Diagnosing glaucoma on imbalanced data with self-ensemble dual-curriculum learning

Article *in* Medical Image Analysis · October 2021

DOI: 10.1016/j.media.2021.102295

4 authors, including:

Zhao Rongchang
Central South University
43 PUBLICATIONS **410** CITATIONS
SEE PROFILE

Zailiang Chen
48 PUBLICATIONS **561** CITATIONS
SEE PROFILE

Shuo Li
The University of Western Ontario
369 PUBLICATIONS **6,116** CITATIONS
SEE PROFILE

Medical Image Analysis (2021)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Diagnosing glaucoma on imbalanced data with self-ensemble dual-curriculum learning

Rongchang Zhao[a], Xuanlin Chen[a], Zailiang Chen[a], Shuo Li[b,*]

[a] *School of Computer Science and Engineering, Central South University, Changsha, 410083, China*
[b] *Western University, London, ON, Canada*

ARTICLE INFO

Article history:
Received 1 May 2013
Received in final form 10 May 2013
Accepted 13 May 2013
Available online 15 May 2013

Communicated by S. Sarkar

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Curriculum learning, Glaucoma diagnosis, Data imbalance, Feature augmentation, Computer-aided diagnosis, Self ensembling

ABSTRACT

Glaucoma diagnosis often suffers from two types of data imbalances: 1) class imbalance, *i.e.,* the non-glaucoma majority cases occupy most of the data; 2) rare cases, *i.e.,* few cases present the uncommon retinopathy *e.g., bayoneting* or *physiologic cupping*. This dual-imbalances make glaucoma diagnosis model easy to be dominated by the majority cases but cannot correctly classify the minority and/or rare ones. In this paper, we propose an adaptive re-balancing strategy in the feature space, Self-Ensemble Dual-Curriculum learning (SEDC), to improve the glaucoma diagnosis on imbalanced data by augmenting feature distribution with feature distilling and feature re-weighting. Firstly, the self-ensembling (SEL) is developed to reinforce the discriminative ability of feature representations for the minority class and rare cases by distilling the features learned from the abundant majority cases. Secondly, the dual-curriculum (DCL) is designed to adaptively re-weight the imbalanced data in the feature space to learn a balanced decision function for accurate glaucoma diagnosis. Benefiting from feature distilling and re-weighting, the proposed SEDC fairly represents fundus images, regardless of the majority or rare cases, by augmenting the feature distribution to obtains the optimal decision boundary for accurate glaucoma diagnosis on the imbalanced dataset.

Experimental results on three challenging glaucoma datasets show that our SEDC successfully delivers accurate glaucoma diagnosis by the adaptive re-balancing strategy, with the average mean value of Accuracy 0.9712, Sensitivity 0.9520, Specificity 0.9816, AUC 0.9928, F2-score 0.9547. Ablation and comparison studies demonstrate that our method outperforms state-of-the-art methods and traditional re-balancing strategies. The experiment also shows that the adaptive re-balancing strategy proposed in our method provides a more effective training approach with optimal convergence performance. It endows our SEDC a great advantage to handle the disease diagnosis on imbalanced data distribution.

1. Introduction

Although computer-aided diagnosis (CAD) makes a sequence of advances on glaucoma diagnosis (Haleem et al., 2013; Zhao and Li, 2020; Fu et al., 2018a; Zhao et al., 2019c), the data imbalance exhibited in fundus images leads to the inaccurate performance of glaucoma diagnosis in clinical application. Fundus images always have imbalanced distribution, *i.e.,* non-glaucoma class claims most of the samples, while other classes have relatively few samples (Fig.1). Data imbalance makes CAD models prefer the dominant samples but perform

*Corresponding author
e-mail: shuoli@gmail.com (Shuo Li)

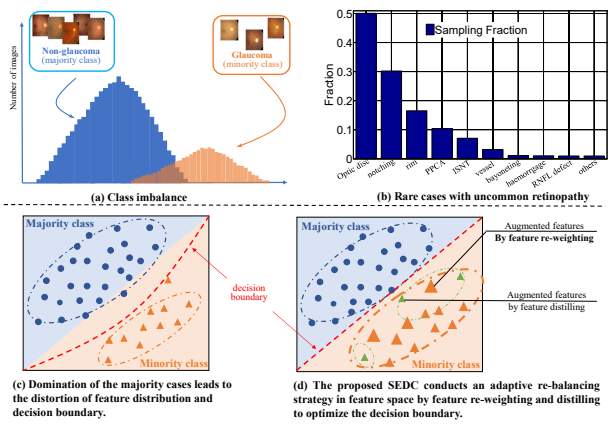

Fig. 1: Fundus images follow two types of data imbalances: **(a) Class imbalance.** Sample frequency exhibits an imbalanced class distribution where non-glaucoma samples dominate while glaucoma class has relatively few samples; **(b) Rare cases with uncommon retinopathy.** Few cases exhibit the uncommon retinopathy features such as *bayoneting*, which makes the accurate diagnosis hard. Exiting methods (c) often leads to the inaccurate performance of glaucoma diagnosis because of the distortion of feature distribution. The proposed SEDC (d) conducts an adaptive re-balancing strategy to obtain the accurate performance by augmenting feature distribution with feature distilling and feature re-weighting.

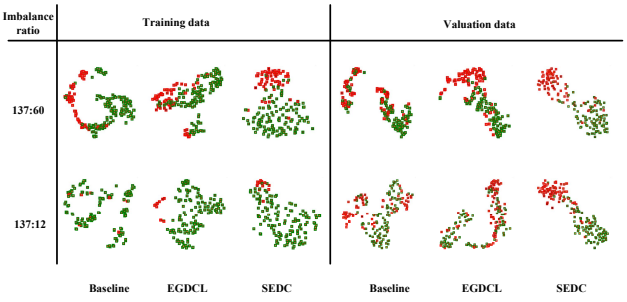

Fig. 2: The imbalanced training data seriously disturbs the learning of feature representation, leading to the distortion of feature distribution. The features, extracted by three different models in the embedding layer, are visualized with t-SNE (Van der Maaten and Hinton, 2008) on the RIM-ONE dataset. For the baseline model, it is hard to separate the glaucoma samples (red) from the normal cases (green) because glaucoma class is narrowly distributed. The narrowed feature distribution leads to the distortion of the original feature distribution. Our SEDC is developed to augment the distorted feature distribution via feature re-weighting and feature distilling, and make the best separation with a distinct margin in the feature distribution

.

poorly on others because the imbalanced measure of the empirical risk minimization (Zhao et al., 2020). When dealing with such fundus data, existing deep learning methods are infeasible to achieve outstanding diagnosis accuracy and miss the accurate diagnosis when facing the cases with rare conditions due to both the data-hungry limitation of deep learning the imbalanced distribution of training datasets.

There are two types of data imbalances for glaucoma diagnosis with fundus images: **1) class imbalance.** The non-glaucoma cases occupy most of the data, whereas glaucomatous have relatively few samples (Fig.1(a)). This class imbalance incurs model-bias classification towards the majority non-glaucoma class and leads to a high false negative. **2) Rare cases.** There are fewly rare cases (defined as *hard sample* in our method) exhibiting the uncommon retinopathy such as *bayoneting* and *physiologic cupping*, which is under-represented by the CAD models. In Fig.1(b), most of glaucomatous fundus images are related to the common retinopathy features of *optic disc changes*, while rarely few cases present the clinical features as *bayoneting* or *haemorrhage*. Therefore, deep learning models are easy to be over-fitted to assess glaucoma based on optic disc appearance but hard to accurately recognize cases with other rare retinopathy features. These imbalanced data distort the overall feature distribution, compromise the discriminative ability of CNN features, and lead to an unaccepted false in some cases (Li et al., 2018).

The impact of data imbalance on the learned feature space is investigated by conducting an empirical study. After the feature learning with baseline and the proposed model, the learned features are visualized with t-SNE (Fig.2), which generates the projection of the features with lower dimension to demonstrate the distortion of feature distribution. As observed in Fig.2, the normal samples and the glaucoma cases present different distribution patterns. The normal samples has a relatively large spatial span, while the glaucoma cases has a significantly small spatial span and is embedded into the large span of the normal class. This uneven distribution between head and tail classes distorts the overall feature distribution, and leads to the challenges of separating the rare glaucoma class from normal class.

The prominent methods for dealing with the data imbalance are class re-balancing strategies (He and Garcia, 2009). Generally, the class re-balancing methods adjust the network training by re-sampling (He and Garcia, 2009; Byrd and Lipton, 2019; More, 2016) or cost-sensitive re-weighting (Lin et al., 2017; Haarburger et al., 2019; Cui et al., 2019; Cao et al., 2019) to promote the performance of classification. Those methods adjust the training strategy of the deep neural network to make it closer to the test distributions. Thus, the class re-balancing methods benefit the updating of classifier's weights and promote the accurate classification of imbalanced datasets.

However, class re-balancing strategy still suffers from three crucial drawbacks when used in glaucoma diagnosis on the imbalanced dataset: **1) Difficult to define the re-balancing weight.** Existing methods often rely on prior knowledge to define the re-balancing weight for each sample in the imbalanced data distribution, which lacks an adaptive instrument to adjust the discriminative ability to obtain the optimal decision-making in glaucoma diagnosis. The trained model is at the risk of blindly over-fitting the rare cases and under-fitting the whole data distribution. **2) Powerless to rare cases.** Methods equipped with a re-balancing strategy can not accurately recognize the rare cases with retinopathy such as bayoneting or hemorrhage. Those rare cases, named as *hard samples* here, are often ignored by the training strategy as noise or outliers for a favorable statistical performance. However, that ignoring leads to low accuracy in practice testing, especially for the glaucoma patient at its early stage. **3) Harmful to representation learning.** The re-balancing strategy enhances the classification but brings unexpected damage to the representation ability of the neural networks due to the distortion of the original distribu-

tion. The latest research (Kang et al., 2019; Zhou et al., 2020) shows the damage of sample re-balancing in computer vision.

Curriculum learning (Bengio et al., 2009) has the potential to address the data imbalances in glaucoma diagnosis by organizing the training data from easy to hard and from imbalanced to balanced (Wang et al., 2019; Zhao et al., 2020). Curriculum learning benefits the effectiveness of model training inspired by the learning proceeds of humans. Curriculum learning highly organizes the training process based on the predefined curricula settings to learn from easy to hard. The learning paradigm has been empirically demonstrated to be effective in achieving outstanding performance for medical image analysis (Haarburger et al., 2019; Jiménez-Sánchez et al., 2019)

However, applying curriculum learning directly into glaucoma diagnosis to deal with the two types of data imbalances is challenging due to **1) The distortion of feature distribution.** The glaucoma samples occupy a relatively narrow span in the feature space compared with the non-glaucoma cases because of the lack of sufficient samples (*e.g.,* Fig.1(c)). This uneven feature distribution compromises the discriminative ability of the learned features and leads to poor classification performance. **2) Pre-defined curriculum.** Pre-defining a fixed curriculum to order the fundus images with its difficultness is not realistic in clinical applications. However, there is no existing method to update the curriculum adaptively along with model training for imbalanced data analysis. **3) Isolated feature learning.** The curriculum learning only organizes the training process by ordering the training sequence while ignores the feature distilling between majority class and minority class. The traditional curriculum learning can not adjust the learning focus in different training stages, leading to isolated feature representation between different classes.

In this paper, we propose a self-ensemble dual-curriculum learning framework (SEDC, Fig.3) to achieve accurate glaucoma diagnosis on the imbalanced dataset. The proposed SEDC innovatively conducts an adaptive re-balancing strategy in the feature space to optimize the decision boundary (red dotted line in Fig.3) by augmenting the feature distribution (*i.e.,* with feature distilling and re-weighting). Firstly, the self-ensembling (SEL) is developed to reinforce the discriminative ability of feature representation for the minority class and rare cases by distilling the feature representations learned from the abundant majority class. Secondly, the dual-curriculum learning (DCL) is designed to learn a balanced decision function by adaptivelly feature re-weighting in different training stages for the optimization of the decision boundary. The re-weighted features distribution moves the decision boundary toward the optimal direction by balancing the training contributions between the majority class, *i.e.,* non-glaucoma and minority class *i.e.,* glaucoma. The proposed SEDC is significant to inherit the advantages of the self-ensemble learning that progressively reinforce the discriminative ability of feature representation for rare cases and the curriculum learning that gradually adjust the re-weighting factors to learn a balanced decision function for accurate glaucoma diagnosis.

The proposed SEDC is capable of achieving effective glaucoma diagnosis on imbalanced dataset due to three advantages:

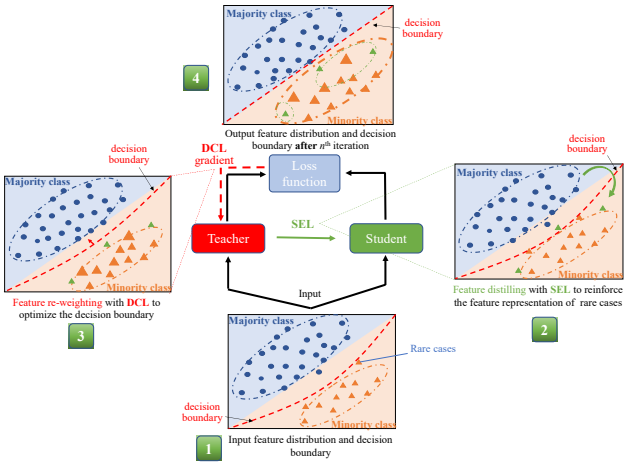

Fig. 3: The proposed SEDC tackles two types of data imbalances in glaucoma diagnosis by augmenting the feature distribution with feature distilling and re-weighting. The feature distribution is augmented by 1) the self-ensemble learning (SEL) to reinforce the discriminative ability of feature representation for the rare cases with the distilled features learned from the majority abundant samples; 2) the dual-curriculum learning (DCL) to conduct feature re-weighting to learn a balanced decision function to move the boundary to the optimal direction.

1) The proposed self-ensemble framework iteratively expands the feature distribution of the minority class and rare cases to represent the imbalanced data in feature space by distilling the elementary knowledge from easy to hard and from imbalanced to balanced. **2)** The dual-curriculum learning helps the model learn a balanced decision function in the feature space by adaptive feature re-weighting in different training stages, which gradually moves the decision boundary toward the optimal direction for accurate glaucoma diagnosis. **3)** The proposed SEDC conducts an adaptive sample selection strategy by progressively paying attention to each sample with different importance weights to learn the discriminative feature representations from majority easy samples and imbalanced hard ones.

Our proposed SEDC achieves top performance on three most competitive glaucoma diagnosis datasets with the two imbalance issues, *i.e.,* LAG (Li et al., 2019), REFUGE (Orlando et al., 2020) and RIM-ONE (Fumero et al., 2011). The proposed self-ensemble dual-curriculum learning paradigm can benefits both accurate classification and effective training of long-tailed recognition in other areas. The main contributions of this work are summarized as follows:

- Dual-curriculum learning paradigm (SEDC) is proposed for the first time to handle data imbalances in glaucoma diagnosis by gradually augmenting the feature distribution via feature distilling and feature re-weighting.
- An effective self-ensemble learning is developed to reinforce the discriminative ability of minority class and rare cases by distilling feature from majority class, which provides a new learning paradigm for balanced deep learning.
- The contrastive re-balanced loss is developed to jointly learn the discriminative representation and the powerful classifier by integrating supervised contrastive loss into the

sample re-balancing strategy.

In this work, we advance our preliminary attempt on ophthalmic disease diagnosis with imbalanced data (Zhao et al., 2020) in the following aspects: (1) conduct self-ensemble framework to boost the performance of glaucoma diagnosis on the rarely hard samples; (2) contrastive re-balanced loss function to learn the optimal discriminative feature representations; (3) carry out more extensive experiments on performance analysis and comparisons.

The rest of this paper is organized as follows: In Section 2 we first introduce the related works, and then we give the detailed presentation of our proposed methodology and the algorithm in Section 3. Experimental configurations and dataset details are introduced in Section 4 and results analysis are presented in Section 5. Section 6 concludes the paper.

2. Related work

Automated glaucoma diagnosis: The success of machine learning has benefited the computer-aided glaucoma diagnosis (Schacknow and Samples, 2010; Zhao and Li, 2020; Zhao et al., 2019c; Fu et al., 2018a). Prior works on computer-aided glaucoma diagnosis devoted to learning a robust classifier by designing the hand-crafted features like texture, higher-order spectra, wavelet-based features. Those methods consider feature embedding and classifier learning individually, thus leads to lower classification accuracy. With the development of deep learning, modern methods shed new light on the automated glaucoma diagnosis with deep learning models in the end-to-end manner(Chen et al., 2015a,b). This type of automated glaucoma diagnosis method employs CNNs and GANs in optic disc segmentation (Fu et al., 2018a; Haleem et al., 2013), medical indices estimation (Zhao and Li, 2020; Zhao et al., 2019a,b) or ONH assessment (Liao et al., 2019; Li et al., 2019) to promote the performance of glaucoma diagnosis. The success of computer-aided glaucoma diagnosis is undoubtedly inseparable to the advantages of deep learning models, which enable the CAD with the power capable of feature representations from collected training datasets.

Balanced data learning: Both image classification (Ren et al., 2018; Sarafianos et al., 2018) and object detection (Lin et al., 2017; Jin et al., 2018) face a massive data imbalance when learning the model from the practical datasets. Data imbalance refers to a disproportionate ratio of observations among the different class or/and disease severity, leading to inefficient and extensive redundancy training due to the imbalanced dataset. Re-balancing training methods fall into two categories: 1) data re-sampling (He and Garcia, 2009; Chawla et al., 2002; Geirhos et al., 2018; Li and Vasconcelos, 2019), which choosing the suitable proportion of data to train a network, including over-sampling adds repeated samples from minor classes, and under-sampling removes random samples. Data re-sampling often leads to either over-fitting or under-fitting. 2) Cost-Sensitive learning (Lin et al., 2017) through elaborately designing training curriculums or learning losses that assigns a weight to each sample and minimizing the weighted loss function (Ren et al., 2018). Besides, hard negative mining samples hard samples during training (Shrivastava et al., 2016). Unfortunately, to our best knowledge, no work has been reported to tackle the special issue of data imbalances in medical diagnosis originating from both class imbalance and rare hard samples.

Mean-teacher mechanism: Mean-teacher framework (Tarvainen and Valpola, 2017) is a self-ensembling model designed for the classification task of natural images. It typically consists of two models, *i.e.,* student model and teacher model, with the same architecture. In the training process, teacher model is trained with the back-propagation algorithm by designing various loss functions, while weights of student model are updated as exponential moving average (EMA) of the weights of teacher model. Therefore, the mean-teacher framework conducts weights ensembling of teacher model at different training process to help build a more reliable student model to produce consistency targets and be adopted in various medical applications (Liu et al., 2020; Huo et al., 2020).

Margin-based softmax loss: Softmax loss is the commonly used loss function in classification problems by combining a fully connected layer, softmax function, and cross-entropy loss. However, the learned features with original softmax loss are not sufficiently discriminative for the classification problems, especially images with the various appearance and mixed features. To directly enhance the feature discrimination, several margin-based softmax loss and contrastive loss functions (Liu et al., 2016; Wang et al., 2018b; Liu et al., 2019) have been proposed where a margin function is carefully designed to enforce greater intra-class compactness and inter-class discrepancy. Recently, angular margin (A-softmax) (Liu et al., 2017), additive margin (AM-Softmax) (Wang et al., 2018a), additive angular margin (ArcSoftmax) (Deng et al., 2019) are proposed and achieved promising results on face recognition.

Curriculum learning: Curriculum learning (Bengio et al., 2009) 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. Its main hypothesis is that the order in which samples are presented to an iterative optimizer is important. Novel variants of curriculum learning include self-paced learning (Jiang et al., 2015) where the curriculum is automated learned. Such approach has been adopted to tackle data imbalance in the medical image analysis, where the curriculum is updated by letting the learner focus on medical knowledge (Jiménez-Sánchez et al., 2019; Jesson et al., 2017).

Contrastive learning: The contrastive learning has recently become a prominent technique in unsupervised learning (He et al., 2020; Grill et al., 2020; Chen et al., 2020), achieving state-of-the-art performance. The contrastive learning learns representation by contrasting positive pairs against negative pairs. Various frameworks use different apporach to study the role of positive pairs to learn the discriminative feature representation, including SimCLR (Chen et al., 2020) uses augmented views of other items in a minibatch as negative samples, Moco (He et al., 2020) uses a momentum updated memory bank of old negative representations to enable the use of very large batches of negatives.

In fundus images, the testing samples are usually disjoint

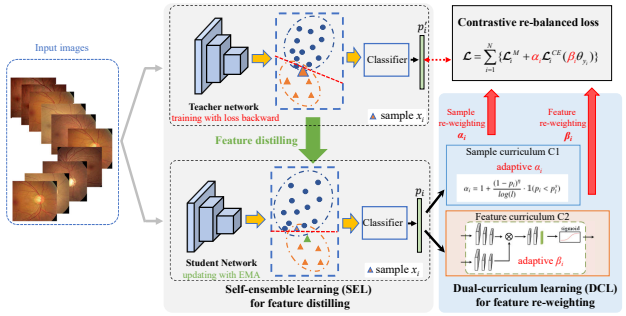

Fig. 4: The proposed SEDC consists of three parts: **self-ensemble learning (SEL)** distills features learned from majority class to reinforce the discriminative representation of minority glaucoma class, **dual-curriculum learning (DCL)** re-weight each given sample x_i in feature space with the adaptive weight factor α_i and β_i, **contrastive re-balanced loss function** exerts cost-sensitive modulation on the loss function with the re-balanced training strategy and re-weighting feature maps for a strong teacher network.

from the training set because of the inhomogeneous appearances of glaucomatous samples in different disease stages. However, those loss functions do not explicitly emphasize each sample according to its diagnosis difficulty and imbalanced class. Therefore, the proposed SEDC learns the discriminative feature presentation of each sample by leveraging the merit of curriculum learning and contrastive learning, which gradually emphasizes the samples with different diagnosis difficulty and imbalanced class.

3. Methodology

The proposed self-ensemble dual-curriculum learning (SEDC, Fig.4) conducts a mean-teacher framework (*teacher* and *student* networks) with curriculum learning for glaucoma diagnosis on imbalanced data. In SEDC, the self-ensemble learning (SEL) distills the feature learned from the majority class to reinforce the discriminative ability of feature representation for minority class and rare cases. In addition, the dual-curriculum learning (DCL) is designed to adaptively calculate the weighting factors online for feature re-weighting, which exerts exact modulation on the imbalanced data in the feature space to learn a balanced cost and optimal separating hypersurface to partition the underlying samples into two classes. Benefited from the momentum updating of the student network, SEDC inherits the advantages of ensemble learning to progressively boost the network's prediction with the newly learned knowledge and encourage the subsequent model to be consistent with the ground truth.

3.1. Overview of the proposed framework

As shown in Fig.4, our SEDC consists of three components: **self-ensemble learning (SEL)** distills the features learned from majority non-glaucoma class to reinforce the discriminative ability of feature representation for the minority glaucoma class and rare cases, **dual-curriculum learning (DCL)** conducts the feature re-weighting with adaptive-updated weighting factors α_i and β_i for the given sample x_i to learn the optimal decision boundary by balancing the training contributions between majority class and minority class, **contrastive re-balanced loss function** exerts cost-sensitive modulation on the loss function with the re-balanced training strategy and re-weighting feature maps for a strong teacher network. The self-ensemble learning is conducted with the mean-teacher structure (Tarvainen and Valpola, 2017), where the *teacher* network acts as the base network trained with the re-balanced loss function whereas *student* network plays a role of ensemble model updated with the teacher network for consistent prediction on the imbalanced data. After each iteration, parameters of the student network are updated with *exponential moving average (EMA)* method to ensemble the model of newly trained teacher and historical student. The teacher network is trained with the re-balanced loss function, while the student network does not participate in the back-propagation. In the inference stage, the test images are inputted into the student network for the prediction.

Owing to the self-ensemble dual-curriculum learning, data imbalances can be counteracted in an adaptive and ensemble manner. In the training, the SEL reinforces the discriminative ability of feature representation for the minority glaucoma class and rare cases by distilling features learned from the majority class, and the DCL updates the weight factor online with the dual-curriculum and learns a balanced optimization of the decision boundary toward the optimal direction by feature re-weighting in the loss function. As the iteration goes on, the model becomes closer to the normal distribution of training data, and the prediction in the ensemble student network gets more accurate. In the test phase, the test images are sent to the student network to achieve the glaucoma diagnosis.

3.2. Self-ensemble learning for feature distilling

Self-ensemble learning (Fig.4) is developed to reinforce the discriminative ability of feature representation for the minority class by reusing the features learned from the abundant majority cases. The self-ensemble network is consisted of two networks with the same architecture (*teacher* and *student* networks). The teacher network is trained with the re-balanced loss function (detailed in Sec.3.4) to gradually optimize the decision boundary. In contrast, the student network plays an ensemble model to enable the discriminative representation for all the samples, regardless of majority non-glaucoma class or minority glaucoma class. Both the teacher and student network share the same architecture, named as *attention network* (Fig.5) (Zhao et al., 2020). After each back-propagation training, the teacher network obtains the optimal feature representation and classification performance on the batch of samples, and then distills the learned feature representation to student network for a well representation of the minority class by updating the parameters with EMA.

3.2.1. Student network architectures

The student network (Fig.5) is a deep classification network with the same architecture while teacher network equipped with channel and spatial attention modules for the accurate feature representation. The student network develops two separate attention pathways, which not only learns the rich contextual fea-

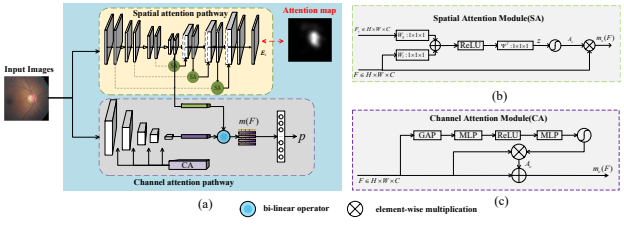

Fig. 5: The student network (a) has the same architecture with teacher network with two separate attention pathways to obtain the discriminative features for accurate glaucoma diagnosis. (b) Spatial attention module (SA). (c) Channel attention module (CA).

tures 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 $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ from previous convolutional layer, the attention modules infer a 3D attention refined feature $\hat{\mathbf{F}} \in \mathbb{R}^{C\times H\times W}$ to enhance the model's discriminative ability.

We adopt a residual learning scheme along with the two separate attention pathways to facilitate the gradient flow. To make the attention modules available in classification network, we first calculate the spatial attention $m_s(\mathbf{F}) \in \mathbb{R}^{C\times H\times W}$ and channel attention $m_c(\mathbf{F}) \in \mathbb{R}^{C\times H\times W}$ at two separate pathways, then integrate them into an unified attention refined feature map $\hat{\mathbf{F}} = m(F) \in \mathbb{R}^{C\times H\times W}$ by a bi-linear operator

$$\hat{\mathbf{F}} = m(F) = \|sqrt(m_s(\mathbf{F}) \otimes m_c(\mathbf{F}))\|_2 \tag{1}$$

where $\otimes$ is a bi-linear operator as suggested in (Lin et al., 2015). For the given feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, its attention refined feature $\hat{\mathbf{F}} \in \mathbb{R}^{C\times H\times W}$ provides a weighted representation for fundus images for each local pixel and each channel.

Given an input image, two attention modules, channel and spatial, are conducted to calculate complementary attention at each location and the networks focus on 'what' and 'where' information respectively. In this work, to apply attention into classification model, the channel and spatial attention modules are placed in a parallel manner.

Spatial attention module (SA). The spatial attention module (Fig.5b) targets to compute the spatial attention coefficient $A_s \in [0, 1]$ to identify salient image regions and prune feature responses to preserve only the activations relevant to the specific task. The output of SA is the element-wise multiplication of input feature map $\mathbf{F}$ and spatial attention coefficient A_s as $M_s(\mathbf{F}) = (1 + A_s) \times \mathbf{F}$. Inspired by the attention Unet (Oktay et al., 2018), the spatial attention coefficient A_s is formulated as Eq.2 by a set of operators with the feature map $\mathbf{F}$ and a gating vector $\mathbf{F}_g$. We adopt soft-attention approach by introducing the gating vector, which is the lower-level feature response in the networks.

$$A_s = \sigma_2(\psi^T(\sigma_1(W_x^T\mathbf{F} + W_g^T\mathbf{F}_g + b_g)) + b_\psi) \tag{2}$$

where linear transformations W_x^T, W_g^T, ψ^T are implemented as a convolution with the kernel of $1 \times 1 \times 1$, b_g and b_ψ are bias. σ_1 is ReLU function, and $\sigma_2(x) = \frac{1}{1+exp(x)}$ corresponds to sigmoid activation function.

Channel attention module (CA). The channel attention module (Fig.5c) aims to compute the channel attention coefficient $A_c \in [0, 1]$ to identify salient feature channels to preserve the activations relevant to the glaucoma diagnosis task. The output of CA is the element-wise multiplication of input feature map $\mathbf{F}$ and the channel attention coefficient A_c as $M_c(\mathbf{F}) = (1+A_c)\times\mathbf{F}$. The channel attention coefficient is formulated as Eq.3 with a simple network as suggested in (Lin et al., 2015). The network is composed of multi-layer perceptron (MLP) with two hidden layers, ReLU function, global average pooling (GAP) and sigmoid function (σ_2).

$$A_c = \sigma_2(MLP(ReLU(MLP(GAP(\mathbf{F}))))) \tag{3}$$

3.2.2. Self-ensemble learning

In our SEDC, the model relies on the self-ensemble of teacher and student networks with the EMA, which allows us to reuse the features learned from the majority class to reinforce the discriminative ability of feature representation for minority class. The representation ability of network is progressively distilled into the student network and updated with the newly learned one after the next iteration. In the beginning, we assume that the data imbalances in the training data are unknown. The model is trained with the standard cross-entropy loss and initial dual-curriculum. After the first iteration, the student network attempts to identify the hard samples from a batch of data and maintains the data imbalances with the help of the adaptive dual-curriculum. The teacher network learns the informative representation to discriminate the disease cases with the re-weighted importance, which introduces easier samples first and focuses on the informative samples to increase the feature margin between different classes. The SEDC model counteracts the imbalances of training data using self-forming ensembles of teacher network and student network. The ensemble student network is evaluated on the entire data and provides updated guidelines to adjust the re-weight factors in the dual-curriculum.

In our framework, the teacher network is trained with the re-balanced loss function updated by the dual-curriculum learning module. Then the feature representation about the majority class is distilled into the student network to preserve the consistency prediction and reinforce the discriminative ability of student network for minority class. Here, the SEDC framework adopts an EMA to address the teacher network's knowledge aggregation to the student network. Specifically, let w_l^t and w_l^s denote the weights of teacher and student network after l^{th} iteration, w_{l-1}^s is the weight of student network after the $(l-1)^{th}$ iteration, we have

$$w_l^s = \gamma w_{l-1}^s + (1-\gamma)w_l^t \tag{4}$$

where $\gamma \in [0, 1)$ is the momentum parameter that controls the weight momentum speed. With the EMA, the parameters of student network is updated with the historical (student network) and newly learned (teacher network) knowledge about data imbalances. The momentum update of student network's

parameters in Eq.4 makes it evolve more smoothly. As a result, though the teacher and student networks face different data imbalances, the adaptive re-balancing strategy in dual-curriculum is adjusted progressively.

3.3. Dual-curriculum learning for feature re-weighting

Innovatively, the dual-curriculum learning module is designed to conduct feature re-weighting for the optimization of decision boundary. The dual-curriculum guides the model to learn from easier samples first and harder samples later, and pay more attention to the important data to move the decision boundary toward the optimal direction. The dual-curriculum is updated along with the training procedure according to what knowledge the model has already learned in each iteration. To effectively deal with the two interwoven imbalances, a dual structure curriculum is designed with sample curriculum C1 and feature curriculum C2.

3.3.1. Sample Curriculum (C1)

The sample curriculum (Fig. 6) is designed to dynamically encode a set of importance 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 SEDC, 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, given a training sample x_i, its weighting factor α_i can be defined as

$$\alpha_i = 1 + \frac{(1-p_i)^\eta}{log(l)} \cdot \mathbb{1}(p_i < p_l^s) \tag{5}$$

where η is a hyperparameter, l is the epoch, and p_i denotes the model's estimated probability for the class with label $y = 1$ based on student network, $p_l^s = \lambda l + p_0^s$ denotes the threshold in the l epoch to identify the hard samples, where λ and p_0^s are the hyperparameter and initial threshold, respectively. The main part of weighting factor α_i consists of two parts: the former, $\frac{(1-p_i)^\eta}{log(l)}$, belonging to $[0, 1]$, represents the modulation for training cost of sample x_i, with tunable focusing parameter $\eta \geq 0$. Whereas the latter, $\mathbb{1}(p_i < p_l^s)$, denotes the identification of hard sample by comparing with an adaptive threshold p_l^s in different training epoch l. $\mathbb{1}(B) \in \{0, 1\}$ is the indicator function that returns 1 if B evaluates as true.

$$\mathbb{1}(p_i < p_l^s) = \begin{cases} 1; & if\ p_i < p_l^s\ (hard\ sample) \\ 0; & elsewise\ (easy\ sample) \end{cases} \tag{6}$$

It should be noted that there are three properties of the weighting factor α_i in model training: **1)** When a sample is misclassified ($p_i < 0.5$) and p_i is small, the weighting factor α_i is equal to $1 + (1 - p_i)^\eta / log(l)$ and the loss is up-weighted. As $p_i \to 1$, the sample is well-classified, so the weighting factor goes to one and the loss is unaffected. **2)** When a sample is misclassified and p_i is small, $\mathbb{1}(p_i < p_l^s) = 1$, it notes that the sample is hard to be correctly classified, so its loss will be modulated with an up-weighted factor α_i to focus the learning on hard samples. As p_i is growing, $\mathbb{1}(p_i < p_l^s)$ becomes invalid, the sample becomes an easy one to be well-classified, then $\alpha_i = 1$, the loss is unaffected. **3)** At the beginning of training, the probability threshold p_l^s is small to introduce more easy samples for the training of the model. As the going of training to epoch l, the probability threshold p_l^s increasing, the harder samples are introduced into the training process. Based on the introduction of identification function $\mathbb{1}(p_i < p_l^s)$, the sample curriculum C1 modulates training contributions of each sample by exerting a weighting factor α on the loss, which makes the model learn from easy to hard and from imbalanced to balanced.

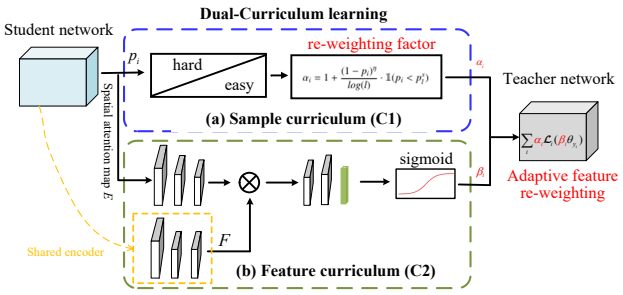

Fig. 6: The dual-curriculum learning module conducts two types of curriculum to simultaneously emphasis the model's focus on the training cost of sample and contributions of spatial pixels. Sample curriculum (a) generates an adaptive importance re-weighting factor α_i to modulate the training cost of sample x_i, whereas feature curriculum (b) models the pixel-level attentions β_i to emphasis the discriminative on the feature maps. The two factor are used to modulate the loss function of teacher network for adaptive feature re-weighting.

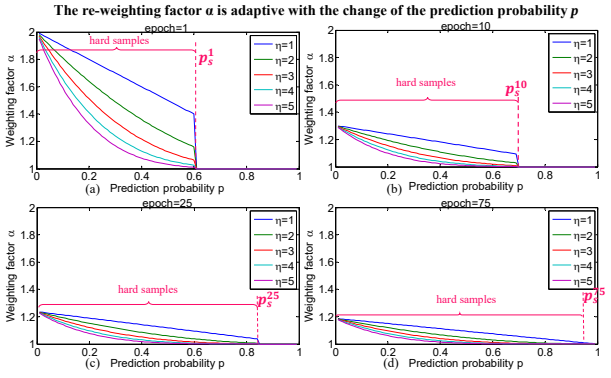

Fig. 7: A weighting factor α is proposed in the sample curriculum (C1) to emphasize the model's focus on each sample corresponding to its importance in different training stages. Specifically, the weighting factor α enables model to pay more attention to the rare hard samples by up-weighting loss contribution with the higher value, when it is mis-classified and its classification probability p is smaller than an adaptive threshold p_l^s.

3.3.2. 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 con-

volutional 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 spatial attention maps $\mathbf{E}$. As shown in Fig. 6, the path shares the input image and spatial attention maps from the student network, and models the feature curriculum as a set of weights β of convolutional features in spatial position

$$\beta_i = UpConv(\sigma(Conv(\mathbf{E}) \times \mathbf{F})) \tag{7}$$

where $\times$ denotes element-wise multiplication, σ is the sigmoid function. *Conv* and *UpConv* indicate the operator of convolution without and with up-sampling, respectively. $\mathbf{F}$ is feature map outputted from the encoder of the student network.

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. 6). 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.

3.4. Contrastive re-balanced loss

The contrastive re-balanced loss is developed to conduct the re-balancing training of the teacher network by assigning adaptive weights to each sample and corresponding feature vectors. Specifically, the contrastive loss is adopted to enhance the features' discriminative power by enforcing greater intra-class compactness and inter-class discrepancy to preserve the beneficial properties of cost re-weighting. The proposed contrastive re-balanced loss is modulated by the dual-curriculum to balance the training contribution of imbalanced data distribution. Here, the training loss of each sample is modulated to balance the training contributions in different iterations. Therefore, the proposed SEDC learns samples weights α_i and feature weights β_i for the input sample x_i.

Given a set of N randomly sampled fundus images and corresponding labels $\{x_i, y_i\}_{i=1,...,N}$, let i represents the indices of an arbitrary images. The contrastive re-balanced loss function is defined as

$$\mathcal{L} = \sum_{i=1}^{N} \{\mathcal{L}_i^M + \alpha_i \mathcal{L}_i^{CE}(\beta_i \theta_{y_i})\} \tag{8}$$

$$\mathcal{L}_i^M = -\frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \mathbb{1}(y_i = y_j) log \frac{e^{M(z_i \cdot z_j/\tau)}}{\sum_{k=1}^{N} \mathbb{1}(k \neq i) e^{M(z_i \cdot z_k/\tau)}} \tag{9}$$

where θ_{y_i} indicates the network parameters for feature representations, $\mathbb{1}(B) \in \{0, 1\}$ is an indicator function that returns 1 if B evaluates as true. $\mathcal{L}_i^{CE}$ represents the cross entropy loss on the samples i with the re-weighting function α_i and feature modulation β_i. $z_i \cdot z_j$ computes an inner product between the normalized feature vectors z_i and z_j for similarity function M.

To enhance the features' discriminative power by enforcing greater intra-class compactness and inter-class discrepancy, the supervised contrastive loss $\mathcal{L}_i^M$ is developed in Eq.9 and Fig.8. The supervised contrastive loss is defined as the similarity of samples based on its labels, where samples belonging to the same class are pulled together in embedding space, while simultaneously pushing apart samples from different classes. Within the context of Eq.9, sample x_i is called as *anchor*, and sample $x_j, j = \{1, ..., N\}$ denotes *positives* if it has the same label as the *anchor*, *i.e.*, $y_i = y_j$, while *negatives* if it has the different label as the *anchor*, *i.e.*, $y_i \neq y_j$. During the training of contrastive loss, the encoder is tuned to maximize the numerator of the log argument in Eq.9 while minimizing the denominator for the greater intra-class compactness and inter-class discrepancy. For each *anchor* x_i, $\{\alpha_i\}_i^N$ and $\{\beta_i\}_i^N$ are encoded in the dual-curriculum (C1 and C2 in Sec.3.2) and adaptively assign importance weights to samples and its features after 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.

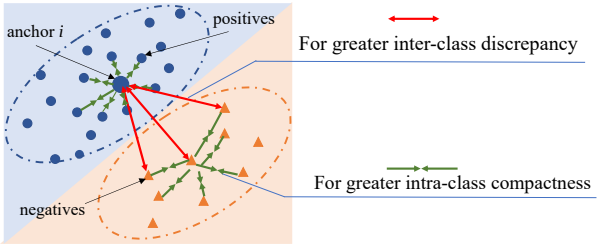

Fig. 8: The supervised contrastive loss $\mathcal{L}_i^M$ is designed to enhance the features' discriminative power by enforcing greater intra-class compactness and inter-class discrepancy with a self-learning manner.

3.5. Algorithm of SEDC

Given a training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^N$, learning a model with our SEDC method described in Sec.3.1 leads to the minimizing of the re-balanced loss function as Eq.8 and then iterative updating of the student network and the dual-curriculum module. The training procedure requires estimating the parameters of the teacher network w^t and updating parameters of the student network w^s with EMA. The learning procedure of the SEDC is iterative with four stages: contrastive learning for discriminative representation, optimizing teacher network with re-balanced loss function, updating the parameters of student network with EMA, updating the dual-curriculum based on the output of student network. Algorithm 1 summarizes the detailed procedure of training for our SEDC.

4. Experiments

4.1. Datasets

In this section, we evaluate the effectiveness of our SEDC on three glaucoma datasets with the two interwoven imbalances, LAG, REFUGE and RIM-ONE (Table. 1). LAG (Li et al., 2019) makes public 4854 fundus images labeled with either glaucoma(1711) or non-glaucoma(3143) obtained from Beijing Tongren Hospital. The dataset is randomly divided into training(2427) and testing(2427) sets. REFUGE challenge (Orlando

Algorithm 1: : Training procedure of SEDC

Input : Training set $\mathcal{S}$; temperature τ; epoch threshold T; momentum γ; similarity function M;
Output: Parameters of SEDC: w^s;

Initialization: $l = 1$; randomly initialize w^s and w^t;
while $l < T$ **do**
  Shuffle the training set $\mathcal{S}$, and fetch mini-batch $\mathcal{S}_n$;
  /* Forward */
  for $i \leftarrow 1$ **to** N **do**
    Obtain the feature vector z_i of the sample x_i;
    Calculate the similarities $M(z_i, z_j)$ of feature vectors between anchor i and positives or negatives j;
    /* contrastive representation learning */;
    Calculate the modulations α_i and β_i by Eq.5 and Eq.7;
    Calculate the loss $\mathcal{L}_i$ by Eq.9;
  end
  Calculate the summarized loss $\mathcal{L}$ by:
  Eq.8: $\mathcal{L} = \sum_{i=1}^{N} \{\mathcal{L}_i^M + \alpha_i \mathcal{L}_i^{CE}(\beta_i \theta_{y_i})\}$;
  /* Backward */
  Compute the gradients and optimize the parameters w^t of teacher network /* optimizing teacher network with SGD */;
  /* Update */
  Update parameters w^s of student network by :
  Eq.4: $w_l^s = \gamma w_{l-1}^s + (1-\gamma) w_l^t$;
  /* updating student network with self-ensembling */;
  Update the dual curriculum α_i and β_i by :
  Fig.4 and Eq.5 : $\alpha_i = 1 + \frac{(1-p_i)^\eta}{log(l)} \cdot \mathbb{1}(p_i < p_l^s)$;
  /* updating dual-curriculum learning */
end
return w^s

et al., 2020) publicly releases a set of 1200 fundus images (120 glaucoma and 1080 non-glaucoma) with clinical ground truth labels, where 800 for training and 400 for test. Furthermore, RIM-ONE (Fumero et al., 2011) dataset is also employed to train our SEDC, where the RIM-ONE dataset makes public 455 fundus images labeled with either glaucoma or normal cases. In our settings, RIM-ONE is randomly split into 273 for training and 182 for testing.

Table 1: The proposed SEDC is evaluated on the three challenging glaucoma datasets with the interwoven imbalances (class imbalance and rare cases).

Dataset	No. of images			Imbalance ratio	Hard samples
	Total	Glaucoma	Non-glaucoma		
LAG	4854	1711	3143	1.8	✓
REFUGE	1200	120	1080	9	✓
RIM-ONE	455	200	255	1.275	✓

Note that, besides the class imbalance, the three datasets also face the rare cases problem (Smirnov et al., 2018), where a rare of samples are hard to be correctly classified owing to its inhomogeneous appearance (disc change, conus, enlarged cupping, pale optic disc, *etc*) and disease severity, especially in the *preperimetric* or *early* stage of glaucoma. In this work, the official splits of training and validation images are utilized for fair comparisons.

4.2. Experimental Settings

SEDC is configured under the mean teacher framework, where the teacher network is adopted only in the training stage, whereas the student network is implemented in both training and inference stages. When training SEDC, the supervision of the diagnosis label is employed for the teacher network to obtain discriminative representation. The loss function of Eq.8 is minimized through the SGD algorithm and 0.9 momentum. The initial learning rate is set to 8×10^{-2}. The initial values of α are set as 1. $\eta = 2$ in Eq.5, $\gamma = 0.5$ in Eq.4 and batch size is set to be 16 in our experiments. The learning rate is decayed at every 10 epoch by 0.5 in our SEDC. Inference involves simply forwarding an image through the trained student network. The predictions from the student network are applied to final evaluations directly.

4.3. Evaluation Criteria

Given the model trained with our method, the results are evaluated in terms of five different metrics: Accuracy $Acc = \frac{TP+TN}{TP+TN+FP+FN}$, Sensitivity $Sen = \frac{TP}{TP+FN}$, Specificity $Spe = \frac{TN}{TN+FP}$, F2-score $F2 = \frac{5TP}{5TP+4FN+FP}$, 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 rare cases. $F2$ 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 backbone classification network without dual-curriculum learning and self-ensemble as experimental ***baseline***, sample curriculum as *C1* and feature curriculum as *C2*.

4.4. Comparison Methods

In our experiments, the SEDC is compared with three groups of methods:

- **Baseline methods.** We employ the basic classification network with attention module and cross-entropy loss as our ***baselines***.
- **Re-balancing strategies.** To prove the effectiveness of our adaptive dual-curriculum learning, we also compare with the re-balancing strategies proposed in state-of-the-art re-sampling and re-weighting works in visual recognition, including focal loss (Lin et al., 2017), CB loss (Cui et al., 2019), hard example mining (Smirnov et al., 2018), LDAM (Cao et al., 2019).

- **State-of-the-art methods.** We compare with state-of-the-art methods on the glaucoma diagnosis (Li et al., 2019; Fu et al., 2018b; Li et al., 2018; Zhao and Li, 2020), which achieve good diagnosis accuracy on these three aforementioned datasets.

5. Results and Analysis

Our SEDC achieves advanced glaucoma diagnosis performance with high classification accuracy and gains excellent precise with high sensitivity and specificity. The effectiveness of our SEDC framework in re-balanced training of glaucoma diagnosis model are validated in three folds. (1) The quantification performance is examined on three challenging datasets with different imbalance ratio: RIM-ONE (1.275), LAG (1.8) and REFUGE (9). (2) The effectiveness of each components in our SEDC is probed to demonstrate its capacity in glaucoma diagnosis. (3) The advantages of the proposed SEDC over existing methods on glaucoma diagnosis are revealed compared with the state-of-the-art methods.

5.1. Diagnosis performance on different dataset

Experimental results on LAG. As shown in Fig. 9 and Table. 2, SEDC delivers accuracy glaucoma diagnosis on the dataset LAG with the top performance on all the evaluation metrics with 0.9712 of *Acc*, 0.9520 of *Sen*, 0.9816 of *Spe*, 0.9547 of *F2* and 0.9928 of *AUC*. The results indicate that our SEDC well handles the imbalances in training data and obtains the accuracy of glaucoma diagnosis with self-ensemble dual-curricular learning. In particular, we need to emphasize the improvement of *Sen* benefited from the accurate assessment of hard samples. Owing to the capability of re-balanced training of the imbalanced dataset, the proposed SEDC effectively learns the discriminative features of hard samples based on the knowledge preserve from easy cases. Therefore, compared with the baseline, our SEDC obtains the highest scores with *Sen* of 0.9520 given the *Spe* of 0.9816. It indicates that more cases with glaucoma are correctly identified by our method, even though the cases with heavy diagnosis difficulty. Besides, the highest *AUC* of 0.9928 indicates our proposed SEDC not only ensures 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. 9 shows the success of our SEDC on glaucoma diagnosis with the ROC curves and AUC values. Evidenced by ROC curves and AUC value (0.9928), the glaucoma diagnosis results indicate that our SEDC achieves a competitive performance by progressively mining the training benefits of different samples from imbalanced classes.

Experimental results on REFUGE. We conduct extensive experiments on REFUGE dataset with an imbalanced ratio of nine. Table. 2 reports the sensitivity and specificity of various methods. The experimental results demonstrate that our SEDC consistently achieves the best performance on glaucoma diagnosis, even though the more serious class imbalance.

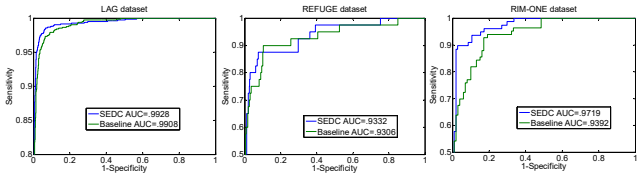

Fig. 9: The ROC curves with AUC values demonstrate the success of our SEDC on glaucoma diagnosis on three challenging datasets.

Table 2: The proposed SEDC achieves advanced glaucoma diagnosis performance with high accuracy and gains excellent precise with high sensitivity and specificity on three challenging datasets with data imbalances.

Dataset	Acc	Sen	Spe	F2	AUC
LAG	0.9712	0.9520	0.9816	0.9547	0.9928
REFUGE	0.9525	0.8000	0.9694	0.7882	0.9332
RIM-ONE	0.9396	0.8974	0.9712	0.9091	0.9719

Especially for dataset with the class imbalanced ratio nine, the SEDC get 0.9525 of *Acc*, 0.8000 of *Sen*, 0.9694 of *Spe*, 0.7882 of *F2* and 0.9332 of *AUC*, which are 1% higher than that of focal loss (Lin et al., 2017). SEDC stands on the ninth of this challenge, and baseline+B+C stands on the third place on REFUGE challenge. Our SEDC obtains comprehensively balanced indicators from *Acc*, *Sen*, *Spe*, *F2* to *AUC*, while the ranking indicator in this challenge is *AUC*. Although the SEDC does not get the first place on the leaderboard based on the AUC value, it targets at the detection of rare cases in imbalanced datasets and improves the performance of glaucoma diagnosis with an average of Acc 2.19%. Benefiting from the newly-designed self-ensemble dual-curriculum learning strategy, our SEDC achieves glaucoma diagnosis on the datasets with different imbalance ratios, especially extremely imbalance. It means that the proposed SEDC framework is capable of dealing with the interwoven issues (class balance and rare cases) existing in the fundus dataset, which is not easy to be handle by the traditional re-balancing methods. Additionally, it can be found that the re-balancing strategies are effective since they obtain a competitive performance comparing with non-balancing methods.

Experimental results on RIM-ONE. An extensive experiment is conduct on the small dataset RIM-ONE, which contains only 455 images and imbalance ratio 1.275. Table. 2 shows the results which demonstrate that our SEDC can obtain the best performance on glaucoma diagnosis with the high 0.9396 of *Acc*, 0.8974 of *Sen*, 0.9712 of *Spe*, 0.9091 of *F2*, and 0.9719 of *AUC*.

Dataset-cross validation. To demonstrate the generalization of our SEDC, the dataset-cross validation is conducted by employing the images in LAG only to train the networks, and employing images in other dataset REFUGE for testing. The results shown in Table 3. demonstrate the outperformance of our SEDC compared with the Baseline model. It should be noticed that, the two methods are all obtain the poor performance on sensitivity because of the distribution shift between the training dataset LAG and testing dataset REFUGE. It becomes another topic about domain adaptive which is out of our discussion in this paper.

Table 3: The dataset-cross validation (training on LAG while testing on REFUGE dataset) demonstrates the generlization ability of our SEDC compared with the Baseline model.

Methods	Acc	Sen	Spe	F2	AUC
Baseline	0.8842	0.08	0.6303	0.096	0.9731
Our SEDC	**0.9258**	**0.30**	**0.8438**	**0.375**	**0.9954**

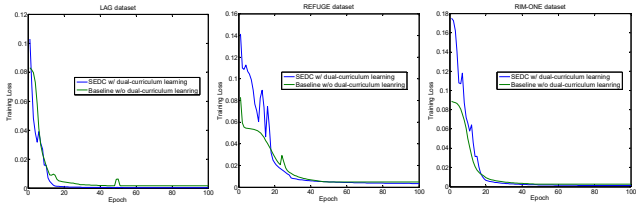

Fig. 10: The curve of train loss along with epoch demonstrates significant improvements of training convergence of diagnosis model.

5.2. Ablation study

We conduct some ablation experiments by comparing against the baselines to analyze the effectiveness of each components in our SEDC. From the results shown in Table. 4, we have the following observations:

1) Effectiveness of dual-curriculum. We conduct two contrast experiments to demonstrate the effectiveness of our newly-designed dual-curriculum learning from two aspects:

Convergence speed. Fig. 10 shows that our SEDC achieves the optimal convergence speed and optimal convergence point due to the dual-curriculum designing. Compared with other configurations, SEDC saves more than half of the training time to get the minimum of optimization. Additionally, we can observe that our proposed dual-curriculum learning paradigm can stably convergence to the minimum after about the 20^{th} epoch. Our proposed training strategy is effective for the medical image classification because the dual-curriculum can adaptively find the optimal training benefit based on sample re-weighting.

The outstanding convergence speed benefits from: **1)** the dual-curriculum learning gradually selects training samples from easy to hard and from imbalanced to balanced to handle the imbalanced data by the feature distilling. The gradual curriculum learning strategy helps find the effective learning direction of training samples, which benefits the searching for better local minimal of a non-convex training criterion. Therefore, our SEDC finds a better local minimum solution compared with previous studies by progressive knowledge distillation. **2)** The sample curriculum emphasizes the training benefits of each sample, no matter it is easy or hard, in the model learning process by weighting the training cost with a nonlinear factor function, 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. 4 and 5. For all the evaluation metrics, SEDC outperforms the baseline models with an average of 0.88%, where no dual-curriculum learning is explored during the training on LAG. It should be noted that the improvement of sensitivity is 0.56% up to 0.952 whereas the improvement of specificity is 1.46% up to 0.982 on LAG, which means our SEDC can not only accurately assess the hard samples (fundus images) with ambiguous conditions, but also effectively reduces the false positives. Benefiting from the strategy of learning from easy to hard and from imbalanced to balanced, the hard samples are correctly assessed by the knowledge distillation learned from the easy samples. These significant improvements attribute the success to hard sample mining with the dual-curriculum learning. We can also obtain this observation from the REFUGE and RIM-ONE datasets in Table. 4 that the integration of dual-curriculum learning provides the optimal advance for glaucoma diagnosis.

To demonstrate the effectiveness for hard samples mining, an experimental analysis of rare samples is conducted. The experimental results show that, compared with the baseline method, the SEDC significantly increases the number of detected hard samples from 183 to 1335 on LAG dataset, from 261 to 272 on RIM-ONE dataset, and from 37 to 86 on REFUGE dataset.

2) Effectiveness of self-ensemble learning. This experiment illustrates the efficacy of our proposed self-ensemble mechanism. Firstly, we individually train student network for curriculum generation and then teacher network for glaucoma diagnosis, and there are no interactions between the two networks (denoted as $+B+C$). Then the proposed method trains the student and teacher network together with the self-ensemble learning in the iterative manner (denoted as $+A+B+C$). The compared experimental results between w/ and w/o self-ensemble learning are shown in Table. 5. It can be observed that the self-ensemble learning achieves an average improvement of F2-score with 1.88% and Accurate with 0.82% on LAG. Owing to the model ensembling, our SEDC gradually discovers the data imbalances, updates the training strategy, and re-balances the training benefits with the spiral promoting manner. In this way, the ensemble student network updates the dual-curriculum by identifying the data imbalances coupled with the knowledge of the teacher network, while the teacher network is trained with the re-balanced loss function modulated by the adaptive dual-curriculum. The self-ensemble learning allows the model's performance to remain consistent across the student and teacher network.

We conducted another quantitative evaluation to analyze the importance of self-ensemble learning by adjusting the number of iteration in the training process. As shown in Fig. 11, an apparent improvement of the performance was observed as the iteration number increased. It should be noted that the proposed self-ensemble learning method achieves almost unchanged Acc before the 20^{th} epoch at REFUGE dataset because the model tends to the majority class of the data distribution in the beginning of the training, and the model swiftly boots the performance at a high level when the knowledge is learned from easy samples.

3) Effectiveness of contrastive re-balanced loss. A quantitative experiment of contrastive re-balanced loss is conducted by setting the different loss functions, where $+A+B+C$ denotes with our newly-designed loss and $+A+C$ denotes with the traditional cross-entropy loss function. The part of modulation coefficients α and β would not be calculated in the cross-entropy loss. To demonstrate the effectiveness of our contrastive re-

Table 4: Performance of our SEDC under different configurations for glaucoma diagnosis with five evaluation criterion. Here, A, B and C denote the self-ensemble, contrastive re-balanced loss function and dual-curriculum, respectively.

Dataset		*Acc*	*Sen*	*Spe*	*F2*	*AUC*
LAG	Baseline	0.9604	0.9467	0.9675	0.9448	0.9908
	+A	0.9637	0.9509	0.9707	0.9500	0.9925
	+B	0.9629	0.9533	0.9682	0.9510	0.9910
	+C	0.9707	**0.9796**	0.9543	**0.9788**	0.9919
	+ABC	**0.9712**	0.9520	**0.9816**	0.9547	**0.9928**
REFUGE	Baseline	0.9450	0.6750	**0.9750**	0.6888	0.9306
	+A	0.9400	0.7500	0.9611	**0.7933**	**0.9391**
	+B	0.9275	0.7500	0.9472	0.7177	0.9321
	+C	0.9475	0.7000	0.9750	0.7407	0.9182
	+ABC	**0.9525**	**0.8000**	0.9694	0.7882	0.9332
RIM-ONE	Baseline	0.8626	0.8193	0.8990	0.8293	0.9392
	+A	0.8956	0.8481	0.9320	0.8590	0.9523
	+B	0.9341	0.8904	0.9633	0.9003	0.9728
	+C	0.8901	0.8250	0.9411	0.8418	0.9706
	+ABC	**0.9396**	**0.8974**	**0.9712**	**0.9091**	**0.9719**

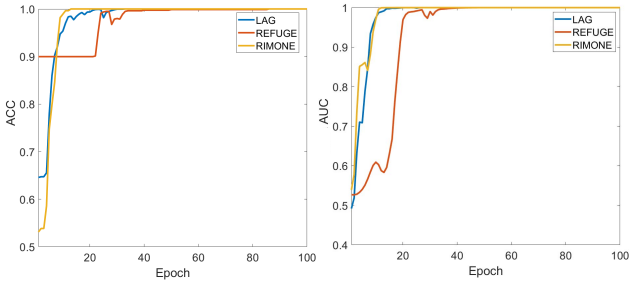

Fig. 11: The changes of AUC/Acc show that the self-ensemble learning effective boosts the diagnosis performance as the iteration number increased. It should be noted that for dataset REFUGE, although the Acc is at a lower lever before the 20^{th} epoch, self-ensemble with EMA swiftly achieves the optimal performance when the knowledge is learned from easy samples.

balanced loss, the loss function proposed the evidence-guided curriculum learning (Zhao et al., 2020) is compared. Table. 5 reports that our proposed contrastive re-balanced loss obtains the best performance on the three datasets. It should be noted that the introduction of supervised contrastive loss leads to a discriminative representation of glaucomatous samples, which give the 0.9% improvement of glaucoma diagnosis.

5.3. Performance with different imbalance ratios

To validate the performance of our SEDC with different imbalance ratio, extensive experiments are conducted by synthetically reducing the imbalance ratio (randomly removing positive samples). Compared with the baseline methods, the proposed SEDC obtains the significant improvements on various imbalance ratios from 1413:1 to 1413:427, even though the advances of our SEDC are validated on the dataset that does not seem to be extremely imbalanced. **1) Visualized comparison with the t-SNE graph.** Fig.2 shows the visualized features with t-SNE, which demonstrates the representation learning performance of our SEDC for glaucoma classification. Compared with the baseline and previous method EGDCL, the proposed SEDC is ability of obtaining the discriminative feature representation with a wide region between two classes, each class can be well distinguished. In the feature space, the feature distribution learned by the baseline method and EGDCL are narrowed, which leads to a type of distortion and make the classification difficulty. **2) Quantitative comparison under different imbalance ratios.** Table.6,7,8 show the advantages of the proposed SEDC on imbalanced glaucoma diagnosis under different imbalance ratio. With the gradual decrease of the imbalance ratio, the performance of SEDC method gradually stabilized. It means the proposed SEDC effectively deals with the extremely data imbalance. In the case of extreme imbalance, the SEDC method has stronger robustness than the baseline, which the imbalance ratio is 1413:1, SEDC brings obvious improvement of glaucoma diagnosis performance with the improvement of 8.53%, 24.06%, 0.07%, 28.34%, and 55.25% in terms of accuracy, sensitivity, specificity, F2-score and AUC, respectively.

5.4. Performance comparison

SEDC reveals great advantages for glaucoma diagnosis over existing method such as state-of-the-art glaucoma diagnosis method (Fu et al., 2018b; Li et al., 2019), loss re-weighting and re-sampling methods (Lin et al., 2017; Cui et al., 2019), hard sample mining (Smirnov et al., 2018), and curriculum learning method with attention labels (Zhao et al., 2020).

Compared with the baseline, it is shown that the SEDC obtains the average improvements of *Acc* 1.12%, 0.79%, 8.87% on three datasets, which achieves the best performance on glaucoma data re-balancing and significantly improves the performance of glaucoma diagnosis. From the compared results we have the following observations.

1) SEDC outperforms the state-of-the-art glaucoma diagnosis methods significantly on three challenging datasets. Comparing the results with others of Table.9, it clearly shows SEDC obtains more accurate glaucoma diagnosis on the LAG dataset than other SOTA CAD and re-balancing methods, which demonstrates the remarkable advantages in glaucoma diagnosis. As far as we know, there is no work reported to design algorithms for the data re-balanced training. The best performance of glaucoma diagnosis is achieved by the general classification methods, such as disc-aware glaucoma diagnosis (Fu

Table 5: Ablation studies by comparing against the baselines to analyze the effectiveness of each components in our SEDC. Here, A, B and C denote the self-ensemble, contrastive re-balanced loss function and dual-curriculum, respectively.

Dataset	Settings	*Acc*	*Sen*	*Spe*	*F2*	*AUC*
LAG	Baseline	0.9604	0.9467	0.9675	0.9448	0.9908
	+B+C	0.9633	0.9287	0.9823	0.9359	0.9905
	+A+C	0.9625	0.9357	0.9771	0.9398	0.9915
	+A+B	0.9658	0.9298	**0.9853**	0.9379	0.9927
	+A+B+C	**0.9712**	**0.9520**	0.9816	**0.9547**	**0.9928**
REFUGE	Baseline	0.9450	0.6750	0.9750	0.6888	0.9306
	+B+C	0.9500	0.7500	0.9722	0.7500	**0.9805**
	+A+C	0.9500	0.7750	0.9694	0.7711	0.9319
	+A+B	0.9525	0.7000	**0.9806**	0.7179	0.9376
	+A+B+C	**0.9525**	**0.8000**	0.9694	**0.7882**	0.9332
RIM-ONE	Baseline	0.8626	0.8193	0.8990	0.8293	0.9392
	+B+C	0.9231	0.8941	0.9485	0.9026	**0.9797**
	+A+C	0.8736	0.9231	0.8365	0.8978	0.9312
	+A+B	0.9341	0.8846	0.9711	0.8984	0.9708
	+A+B+C	**0.9396**	**0.8974**	**0.9712**	**0.9091**	0.9719

Table 6: Experimental results under different imbalance ratios indicate that the proposed SEDC outperforms the baseline and SOTA methods on LAG dataset.

Method	Imbalance Ratio	*Acc*	*Sen*	*Spe*	*F2*	*AUC*
Baseline	1413:1	0.6470	0.0012	0.9987	0.0015	0.4323
Baseline+Focal loss (Lin et al., 2017)	1413:1	0.6470	0.0000	0.9994	0.0000	0.5762
Baseline+Hard mining (Smirnov et al., 2018)	1413:1	0.6474	0.0000	**1.0000**	0.0000	0.5314
Our SEDC	1413:1	**0.7323**	**0.2418**	0.9994	**0.2849**	**0.9848**
Baseline	1413:4	0.6499	0.0139	**1.0000**	0.0173	0.5638
Baseline+Focal loss (Lin et al., 2017)	1413:4	0.6483	0.0023	1.0000	0.0029	0.7655
Baseline+Hard mining (Smirnov et al., 2018)	1413:4	0.6483	0.0047	0.9987	0.0058	0.6653
Our SEDC	1413:4	**0.8801**	**0.6612**	0.9994	**0.7091**	**0.9729**
Baseline	1413:12	0.6606	0.0409	**0.9981**	0.0506	0.7495
Baseline+Focal loss (Lin et al., 2017)	1413:12	0.6676	0.0619	0.9975	0.0761	0.7292
Baseline+Hard mining (Smirnov et al., 2018)	1413:12	0.6615	0.0432	**0.9981**	0.0534	0.7528
Our SEDC	1413:12	**0.9106**	**0.754**7	0.9955	**0.7923**	**0.9842**
Baseline	1413:85	0.8064	0.4579	0.9962	0.5128	0.9283
Baseline+Focal loss (Lin et al., 2017)	1413:85	0.8386	0.5666	0.9866	0.6171	0.9285
Baseline+Hard mining (Smirnov et al., 2018)	1413:85	0.8480	0.6098	0.9778	0.6556	0.9375
Our SEDC	1413:85	**0.9143**	**0.7629**	**0.9968**	**0.9289**	**0.984**9
Baseline	1413:171	0.8826	0.6928	0.9860	0.7341	0.9603
Baseline+Focal loss (Lin et al., 2017)	1413:171	0.8752	0.6998	0.9707	0.7361	0.9333
Baseline+Hard mining (Smirnov et al., 2018)	1413:171	0.8410	0.5970	0.9739	0.6426	0.9375
Our SEDC	1413:171	**0.9518**	**0.8832**	**0.9892**	**0.9569**	**0.9910**
Baseline	1413:256	0.8983	0.7710	0.9676	0.7980	0.9608
Baseline+Focal loss (Lin et al., 2017)	1413:256	0.8752	0.7301	0.9542	0.7583	0.9389
Baseline+Hard mining (Smirnov et al., 2018)	1413:256	0.8727	0.7138	0.9593	0.7453	0.9305
Our SEDC	1413:256	**0.9522**	**0.8808**	**0.9911**	**0.9580**	**0.9894**
Baseline	1413:342	0.9337	0.854	0.9771	0.8721	0.9803
Baseline+Focal loss (Lin et al., 2017)	1413:342	0.8909	0.7512	0.9669	0.7806	0.9473
Baseline+Hard mining (Smirnov et al., 2018)	1413:342	0.8781	0.7395	0.9536	0.7664	0.9433
Our SEDC	1413:342	**0.9592**	**0.9054**	**0.9885**	**0.9628**	**0.9927**
Baseline	1413:427	0.9382	0.8762	0.9720	0.8891	0.9824
Baseline+Focal loss (Lin et al., 2017)	1413:427	0.8921	0.7780	0.9542	0.8001	0.9515
Baseline+Hard mining (Smirnov et al., 2018)	1413:427	0.8917	0.8096	0.9364	0.8217	0.9541
Our SEDC	1413:427	**0.9600**	**0.9194**	**0.9822**	**0.9611**	**0.9891**

et al., 2018b). It should be noted that SEDC obtains comparable results with EGDCL methods, which got slightly better results on *Sen* because of its pixel-level annotation of attentions. At the same time, our SEDC is trained with only the image-level annotations. Our SEDC outperforms the state-of-the-art method with *Acc* of 0.936% and *AUC* of 1.43%. We can see from the loss function curve that our SEDC reduces half of the training time. It is easy to understand that our SEDC leverages the advantages of self-ensemble and curriculum learning to adaptive find the optimal training strategy.

Table 7: Experimental results under different imbalance ratios indicate that the proposed SEDC outperforms the baseline and SOTA methods on RIM-ONE dataset.

Dataset	imbalance ratio	Acc	Sen	Spe	$F2$	AUC
Baseline	137:1	0.5714	0.0250	1.0000	0.0311	0.6241
Baseline+Focal loss (Lin et al., 2017)	137:1	0.5659	0.0125	1.0000	0.0156	0.7119
Baseline+Hard mining (Smirnov et al., 2018)	137:1	0.6374	0.1875	0.9902	0.2232	0.6614
Our SEDC	137:1	**0.6758**	**0.2625**	**1.0000**	**0.3079**	**0.9616**
Baseline	137:2	0.5769	0.0375	1.0000	0.0464	0.7161
Baseline+Focal loss (Lin et al., 2017)	137:2	0.5714	0.0250	1.0000	0.0311	0.7725
Baseline+Hard mining (Smirnov et al., 2018)	137:2	0.5604	0.0125	0.9902	0.0155	0.5939
Our SEDC	137:2	**0.7308**	**0.3875**	**1.0000**	**0.4416**	**0.9560**
Baseline	137:12	0.6648	0.3000	0.9510	0.3438	0.8056
Baseline+Focal loss (Lin et al., 2017)	137:12	0.6868	0.3000	0.9902	0.3478	0.7466
Baseline+Hard mining (Smirnov et al., 2018)	137:12	0.6484	0.2250	0.9804	0.2647	0.6349
Our SEDC	137:12	**0.8846**	**0.7500**	**0.9902**	**0.9028**	**0.9674**
Baseline	137:24	0.7418	0.4625	0.9608	0.5125	0.8172
Baseline+Focal loss (Lin et al., 2017)	137:24	0.6429	0.2875	0.9216	0.3276	0.6967
Baseline+Hard mining (Smirnov et al., 2018)	137:24	0.6264	0.2750	0.9020	0.3125	0.6233
Our SEDC	137:24	**0.8901**	**0.8000**	**0.9608**	**0.8999**	**0.9409**
Baseline	137:36	0.7582	0.5875	0.8922	0.6217	0.8678
Baseline+Focal loss (Lin et al., 2017)	137:36	0.6044	0.2875	0.8529	0.3212	0.6140
Baseline+Hard mining (Smirnov et al., 2018)	137:36	0.7253	0.4750	0.9216	0.5191	0.7688
Our SEDC	137:36	**0.9066**	**0.8125**	**0.9804**	**0.9186**	**0.9673**
Baseline	137:48	0.7637	0.5875	0.9020	0.6233	0.8195
Baseline+Focal loss (Lin et al., 2017)	137:48	0.6648	0.5875	0.7255	0.5949	0.6940
Baseline+Hard mining (Smirnov et al., 2018)	137:48	0.6099	0.3625	0.8039	0.3930	0.5469
Our SEDC	137:48	**0.9121**	**0.8625**	**0.9510**	**0.9161**	**0.9551**
Baseline	137:60	0.7857	0.7250	0.8333	0.7342	0.8452
Baseline+Focal loss (Lin et al., 2017)	137:60	0.6923	0.5375	0.8137	0.5628	0.7748
Baseline+Hard mining (Smirnov et al., 2018)	137:60	0.7967	0.6875	0.8824	0.7106	0.8605
Our SEDC	137:60	**0.9340**	**0.8750**	**0.9804**	**0.9289**	**0.9634**

2) SEDC outperforms the best of existing re-balancing methods on glaucoma diagnosis. In contrast to the existing re-balancing methods that only deal with the class imbalance problem, our SEDC proposes a novel framework to simultaneously handle the two interwoven issues (class balance and rare cases), widespread in medical image analysis. α is calculated in focal loss and class-balanced loss based on the corresponding works (Lin et al., 2017; Cui et al., 2019). The proposed SEDC framework effectively trains the CAD model with an imbalanced dataset by leveraging the merit of self-ensemble learning and curriculum learning, which gradually adapts the training strategy and absorbs the suitable data to promote the capability of the CAD model in each iteration. The adaptive training strategy brings obvious improvement of glaucoma diagnosis performance with the improvement of 1.38%, 1.6%, 1.3%, 0.18%, and 1.78% in terms of accuracy, sensitivity, specificity, AUC, and F2-score, respectively.

3) SEDC obviously improves the glaucoma diagnosis by simultaneously obtains the discriminative feature representation and excellent classifier. As pointed in (Zhou et al., 2020; Kang et al., 2019), sample re-balancing often brings unexpected damage of the representation ability of the neural networks. Our SEDC proposes a balanced contrastive loss to improve the ability of feature representation by exploring the margin between data distributions. The SEDC simultaneously learns the discriminative feature representation and excellent classifier to improves the glaucoma diagnosis. It is easy to see from Table.9 that this strategy brings obvious improvements. Compared the EGDCL method with pixel-level annotations of attentions, our SEDC is only trained with the image-level labels and obtains comparable results with the EGDCL method.

4) The gradual curriculum learning is capable of effective utilizing the training data to improving the diagnosis performance on small dataset. The dataset RIM-ONE (Fumero et al., 2011) only contains 455 images with 200 glaucomatous and 255 normal eyes. Experiment results on RIM-ONE show that our SEDC achieves outperforms with this small data. Benefiting from the contrastive re-balanced loss function, our SEDC learns the discriminative representation with contrastive learning and classification knowledge about the fundus images by coupling the feature representation and classification in an unified framework.

5.5. Significant difference analysis

Statistical significance of the proposed SEDC versus *baseline* model and previous EGDCL are examined by the paired *Mann-Whitney U*-test with significance level of 0.1%. The *p*-value for each pair of settings are computed to demonstrate the significance improvement of our SEDC. A lower *p*-value than 0.001 indicates that the method achieves the significant difference performance than the baseline model and previously proposed EGDCL. From the results in Table.10 indicate that the proposed SEDC significantly outperforms baseline model and previously EGDCL method. In addition, the statistical signif-

Table 8: Experimental results under different imbalance ratios indicate that the proposed SEDC outperforms the baseline and SOTA methods on REFUGE dataset.

DMethod	Imbalance Ratio	*Acc*	*Sen*	*Spe*	*F2*	*AUC*
Baseline	648:1	0.7600	0.0750	0.8361	0.0676	0.5110
Baseline+Focal loss (Lin et al., 2017)	648:1	0.8775	0.0000	0.9750	0.0000	0.7769
Baseline+Hard mining (Smirnov et al., 2018)	648:1	0.5275	**0.7750**	0.5000	0.4178	0.7019
Our SEDC	648:1	**0.9725**	0.7500	**0.9972**	**0.7853**	**0.9977**
Baseline	648:4	0.7425	0.3000	0.7917	0.2429	0.6013
Baseline+Focal loss (Lin et al., 2017)	648:4	0.8775	0.0500	0.9694	0.0578	0.6695
Baseline+Hard mining (Smirnov et al., 2018)	648:4	0.7550	**0.7250**	0.7583	0.5254	0.8432
Our SEDC	648:4	**0.9525**	0.5500	**0.9972**	**0.6011**	**0.9521**
Baseline	648:8	0.9025	0.0500	0.9972	0.0613	0.6357
Baseline+Focal loss (Lin et al., 2017)	648:8	0.9000	0.0750	0.9917	0.0904	0.7282
Baseline+Hard mining (Smirnov et al., 2018)	648:8	0.6900	**0.7750**	0.6806	**0.5065**	0.8234
Our SEDC	648:8	**0.9275**	0.2750	**1.0000**	0.3216	**0.9856**
Baseline	648:16	0.8975	0.2250	0.9722	0.2514	0.8711
Baseline+Focal loss (Lin et al., 2017)	648:16	0.8975	0.0250	**0.9944**	0.0307	0.9383
Baseline+Hard mining (Smirnov et al., 2018)	648:16	0.6050	0.7000	0.5944	0.4192	0.7441
Our SEDC	648:16	**0.9775**	**0.9250**	0.9833	**0.9113**	**0.9908**
Baseline	648:24	0.8575	0.3500	0.9139	0.3415	0.8551
Baseline+Focal loss (Lin et al., 2017)	648:24	0.9050	0.1750	0.9861	0.2035	0.9137
Baseline+Hard mining (Smirnov et al., 2018)	648:24	0.8725	0.7250	0.8889	0.6332	0.9357
Our SEDC	648:24	**0.9700**	**0.8250**	**0.9861**	**0.8333**	**0.9788**
Baseline	648:32	0.9200	0.2750	0.9917	0.3161	0.9359
Baseline+Focal loss (Lin et al., 2017)	648:32	0.8000	0.4250	0.8417	0.3632	0.7139
Baseline+Hard mining (Smirnov et al., 2018)	648:32	0.8925	0.7500	0.9083	0.6726	0.9528
Our SEDC	648:32	**0.9850**	**0.9250**	**0.9917**	**0.9250**	**0.9924**
Baseline	648:40	0.9400	0.6000	0.6667	0.6502	0.9808
Baseline+Focal loss (Lin et al., 2017)	648:40	0.8650	0.2250	0.9361	0.2344	0.7317
Baseline+Hard mining (Smirnov et al., 2018)	648:40	0.8000	0.5250	0.8306	0.4339	0.7086
Our SEDC	648:40	**0.9675**	**0.7250**	**0.9944**	**0.7592**	**0.9914**

Table 9: Comparison with state-of-the-art methods for glaucoma diagnosis on LAG dataset. Compared with the optimal re-balanced method (hard sample mining), SEDC achieves the best performance with the improvement of 1.38%, 1.6%, 1.3%, 0.18% and 1.78% in terms of accuracy, sensitivity, specificity, AUC and F2-score, respectively.

Method	*Acc*	*Sen*	*Spe*	*AUC*	*F2*
GON (Li et al., 2018)	0.897	0.914	0.884	0.960	0.901
DCNN (Chen et al., 2015a)	0.892	0.906	0.882	0.956	0.894
MCL-Net (Zhao and Li, 2020)	0.962	0.964	0.957	0.979	0.958
DENet (Fu et al., 2018b)	0.756	0.631	0.843	0.822	0.650
AG-CNN (Li et al., 2019)	0.953	0.954	0.952	0.975	0.951
Focal loss (Lin et al., 2017)	0.951	0.908	0.973	0.986	0.915
Class-balance (Cui et al., 2019)	0.949	0.915	0.968	0.986	0.919
Hard mining (Smirnov et al., 2018)	0.958	0.937	0.969	0.991	0.938
EGDCL (Zhao et al., 2020)	0.971	**0.972**	0.971	0.993	**0.967**
Our SEDC	**0.971**	0.952	**0.982**	**0.993**	0.955

Table 10: Statistical significance of the proposed SEDC versus baseline model is examined by the paired *Mann-Whitney U*-test with significance level of 0.1%. A lower p-value than 0.001 indicates that the method achieves the significant difference performance than the baseline model.

Dataset	LAG	REFUGE	RIM-ONE
SEDC/Baseline	1.17×10^{-48}	8.20×10^{-98}	1.04×10^{-5}
SEDC/EGDCL	2.78×10^{-102}	3.17×10^{-71}	7.90×10^{-6}
SEDC/Baseline+A	6.85×10^{-56}	1.07×10^{-68}	1.99×10^{-5}
SEDC/Baseline+B	2.39×10^{-7}	1.88×10^{-83}	1.98×10^{-5}
SEDC/Baseline+C	6.80×10^{-64}	8.76×10^{-56}	1.80×10^{-6}
SEDC/Baseline+AB	2.76×10^{-9}	3.31×10^{-76}	6.54×10^{-5}
SEDC/Baseline+AC	1.54×10^{-3}	7.50×10^{-83}	2.06×10^{-6}
SEDC/Baseline+BC	2.35×10^{-8}	1.65×10^{-43}	7.08×10^{-7}

icance versus previous EGDCL with a value small than 0.001 indicates the significant improvement of the proposed SEDC.

6. Conclusion

In this paper, we propose the SEDC framework to deal with the data imbalances in glaucoma diagnosis. The proposed SEDC designs an adaptive feature re-balancing strategy to move the decision boundary to the optimal direction by augmenting the feature distribution. In the proposed SEDC, dual-curriculum and self-ensemble learning are developed to augment the distorted feature distribution via feature re-weighting and feature distilling. Benefiting from the feature augmenting, the representation learning on imbalanced data is well done to promote the glaucoma diagnosis performance. Extensive experiments in terms of three datasets demonstrate the superiority of the proposed SEDC that improves the state-of-the-art glaucoma diagnosis results.

CRediT authorship contribution statement

Rongchang Zhao: Conceptualization, Methodology, Formal analysis, Writing - original draft. **Xuanlin Chen:** Soft-

ware, Formal analysis. **Zailiang Chen:** Conceptualization, Methodology - review & editing. **Shuo Li:** review & editing, Supervision.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61702558, 61972419), Hunan Science and Technology Project (2017WK2074, 2020SK2055), National Key R&D Program of China (2017YFC0840104, 2020YFC2008500), Natural Science Foundation of Hunan Province of China (2021JJ30879, 2020JJ4120) and Fundamental Research Funds for the Central Universities of Central South University (2020zzts610).

References

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning, in: ICML, ACM. pp. 41–48.

Byrd, J., Lipton, Z., 2019. What is the effect of importance weighting in deep learning?, in: International Conference on Machine Learning, pp. 872–881.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T., 2019. Learning imbalanced datasets with label-distribution-aware margin loss, in: Advances in Neural Information Processing Systems, pp. 1565–1576.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chen, X., Xu, Y., Wong, D.W.K., Wong, T.Y., Liu, J., 2015a. Glaucoma detection based on deep convolutional neural network, in: EMBC, IEEE. pp. 715–718.

Chen, X., Xu, Y., Yan, S., Wong, D.W.K., Wong, T.Y., Liu, J., 2015b. Automatic feature learning for glaucoma detection based on deep learning, in: MICCAI, Springer. pp. 669–677.

Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S., 2019. Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9268–9277.

Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4690–4699.

Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X., 2018a. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE TMI 37, 1597–1605.

Fu, H., Cheng, J., Xu, Y., Zhang, C., Wong, D.W.K., Liu, J., Cao, X., 2018b. Disc-aware ensemble network for glaucoma screening from fundus image. IEEE TMI 37, 2493–2501.

Fumero, F., Alayón, S., Sanchez, J.L., Sigut, J., Gonzalez-Hernandez, M., 2011. Rim-one: An open retinal image database for optic nerve evaluation, in: 2011 24th international symposium on computer-based medical systems (CBMS), IEEE. pp. 1–6.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W., 2018. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 .

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al., 2020. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 .

Haarburger, C., Baumgartner, M., Truhn, D., Broeckmann, M., Schneider, H., Schrading, S., Kuhl, C., Merhof, D., 2019. Multi scale curriculum cnn for context-aware breast mri malignancy classification .

Haleem, M.S., Han, L., Van Hemert, J., Li, B., 2013. Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: a review. CMIG 37, 581–596.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering 21, 1263–1284.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738.

Huo, J., Si, L., Ouyang, X., Xuan, K., Yao, W., Xue, Z., Zhang, L., Wang, Q., 2020. A self-ensembling framework for semi-supervised knee osteoarthritis localization and classification with dual-consistency. arXiv preprint arXiv:2005.09212 .

Jesson, A., Guizard, N., Ghalehjegh, S.H., Goblot, D., Soudan, F., Chapados, N., 2017. Cased: curriculum adaptive sampling for extreme data imbalance, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 639–646.

Jiang, L., Meng, D., Zhao, Q., Shan, S., Hauptmann, A., 2015. Self-paced curriculum learning, in: Proceedings of the AAAI Conference on Artificial Intelligence.

Jiménez-Sánchez, A., Mateus, D., Kirchhoff, S., Kirchhoff, C., Biberthaler, P., Navab, N., Ballester, M.A.G., Piella, G., 2019. Medical-based deep curriculum learning for improved fracture classification, in: MICCAI, Springer. pp. 694–702.

Jin, S., RoyChowdhury, A., Jiang, H., Singh, A., Prasad, A., Chakraborty, D., Learned-Miller, E., 2018. Unsupervised hard example mining from videos for improved object detection, in: ECCV, pp. 307–324.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y., 2019. Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217 .

Li, L., Xu, M., Wang, X., Jiang, L., Liu, H., 2019. Attention based glaucoma detection: A large-scale database and cnn model, in: CVPR, pp. 10571–10580.

Li, Y., Vasconcelos, N., 2019. Repair: Removing representation bias by dataset resampling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9572–9581.

Li, Z., He, Y., Keel, S., Meng, W., Chang, R.T., He, M., 2018. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology 125, 1199–1206.

Liao, W., Zou, B., Zhao, R., Chen, Y., He, Z., Zhou, M., 2019. Clinical interpretable deep learning model for glaucoma diagnosis. IEEE JBHI .

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: ICCV, pp. 2980–2988.

Lin, T.Y., RoyChowdhury, A., Maji, S., 2015. Bilinear cnn models for fine-grained visual recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Liu, H., Zhu, X., Lei, Z., Li, S.Z., 2019. Adaptiveface: Adaptive margin and sampling for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11947–11956.

Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A., 2020. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Transactions on Medical Imaging .

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L., 2017. Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 212–220.

Liu, W., Wen, Y., Yu, Z., Yang, M., 2016. Large-margin softmax loss for convolutional neural networks., in: ICML, p. 7.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9.

More, A., 2016. Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv preprint arXiv:1608.06048 .

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 .

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., 2020. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis 59, 101570.

Ren, M., Zeng, W., Yang, B., Urtasun, R., 2018. Learning to reweight examples for robust deep learning, in: ICML, pp. 4331–4340.

Sarafianos, N., Xu, X., Kakadiaris, I.A., 2018. Deep imbalanced attribute classification using visual attention aggregation, in: ECCV, pp. 680–697.

Schacknow, P.N., Samples, J.R., 2010. The glaucoma book: a practical, evidence-based approach to patient care. Springer Science & Business Media.

Shrivastava, A., Gupta, A., Girshick, R., 2016. Training region-based object detectors with online hard example mining, in: CVPR, pp. 761–769.

Smirnov, E., Melnikov, A., Oleinik, A., Ivanova, E., Kalinovskiy, I., Luckyanets, E., 2018. Hard example mining with auxiliary embeddings, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 37–46.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Advances in neural information processing systems, pp. 1195–1204.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W., 2018a. Cosface: Large margin cosine loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274.

Wang, X., Zhang, S., Lei, Z., Liu, S., Guo, X., Li, S.Z., 2018b. Ensemble soft-margin softmax loss for image classification. arXiv preprint arXiv:1805.03922 .

Wang, Y., Gan, W., Yang, J., Wu, W., Yan, J., 2019. Dynamic curriculum learning for imbalanced data classification, in: Proceedings of the IEEE international conference on computer vision, pp. 5017–5026.

Zhao, R., Chen, X., Chen, Z., Li, S., 2020. Egdcl: an adaptive curriculum learning framework for unbiased glaucoma diagnosis, in: Proceedings of European Conference on Computer Vision.

Zhao, R., Chen, X., Xiyao, L., Zailiang, C., Guo, F., Li, S., 2019a. Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning. IEEE JBHI .

Zhao, R., Chen, Z., Liu, X., Zou, B., Li, S., 2019b. Multi-index optic disc quantification via multitask ensemble learning, in: MICCAI, Springer. pp. 21–29.

Zhao, R., Li, S., 2020. Multi-indices quantification of optic nerve head in fundus image via multitask collaborative learning. Medical Image Analysis 60, 101593.

Zhao, R., Liao, W., Zou, B., Chen, Z., Li, S., 2019c. Weakly-supervised simultaneous evidence identification and segmentation for automated glaucoma diagnosis, in: AAAI, AAAI. pp. 809–816.

Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M., 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9719–9728.