze the confusion matrices of state-of-the-art methods and our method, as shown in Figure 5. From confusion matrices, we observe that it is difficult for a model to distinguish adjacent ACA levels. For example, N1 is easily mis-classified as N2 because the CBB structure is difficult to recognize. The difference between N1 and N2 is that N1 has the CBB structure while N2 does not, as shown in Table 1. The same pattern appears again in N2/N3 and N3/N4. Thus, we conclude that two closely related categories are also the easiest to confuse with each other.

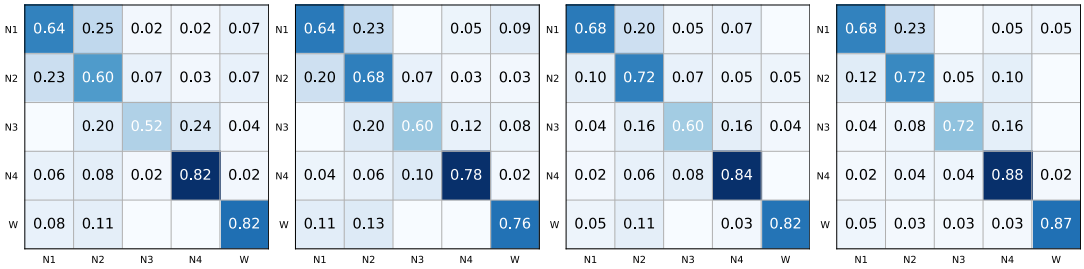
We plot the ROC curves of all methods as shown in Figure 6. Our proposed **GCNet** achieves good ROC performance and the best AUC value when compared to the competing methods. These results further suggest the efficiency of **GCNet** in ACA evaluation.

4.3.2 Ablation studies

To evaluate the contributions of weakly-supervised metric learning and cross-module communications, we further compare **GCNet** with its four types of variants, including **GCNet-C**, **GCNet-CE**, **GCNet-CEP**, **GCNet-CEPS** on both datasets. Each variant corresponds to the proposed model removing one or more proposed modules. We remove modules one by one from top to bottom in Table 4 based on the dependence between modules. Module "P" (Pseudo Labeling) would not be possible without "S" (Segmentation). Therefore, we remove more than one element of our method at once, rather than comparing the effectiveness of each permutation of element removals. The experimental results are shown in Table 4.

First, we analyze the prediction performance of weakly-supervised metric learning. From Table 4, **GCNet-C** achieves the second-highest ACC/AUC without cross-module communications in the ACA dataset and the REFUGE dataset respectively, indicating that weakly-supervised metric learning has improved on the proposed network. Note that we used ACC as the main evaluation criteria for ACA dataset because of the relatively balanced samples across 5 levels and AUC for the REFUGE dataset because of the imbalance of positive and negative cases. Then, compared to **GCNet-CEPS**, the ACC of **GCNet-CE** increases from 74.82% to 76.39%. Besides, the ACC of **GCNet-C** increases from 76.39% to 76.95% due to metric learning. The performance of the proposed model on the REFUGE dataset has similar improvements on AUC. The results show that **GCNet** finds useful information based on domain knowledge utilizing weakly-supervised metric learning in ACA images, e.g., structural details and discriminative information.

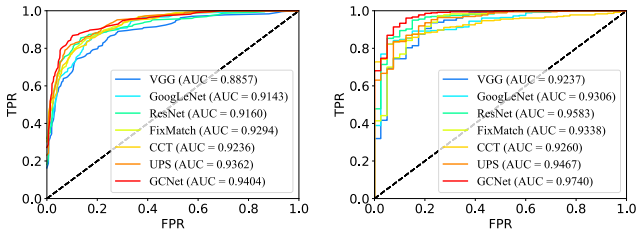
Next, **GCNet** achieves better prediction scores compared with **GCNet-C** which uses weakly-supervised metric learning only. It suggests that **CMANet** with information communication in picture levels as well as pixel levels improves classification results. This confirms that spatial details of structures in dense predictions contribute to image-level classifications.



(a) VGG. (b) GoogLeNet. (c) ResNet. (d) **GCNet-C**.

(e) FixMatch. (f) CCT. (g) UPS. (h) **GCNet**.

Figure 5. Confusion matrices of all methods on five classes. Abscissa is ground-truth and ordinate is prediction.



(a) ACA. (b) REFUGE.

Figure 6. The ROC curves with AUC scores of multiple baselines from the classification results

5. Societal Impact and Limitations

Glaucoma is an irreversible blindness disease and ACA evaluation is an important basis to judge the severity of glaucoma. The proposed **GCNet** is, to our best knowledge, the first to use deep learning for ACA evaluations. Our research can be used as auxiliary means to help the prognosis and treatment of glaucoma. In addition, the proposed research can also improve the efficiency of ophthalmologists.

This study has some potential limitations. For instance, all of our ACA images are collected from one race. The four structures will be different due to individual differences, affecting the evaluation of the ACA levels.

6. Conclusion

In this paper, we propose **GCNet**, an end-to-end DNN framework to overcome the challenges in computer-aided ACA classification. We introduce a weakly-supervised metric learning convolution network to mine spatial and structural details and fuse them into image-level classifications using a cross-module communication strategy. Experiments on the ACA and REFUGE datasets show that **GCNet** outperforms other state-of-the-art DNN baseline models. In the current work, we only focus on using image data to capture the entire context and spatial information of structures. In the future, we plan to investigate communication strategies such as self-supervised frameworks to find complementary information including other modalities (e.g., domain knowledge). We plan to examine structure correlations to learn effective semantic information for the critical structures.

Table 4. Ablation study on the ACA and REFUGE datasets. List of abbreviations: **S**, Segmentation module; **P**, Pseudo label; **E**, Embedding module; **C**, Cross-module communication strategy.

Methods	ACC(%)	AUC(%)	F1(%)
ACA			
GCNet	**79.19±0.72**	**94.32±0.42**	**78.53±0.29**
-C	76.95±0.76	93.52±0.72	76.43±0.77
-CE	76.39±0.26	93.10±0.63	75.70±0.04
-CEP	75.97±0.24	93.27±0.70	74.61±1.10
-CEPS	74.82±0.52	91.64±0.41	73.03±1.07
REFUGE			
GCNet	**96.30±0.43**	**97.20±0.33**	**97.97±0.23**
-C	95.58±0.24	96.70±0.53	97.58±0.14
-CE	95.83±0.31	95.60±0.42	97.71±0.19
-CEP	95.33±0.12	94.78±0.22	97.44±0.08
-CEPS	95.25±0.57	94.00±1.92	97.37±0.32

References

[1] W. L. Alward and R. A. Longmuir. *Color atlas of gonioscopy*, volume 29(2):. American Academy of Ophthalmology, 2008. 3

[2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 2

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32:5049–5059, 2019. 2

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[5] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *European Conference on Computer Vision*, pages 695–714. Springer, 2020. 2

[6] Yifeng Chen, Guangchen Lin, Songyuan Li, Omar Bourahla, Yiming Wu, Fangfang Wang, Junyi Feng, Mingliang Xu, and Xi Li. Banet: Bidirectional aggregation network with occlusion handling for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, 2020. 3

[7] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, July 2017. 5, 6

[8] A. Dellaporta. Historical notes on gonioscopy. *Survey of ophthalmology*, 20(2):137–149, 1975. 1

[9] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomedical optics express*, 8(5):2732–2744, 2017. 1, 2

[10] Marcos Melo Ferreira, Giovanna Pavani Esteve, Geraldo Braz Junior, João Dallyson Sousa de Almeida, Anselmo Cardoso de Paiva, and Rodrigo Veras. Multilevel cnn for angle closure glaucoma detection using as-oct images. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 105–110. IEEE, 2020. 2

[11] P. J. Foster, R. Buhrmann, and G. J Quigley, H. A.and Johnson. The definition and classification of glaucoma in prevalence surveys. *British journal of ophthalmology*, 86(2):238–242, 2002. 1

[12] Huazhu Fu, Yanwu Xu, Stephen Lin, Damon Wing Kee Wong, Mani Baskaran, Meenakshi Mahesh, Tin Aung, and Jiang Liu. Angle-closure detection in anterior segment oct based on multilevel deep network. *IEEE transactions on cybernetics*, 2019. 2

[13] Huazhu Fu, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, Mani Baskaran, Shamira A Perera, and Tin Aung. Automatic anterior chamber angle structure segmentation in as-oct image based on label transfer. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1288–1291. IEEE, 2016. 2

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 3

[15] Huaying Hao, Yitian Zhao, Huazhu Fu, Qiaoling Shang, Fei Li, Xiulan Zhang, and Jiang Liu. Anterior chamber angles classification in anterior segment oct images via multi-scale regions convolutional neural networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 849–852. IEEE, 2019. 2

[16] Jinkui Hao, Huazhu Fu, Yanwu Xu, Yan Hu, Fei Li, Xiulan Zhang, Jiang Liu, and Yitian Zhao. Reconstruction and quantification of 3d iris surface for angle-closure glaucoma detection in anterior segment oct. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 704–714. Springer, 2020. 2

[17] 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. 7

[18] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 429–445. Springer, 2020. 2

[19] Jichang Li, Guanbin Li, Yemin Shi, and Yizhou Yu. Cross-domain adaptive clustering for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2505–2514, 2021. 2

[20] 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, 6

[21] W. Li, Q. Chen, Z. Jiang, G. Deng, Y. Zong, G. Shi, C. Jiang, and X. Sun. Automatic anterior chamber angle measurement for ultrasound biomicroscopy using deep learning. *Journal of Glaucoma*, 29(2)::81–85, 2020. 1

[22] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. Contextual transformer networks for visual recognition. *arXiv e-prints*, pages arXiv–2107, 2021. 2

[23] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 4

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid

networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 4

[25] H. Liu, L. Li, I. M. Wormstone, C. Qiao, C. Zhang, P. Liu, S. Li, H. Wang, D. Mou, R. Pang, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmology*, 137(12):1353–1360, 2019. 1, 2

[26] Wenfeng Luo and Meng Yang. Semi-supervised semantic segmentation via strong-weak dual-branch network. In *European Conference on Computer Vision*, pages 784–800. Springer, 2020. 2

[27] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8801–8809, 2021. 2

[28] Shishir Maheshwari, Ram Bilas Pachori, and U Rajendra Acharya. Automated diagnosis of glaucoma using empirical wavelet transform and correntropy features extracted from fundus images. *IEEE journal of biomedical and health informatics*, 21(3):803–813, 2016. 2

[29] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[30] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 286–291. IEEE, 2018. 5, 6

[31] Dong Nie, Jia Xue, and Xiaofeng Ren. Bidirectional pyramid networks for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[32] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. 2, 6

[33] José Ignacio Orlando, Philipp Seeböck, Hrvoje Bogunović, Sophie Klimscha, Christoph Grechenig, Sebastian Waldstein, Bianca S Gerendas, and Ursula Schmidt-Erfurth. U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. In *2019 IEEE 16th International Symposium on Biomedical Imaging*, pages 1441–1445. IEEE, 2019. 2

[34] 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. 2, 6

[35] Mike Pekala, Neil Joshi, TY Alvin Liu, Neil M Bressler, D Cabrera DeBuc, and Philippe Burlina. Deep learning based retinal oct segmentation. *Computers in Biology and Medicine*, 114:103445, 2019. 2

[36] A. Peroni, C. A. Cutolo, L. A. Pinto, A. Paviotti, M. Campigotto, C. Cobb, J. Gong, S. Patel, A. Tatham, S. Gillan, et al. A deep learning approach for semantic segmentation of gonioscopic images to support glaucoma categorization. In *Annual Conference on Medical Image Understanding and Analysis*, pages 373–386. Springer, 2020. 1

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 6

[38] 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. 2, 7

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[40] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch. A survey on semi-, self-and unsupervised learning for image classification. *arXiv preprint arXiv:2002.08721*, 2, 2020. 2

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[42] 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, 6

[43] 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. 7

[44] Yih-Chung Tham, Xiang Li, Tien Y Wong, Harry A Quigley, Tin Aung, and Ching-Yu Cheng. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*, 121(11):2081–2090, 2014. 1

[45] Haoran Wang, Licheng Jiao, Shuyuan Yang, Lingling Li, and Zexin Wang. Simple and effective: Spatial rescaling for person reidentification. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 3

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 3

[47] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[48] Junde Wu, Shuang Yu, Wenting Chen, Kai Ma, Rao Fu, Hanruo Liu, Xiaoguang Di, and Yefeng Zheng. Leveraging undiagnosed data for glaucoma classification with teacher-student learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 731–740. Springer, 2020. 2

[49] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[50] Rongchang Zhao, Xuanlin Chen, Xiyao Liu, Zailiang Chen, Fan Guo, and Shuo Li. Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning. *IEEE Journal of Biomedical and Health Informatics*, 24(4):1104–1113, 2019. 2

[51] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4090, 2021. 2

[52] Yi Zhu, Zhongyue Zhang, Chongruo Wu, Zhi Zhang, Tong He, Hang Zhang, R Manmatha, Mu Li, and Alexander Smola. Improving semantic segmentation via self-training. *arXiv e-prints*, pages arXiv–2004, 2020. 2

[53] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems*, 33, 2020. 2

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

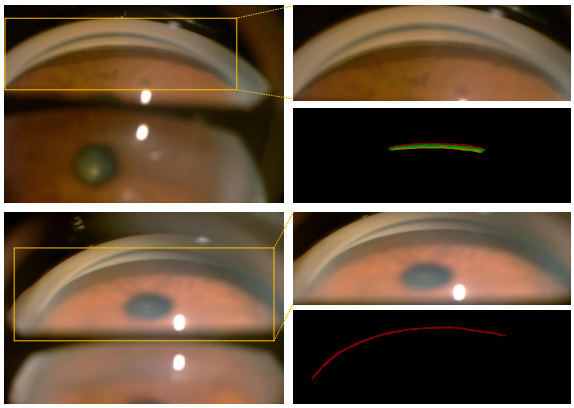


Figure 1. Example of cropped ACA images by YoLo detector. The top row is N2 and the bottom row is N4.

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:

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

The total loss of this segmentation module is:

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

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