A Clinical-oriented Multi-level Contrastive Learning Method for Disease Diagnosis in Low-quality Medical Images

Qingshan Hou[1,2], Shuai Cheng[1,2], Peng Cao[1,2,3,(✉)], Jinzhu Yang[1,2,3,(✉)], Xiaoli Liu[1,2], Yih Chung Tham[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
[3] National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Shenyang 110819, China
caopeng@mail.neu.edu.cn
yangjinzhu@cse.neu.edu.cn
[4] Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[5] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Canada

Abstract. Representation learning offers a conduit to elucidate distinctive features within the latent space and interpret the deep models. However, the randomness of lesion distribution and the complexity of low-quality factors in medical images pose great challenges for models to extract key lesion features. Disease diagnosis methods guided by contrastive learning (CL) have shown significant advantages in lesion feature representation. Nevertheless, the effectiveness of CL is highly dependent on the quality of the positive and negative sample pairs. In this work, we propose a clinical-oriented multi-level CL framework that aims to enhance the model's capacity to extract lesion features and discriminate between lesion and low-quality factors, thereby enabling more accurate disease diagnosis from low-quality medical images. Specifically, we first construct multi-level positive and negative pairs to enhance the model's comprehensive recognition capability of lesion features by integrating information from different levels and qualities of medical images. Moreover, to improve the quality of the learned lesion embeddings, we introduce a dynamic hard sample mining method based on self-paced learning. The proposed CL framework is validated on two public medical image datasets, EyeQ and Chest X-ray, demonstrating superior performance compared to other state-of-the-art disease diagnostic methods.

Keywords: Contrastive learning · Disease diagnosis · Low-quality medical images.

Qingshan Hou and Shuai Cheng contribute equally to this work.

1 Introduction

Medical image classification plays a crucial role in clinical disease diagnosis. Automatically identifying whether medical images indicate health or disease, and even pinpointing specific illnesses, can alleviate the repetitive burden on clinicians and increase the efficiency of diagnoses. Though recent studies have shown that deep learning techniques hold promise for medical imaging applications [12], these techniques are often constrained by limited data annotations and insufficient supervision.

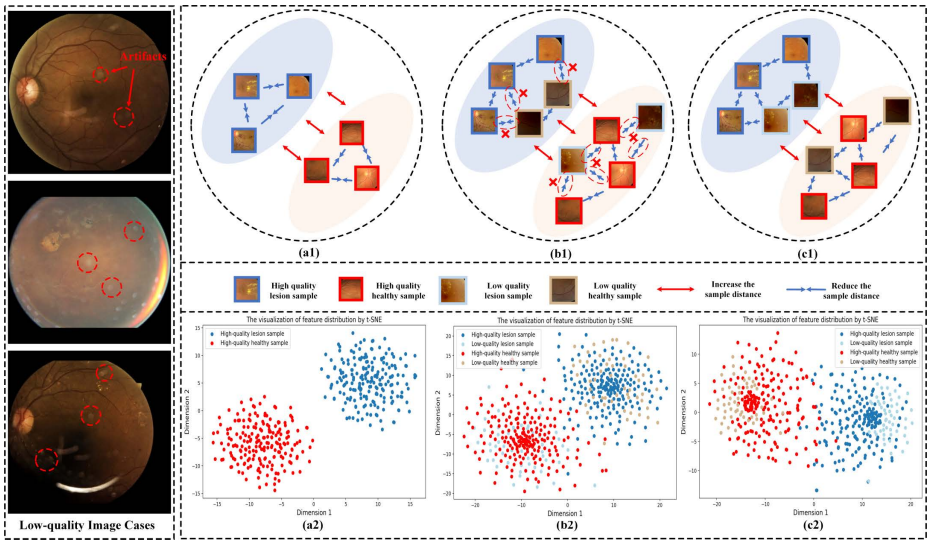

Fig. 1. *Left*: Some cases of low-quality image. *Right*: We use t-SNE to visualize the feature distribution, with 200 high-quality lesion samples, 200 high-quality healthy samples, 50 low-quality lesion samples, and 50 low-quality healthy samples. (a1)&(a2)-General contrastive learning with only high-quality images and the visualization of corresponding feature distribution. (b1)&(b2)-The impact of low-quality factors on general contrastive learning and the visualization of feature distribution. (c1)&(c2)-The proposed contrastive learning and the visualization of feature distribution.

To address these challenges in real clinical settings and fully exploit medical images without pixel-level annotations, some existing studies [11, 15, 24] proactively explore the impact of contrastive learning (CL) [4, 8] on automated disease diagnosis models. However, they do not fully consider the common quality variations in medical images, which limits their effectiveness in eliminating the interference of low-quality factors on disease diagnosis. As shown in Figure 1 *left*, the medical images often suffer from various low-quality factors such as artifacts, blurring and so on, leading to a quality degradation [1, 16]. Ideally, CL guides diagnostic models to effectively distinguish between lesion samples and healthy

samples in Figure 1(a1)&(a2). However, as illustrated in Figure 1(b1)&(b2), low-quality factors may cause CL to incorrectly pull the distance in the embedding space between lesion samples and low-quality healthy samples, or between healthy samples and low-quality lesion samples, thereby degrading the diagnostic performance of diseases.

The challenge mentioned above motivates us to develop a **C**linical-oriented **M**ulti-level **C**ontrastive **L**earning method, named CoMCL, tailored for automatic disease diagnosis on low-quality medical images. In the deployment of CL for disease diagnosis, our objective is to mitigate the effects of low-quality factors and false negative samples [20, 26], and to ensure that samples with similar semantic information remain close in the joint embedding space, as shown in Figure 1(c1)&(c2). To achieve this, we construct multi-level positive and negative pairs for the following three purposes: 1) Enhancing the ability of CL to distinguish low-quality factors from lesions in low-quality medical images; 2) Improving the ability of CL to discriminate between lesion and non-lesion areas; 3) Enhancing the awareness of CL to identify lesion characteristics in low-quality images. These abilities facilitate the modeling of the potential lesion-related embeddings, hereby enhancing the performance of the downstream diagnostic tasks. In summary, our contributions can be summarized as follows. (1) We propose a multi-level CL framework that focuses on alleviating the impact of low-quality factors on lesion feature extraction. Besides, it also alleviates the issue of false negatives that occur when CL is introduced into the diagnosis of low-quality medical images. (2) To improve the capability of CL in extracting lesion-related embeddings in low-quality medical images, we introduce a self-paced learning strategy to fully exploit and leverage hard negatives. (3) CoMCL is evaluated on the large-scale EyeQ and Chest X-ray datasets. Experimental results show that CoMCL significantly outperforms the state-of-the-art automatic disease diagnosis methods.

2 Method

Figure 2 shows the overview of CoMCL. In the first phase, we construct multi-level positive and negative pairs based on a lesion detector [11] pre-trained on an auxiliary dataset (IDRiD [17]) with pixel-level lesion annotations. This framework focuses on representing shared information among multi-level positive and negative pairs, alleviating the influence of imaging quality factors. In the second phase, a self-paced learning-based dynamic sampling method effectively leverages hard negatives [3, 18] and enhances the learned feature embedding quality.

2.1 Construction of multi-level positive and negative pairs

In this section, we provide a comprehensive description of the multi-level construction of positive and negative pairs. The incorporation of the multi-level positive and negative pairs serves the dual purpose of aligning samples with similar semantic features and mitigating the effects of false negatives on the

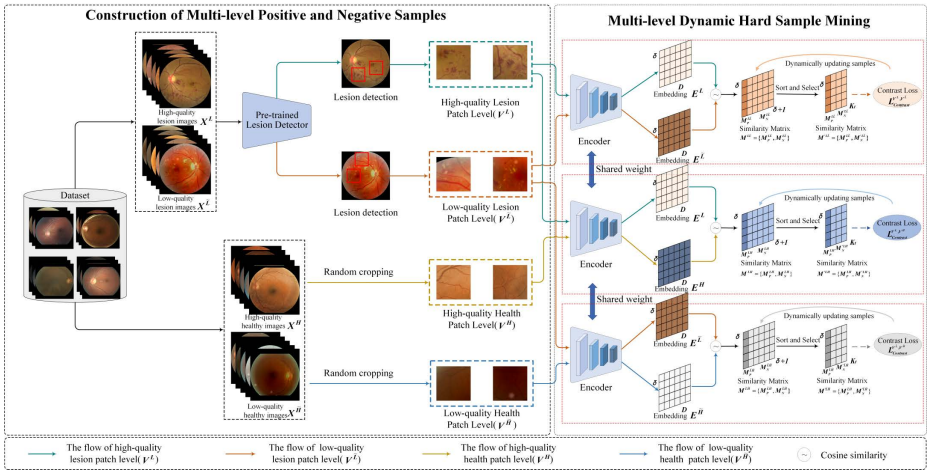

Fig. 2. The overall architecture of CoMCL comprises two components. 1) Construction of multi-level positive and negative pairs for mitigating the effects of low-quality factors and false negatives. 2) Multi-level dynamic hard sample mining for improving the quality of the learned lesion-related embeddings.

feature embeddings learned by the model. Specifically, given a dataset X annotated with disease and quality annotations, we segregate X into four subsets, including high-quality lesion images subset X^L, low-quality lesion images subset $X^{\widetilde{L}}$, high-quality healthy image subset X^H and low-quality healthy image subset $X^{\widetilde{H}}$. Then, we apply a pre-trained detector $f_{\text{det}}(\cdot)$ on $X^L/X^{\widetilde{L}}$ and obtain high-confidence detection regions. Finally, the different level samples, denoted as $V^m = \{m \in \{L, \widetilde{L}, H, \widetilde{H}\} | v_1^m, v_2^m \dots v_i^m\}$, includes two parts: $\Omega(f_{\text{det}}(X^m) > conf)$ $(m \in \{L, \widetilde{L}\})$ and $RandC(X^m)$ $(m \in \{H, \widetilde{H}\})$, 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 to ensure the inclusion of lesions as much as possible, and $RandC(\cdot)$ indicates randomly cropping images into patches with 128*128 from the healthy images.

Given a patch v_i^m generated from V^m, we consider $\widetilde{v_i^m}$ that is an augmented version from v_i^m as a positive sample and every patch v_k^n in the $V^{n|n!=m}$ as negatives. Upon encoding each positive and negative sample, we acquire the feature embedding matrixes $e_i^m / \widetilde{e_i^m} \in E^m$ and $e_k^n \in E^n$, respectively. Subsequently, the contrastive loss can be defined as:

$$L_{\text{Contrast}}^{V^m, V^n} = -\sum_i \log \left(\frac{\exp\left(\text{sim}\left(e_i^m, \widetilde{e_i^m}\right)/\tau\right)}{\exp(\text{sim}(e_i^m, \widetilde{e_i^m})/\tau) + \sum_k \exp\left(\text{sim}\left(e_i^m, e_k^n\right)/\tau\right)} \right). \tag{1}$$

The primary goal of CoMCL is to enhance the model's ability to handle challenges arising from variations in image quality and concentrate on the accurate extraction of lesion-related features in the complicated clinical setting. To this

end, V^L and its augmented versions are designated as positives, while $V^{\widetilde{L}}$ is identified as negatives, thereby deriving a CL loss, named $L_{\text{Contrast}}^{V^L,V^{\widetilde{L}}}$. This constraint enables the model to better discriminate low-quality factors in medical images and minimize the influence of low-quality factors, leading to a more accurate embedding of lesion-related features. Then, to further facilitate the extraction of lesion features and ensure the model's capability to differentiate between the lesion and healthy patches, we regard samples from V^L as positive samples while V^H as negatives, and define a CL loss, named $L_{\text{Contrast}}^{V^L,V^H}$. Finally, to further improve the model's robustness and enhance its ability to distinguish lesion from non-lesion regions in conditions of poor image quality, we treat samples from $V^{\widetilde{L}}$ as positives and samples from $V^{\widetilde{H}}$ as negatives, and devise a CL loss, named $L_{\text{Contrast}}^{V^{\widetilde{L}},V^{\widetilde{H}}}$.

2.2 Multi-level dynamic hard negatives mining

In various supervised or unsupervised algorithms [2,13] based on metric learning, research on the impact of mining hard negatives on training suggests that not all negatives hold equal value in CL. Moreover, hard negatives are semantically more similar to positives than regular negatives, indicating that hard negatives contain features of higher learning value, offering more potential beneficial information for CL. Based on this finding, we introduce self-paced learning into CL. Given the model parameters w at the current training step t, when updating the parameters of the model, we incorporate a binary variable s_i based on the previous loss $L_{\text{Contrast}}^{V^m,V^n}$, to decide whether each sample is selected. According to the similarity matrix $M^{mn} = \{M_P^{mn}, M_N^{mn}\}$, the resampled sample matrix $M^{'mn} = \{M_P^{mn}, M_N^{'mn}\}$ can be defined as:

$$M^{'mn} = \left\{ z_i^m \mid z_i^m \in Sort(M^{mn}), \text{sim}\left(z_i^m, z_k^n\right) \geq \text{sim}(z_i^m, z_{K_t}^n) \right\}, \tag{2}$$

where K_t is an adaptive parameter, determining the number of hard negatives to be considered. Specifically, $K_t = \lfloor \delta * cos(\frac{\pi t}{2T_{max}}) \rfloor$, where δ indicates the total number of negatives and T_{max} is the maximum training step. Therefore, for the update of multi-level CL models, we define the following optimization objective:

$$(w_{t+1}, v_{t+1}) = \text{argmin} \left(r(w_t) + \sum_{i=1}^{n} s_i L_{\text{Contrast}}^{V^m,V^n}(M^{'mn}) - \frac{1}{K_t} \sum_{i=1}^{n} s_i \right), \tag{3}$$

where $r(\cdot)$ denotes a regularization item, preventing the model from overfitting. By adjusting the value of K_t, we can adjust the number of hard negatives that affect the training procedure and model's generalization. If $L_{\text{Contrast}}^{V^m,V^n}(M^{'mn}) < \frac{1}{K_t}$, then $s_i = 1$ indicates that the sample is chosen for model fine-tuning. Otherwise, $s_i = 0$ indicates that the sample is not selected. To validate the effectiveness of CoMCL, the parameters obtained from the multi-level CL phase are transferred to downstream disease diagnostic models, and then fine-tuned in a supervised learning setting to adapt to specific disease diagnostic tasks.

3 Experiments

3.1 Datasets and Implementation Details

EyeQ dataset [6] is a large public fundus image benchmark for diabetic retinopathy (DR) grading and quality assessment, containing 12,543 training and 16,249 testing images. Based on image quality and DR severity, the images are classified into three quality categories and five severity levels.
Chest X-ray dataset [22] is obtained from the public NIH-ChestXray14 multi-label dataset. The training and test sets contain 8,573 and 7,007 frontal X-ray images, respectively. Based on image quality and chest diseases, the dataset includes two quality labels: high and low, and eight disease types. We validate the promotive effect of CoMCL on multi-class disease diagnosis using this dataset.
Implementation Details. ResNet50 [9] is used as the backbone network for feature extraction, with the global average pooling and fully connected layers removed. For the construction of multi-level positive and negative pairs, all patches are cropped to 128×128 due to varying original image sizes. The temperature parameter τ in Equation 1 is set to 0.07. During multi-level dynamic hard sample mining, parameters are optimized using the Adam optimizer (momentum=0.9) over 800 epochs. Training starts with a learning rate of 1×10^{-3} and a batch size of 400.

3.2 Comparison with the State-of-the-Art

This section presents both quantitative and qualitative comparisons with various recent disease diagnostic methods on the EyeQ and Chest X-ray datasets, showcasing the effectiveness of the CoMCL framework for single-label (DR grading) as well as multi-label chest disease diagnosis. We compare CoMCL with several comparable methods, including Resnet50, Inception-v3, DenseNet-121, MMCNN [25], Zoom-in-Net [23], Lesion-base CL [11], CABNet [7], DeepMT-DR [21], Lesion-aware CL [5], and LANet [10]. For all comparable methods, we follow the same experimental setup described in their original papers to ensure the fairness and competitiveness of each competing approach.

As shown in Table 1, CoMCL significantly outperforms comparable methods for both DR grading and multi-label chest disease diagnosis, achieving higher Kappa and Accuracy scores. The results highlight several interesting observations: (1) We first explore the impact of the proportion of low-quality images in the EyeQ and Chest X-ray datasets on CoMCL and comparable methods. As the proportion of low-quality images increases from the original 33.4%/43.2% (the proportion of low-quality images in the original dataset) to 100% (by degrading all images using [14, 19]), the performance of all diagnostic methods decreases across datasets, indicating the negative impact of low-quality factors. However, CoMCL exhibits a stronger ability to avoid interference from low-quality factors compared to other methods. (2) Compared to previous CL methods (Lesion-base CL and Lesion-aware CL), CoMCL exhibits significant advantages across datasets under different proportions of low-quality images. This

Table 1. The comparison between CoMCL and the comparable methods in DR grading and multi-label chest disease diagnosis.

Methods	EyeQ Dataset						Chest X-ray Dataset					
	33.4%		70%		100%		43.2%		70%		100%	
	Kappa	ACC	Kappa	ACC	Kappa	ACC	Kappa	ACC	Kappa	ACC	Kappa	ACC
Resnet50	0.804	0.783	0.743	0.715	0.674	0.662	0.619	0.626	0.594	0.613	0.564	0.589
Inception-v3	0.798	0.776	0.728	0.706	0.652	0.637	0.615	0.624	0.587	0.609	0.558	0.576
DenseNet-121	0.813	0.794	0.756	0.732	0.683	0.668	0.624	0.635	0.603	0.658	0.572	0.607
MMCNN [25]	0.862	0.841	0.795	0.778	0.725	0.704	0.657	0.672	0.632	0.644	0.604	0.626
Zoom-in-Net [23]	0.873	0.854	0.812	0.784	0.736	0.713	0.662	0.684	0.648	0.653	0.615	0.631
Lesion-base CL [11]	0.848	0.832	0.783	0.761	0.694	0.672	0.636	0.652	0.617	0.628	0.584	0.619
CABNet [7]	0.865	0.847	0.797	0.782	0.731	0.709	0.660	0.675	0.643	0.651	0.607	0.628
DeepMT-DR [21]	0.857	0.839	0.791	0.768	0.712	0.694	0.649	0.665	0.621	0.639	0.592	0.624
Lesion-aware CL [5]	0.876	0.857	0.824	0.816	0.757	0.724	0.673	0.691	0.653	0.664	0.626	0.645
LANet [10]	0.854	0.835	0.786	0.765	0.706	0.683	0.642	0.658	0.619	0.634	0.587	0.622
CoMCL(Ours)	**0.884**	**0.872**	**0.852**	**0.837**	**0.793**	**0.776**	**0.682**	**0.708**	**0.665**	**0.672**	**0.641**	**0.663**

advantage mainly benefits from the fact that CoMCL incorporates information from multiple levels and different qualities of medical images. Therefore, CoMCL acquires lesion features that are not obvious in low-quality patches, mitigating the influence of low-quality factors on the learned lesion embeddings, and enhancing the model's ability to comprehensively identify lesion-related features.

3.3 Ablation Study

To comprehensively investigate the contribution of multi-level positive and negative pairs and self-paced learning to disease diagnosis, we compare CoMCL with several variants: (1) Baseline (Resnet 50): Training a basic classification model on the EyeQ dataset. (2) Basic CL: Pre-training a basic CL model [8] and fine-tuning the downstream classification model on EyeQ. (3) CoMCL w/o Multi-level: Constructing positives and negatives using only lesion and healthy samples respectively, without the construction of multi-level positive and negative pairs. (4) CoMCL w/o SPL: Training without considering self-paced learning, i.e., without considering hard negatives.

Table 2. Ablation study of CoMCL on EyeQ dataset.

Methods	ACC	Kappa
Baseline(Resnet 50)	0.804	0.783
Basic CL	0.842	0.827
CoMCL w/o Multi-level	0.859	0.836
CoMCL w/o SPL	0.868	0.860
CoMCL	**0.884**	**0.872**

Table 3. Performance of CoMCL on EyeQ under different combinations.

Method	V^L	$V^{\tilde{L}}$	V^H	$V^{\tilde{H}}$	ACC	Kappa
CoMCL_v1	✓	✓			0.846	0.835
CoMCL_v2	✓	✓	✓		0.867	0.856
CoMCL_v3	✓	✓	✓	✓	**0.884**	**0.872**

The experimental results are shown in Table 2. The following aspects can be revealed: 1) The baseline model exhibits the lowest performance, underscoring the importance of CL in modeling specific lesion embeddings. Furthermore, compared to other variants of CoMCL, the baseline CL performs the worst, reflecting

the adverse impact of low-quality factors on contrastive learning and the significance of hard negatives to contrastive learning. 2) The performance of CoMCL w/o Multi-level is lower than that of CoMCL. The construction of multi-level positive and negative pairs is crucial for improving diagnostic performance when introducing CL for automatic disease diagnosis. By considering the impact of low-quality factors on lesion-related embedding extraction, CoMCL can learn relevant lesion embeddings more effectively, thereby achieving improved disease diagnostic performance. 3) The performance of CoMCL w/o SPL is lower than CoMCL. By dynamically mining hard negatives through self-paced learning, the quality of lesion embeddings is further improved, thereby enhancing the performance of disease diagnostic tasks.

Subsequently, we investigate the impact of incorporating different levels on the final result, as depicted in Table 3, which shows the performance comparison of the CoMCL method under varying level combinations. From the results, it is clear that the model CoMCL_v3, by considering all level patches, achieves the best results in terms of accuracy and kappa. The experimental results show that simultaneously considering the learning of discriminative embeddings for lesions and the identification of quality factors achieves a more significant performance boost after fine-tuning the downstream tasks. Additionally, incorporating low-quality levels can enhance the model's ability to discern lesion features under limited imaging conditions, ensuring that the model's diagnostics do not solely rely on high-quality image features. This approach ensures the effectiveness of the algorithm under complex clinical imaging conditions.

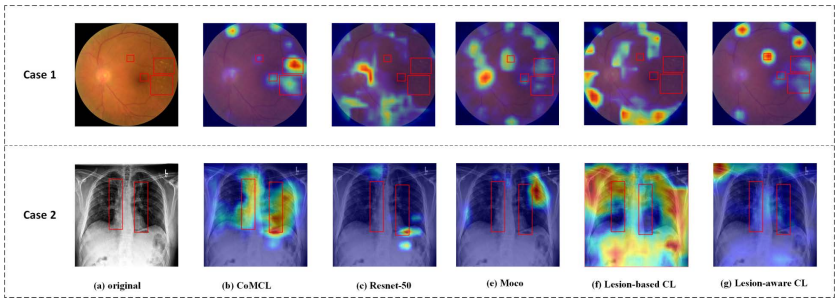

Fig. 3. Visualization results of Regions of Interest (RoIs) across the ResNet and the representative CL methods.

Figure 3 illustrates the visualization results of baseline(Resnet50) and different contrastive learning-based diagnostic methods (Moco, Lesion-base CL and Lesion-aware CL) for two medical cases: a fundus image indicative of proliferative diabetic retinopathy (Case 1) and a chest X-ray showing potential pulmonary pathologies (Case 2). While the other methods are capable of capturing lesions to a certain extent, they are distracted by more prominent physiological features and high-contrast edges. In contrast, CoMCL more clearly emphasizes lesion re-

gions. This indicates that CoMCL has the ability to distinguish lesion-relevant features from structural aspects and complex low-quality factors of the medical image, which is crucial for accurate disease diagnosis.

4 Conclusion

In this study, we propose a clinical-oriented multi-level contrastive learning framework for disease diagnosis in low-quality medical images. The proposed framework, by constructing multi-level positive and negative pairs, can explore lesion features from different levels, thereby mitigating the influence of low-quality factors on the model's extraction of lesion features. Additionally, we design a dynamic hard negative mining scheme based on self-paced learning to fully utilize hard negative samples, significantly improving the quality of feature embedding. Experimental results show that CoMCL significantly improves the accuracy of disease diagnosis in low-quality medical images, which is crucial for conserving medical resources and enhancing the efficiency of medical services.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (No.62076059), the Science and Technology Joint Project of Liaoning province (2023JH2/101700367, ZX20240193) and the China Scholarship Council(202306080125).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anand, S., Roshan, R., et al.: Chest x ray image enhancement using deep contrast diffusion learning. Optik **279**, 170751 (2023)
2. Birodkar, V., Mobahi, H., Bengio, S.: Semantic redundancies in image-classification datasets: The 10% you don't need. arXiv preprint arXiv:1901.11409 (2019)
3. 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)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020)
5. Cheng, S., Hou, Q., Cao, P., Yang, J., Liu, X., Zaiane, O.R.: Lesion-aware contrastive learning for diabetic retinopathy diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 671–681. Springer (2023)
6. Fu, H., Wang, B., Shen, J., Cui, S., Xu, Y., Liu, J., Shao, L.: Evaluation of retinal image quality assessment networks in different color-spaces. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. pp. 48–56. Springer (2019)
7. He, A., Li, T., Li, N., Wang, K., Fu, H.: Cabnet: Category attention block for imbalanced diabetic retinopathy grading. IEEE Transactions on Medical Imaging **40**(1), 143–153 (2021). https://doi.org/10.1109/TMI.2020.3023463

8. 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)
9. 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)
10. Hou, J., Xiao, F., Xu, J., Feng, R., Zhang, Y., Zou, H., Lu, L., Xue, W.: Diabetic retinopathy grading with weakly-supervised lesion priors. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
11. Huang, Y., Lin, L., Cheng, P., Lyu, J., Tang, X.: Lesion-based contrastive learning for diabetic retinopathy grading from fundus images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 113–123. Springer (2021)
12. Jiang, H., Diao, Z., Shi, T., Zhou, Y., Wang, F., Hu, W., Zhu, X., Luo, S., Tong, G., Yao, Y.D.: A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. Computers in Biology and Medicine p. 106726 (2023)
13. Kaya, M., Bilge, H.Ş.: Deep metric learning: A survey. Symmetry **11**(9), 1066 (2019)
14. Li, H., Liu, H., Fu, H., Shu, H., Zhao, Y., Luo, X., Hu, Y., Liu, J.: Structure-consistent restoration network for cataract fundus image enhancement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 487–496. Springer (2022)
15. 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 Transactions on Medical Imaging **39**(12), 4023–4033 (2020)
16. Philip, S., Cowie, L., Olson, J.: The impact of the health technology board for scotland's grading model on referrals to ophthalmology services. The British Journal of Ophthalmology **89**(7), 891 (2005)
17. 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)
18. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. In: International Conference on Learning Representations (ICLR) (2021)
19. Shen, Z., Fu, H., Shen, J., Shao, L.: Modeling and enhancing low-quality retinal fundus images. IEEE transactions on medical imaging **40**(3), 996–1006 (2020)
20. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? Advances in neural information processing systems **33**, 6827–6839 (2020)
21. 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)
22. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)

23. Wang, Z., Yin, Y., Shi, J., Fang, W., Li, H., Wang, X.: Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. pp. 267–275. Springer (2017)
24. 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)
25. Zhou, K., Gu, Z., Liu, W., Luo, W., Cheng, J., Gao, S., Liu, J.: 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)
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)