Lesion-Aware Contrastive Learning for Diabetic Retinopathy Diagnosis

Shuai Cheng[1,2], Qingshan Hou[1,2], Peng Cao[1,2,3(✉)], Jinzhu Yang[1,2,3(✉)], Xiaoli Liu[4], and Osmar R. Zaiane[5]

[1] Computer Science and Engineering, Northeastern University, Shenyang, China
[2] Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang, China
caopeng@mail.neu.edu.cn, yangjinzhu@cse.neu.edu.cn
[3] National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Shenyang 110819, China
[4] DAMO Academy, Alibaba Group, Hangzhou, China
[5] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Canada

Abstract. Early diagnosis and screening of diabetic retinopathy are critical in reducing the risk of vision loss in patients. However, in a real clinical situation, manual annotation of lesion regions in fundus images is time-consuming. Contrastive learning(CL) has recently shown its strong ability for self-supervised representation learning due to its ability of learning the invariant representation without any extra labelled data. In this study, we aim to investigate how CL can be applied to extract lesion features in medical images. However, can the direct introduction of CL into the deep learning framework enhance the representation ability of lesion characteristics? We show that the answer is no. Due to the lesion-specific regions being insignificant in medical images, directly introducing CL would inevitably lead to the effects of false negatives, limiting the ability of the discriminative representation learning. Essentially, two key issues should be considered: (1) How to construct positives and negatives to avoid the problem of false negatives? (2) How to exploit the hard negatives for promoting the representation quality of lesions? In this work, we present a lesion-aware CL framework for DR grading. Specifically, we design a new generating positives and negatives strategy to overcome the false negatives problem in fundus images. Furthermore, a dynamic hard negatives mining method based on knowledge distillation is proposed in order to improve the quality of the learned embeddings. Extensive experimental results show that our method significantly advances state-of-the-art DR grading methods to a considerable 88.0%ACC/86.8% Kappa on the EyePACS benchmark dataset. Our code is available at https://github.com/IntelliDAL/Image.

S. Cheng and Q. Ho—Contribute equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_63.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 671–681, 2023.
https://doi.org/10.1007/978-3-031-43990-2_63

Keywords: Diabetic Retinopathy · Contrastive Learning · Hard Negative Mining · Knowledge Distillation

1 Introduction

Diabetic retinopathy(DR) is a common long-term complication of diabetes that can lead to impaired vision and even blindness as the disease worsen [13,14]. Hence, conducting large-scale screening for early DR is an essential step to prevent visual impairment in patients. Screening fundus images by the ophthalmologist alone is not sufficient to prevent DR on a large scale, and the diagnosis of DR heavily relies on the experience of the ophthalmologist [1]. Therefore, the automatic DR diagnosis on retinal fundus images is urgently needed [3,25]. Recently, in light of the powerful feature extraction and representation capabilities of convolutional neural networks, deep learning technology has developed rapidly in medical image analysis [5,22]. However, leveraging only the image-level grading annotation hinders deep learning algorithms from extracting features of suspicious lesion regions, which further affects the diagnosis of diseases. For these reasons, some previous work [17,19] considers the introduction of pixel-level lesion annotation to improve the model's feature extraction capability for lesion regions. Despite the methods have achieved promising results, the large-scale pixel-level annotation process is time-intensive and error-prone which imposes a heavy burden on the ophthalmologist. To address this problem, contrastive learning(CL) [10,11,23] has received a great deal of attention in medical images, but how to harness the power of CL in the medical applications remains unclear.

The challenges mainly lie in: (I) The diagnosis of fundus diseases relies more on local pathological features (haemorrhages, microaneurysms, etc.) than on the global information. How can contrastive learning enable models to extract features of lesion information more effectively on the large datasets with only image-level annotation? (II) The false negatives tend to disrupt the feature extraction of contrastive learning [26], resulting in the issue of inaccurate alignment of feature distributions [18] (i.e. similar samples have dissimilar features). How to address the issue of false negatives caused by introducing contrastive learning into automatic disease diagnosis? (III) The performance of contrastive learning benefits from the hard negatives [2,16]. How to effectively exploit hard negatives for improving the quality of the learned feature embeddings?

To address the aforementioned issues, we propose the lesion-aware CL framework for DR grading. Specifically, to eliminate false negatives during contrastive learning introduced in automatic disease diagnosis and ensure that samples having similar semantic information stay close in the joint embedding space, we first capture lesion regions in fundus images using a pre-trained lesion detector. Based on the detected regions, we construct a lesion patch set and a healthy patch set, respectively. Then, we develop an encoder and a momentum encoder [6] for extracting the features of positives (lesion patches) and negatives (healthy patches). The introduced momentum encoder enables the contrastive learning to maintain consistency in critical features while creating different perspectives for

the positive samples. Secondly, considering the critical role of hard negatives in the contrastive learning, we formulate a two-stage scheme based on knowledge distillation [8,21] to dynamically exploit hard negatives, which further enhances the lesion-aware capability of the diagnosis models, and further improves the quality of the learned feature embeddings. Finally, we fine-tune the proposed framework in the DR grading task to demonstrate its effectiveness.

To the best of our knowledge, this is the first work to rethink the potential issues of contrastive learning for medical image analysis. In summary, our contributions can be summarized as follows. (1) A new scheme of constructing positives and negatives is proposed to prevent false negatives from disrupting the extraction of lesion features. This design can be easily extended to other types of medical images with less prominent physiological features to achieve better lesion representation. (2) To enhance the capability of CL in extracting lesion features for medical fundus image analysis and improve the quality of learned feature embeddings, a lesion-aware CL framework is proposed for sufficiently exploiting hard negatives. (3) We evaluate our framework on the large-scale EyePACS dataset for DR grading. The experimental results indicate the proposed method leads to a performance boost over the state-of-the-art DR grading methods.

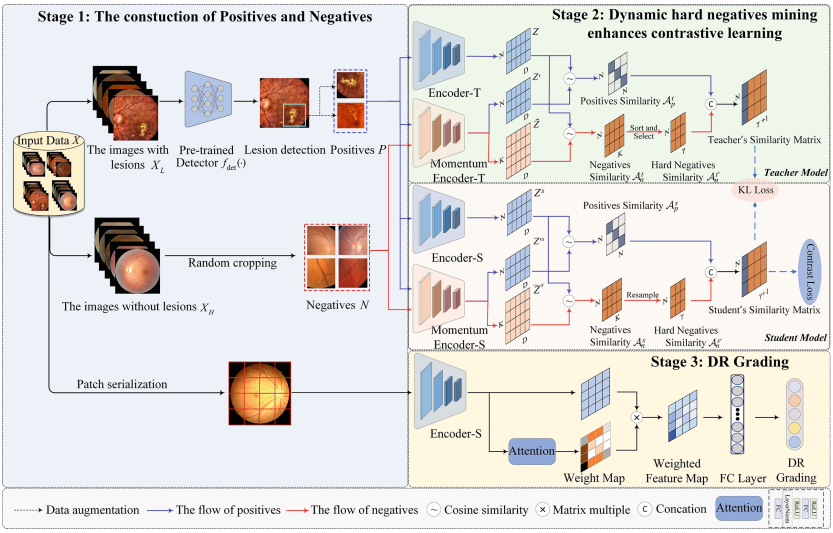

Fig. 1. The overall architecture of the proposed framework. **Stage 1:** The construction of positives and negatives based on the pre-trained lesion detector. **Stage 2:** Dynamic hard negatives mining enhances contrastive learning. **Stage 3:** Fine-tuning our model on the downstream diabetic retinopathy grading task.

2 Methodology

Fig. 1 shows the illustration of the proposed framework. In stage 1, we construct positives and negatives based on a pre-trained lesion detector pre-trained on a auxiliary dataset (IDRiD [15]) with pixel lesion annotation, to avoid the effect of false negatives on the learned feature embeddings while aligning samples with similar semantic features. In stage 2, a dynamically sampling method is developed based on knowledge distillation to effectively exploit hard negatives and improve the quality of the learned feature embeddings. In the last stage, we finetune our model on the downstream DR grading task. Remarkably, to bridge the gap between local patches in the pretext task and global images in the downstream task, we introduce an attention mechanism on the fragmented patches to highlight the contributions of different patches on the grading results.

2.1 Construction of Positives and Negatives

In this section, we provide a detailed description regarding the construction of positives and negatives. As opposed to traditional CL working on the whole medical images, it is essential to enable the model to focus more on the lesion regions in the images. Our goal is to eliminate the effect of false negatives on contrastive learning for obtaining a better representation of the lesion features. Specifically, given a training dataset X with five labels (1-4 indicating the increasing severity of DR, 0 indicating healthy). We first divide dataset X into lesion subset X_L and healthy subset X_H based on the disease grade labels of X. Then, we apply a pre-trained detector $f_{\text{det}}(\cdot)$ only on X_L and obtain high-confidence detection regions. Finally, the construction process of positives $P = \{p_1, p_2, \ldots, p_j\}$ and negatives $N = \{n_1, n_2, \ldots, n_k\}$ can be represented as $P = \Omega(f_{\text{det}}(X_L) > conf)$ and $N = Randcrop(X_H)$, where $conf$ denotes the confidence threshold of detection results, $\Omega(\cdot)$ indicates the operation of expanding the predicted boxes of $f_{\text{det}}(\cdot)$ to 128*128 for guaranteeing that the lesions are included as much as possible, and $Randcrop(\cdot)$ indicates randomly cropping images into patches with 128*128 from the healthy images.

2.2 Dynamic Hard Negatives Mining Enhances Contrastive Learning

Given the constructed positives P and negatives N, a negatives sampling scheme based on offline knowledge distillation is developed to enable contrastive learning to dynamically exploit hard negatives, and we adjust the update mechanism of the negatives queue(i.e. only enqueue and dequeue N to avoid confusion with P) to better adapt contrastive learning to the medical image analysis task.

Training the Teacher Network. With the positives P, we obtain two views $\tilde{P} = \{\tilde{p}_1, \tilde{p}_2, \tilde{p}_3 ... \tilde{p}_j\}$ and $\tilde{P}' = \{\tilde{p}'_1, \tilde{p}'_2, \tilde{p}'_3 ... \tilde{p}'_j\}$ by data augmentation(i.e. color

distortion rotation, cropping followed by resize). Correspondingly, with the negatives N, to increase the diversity of the negatives, we apply a similar data augmentation strategy to obtain the augmented negatives $\tilde{N} = \{\tilde{n}_1, \tilde{n}_2, \tilde{n}_3 ... \tilde{n}_k\}$ (where $k \gg j$). We feed $\tilde{P}$ and $\tilde{P}' + \tilde{N}$ to the encoder $En(\cdot)$ and the momentum encoder $MoEn(\cdot)$ to obtain their embeddings $Z = \{z_1, ..., z_j | z_j = En(\tilde{p}_j)\}$, $Z' = \{z'_1, ..., z'_j | z'_j = MoEn(\tilde{p}'_j)\}$ and $\tilde{Z} = \{\tilde{z}_1, ..., \tilde{z}_k | \tilde{z}_k = MoEn(\tilde{n}_k)\}$. Then, we calculate the positive and negative similarity matrix by the samples of Z, Z' and $\tilde{Z}$. According to the similarity matrix, the contrastive loss $L_{\text{cl-t}}$ of the teacher model training process can be defined as:

$$\begin{aligned} L_{\text{cl-t}} &= -\sum \log \left(\frac{\exp\left(\mathcal{A}_p^t/\tau\right)}{\exp(\mathcal{A}_p^t/\tau) + \sum \exp\left(\mathcal{A}_n^t/\tau\right)} \right) \\ &= -\sum_j \log \left(\frac{\exp\left(\text{sim}\left(z_j, z'_j\right)/\tau\right)}{\exp(\text{sim}(z_j, z'_j)/\tau) + \sum_k \exp\left(\text{sim}\left(z_j, \tilde{z}_k\right)/\tau\right)} \right), \end{aligned} \tag{1}$$

where $\text{sim}\left(z_j, z'_j/\tilde{z}_k\right) = \frac{\text{dot}\left(z_j, z'_j/\tilde{z}_k\right)}{\|z_j\|_2 \|z'_j/\tilde{z}_k\|_2}$, τ denotes a temperature parameter, $\mathcal{A}_p^t$ and $\mathcal{A}_n^t$ represent the similarity matrix of positives and negatives, respectively. In order to create a positive sample view different from that of $En(\cdot)$, it should be noted that the parameters θ_{q} of $En(\cdot)$ are updated using gradient descent, while $MoEn(\cdot)$ introduces an extra momentum coefficient $m = 0.99$ to update its parameters $\theta_{\text{k}} \leftarrow m\theta_{\text{k}} + (1 - m)\theta_{\text{q}}$.

Training the Student Network. Previous works [2,16] reveal that not all negatives are useful for the contrastive learning. Moreover, the hard negatives may exhibit more semantically similar to the positives than the normal negatives, indicating that hard negatives provide more potentially useful information for facilitating the following DR grading. Meanwhile, the number of hard negatives significantly affects the difficulty of training the model, in other words, the network should be capable of dynamically adjust the optimisation process by controlling the number of hard negatives. In light of the above two points, we formulate and introduce a well-balanced strategy of hard negatives during the training phase of the student model. Specifically, based on the trained teacher model, we first input P and N into both the teacher and student models to generate similarity matrices A_p^t, A_n^t and A_p^s, A_n^s, respectively. According to the negative similarity matrix A_n^t produced by the teacher model, we prioritise the negatives that are likely to be confused with the positives in descending order and only select the top δ samples for distillation learning during the student model's training phase. For each negative $\tilde{z}_k$ in A_n^t, the resampled negative set $\mathcal{A}_n^{t'}$ can be defined as:

$$\mathcal{A}_n^{t'} = \left\{ \tilde{z}_k \mid \tilde{z}_k \in Sort(\mathcal{A}_n^t), \text{sim}\left(z_j, \tilde{z}_k\right) \geq \text{sim}(z_j, \tilde{z}_\gamma) \right\}, \tag{2}$$

where $\gamma = \delta/(cos(\frac{\pi s}{2S}) + 1)$ represents the number of the current hard negatives, s and S denotes the current and maximum training step, respectively. As s

increases during the training process, we dynamically adjust the number of hard negatives such that the difficulty of distillation learning proceeds from easy to hard. Based on the index in $A_n^{t'}$, the elements at the corresponding positions in A_n^s are obtained and a resampled negatives similarity matrix $A_n^{s'}$ is constructed. Hence, the CL loss $L_{\text{cl-s}}$ in training process of student can be formulated as:

$$L_{\text{cl-s}} = -\sum \log\left(\frac{\exp\left(\mathcal{A}_p^s/\tau\right)}{\exp(\mathcal{A}_p^s/\tau) + \sum_k \exp\left(\mathcal{A}_n^{s'}/\tau\right)}\right), \tag{3}$$

In addition, to improve the quality of embeddings learned by the student model, we leverage the generated similarity matrices to facilitate the richer knowledge distilled from the teacher to the student. Formally, the KL-divergence loss L_{kd} between $\mathcal{A}_p^t$, $\mathcal{A}_n^{t'}$ and $\mathcal{A}_p^s$, $\mathcal{A}_n^{s'}$ is represented as follows.

$$L_{\text{kd}} = -\tau^2 \sum C(\mathcal{A}_p^t, \mathcal{A}_n^{t'}) \log\left(C(\mathcal{A}_p^s, \mathcal{A}_n^{s'})\right), \tag{4}$$

where $C(\cdot)$ denotes the matrix concatenation. The final loss of the student model is $L = L_{\text{cl-s}} + \lambda_1 L_{\text{kd}}$, where the λ_1 is a positive parameter controlling the weight of the knowledge distillation loss L_{kd}.

2.3 DR Grading Task

To evaluate the effectiveness of the proposed method, we take the encoder of the pre-trained student model as a backbone and fine-tune it for the downstream DR grading task. Considering that the proposed contrastive learning framework is trained with patches, whereas the downstream grading task relies on entire fundus images, an additional attention mechanism is incorporated to break the gap between the inputs of pretext and downstream tasks. Specifically, we first fragment the entire fundus image into patches $x = \{x_{p1}, \ldots, x_{p_i}\}$. Then, feature embedding v_i of x_{p_i} is generated by the encoder. Meanwhile, an attention module with two linear layers is utilized in the DR grading task to obtain the attention weight α_i of each patch x_{p_i}.

$$\alpha_i = softmax\left(W_2^T max\left(LayerNorm\left(W_1 v_i^T\right), 0\right)\right), \tag{5}$$

where W_1, W_2 are the parameters of the two linear layers , $LayerNorm$ is the layernorm function. Finally, α_i is assigned to the corresponding patch's embedding v_i to highlight the contribution of patch x_{pi}, and the predicted results of DR obtained by $\hat{y} = W_3^T \cdot \sum_{i=1}^N \alpha_i v_i$, and W_3^T is parameter of the grading layer.

3 Experiments

3.1 Datasets and Implementation Details

EyePACS [4]**.** EyePACS is the largest public fundus dataset which contains 35,126 training images and 53,576 testing images with only image-level DR grading labels. According to the severity of DR, images are classified into five grades: 0 (normal), 1 (mild), 2 (moderate), 3 (severe), and 4 (prolifera-tive).

Implementation Details. The proposed framework is implemented by Pytorch on two Tesla T4 Tensor core GPUs. We employ the IDRiD dataset [15] for the pre-training of the lesion detector $f_{\text{det}}(\cdot)$. During the sample construction stage, considering the diversity of sizes of the original fundus images, all images are resized to 768×768, and the data enhancement strategies include random rotation, flipping and color distortion. During the phase of dynamically mining hard negatives, the Adam optimizer with momentum 0.9 is applied to train and update the parameters of the framework with 800 epochs, the initial learning rate of 1×10^{-3} and the batch size of 400. The designed hyper-parameter δ is set to be 4,000 after extra experiments (*please refer to the supplementary materials for more details*). In the downstream DR grading task, we fine-tune the encoder(i.e. ResNet50) for 25 epochs with an initial learning rate of 1×10^{-4} and a batch size of 32. In addition to the normal classification accuracy, we also introduce the quadratic weighted kappa metric to reflect the performance of the proposed method and a range of comparable methods.

3.2 Comparison with the State-of-the-Art

In this section, we provide qualitative and quantitative comparisons with various DR grading methods and demonstrate the effectiveness of the proposed method. As shown in Table 1, we conduct a comprehensive comparison of the proposed method with three types of comparable methods: covering the popular backbone network [7], the top two places of Kaggle challenge [4] and the current SOTA DR grading methods [9,10,12,19,20,24].

Table 1. The comparison between our method and the SOTA methods in DR grading task on EyePACS dataset

Model	Kappa	Accuracy	Model	Kappa	Accuracy
Resnet50 [7]	0.823	0.845	AFN(2019) [12]	0.859	–
Min-pooling [4]	0.849	–	DeepMT-DR(2021) [19]	0.839	0.857
o_O [4]	0.844	–	CL-DR(2021) [10]	0.832	0.848
Zoom-in-Net(2017) [20]	0.854	0.873	CLEAQ-DR(2022) [9]	0.863	–
MMCNN(2018) [24]	0.841	0.862	**Lesion-aware CL (Ours)**	**0.868**	**0.880**

From the Table 1, it can be observed that our method consistently achieves the best results with respect to both the Kappa and Accuracy. The results show that our framework presents a notably better DR grading performance than the SOTA methods due to improve quality of the learned lesion embeddings by eliminating the false negatives and dynamically mining hard negatives, and in turn enhancing the lesion-awareness of CL, which is beneficial for DR grading.

Table 2. The ablation experiment results of the proposed framework on EyePACS

Method	Kappa	Accuracy
CL	0.844	0.858
Lesion-aware CL *w/o* CPN	0.853(+**0.9**%)	0.864(+**0.6**%)
Lesion-aware CL *w/o* DHM	0.857 (+**1.3**%)	0.871(+**1.3**%)
Lesion-aware CL	0.868(+**2.4**%)	0.880(+**2.2**%)

3.3 Ablation Study

To more comprehensively evaluate the Lesion-aware CL, we conduct ablation studies to analyze the correlation among DR grading, the construction of positives and negatives(CPN) and dynamic hard negatives mining(DHM). We compare the proposed method with its several variants. (1) CL: the proposed model is trained without CPN and DHM, it indicates a basic CL method. (2) Lesion-aware CL *w/o* CPN: the model is trained without CPN. (3) Lesion-aware CL *w/o* DHM: the model is trained without DHM.

The results of ablation study are reported in Table 2. We can draw conclusions from several aspects: (1) CL shows the worst performance and the performance of Lesion-aware CL *w/o* CPN is obviously degraded compared to Lesion-aware CL (i.e. kappa reduces 1.5%). The results suggest that CPN is critical for improving the performance when contrastive learning is introduced in fundus images. Without false negatives disrupting the feature extraction procedure of lesions, the model is able to extract a better representation for the regions of lesions and thus achieve better DR grading performance. (2) Lesion-aware CL *w/o* DHM performs worse than Lesion-aware CL. As opposed to the common CL methods which uses all negative samples, our model takes into account the difference of the negatives with difficulty level. The teacher network is able to dynamically exploit the hard negatives and transfer the learned knowledge to the student, thereby improving the quality of the feature embeddings in subsequent

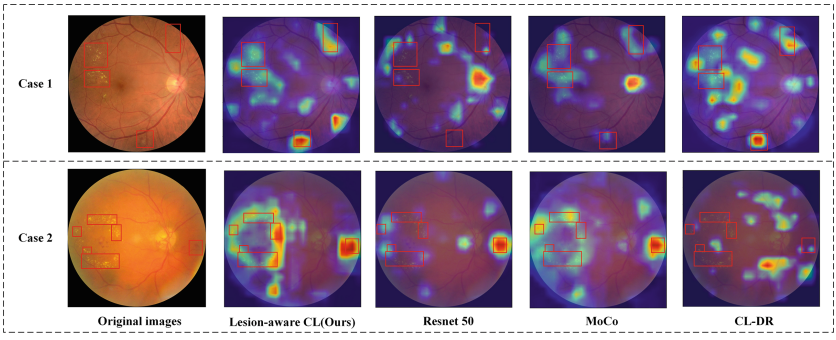

Fig. 2. Visualization results from GradCAM between the four representative methods

contrastive learning. Figure 2 shows the visualization results from GradCAM of four representative methods including Resnet50, the common CL methods (MoCo, CL-DR), and the Lesion-aware CL. Two cases with proliferate DR(DR-4) are visualized by four representative models. The intensity of the heatmap indicates the importance of each pixel in the corresponding image for making the prediction. In case 1, both Resnet and typical CL methods focus on the optic disc where has obvious physiological characteristics, while our method focuses more on the lesion regions and less on the structural aspects of the fundus image. In case 2, our method provides a promising perception of the lesion regions than other methods, suggesting that our approach allows the DR grading model to learn better representation of lesion and thus be sensitive to the DR grading.

4 Conclusion

In this paper, we propose a novel lesion-aware CL framework for DR grading. The proposed method first overcomes the false negatives problem by reconstructing positives and negatives. Then, to improve the quality of learned feature embeddings and enhance the awareness for lesion regions, we design the dynamic hard negatives mining scheme based on knowledge distillation. The experimental results demonstrate that the proposed framework significantly improves the latest results of DR grading on the benchmark dataset. Furthermore, our approaches are migratable and can be easily applied to other medical image analysis tasks.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (No.62076059), the Science Project of Liaoning province under Grant (2021-MS-105) and the 111 Project (B16009).

References

1. Ayhan, M.S., Kühlewein, L., Aliyeva, G., Inhoffen, W., Ziemssen, F., Berens, P.: Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. Med. Image Anal. **64**, 101724 (2020)
2. Cai, T.T., Frankle, J., Schwab, D.J., Morcos, A.S.: Are all negatives created equal in contrastive instance discrimination? arXiv preprint arXiv:2010.06682 (2020)
3. Cao, P., Hou, Q., Song, R., Wang, H., Zaiane, O.: Collaborative learning of weakly-supervised domain adaptation for diabetic retinopathy grading on retinal images. Comput. Biol. Med. **144**, 105341 (2022)
4. Emma Dugas, Jared, J.W.C.: Diabetic retinopathy detection (2015). https://kaggle.com/competitions/diabetic-retinopathy-detection
5. Fu, H., et al.: Evaluation of retinal image quality assessment networks in different color-spaces. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 48–56. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_6
6. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)

7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. Comput. Sci. **14**(7), 38–39 (2015)
9. Hou, Q., Cao, P., Jia, L., Chen, L., Yang, J., Zaiane, O.R.: Image quality assessment guided collaborative learning of image enhancement and classification for diabetic retinopathy grading. IEEE J. Biomed. Health Inform. 1–12 (2022). https://doi.org/10.1109/JBHI.2022.3231276
10. Huang, Y., Lin, L., Cheng, P., Lyu, J., Tang, X.: Lesion-based contrastive learning for diabetic retinopathy grading from fundus images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 113–123. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_11
11. Li, X., Jia, M., Islam, M.T., Yu, L., Xing, L.: Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. IEEE Trans. Med. Imaging **39**(12), 4023–4033 (2020)
12. Lin, Z., et al.: A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 74–82. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_9
13. Liu, X., et al.: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit. Health **1**(6), e271–e297 (2019)
14. Ogurtsova, K., et al.: IDF diabetes atlas: global estimates of undiagnosed diabetes in adults for 2021. Diabetes Res. Clin. Pract. **183**, 109118 (2022)
15. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)
16. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. In: International Conference on Learning Representations (ICLR) (2021)
17. Shen, Z., Fu, H., Shen, J., Shao, L.: Modeling and enhancing low-quality retinal fundus images. IEEE Trans. Med. Imaging **40**(3), 996–1006 (2020)
18. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? Adv. Neural. Inf. Process. Syst. **33**, 6827–6839 (2020)
19. Wang, X., Xu, M., Zhang, J., Jiang, L., Li, L.: Deep multi-task learning for diabetic retinopathy grading in fundus images. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2826–2834 (2021)
20. Wang, Z., Yin, Y., Shi, J., Fang, W., Li, H., Wang, X.: Zoom-in-Net: deep mining lesions for diabetic retinopathy detection. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 267–275. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_31
21. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 588–604. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_34

22. Yu, S., et al.: MIL-VT: multiple instance learning enhanced vision transformer for fundus image classification. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 45–54. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_5
23. Zeng, X., Chen, H., Luo, Y., Ye, W.: Automated detection of diabetic retinopathy using a binocular Siamese-like convolutional network. In: 2019 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5. IEEE (2019)
24. Zhou, K., et al.: Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2724–2727. IEEE (2018)
25. Zhou, Y., Wang, B., Huang, L., Cui, S., Shao, L.: A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. IEEE Trans. Med. Imaging **40**(3), 818–828 (2020)
26. Zolfaghari, M., Zhu, Y., Gehler, P., Brox, T.: CrossCLR: cross-modal contrastive learning for multi-modal video representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1450–1459 (2021)