EGDCL: An Adaptive Curriculum Learning Framework for Unbiased Glaucoma Diagnosis

Rongchang Zhao[1][0000−0002−5171−4121], Xuanlin Chen[1][0000−0003−1442−3349], Zailiang Chen[1], and Shuo Li[2,*][0000−0002−5184−3230]

[1] School of Computer Science, Central South University, Changsha, China
zhaorc@csu.edu.cn
[2] Western University, London, ON, Canada
slishuo@gmail.com

Abstract. Today's computer-aided diagnosis (CAD) model is still far from the clinical practice of glaucoma detection, mainly due to the training bias originating from 1) the normal-abnormal class imbalance and 2) the rare but significant hard samples in fundus images. However, debiasing in CAD is not trivial because existing methods cannot cure the two types of bias to categorize fundus images. In this paper, we propose a novel curriculum learning paradigm (EGDCL) to train an unbiased glaucoma diagnosis model with the adaptive dual-curriculum. Innovatively, the dual-curriculum is designed with the guidance of evidence maps to build a training criterion, which gradually cures the bias in training data. In particular, the dual-curriculum emphasizes unbiased training contributions of data from easy to hard, normal to abnormal, and the dual-curriculum is optimized jointly with model parameters to obtain the optimal solution. In comparison to baselines, EGDCL significantly improves the convergence speed of the training process and obtains the top performance in the test procedure. Experimental results on challenging glaucoma datasets show that our EGDCL delivers unbiased diagnosis (0.9721 of Sensitivity, 0.9707 of Specificity, 0.993 of AUC, 0.966 of F2-score) and outperform the other methods. It endows our EGDCL a great advantage to handle the unbiased CAD in clinical application.

Keywords: Curriculum Learning · Unbiased diagnosis · Sample Imbalance · Hard sample · Computer-aided Diagnosis

1 Introduction

Ophthalmic disease seriously affects the visual health of people. For example, as the common irreversible blinding ophthalmopathy, glaucoma will attack about 76 million people in the world by 2020 [27]. Computer-aided diagnosis (CAD) plays a significant role in early detection to prevent vision loss of patients with glaucoma [28]. Currently, the success of machine learning model has benefited

*Corresponding Author.

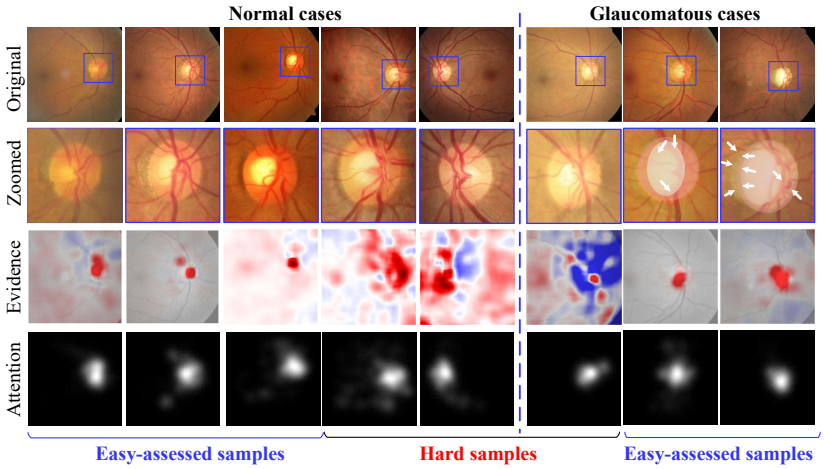

Fig. 1: Training bias is an essential yet challenging problem which seriously impedes the clinical application of CAD algorithms. The challenges originate from extreme normal-abnormal class imbalance and rare hard samples. we observe that, from both sides to the middle, glaucoma identification becomes more and more hard, whereas training samples become rarer.

ophthalmic disease diagnosis with automated algorithm [9, 12], in particular, automatically detecting of glaucoma in fundus images [12, 30, 28, 29, 6, 31]. Through a sequence of advances, those automated diagnosis methods achieves compared accuracy with less time consuming on the challenging benchmarks.

However, training bias seriously impedes clinical applications of existing models due to the introduced false positives. In practice, there are two properties of bias encountered during training a CAD model: 1) the normal-abnormal class imbalance is suffered during collecting the training dataset in the clinic because healthy cases account for the vast majority of populations; 2) A rare of hard samples exists in abnormal cases that are clinically significant for population screening and diagnosis. Therefore, the CAD model is confronted with a great challenge, where the overwhelming majority of training data is composed of normal cases, but the trained models need to robustly recognize the hard abnormal cases, e.g., patients in the early stage of glaucoma (Fig. 1). Obviously, the biased models will misdiagnose those hard abnormal and cannot be absorbed by current healthcare infrastructures because of its limitations on the reliable assessment of hard samples [16, 18] and unacceptable sensitivity.

Training bias can be potentially addressed by curriculum learning with the idea of data reweighting that assigns a weight to each sample and minimizes the weighted loss [22]. Curriculum learning [1] benefits to start with easier samples and gradually takes more complex samples into consideration. Curriculum learning highly organizes the training process by introducing different concepts at different times in curriculum to exploit previously learned concepts to ease the

learning of complex one. This learning paradigm has been empirically demonstrated to be effective in achieving better generalization results for medical image analysis [10, 11, 13].

However, existing curriculum learning methods suffer from two crucial drawbacks when used in unbiased glaucoma diagnosis: 1) the fixed curriculum cannot adaptively represent the training criteria of the developing CAD model to deal with the biased training data, which result in inconsistency between the fixed training criteria and the biased data distribution. There is no guarantee that the fixed curriculum leads to a converged solution for training bias. 2) Curriculum learning often discards the rare hard samples as noise or outliers in the training process, which leads to a serious ineffectiveness and imbalance of training benefits. In gradient optimization, frequent easy samples contribute more loss gradients during training while hard samples are not focused. This results in poor sensitivity and biased models that cannot deal with the hard samples in glaucoma negatives.

In this paper, we propose a novel evidence-guided dual-curriculum learning (EGDCL) to train an unbiased CAD model with the adaptive dual-curriculum. The adaptive dual-curriculum is innovatively developed with the guidance of evidence maps to gradually cure the bias in training data. Therefore, the dual-curriculum can be considered as a novel adaptive training criterion to balance the training benefits of biased dataset from easy to hard, from normal to abnormal. In our EGDCL, evidence maps quantitatively provide the discriminative local features and diagnosis difficulty of each sample as the prior knowledge to identify the bias of training data. The dual-curriculum not only inherits the advantages of curriculum learning that select gradually training samples for effective training, but also adaptively learns the effective weights to balance training benefits by feature reweighting and loss reweighting.

Our EGDCL is a teacher-student framework where the student model provides prior knowledge for dual-curriculum generation by identifying the bias of the decision procedure, while the teacher model learns the CAD model for unbiased glaucoma diagnosis by resampling the data distribution with the newly-designed dual-curriculum. EGDCL is capable of achieving effective unbiased glaucoma diagnosis due to two advantages: **1)** The proposed dual-curriculum adaptively encodes training criteria of sample reweighting as sample weights and feature weights to deal with the training bias. **2)** The proposed teacher-student framework jointly optimizes the dual-curriculum designing and glaucoma classifying in a unified model to obtain the optimal solution of curriculum learning.

Our proposed EGDCL achieves top performance on two most competitive glaucoma diagnosis dataset, i.e., LAG [16] and RIM-ONE [8]. The proposed dual-curriculum learning paradigm can benefit both unbiased classification and effective training in other areas. The main contributions of this work are summarized as follows:

- A novel dual-curriculum learning paradigm (EGDCL) is proposed to tackle the issue of training bias for unbiased glaucoma diagnosis consisting of class imbalance and hard sample mining.

- An effective learning method is proposed to jointly optimize dual-curriculum designing and glaucoma classifying for the optimal solution of training bias, which provides a new learning paradigm for deep embedding learning.
- Our EGDCL achieves top performance on various competitive glaucoma datasets, demonstrating its classification effectiveness and optimal convergence speed on unbiased glaucoma diagnosis.

2 Related Work

Computer-Aided Glaucoma Diagnosis: The success of machine learning has benefited CAD applications [20], especially glaucomatous disease classification [24, 30, 31, 6]. Prior works on glaucoma diagnosis devoted to classifier designing with hand-crafted features like texture, higher-order spectra, wavelet-based features. Those methods consider feature representation and classifier design individually, thus leads to lower classification accuracy. Along with the development of deep learning, [3, 4] reports their work on automated glaucoma detection based on deep learning models. This type of diagnosis methods employs CNNs and GANs in optic disc segmentation [6, 12], medical indices estimation [30, 28, 29] or ONH assessment [18, 16] to promote the performance of glaucoma diagnosis.

Unbiased Classification: Both image classification [22, 23] and object detection [19, 14] face a large training bias. Training bias refers to a disproportionate ratio of observations among the different class, which leads to inefficient training because large redundancy of training samples exist in the biased dataset that have no contributions to model training. There have two types of methods developed to tackle the training bias: data resampling [2], which choosing the suitable proportion of data to train a network, and data reweighting that assigns a weight to each sample and minimizing the weighted loss function [22]. Curriculum learning [1] and self-paced learning [15] represents a learning regime inspired by the learning proceeds of humans that gradually proceeds from easy to more complex or hard to deal with the samples imbalance. Besides, hard negative mining samples hard samples during training [25]. Recently, the focal loss is proposed to address the class imbalance in one-stage objection detection [19]. Unfortunately, to our best knowledge, no work has been reported to tackle the special issue of training bias in disease diagnosis originating from both class imbalance and rare hard sample.

3 Methodology

As shown in Fig. 2, our EGDCL consists of three tightly integrated parts: **1)** A self-attention student network $S(\Theta)$ is proposed with an evidence identification algorithm to learn evidence maps E for the representation of training bias, *e.g.,* diagnosis difficulty and discriminative features. **2)** A curriculum generation module is innovatively designed with the help of evidence maps to learn two adaptive sequences of training criteria ($C1$ and $C2$) for training benefits balancing. **3)** A

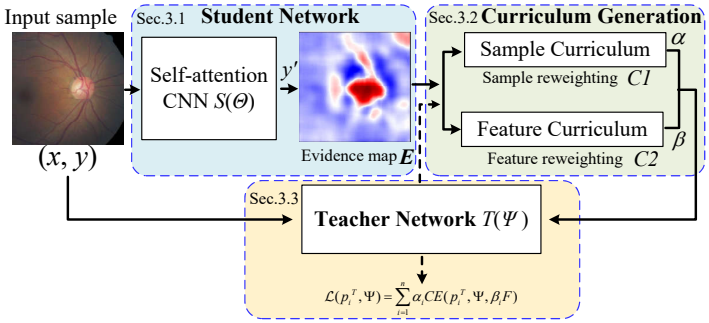

Fig. 2: The proposed evidence-guided dual-curriculum learning (EGDCL) consists of: **Student Network** for evidence identification, **Curriculum Generation** for adaptive training criteria to balance training benefits of biased data, and **Teacher Network** for unbiased glaucoma diagnosis with the regulation of dual-curriculum.

reweighted loss function is constructed for teacher network $T(\Psi)$ according to the dual-curriculum outputs (α and β) to train the unbiased diagnosis model.

3.1 Student Network for Spatial Evidence Identification

Student network $S(\Theta)$ is constructed with two self-attention modules and an evidence identification algorithm to quantitatively identify evidences E of the decision procedure. The student network discovers evidence maps to represent the diagnosis difficulty of samples and highlight the discriminative local features supporting the disease classification, which provides prior knowledge of training bias for the dual-curriculum generation.

Student Network. The student network is a self-attention deep nets with an evidence identification algorithm for the generation of evidence maps. The self-attention structure develops two separate attention pathways, which not only learns the rich contextual features by inferring the feature interdependencies along two separate attention pathways, but also learns to focus on specific structures and contexts of the varying shapes and appearance to capture reliable biomarkers.

Given the input feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, the self-attention modules infer a 3D attention map $m(\mathbf{F}) \in \mathbb{R}^{C \times H \times W}$. The refined feature $\tilde{\mathbf{F}}$ can be computed as

$$\tilde{\mathbf{F}} = \mathbf{F} \otimes (\mathbf{1} + m(\mathbf{F})) = \mathbf{F} \otimes (\mathbf{1} + m_s(\mathbf{F}) \cdot m_c(\mathbf{F})) \tag{1}$$

where $\otimes$ denotes element-wise multiplication. We adopt a residual learning scheme along with the two separate attention pathways to facilitate the gradient flow. To apply the attention modules in classification network, we first compute the spatial attention $m_s(\mathbf{F}) \in \mathbb{R}^H \times W$ and channel attention $m_c(\mathbf{F}) \in \mathbb{R}^C$ at two

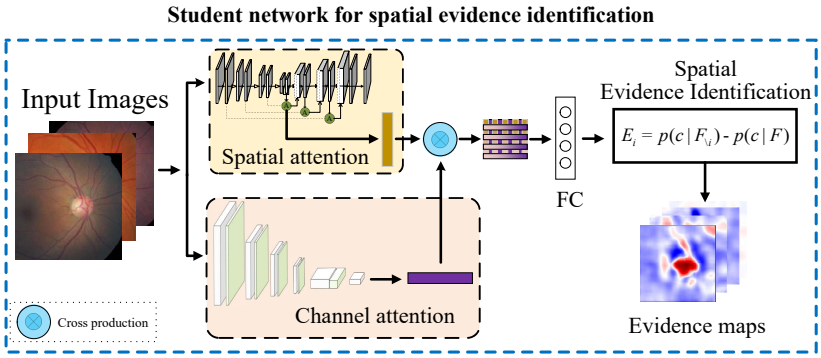

Fig. 3: Student network learns the feature independencies and then provides the quantitative evidence maps with two self-attention modules and an evidence identification algorithm. The evidence maps provide prior knowledge of training bias about diagnosis difficulty and local features for dual-curriculum generation.

separate pathways, then integrate them into attention map $m(\mathbf{F}) \in \mathbb{R}^{C\times H\times W}$ by a bilinear operator

$$m(\mathbf{F}) = \|sqrt(m_s(\mathbf{F}) \otimes m_c(\mathbf{F}))\|_2 \tag{2}$$

where $\otimes$ is the cross production.

Spatial Evidence Identification. Once the student network captures rich contextual features, the prediction difference analysis[32] can be adopted to estimate the spatial evidence maps by producing a relevance matrix $\mathbf{E}$, which reflects the relative importance of all features.

The relevance of a feature $\mathbf{F}_i$ can be estimated by measuring the difference between $p(c|\mathbf{F})$ and $p(c|\mathbf{F}_{\setminus i})$, where $\mathbf{F}_{\setminus i}$ denotes the set of all features except F_i. Here, i indicates the location index of the feature map or evidence map, and $c \in [0, 1]$ represents the class label where 0 indicates normal and 1 is glaucoma. The difference represents how the prediction changes if the feature is unknown.

$$E_i = p(c|\mathbf{F}_{\setminus i}) - p(c|\mathbf{F}) \tag{3}$$

The prediction $p(c|\mathbf{F}_{\setminus i})$ if feature $\mathbf{F}_i$ unknown can be simulated by marginalizing

$$p(c|\mathbf{F}_{\setminus i}) = \sum_{F_i} p(F_i|\mathbf{F}_{\setminus i})p(c|\mathbf{F}) \tag{4}$$

In Eq.(4), the conditional probability $p(\mathbf{F}_i|\mathbf{F}_{\setminus i})$ of feature $\mathbf{F}_i$ is infeasible to be modeled because pixel value is highly dependent on other pixels in medical image. However, there exits an underlying assumption that the conditional of a pixel given its neighborhood does not depend on the position of the pixel in the image, even though a pixel often depends strongly on its small neighborhood. Therefore, the conditional probability $p(\mathbf{F}_i|\mathbf{F}_{\setminus i})$ can be approximated by assuming that

feature $\mathbf{F}_i$ is independent of others $\mathbf{F}_{\setminus i}$ by finding a patch that contains $\mathbf{F}_i$. The prediction can be computed as

$$p(c|\mathbf{F}_{\setminus i}) \approx \sum_{F_i} p(\mathbf{F}_i)p(c|\mathbf{F}) \tag{5}$$

Based on the Eq.(3) and (5), we can estimate the relevance matrix E_i of the same size as the input image. In the matrix, a large value means that the feature contributed substantially to the classification, whereas a small one indicates the feature was not important for the decision. Therefore, we can employ the relevance matrix E_i as evidence maps guidance for the dual-curriculum generation in the succeeding iterative steps described in the next section.

Summarized Advantages: Student network is developed with two self-attention pathways coupled with an evidence identification algorithm to explore prior knowledge of training bias for dual-curriculum generation.

3.2 Curriculum Generation

Innovatively, the dual-curriculum is designed to exploit a novel training criteria to gradually tackle training bias. The dual-curriculum not only to adaptively balance training benefits of biased samples, but also to emphasize the training contribution of rare hard samples (Fig. 5), with the help of two types of weights α and β. The weights are updated along with the training procedure of the diagnosis model according to what knowledge the model has already learned in each iteration as described in Sec. 3.1.

Sample Curriculum ($C1$). The sample curriculum (Fig. 4(a)) is designed to dynamically encode a set of weights on the loss function to balance the training contributions. Initially, the weights favor easily diagnosed samples, and then gradually involve an adaptive change of weights to increase the training focus of rare hard samples. In EGDCL, we propose to reshape the loss function with a weighting factor α not only to adjust the training benefits of each sample from easy to hard, but also to focus training on rare hard negatives.

Formally, the weighting factor α of samples is defined as

$$\alpha_i = \gamma(\frac{1}{1-p_i^E})bool_i + (1-\gamma)(\frac{1}{p_i^T}) \tag{6}$$

where γ is a hyperparameter, and p_i^T and p_i^E denote the model's estimated probability for the class with label $y = 1$ based on teacher network and evidence maps. The weighting factor consists of two parts: the former represents the contribution from evidence maps, whereas the latter denotes contribution from the training model.

For the former, a compact classifier *correct sub-network* is adopted to assess the sample x_i based on the evidence maps E_i, and give the classification probability p_i^E for the class with label $y = 1$ and recognition results y', then a *bool* function is defined to validate the effectiveness of the evidence maps for disease

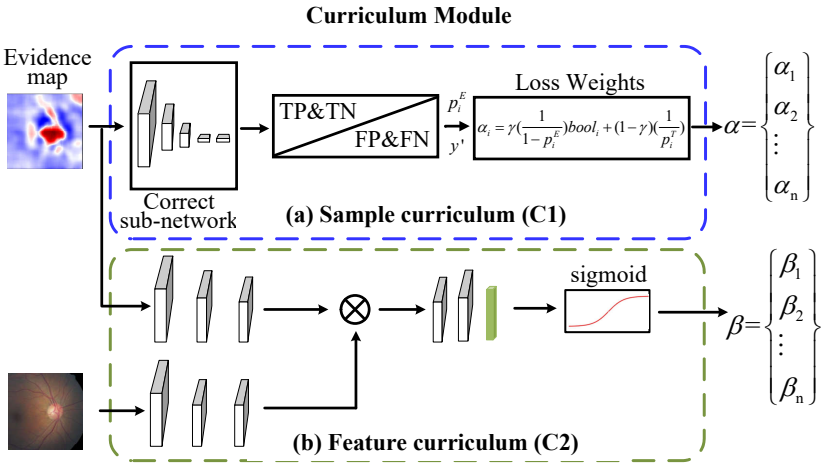

Fig. 4: The dual-curriculum adaptively provides two types of weights α and β along with the training procedure as the sample curriculum (a) and feature curriculum (b), respectively. The weights α and β are adopted in teacher network to balance the training benefits for unbiased glaucoma diagnosis.

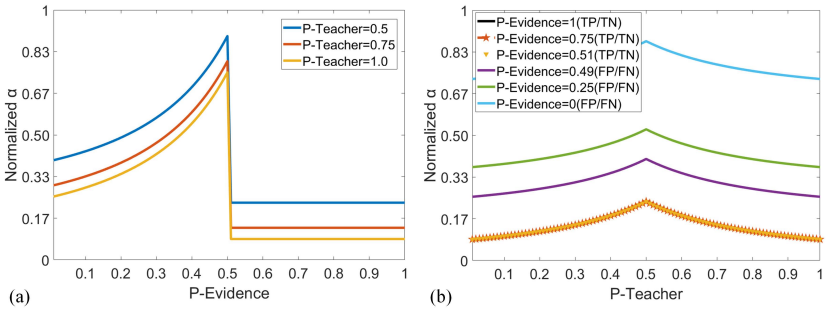

Fig. 5: A weighting factor α is proposed to balance training benefits of samples from normal to abnormal, form easy to hard. Specifically, the weighting factor enables model to focus on rare hard samples by reshaping the loss function. The factor up-weights loss contribution of rare hard samples with the greatest value when it is misclassified and its classification probability p is near to 0.5.

diagnosis and then determine its contribution to sample reweighting. So the *bool* function is defined as

$$bool_i = \begin{cases} 0 & y_i' == y_i \\ 1 & y_i' \neq y_i \end{cases} \tag{7}$$

It should be noted that there are three properties of the weighting factor α_i in Eq.(6): **1)** When a sample is misclassified based on the evidence maps

($p_i^E < 0.5$), $bool_i = 1$ and the weighting factor α_i is regulated by p_i^E, whereas when the sample is classified correctly, the weighting factor α_i is unaffected by p_i^E (Fig. 5(a)). This setting balances the training contribution of samples from both positives and negatives. **2)** When a sample is misclassified based on the evidence maps ($p_i^E < 0.5$), as p_i^E gets closer to 0.5, the weighting factor α_i becomes larger and the loss is up-weighted (Fig. 5(a)). **3)** The weighting factor α_i is also regulated by p_i^T to focus on the hard samples. As p_i^T gets closer to 0.5, the latter part of weighting factor α_i becomes larger, whereas as p_i^T gets father to 0.5, the latter part of weighting factor α_i becomes smaller (Fig. 5(b)). Based on the regulation of p_i^T, the model well focuses on hard samples ($p_i^T \approx 0.5$)

Feature Curriculum ($C2$). Feature curriculum is designed to encode the importance of local features by a set of spatial weights β on each sample. The feature curriculum is created by up-weighting highly discriminative regions and corresponding disease-specific evidential features that potentially contribute to the final disease recognition. The evidential regions represent visual attention and diagnosis focus of disease patterns. In our work, a nonlinear weighting is designed to enforce the curriculum learning of better convolutional features, which not only generate potential disease biomarkers but also abstract more semantic classification.

A CNN-based path is designed to guide the learning of better spatial features using the evidence maps E_i. As shown in Fig. 4(b), the path shares the input image and evidence maps from the student network, and models the feature curriculum as a set of weights β of convolutional features in spatial position

$$\beta_i = UpConv(\sigma(MLP(\mathbf{E}_i) \otimes MLP(\mathbf{F}_i))) \tag{8}$$

where $\otimes$ denotes element-wise multiplication, σ is the sigmoid function. *MLP* and *UpConv* indicate the operator of multi-layer perceptron and convolution with up-sampling, respectively. $\mathbf{F}_i$ is feature map outputted from *MLP*.

The convolutional layer with 1×1 kernel is designed to transform the multiple dimensional matrix into single channel. Sigmoid function is used to shape the value to a range of [0,1] and *UpConv* operator up-samples the matrix as the same size of the original image (Fig. 4(b)). Sigmoid function is used to reshape the value to a range of [0,1] and *UpConv* operator up-samples the matrix as the same size of the original image and exerts one weight on each feature of the position.

Summarized Advantages: The dual-curriculum is innovatively designed to encode two sequences of training criteria $C1$ and $C2$ with weighting factors α and β to balance the training benefits of biased data distribution.

3.3 Teacher Network for Glaucoma Diagnosis

Teacher network is a CNN-based classification model with two distinguished characteristics: **1)** an effective training objective is defined with the help of sample weights α, which is updated in each iteration towards a uniform distribution; **2)** a sophisticated feature attention is designed with the renovation of feature

curriculum β to guide the teacher network capture the discriminative feature which is meaningful for glaucoma diagnosis in an iterative training process.

In standard training, the model often minimizes the expected loss for the training set and equally weight each input sample in the loss function. However, training contributions from samples with a different disease severity and distribution are unequal because of difference of gradient flow from the biased dataset. To balance the contributions, here, the proposed EGDCL learns samples weights α_i and feature weights β for the input sample x_i. Therefore, we minimize the newly-designed loss function as

$$\Theta^*, \Psi^* = \arg\min_{\Theta,\Psi} \sum_{i}^{N} \alpha_i CE(p_i^T, \Theta, \Psi, \beta_i \mathbf{F}) \tag{9}$$

where α_i, β_i are the loss and feature weights of the i^{th} sample, respectively. $CE(p_i^T, \Theta, \Psi, \beta_i \mathbf{F})$ denotes the standard cross-entropy loss on the sample i with the reweighted feature maps $\beta_i \mathbf{F}$. Note that $\{\alpha_i\}_i^N$ and $\{\beta_i\}_i^N$ are encoded in the dual-curriculum and adaptively assign importance weights to samples and its features in each iteration. The loss function is defined not only on the learning contribution of each sample, but also on the feature aggregation at each position.

Summarized Advantages: A novel reweighted loss function and local feature aggregation are proposed to train the unbiased diagnosis model with the debiasing training criteria (dual-curriculum).

4 Experiments and Results

To demonstrate the superiority of the proposed EGDCL, we conduct some experiments on the unbiased glaucoma diagnosis problem and compare the results with baselines and the state-of-the-art methods.

4.1 Dataset and Evaluation

Dataset. Our EGDCL is validated with the challenging dataset LAG [16], which makes public 4854 fundus images labeled with either positive glaucoma (1711) or negative glaucoma (3143). The dataset is randomly divided into training (2427) and testing (2427) sets. Furthermore, the EGDCL is also validated on other challenging dataset RIM-ONE [8] with 51 glaucomatous and 118 normal eyes. To compare with the baselines, fundus images are all resized to 224×224 before inputting to EGDCL.

Evaluation Metrics. Given the model trained with our method, the results are evaluated in terms of five different metrics: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, $Sensitivity = \frac{TP}{TP+FN}$, $Specificity = \frac{TN}{TN+FP}$, $F2-score = \frac{5TP}{5TP+4FN+FP}$, and AUC. Here, TP, TN, FP, and FN are the numbers of true positive, true negative, false positive and false negative, respectively. It should be noted that the sensitivity measures the performance at detecting the positives, which is significant to evaluate how good a model is at classifying disease cases, especially

hard samples. F2-score is adopted to emphasize the significance of sensitivity because a high sensitivity indicates rare overlooks of the actual positive.

In addition, the receiver operating characteristic curve (ROC) and area under ROC (AUC) are adopted in our experiments. We indicate the teacher network without dual-curriculum learning as experimental *Baseline*, sample curriculum as *C1* and feature curriculum as *C2*.

4.2 Training and Inference

EGDCL is configured under the teacher-student framework where student network is adopted only in training stage, whereas teacher network is implemented in both training and inference stages. When training EGDCL, the supervision of diagnosis label and attention maps are simultaneously employed for student network to obtain evidence maps. The loss function of Eq.(9) is minimized through the SGD algorithm with Adam optimizer and 0.9 momentum. The initial learning rate is set to 4×10^{-4}. The initial values of α are set as 1. $\gamma = 0.5$ in Eq.(6) and batch size is set to be 8 in our experiments. Inference involves simply forwarding an image through the trained teacher network. The predictions from teacher network are applied to final evaluations directly.

4.3 Performance of Unbiased Glaucoma Diagnosis

As shown in Fig. 6 and Table. 1, EGDCL delivers unbiased glaucoma diagnosis and hard sample mining on LAG dataset [16] with the top performance on all the evaluation metrics with 0.9712 of *Accuracy*, 0.9721 of *Sensitivity*, 0.9707 of *Specificity*, 0.9665 of *F2-score* and 0.9931 of *AUC*. The results indicate that our EGDCL well handles the training bias and obtains the accuracy of glaucoma diagnosis with the help of the dual-curriculum. In particular, we need to emphasize the improvement of *Sensitivity* benefited from the accurate assessment of hard samples. Compared with the baselines, our EGDCL obtains the highest scores with *Sensitivity* of 0.9721 given the *Specificity* of 0.9707, which indicates that more cases with glaucoma are correctly identified by our method, even though the cases with heavy diagnosis difficulty. Besides, the highest *F2-score* of 0.9665 indicates our proposed EGDCL not only ensures the specificity by identifying the true negatives, but also obtains excellent sensitivity by correctly finding the true positives. This means our method can help clinicians find more of hard glaucomatous cases.

Fig. 7 shows the success of our EGDCL on glaucoma diagnosis with the ROC curves and AUC values. Evidenced by ROC curves and AUC value (0.9931), the glaucoma diagnosis results indicate that our EGDCL achieves a competitive performance by mining the hard positives and negatives cases.

In addition, we conduct extensive experiments on other glaucoma dataset (RIM-ONE [8]) to demonstrate the effectiveness of the EGDCL. This experiment adopts 169 cases for training and testing, and the results show promised performance with 0.951 of *Accuracy*, 0.916 of *Sensitivity*, 0.979 of *Specificity*, 0.976 of *F2-score* and 0.927 of *AUC*.

Table 1: Performance of our EGDCL on LAG under different configurations for glaucoma diagnosis with five evaluation criterion. Each cell contains the corresponding value and its improvement versus baseline.

Method	Accuracy	Sensitivity	Specificity	AUC	F2-score
Baseline	0.9604	0.9467	0.9675	0.9908	0.9448
Baseline+C1	0.9662 (↑ 0.58%)	0.9709 (↑ 0.96%)	0.9638 (↓ 0.37%)	**0.9945** **(↑ 0.37%)**	0.9630 (↑ 1.84%)
Baseline+C2	0.9571 (↓ 0.33%)	0.9345 (↓ 1.22%)	0.9688 (↑ 0.13%)	0.9907 (↓ 0.01%)	0.9355 (↓ 0.93%)
Baseline+C1+C2	**0.9712** **(↑ 1.08%)**	**0.9721** **(↑ 2.54%)**	**0.9707** **(↑ 0.32%)**	0.9931 (↑ 0.23%)	**0.9665** **(↑ 2.17%)**

4.4 Effectiveness of Dual-Curriculum Learning

Convergence Speed. Our EGDCL achieves the optimal convergence speed due to the help of dual-curriculum design. Compared with other configurations, EGDCL saves more than half of the training time to get the minimum of the loss. From Fig. 6 we can observe that our proposed learning paradigm can stably convergence to the minimum after about the 15^{th} epoch. It should be noted that the feature curriculum *C2* introduces improvement of diagnosis effectiveness despite there exists a weak disturbance of the convergence curve between *Baseline+C1* and *Baseline+C1+C2*.

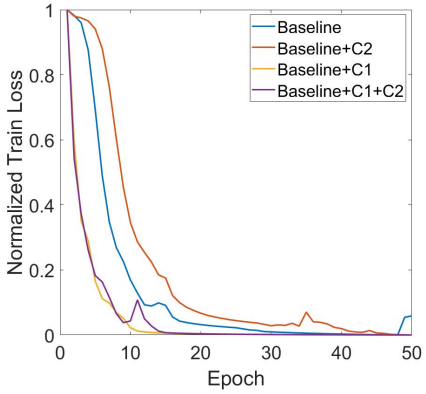

Fig. 6: The curve of train loss along with epoch demonstrates significant improvements of training convergence of diagnosis model.

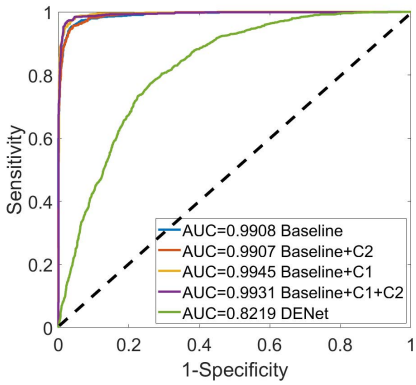

Fig. 7: The ROC curves with AUC scores for glaucoma diagnosis based on the different configurations of the proposed EGDCL.

The outstanding convergence speed benefits from: **1)** the dual-curriculum gradually selects training samples from easy to hard by sample reweighting.

This strategy helps to find better local minimal of a non-convex training criterion by loss reweighting, hence, our EGDCL achieves the global optimization by knowledge accumulation from easy samples. **2)** The dual-curriculum learning emphasizes the training contributions of different samples with different diagnosis difficulty, which gives rise to improved generalization and faster convergence.

Effectiveness for Hard Sample Mining. The effectiveness of the dual curriculum learning on hard sample mining can be proven by Table. 1. For all the evaluation metrics, EGDCL outperforms the baseline models with an average of 1.27%, where no dual-curriculum learning is explored during the training. It should be noted that the improvement of sensitivity is 2.54% up to 0.972 whereas the improvement of specificity is 0.32% up to 0.971, which means our EGDCL can accurately discover more patients who have the condition in the early stage. These significant improvements attribute the success to hard negatives mining with the dual-curriculum learning. We can also observe from Table. 1 that the integration of two types of curriculum (sample curriculum and feature curriculum) provides the optimal advance for unbiased glaucoma diagnosis.

4.5 Performance Comparison

Comparisons reveal the great advantages of our EGDCL for unbiased glaucoma diagnosis over existing methods [30, 16, 17, 7], as shown in Table. 2. We compare our EGDCL with all the methods that have been tested on the LAG dataset, namely the GON [17], DCNN [3], MCL-Net [30], DENet [7] and AG-CNN [16].

Compared results show that EGDCL achieves the best performance on glaucoma diagnosis, and obtains the average improvement of 1.08%, 1.81%, 1.77%, 1.81% and 1.55% in terms of accuracy, sensitivity, specificity, AUC and F2-score, respectively. Specifically, the EGDCL significantly improves the sensitivity of glaucoma diagnosis to 97.21% by accurately identifying the potential glaucomatous cases, which is crucial to identify potential positives in clinical diagnosis. The above results indicate that the proposed EGDCL significantly outperforms other state-of-the-art methods in all metrics.

Fig. 7 plots the ROC curves of our method and others, for visualizing the trade-off between sensitivity and specificity. It is easily seen in the plot that the ROC curve of our EGDCL is closer to the upper-left corner, which means that the sensitivity of our EGDCL is always higher than other methods given the same specificity value. Further quantification evaluation is reported in Table. 2, which shows the great advantages of our method in terms of AUC value.

To demonstrate the advantaged performance, our EGDCL is compared with other state-of-the-art methods on RIM-ONE dataset, which suffers more serous class balance between positives and negatives. The evaluation metrics in Table. 3 indicates that our method obtains a significant improvement of 8.8% compared with the state-of-the-art, which outperforms other methods in all metrics. It is worth noted that our EGDCL performs significantly better than other methods in terms of sensitivity.

Table 2: Comparison with state-of-the-art methods for glaucoma diagnosis on LAG dataset. EGDCL achieves the best performance with the average improvement of 1.08%, 1.81%, 1.77%, 1.81% and 1.55% in terms of accuracy, sensitivity, specificity, AUC and F2-score, respectively, comparing with AG-CNN [16].

Method	Accuracy	Sensitivity	Specificity	AUC	F2-score
GON [17]	0.897	0.914	0.884	0.960	0.901
DCNN [3]	0.892	0.906	0.882	0.956	0.894
MCL-Net [30]	0.962	0.964	0.957	0.979	0.958
DENet [7]	0.756	0.631	0.843	0.822	0.650
AG-CNN [16]	0.953	0.954	0.952	0.975	0.951
Focal loss [19]	0.951	0.908	0.973	0.986	0.915
Class-balance [5]	0.949	0.915	0.968	0.986	0.919
Hard mining [26]	0.958	0.937	0.969	0.991	0.938
Our EGDCL	**0.9712** (↑ **1.08**%)	**0.9721** (↑ **2.54**%)	**0.9707** (↑ **0.32**%)	**0.9931** (↑ **0.23**%)	**0.9665** (↑ **2.17**%)

Table 3: Comparison with state-of-the-art methods for glaucoma diagnosis on RIM-ONE dataset. The proposed EGDCL outperforms others with a significant improvement of 8.8%, comparing with AG-CNN [16].

Method	Accuracy	Sensitivity	Specificity	AUC	F2-score
GON [17]	0.661	0.717	0.623	0.681	0.679
DCNN [3]	0.800	0.696	0.870	0.831	0.711
MCL-Net [30]	0.824	0.786	0.823	0.803	0.721
DENet [7]	0.558	0.492	0.569	0.574	0.338
AG-CNN [16]	0.852	0.848	0.855	0.916	0.837
Our EGDCL	**0.951** (↑ **9.85**%)	**0.916** (↑ **6.77**%)	**0.979** (↑ **12.48**%)	**0.976** (↑ **6.01**%)	**0.927** (↑ **8.98**%)

5 Conclusions

The proposed novel curriculum learning paradigm (EGDCL) performs unbiased glaucoma diagnosis by designing an adaptive dual-curriculum. Innovatively, the dual-curriculum is designed with the guidance of evidence maps to build a training criterion, which gradually cures the bias in training data. The dual-curriculum balances training benefits of biased data and gradually cures the training bias from easy to hard, from normal to abnormal. Generally, the dual-curriculum is designed to represent the training criteria of sample reweighting, which simultaneously encodes the feature and sample weights with the guidance of evidence maps. Experimental results indicate that our EGDCL outperforms the baselines and the state-of-the-art methods. The proposed EGDCL not only gives rise to improved faster convergence, but also obtains the top performance on unbiased glaucoma diagnosis. It endows our EGDCL a great advantage to handle the special issue of training bias in clinical applications.

References

1. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML. pp. 41–48. ACM (2009)
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
3. Chen, X., Xu, Y., Wong, D.W.K., Wong, T.Y., Liu, J.: Glaucoma detection based on deep convolutional neural network. In: EMBC. pp. 715–718. IEEE (2015)
4. Chen, X., Xu, Y., Yan, S., Wong, D.W.K., Wong, T.Y., Liu, J.: Automatic feature learning for glaucoma detection based on deep learning. In: MICCAI. pp. 669–677. Springer (2015)
5. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9268–9277 (2019)
6. Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE TMI **37**(7), 1597–1605 (2018)
7. Fu, H., Cheng, J., Xu, Y., Zhang, C., Wong, D.W.K., Liu, J., Cao, X.: Disc-aware ensemble network for glaucoma screening from fundus image. IEEE TMI **37**(11), 2493–2501 (2018)
8. Fumero, F., Alayón, S., Sanchez, J.L., Sigut, J., Gonzalez-Hernandez, M.: Rim-one: An open retinal image database for optic nerve evaluation. In: 2011 24th international symposium on computer-based medical systems (CBMS). pp. 1–6. IEEE (2011)
9. Gargeya, R., Leng, T.: Automated identification of diabetic retinopathy using deep learning. Ophthalmology **124**(7), 962–969 (2017)
10. Guo, S., Huang, W., Zhang, H., Zhuang, C., Dong, D., Scott, M.R., Huang, D.: Curriculumnet: Weakly supervised learning from large-scale web images. In: ECCV. pp. 135–150 (2018)
11. Haarburger, C., Baumgartner, M., Truhn, D., Broeckmann, M., Schneider, H., Schrading, S., Kuhl, C., Merhof, D.: Multi scale curriculum CNN for context-aware breast MRI malignancy classification (2019)
12. Haleem, M.S., Han, L., Van Hemert, J., Li, B.: Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: a review. CMIG **37**(7-8), 581–596 (2013)
13. Jiménez-Sánchez, A., Mateus, D., Kirchhoff, S., Kirchhoff, C., Biberthaler, P., Navab, N., Ballester, M.A.G., Piella, G.: Medical-based deep curriculum learning for improved fracture classification. In: MICCAI. pp. 694–702. Springer (2019)
14. Jin, S., RoyChowdhury, A., Jiang, H., Singh, A., Prasad, A., Chakraborty, D., Learned-Miller, E.: Unsupervised hard example mining from videos for improved object detection. In: ECCV. pp. 307–324 (2018)
15. Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: NeurIPS. pp. 1189–1197 (2010)
16. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: A large-scale database and cnn model. In: CVPR. pp. 10571–10580 (2019)
17. Li, Z., He, Y., Keel, S., Meng, W., Chang, R.T., He, M.: Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology **125**(8), 1199–1206 (2018)

18. Liao, W., Zou, B., Zhao, R., Chen, Y., He, Z., Zhou, M.: Clinical interpretable deep learning model for glaucoma diagnosis. IEEE JBHI (2019)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
20. Mookiah, M.R.K., Acharya, U.R., Chua, C.K., Lim, C.M., Ng, E., Laude, A.: Computer-aided diagnosis of diabetic retinopathy: A review. Computers in biology and medicine **43**(12), 2136–2155 (2013)
21. Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis **59**, 101570 (2020)
22. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: ICML. pp. 4331–4340 (2018)
23. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Deep imbalanced attribute classification using visual attention aggregation. In: ECCV. pp. 680–697 (2018)
24. Schacknow, P.N., Samples, J.R.: The glaucoma book: a practical, evidence-based approach to patient care. Springer Science & Business Media (2010)
25. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: CVPR. pp. 761–769 (2016)
26. Smirnov, E., Melnikov, A., Oleinik, A., Ivanova, E., Kalinovskiy, I., Luckyanets, E.: Hard example mining with auxiliary embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 37–46 (2018)
27. Tham, Y.C., Li, X., Wong, T.Y., Quigley, H.A., Cheng, C.Y.: Global prevalence of glaucoma and projections of glaucoma burden through 2040 a systematic review and meta-analysis. Ophthalmology **121**(11), 2081–2090 (2014)
28. Zhao, R., Chen, X., Xiyao, L., Zailiang, C., Guo, F., Li, S.: Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning. IEEE JBHI (2019)
29. Zhao, R., Chen, Z., Liu, X., Zou, B., Li, S.: Multi-index optic disc quantification via multitask ensemble learning. In: MICCAI. pp. 21–29. Springer (2019)
30. Zhao, R., Li, S.: Multi-indices quantification of optic nerve head in fundus image via multitask collaborative learning. Medical Image Analysis **60**, 101593 (2020)
31. Zhao, R., Liao, W., Zou, B., Chen, Z., Li, S.: Weakly-supervised simultaneous evidence identification and segmentation for automated glaucoma diagnosis. In: AAAI. vol. 33, pp. 809–816. AAAI (2019)
32. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595 (2017)