Two Eyes Are Better Than One: Exploiting Binocular Correlation for Diabetic Retinopathy Severity Grading

Peisheng Qian[*1], Ziyuan Zhao[*1], Cong Chen[1,2], Zeng Zeng[†1], Xiaoli Li[1]

Abstract— Diabetic retinopathy (DR) is one of the most common eye conditions among diabetic patients. However, vision loss occurs primarily in the late stages of DR, and the symptoms of visual impairment, ranging from mild to severe, can vary greatly, adding to the burden of diagnosis and treatment in clinical practice. Deep learning methods based on retinal images have achieved remarkable success in automatic DR grading, but most of them neglect that the presence of diabetes usually affects both eyes, and ophthalmologists usually compare both eyes concurrently for DR diagnosis, leaving correlations between left and right eyes unexploited. In this study, simulating the diagnostic process, we propose a two-stream binocular network to capture the subtle correlations between left and right eyes, in which, paired images of eyes are fed into two identical subnetworks separately during training. We design a contrastive grading loss to learn binocular correlation for five-class DR detection, which maximizes inter-class dissimilarity while minimizing the intra-class difference. Experimental results on the EyePACS dataset show the superiority of the proposed binocular model, outperforming monocular methods by a large margin.

Clinical relevance— Compared to conventional DR grading methods based on monocular images, our approach can provide more accurate predictions and extract graphical patterns from retinal images of both eyes for clinical reference.

I. INTRODUCTION

Diabetic retinopathy (DR) is one of the most prevailing eye diseases among patients with diabetes. It has become the primary cause of blindness in the working-age population of the developed world [1]. In Singapore, over 40% of diabetic patients suffer from DR at various stages from mild to severe [2]. Prevention of DR is challenging because the symptoms of DR are hardly recognizable at the early stage. The gold standard for diagnosis of DR is digital color fundus photography. Digital color fundus photography is the gold standard for diagnosing DR. However, observing and evaluating fundus images is time-consuming and labor-intensive, requiring experienced ophthalmologists.

Deep learning approaches have achieved great success in DR grading based on retinal images [3]. Different from conventional machine learning methods, which rely on hand-crafted features, *e.g.*, the retinal blood vessels and the optic disc, deep neural networks can effectively extract high-level features and learn complex representations from retinal images, which better facilitates the clinical process and
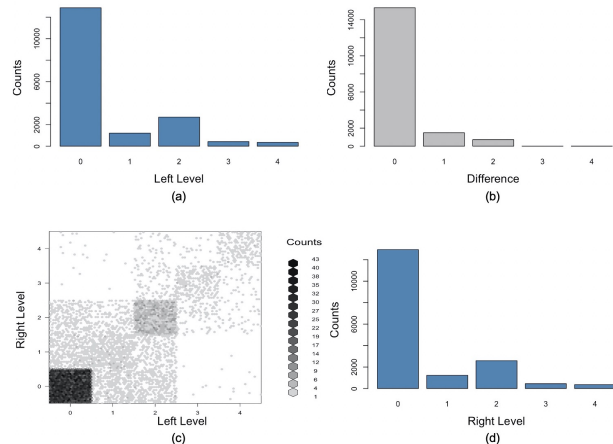

Fig. 1. (a) The distribution of C_l (disease severity levels of left eyes). (b) The distribution of $|C_l - C_r|$ (difference between disease severity levels of each patient's left and right eyes). (c) The scatter plot of disease severity levels between each patient's left and right eyes. (d) The distribution of C_r (disease severity levels of right eyes).

eliminates human errors. Most of the existing methods take monocular images as the model inputs, regardless of the difference between left and right eyes [2], [4], [5]. However, in the clinical diagnosis of eye diseases, both the left and right eyes are taken into consideration [6], [7]. In other words, the correlation between left and right eyes can be used for DR grading in clinical practice. As shown in Fig. 1, we performed exploratory data analysis on the Kaggle dataset provided by EyePACS [8]. For better visualization, we add random variance to each level of two eyes concurrently in Fig. 1 (c). We can see that both eyes of the same patient are highly correlated and the Pearson's correlation coefficient ρ of them is 0.85 [8]. Motivated by the clinical process and our analysis, we hypothesize that the left-right correlation can be used in deep learning for better DR grading.

In this work, we propose a two-stream binocular network. The network consists of 2 convolutional neural networks (CNNs) that share the same weights and take the left and right eyes of the same patient as inputs, respectively. The model is simultaneously updated by both left and right eye images. With this learning framework, the model can recognize individual patterns in each eye as well as the similarities between the left and right eyes for DR grading. We propose a hybrid loss to optimize the network. The loss consists of a contrastive grading loss and a weighted cross-entropy loss. More specifically, we introduce the contrastive grading loss to optimize the network in accordance with the

left-right similarity, in which we scale the contrastive loss proportional to the discrepancies in DR grading for finer classification granularity. The contribution of this paper are summarized as follows:

- We construct a two-stream binocular network that consists of two identical networks with shared weights. We design a novel training strategy that takes both left and right eyes of the same patient as inputs.
- We propose a hybrid loss function which consists of the contrastive grading loss and the weighted cross-entropy loss. The contrastive grading loss optimizes the network based on the similarity between the left and right eyes. Experiments show that with the proposed loss function, the network can recognize similarities among left and right eyes, and obtain superior results than baselines.

II. RELATED WORK

Early studies relied heavily on experts manually extracting features and certain textural properties for DR classification [9]. In recent years, deep learning techniques, such as CNNs, have been proved effective in DR grading. Bravo *et al.* designed a VGG-based network architecture and combine it with various pre-processing images [10], reaching 50.5% classification accuracy on a balanced dataset. Zhao *et al.* described a bilinear model with an attention mechanism for fine-grained classification of DR [5]. Wang *et al.* implemented Zoom-in-Net with multiple sub-networks for DR grading and localization of suspicious regions [4]. Zhao *et al.* further investigated the subtle differences between different DR severity levels and developed a new network architecture, SEA-Net, in which spatial and channel attention are alternatively stacked [2]. The aforementioned methods enhance model architectures and prove their effectiveness in DR grading. However, they do not leverage the left-right eye correlations.

Existing research suggests that the correlation between left and right eyes could be explored on eye symptoms [6], [7], [11]. We are not the first one to address the correlation between the two eyes. Zeng *et al.* presented promising results with binocular inputs to siamese-like deep learning models for DR classification [2]. While this method introduces weights sharing between CNNs in their model, it omits the calculation of similarities or variances among different DR grades. There is also no modification to the original contrastive loss. To overcome the above-mentioned shortcomings and clearly reflect the differences between DR grades in the loss function, we propose a two-stream binocular network and a novel contrastive grading loss, which are illustrated in Section III.

III. METHODOLOGY

The proposed two-stream binocular network and training strategy for DR grading is illustrated in Fig. 2, in which, a pair of left and right eye images X_l, X_r are taken as inputs to two identical sub-networks respectively. The two sub-networks with shared weights extract features from the two eyes and classify their DR grading separately. To leverage
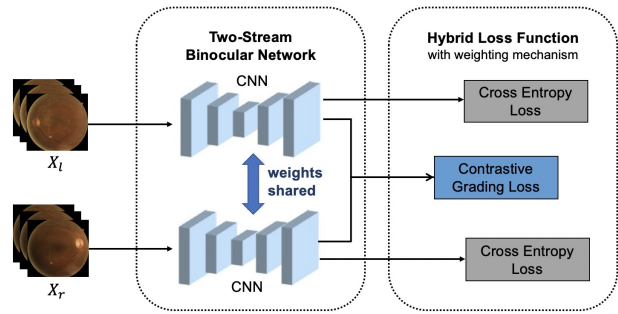

Fig. 2.   The architecture of the two-stream binocular network.

the left-right eye correlations, we apply a novel loss function during training, which includes a contrastive grading loss and a weighted cross-entropy loss.

A. Two-stream Binocular Network

The proposed framework consists of two convolutional neural networks (CNNs) with shared weights. In this study, we use ResNet-50 [12] and BiRA-Net [5] as the backbones of the network to study the effectiveness of the proposed architecture. Both eyes from the same patient are paired and passed to each of the CNNs. The network captures features from both left and right eyes as well as the similarities between them, based on which the network classifies DR grading for both eyes. The process simulates the real-life clinical DR diagnosis on both eyes and utilizes the correlations between them for diagnosis.

B. The Proposed Hybrid Loss

1) Contrastive grading loss: To optimize the network and extract similarities between left and right eyes, we propose the contrastive grading loss. The loss function is improved from the contrastive loss, which is commonly used in conjunction with siamese networks [13]. The loss calculates the similarity between eyes based on Euclidean distances between hidden features in the last layers of the 2 CNNs. Assuming that the disease severity levels of paired images X_l and X_r are C_l and C_r respectively, the grading difference between the paired images can be represented as $|C_l - C_r|$. The contrastive grading loss function is defined as:

$$\mathcal{L}_{cg} = F_c d^2 + |C_l - C_r| max(margin - d, 0)^2 \quad (1)$$

where F_c is the binary control factor for the first term in the loss function. If the disease severity level in both images is the same, $F_c = 1$. Otherwise, $F_c = 0$. d represents the Euclidean distance between outputs from the two sub-networks in the two-stream binocular network. $margin$ is the threshold, which is adjusted empirically and set to 2 in this study [13]. Compared to the existing contrastive loss, Eqn. 1 scales the second term in the loss function proportional to the differences between left and right features and therefore optimizes the network to recognize similarities between left and right eyes in the feature space.

2) Weighted Cross-Entropy Loss: To alleviate the overfitting problem due to the imbalance of the DR dataset, we add the weighting mechanism in our hybrid loss function. $x[0], x[1], x[2], \ldots, x[C-1]$ denotes the class probability of input x, and the class index y is in the range $[0, C-1]$, where C is the number of classes. Each sample is scaled by the weight proportional to the inverse of the percentage of the class of the sample in the training set, denoted as $weight[y]$ in Eqn. 2. The weighted cross-entropy loss [14] is formulated as:

$$\mathcal{L}_{ce} = weight[y] \left(-x[y] + \log \left(\sum_{i=0}^{C-1} \exp(x[i]) \right) \right) \quad (2)$$

Finally, the proposed hybrid loss function is a weighted sum of the contrastive grading loss and the cross-entropy loss, which is defined in Eqn. 3:

$$\mathcal{L} = \lambda \mathcal{L}_{cg} + \mathcal{L}_{ce} \quad (3)$$

where λ is the factor controlling the scale of the contrastive grading loss in the hybrid loss. In the experiments, λ is empirically set to 0.1 for model optimization.

IV. EXPERIMENTS

A. Dataset and Implementation

We collate the dataset provided by EyePACs, hosted on Kaggle [8]. The dataset is labeled with a set of definitions to ensure label consistency. The grades of DR are categorized into 5 classes from 0 to 4 with increasing disease severity. Grade 0 indicates non-diabetic, while grade 4 indicates the most severe diabetes. We randomly split the retinal images from the dataset into $33,566$ images as the training set and $1,560$ images as the test set. The test set is balanced.

The implementation details are described as follows. Left and right retina images of the same patients are selected in pairs as the input. To augment the training set, random horizontal and vertical flipping, and random rotation of ±10 degrees are applied to the input images. The images are then resized to 224×224. Finally, the images are standardized across the RGB channels by subtracting the mean and dividing by the standard deviation of each channel.

We load the ImageNet pre-trained weights into the network before starting the training process [12]. The network is trained using the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.001 and a weight decay factor of $1e-8$. The learning rate is multiplied by 0.1 when the model performance on the test set does not improve for 10 consecutive epochs. The network is trained for 100 epochs with a batch size of 32. The experiments are implemented on NVIDIA RTX 2080Ti GPUs with Pytorch 1.7.1.

B. Performance Metrics

For a comprehensive evaluation, we employ 3 commonly-used metrics to quantitatively evaluate the performance, which have also been used in previous research [2], [5]. They are:

- ACA: Average classification accuracy.
- F1: Averaged Macro-F1 score of the 5 classes.
- AUC: The area under the receiver operating characteristics (ROC) curve.

C. Baseline Methods

We compare our framework with several baselines. In [10], a VGG-based classifier is trained on the dataset with preprocessing techniques including circular RGB, grayscale and color-centered sets. Zhao *et al.* report results of ResNet-50 on the Kaggle dataset, and we re-implement ResNet-50 with slightly higher ACA [5]. We combine ResNet-50 with mean squared error (MSE) in the loss function as another baseline. In [5], BiRA-Net is invented, which features a bilinear learning strategy together with a grading loss for fine-grained classification. Our methods with different backbones are shown as follows,

- TSBN (ResNet-50): two-stream binocular network, with ResNet-50 as the backbone.
- TSBN (BiRA-Net): two-stream binocular network, with BiRA-Net as the backbone. The network consists of 2 BiRA-Net models that share weights.

D. Results and Discussion

TABLE I
EXPERIMENTAL RESULTS ON DR GRADING.

Method	ACA	Macro-F1	AUC
Bravo *et al.* [10]	0.5051	0.5081	-
ResNet-50 [12]	0.4820	0.4877	0.8091
ResNet-50, MSE	0.4985	0.4995	0.8144
TSBN (ResNet-50)	**0.5212**	**0.5242**	**0.8218**
BiRA-Net [5]	0.5431	0.5723	0.8448
TSBN (BiRA-Net)	**0.5513**	**0.5792**	**0.8490**

TABLE II
GRADING ACCURACIES IN EACH DR LEVEL.

Class	ResNet-50	**TSBN (ResNet-50)**
Normal (0)	0.6250	**0.6442**
Mild (1)	0.3814	**0.4327**
Moderate (2)	0.4327	**0.4423**
Severe (3)	0.4103	**0.4615**
Proliferative (4)	0.5609	**0.6250**

In Table I, the experimental results of our approach are compared with baseline methods. When using ResNet-50 as the backbone, our method has a clear advantage of 4% increase in classification accuracy. It proves that our training strategy and loss function can better optimize the model. We also outperform the original BiRA-Net, a more sophisticated architecture engineered by a dedicated grading loss [5]. The results confirm that our approach is widely applicable to different backbone models for DR grading.

In Table. II, it is clear that our approach reaches higher classification accuracy in all levels of DR grading, especially in higher levels where the training samples are sparse. Fig. 4 represents confusion matrices for ResNet-50 and our method respectively. Our model distinguishes more cases among

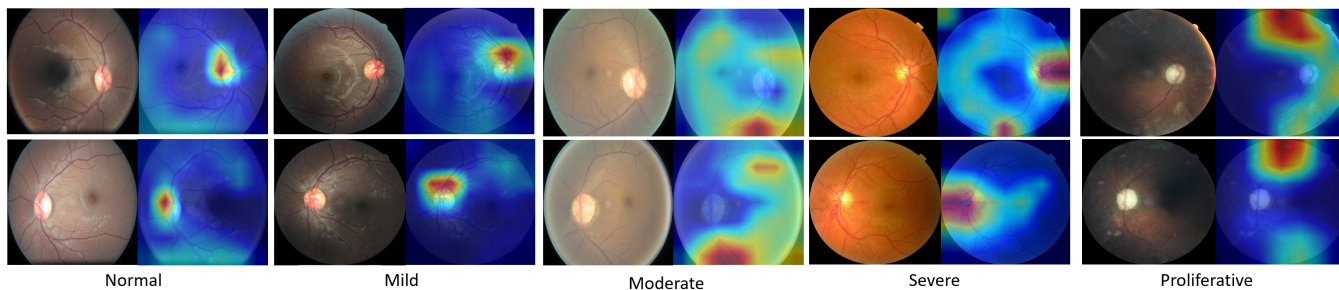| Normal | Mild | Moderate | Severe | Proliferative |

Fig. 3. Illustrative examples of original images and saliency maps from TSBN (ResNet-50). Each column contains left (top) and right (bottom) retina images from the same patient who has the same DR level in both eyes.
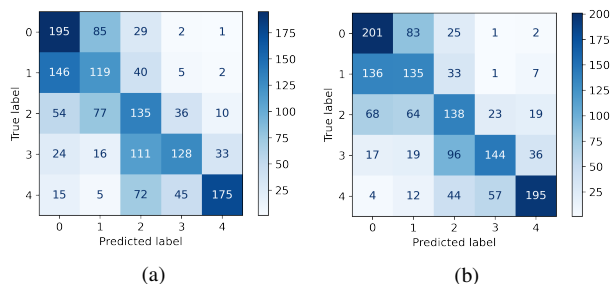

(a)   (b)

Fig. 4. Confusion matrices of (a) ResNet-50, (b) TSBN (ResNet-50).

moderate to serious DR grades (grade 2 to 4). In other words, our model identifies patients who require clinical attention most, proving its clinical significance.

Examples of saliency maps on the test data are illustrated in Fig. 3 with Grad-CAM [15]. Comparing the left and right saliency maps of the same patient, it is evident that our model is activated at symmetrical positions with similar intensities. It confirms that the similarities of DR symptoms in both eyes are informative for DR grading. It is also validated that our approach effectively learns such similarities for classification. Besides, it is observed that there are some discrepancies between the saliency maps of left and right eyes, due to variances of the DR symptoms and limitations in the model capability.

V. CONCLUSIONS AND FUTURE WORK

We propose a two-stream binocular network, which explores similarities and correlations between left and right eyes for DR grading. The framework consists of 2 CNNs with shared weights and classifies the left and right eyes of the same patient respectively. To capture the similarities between left and right eyes, a hybrid loss function is proposed, which combines a contrastive grading loss and a weighted cross-entropy loss. Extensive experiments has shown that our approach is effective. The left-right eye similarities are visualized in the saliency maps of our model. In the future, we can further improve the model performance by exploring left-right eye correlations with domain knowledge.

REFERENCES

[1] Sneha Das and C Malathy, "Survey on diagnosis of diseases from retinal images," *Journal of Physics: Conference Series*, vol. 1000, pp. 012053, apr 2018.

[2] Ziyuan Zhao, Kartik Chopra, Zeng Zeng, and Xiaoli Li, "Sea-net: Squeeze-and-excitation attention net for diabetic retinopathy grading," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2496–2500.

[3] Tao Li, Wang Bo, Chunyu Hu, Hong Kang, Hanruo Liu, Kai Wang, and Huazhu Fu, "Applications of deep learning in fundus images: A review," *Medical Image Analysis*, vol. 69, pp. 101971, 2021.

[4] Zhe Wang, Yanxin Yin, Jianping Shi, Wei Fang, Hongsheng Li, and Xiaogang Wang, "Zoom-in-net: Deep mining lesions for diabetic retinopathy detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 267–275.

[5] Ziyuan Zhao, Kerui Zhang, Xuejie Hao, Jing Tian, Matthew Chin Heng Chua, Li Chen, and Xin Xu, "Bira-net: Bilinear attention net for diabetic retinopathy grading," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1385–1389.

[6] John HK Liu, Arthur J Sit, and Robert N Weinreb, "Variation of 24-hour intraocular pressure in healthy individuals: right eye versus left eye," *Ophthalmology*, vol. 112, no. 10, pp. 1670–1675, 2005.

[7] T. Eppig, C. Spira-Eppig, S. Goebels, B. Seitz, M. El-Husseiny, M. Lenhart, K. Papavasileiou, N. SzentmÃ¡ry, and A. Langenbucher, "Asymmetry between left and right eyes in keratoconus patients increases with the severity of the worse eye," *Current Eye Research*, vol. 43, no. 7, pp. 848–855, 2018, PMID: 29558197.

[8] Ben Graham, "Kaggle diabetic retinopathy detection competition report," *University of Warwick*, 2015.

[9] Arslan Ahmad, Atif Bin Mansoor, Rafia Mumtaz, Mukaram Khan, and SH Mirza, "Image processing and classification in diabetic retinopathy: A review," in *2014 5th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2014, pp. 1–6.

[10] María A. Bravo and Pablo A. Arbeláez, "Automatic diabetic retinopathy classification," in *13th International Symposium on Medical Information Processing and Analysis*, 2017.

[11] Xianglong Zeng, Haiquan Chen, Yuan Luo, and Wenbin Ye, "Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network," *IEEE Access*, vol. 7, pp. 30744–30753, 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] Matthew D Li, Ken Chang, Ben Bearce, Connie Y Chang, Ambrose J Huang, J Peter Campbell, James M Brown, Praveer Singh, Katharina V Hoebel, Deniz Erdoğmuş, et al., "Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–9, 2020.

[14] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European conference on computer vision*. Springer, 2016, pp. 354–370.

[15] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.