2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871507

Understanding How Fundus Image Quality Degradation Affects CNN-based Diagnosis

Haofeng Liu[1], Haojin Li[1], Xiaoxuan Wang[1], Heng Li[1,*], Mingyang Ou[1], Luoying Hao[1], Yan Hu[1,*], Jiang Liu[1,2]

***Abstract*— Quality degradation (QD) is common in the fundus images collected from the clinical environment. Although diagnosis models based on convolutional neural networks (CNN) have been extensively used to interpret retinal fundus images, their performances under QD have not been assessed. To understand the effects of QD on the performance of CNN-based diagnosis model, a systematical study is proposed in this paper. In our study, the QD of fundus images is controlled by independently or simultaneously importing quantified interferences (e.g., image blurring, retinal artifacts, and light transmission disturbance). And the effects of diabetic retinopathy (DR) grading systems are thus analyzed according to the diagnosis performances on the degraded images. With images degraded by quantified interferences, several CNN-based DR grading models (e.g., AlexNet, SqueezeNet, VGG, DenseNet, and ResNet) are evaluated. The experiments demonstrate that image blurring causes a significant decrease in performance, while the impacts from light transmission disturbance and retinal artifacts are relatively slight. Superior performances are achieved by VGG, DenseNet, and ResNet in the absence of image degradation, and their robustness is presented under the controlled degradation.**

I. INTRODUCTION

Due to the safety, high efficiency, and low-cost of observing retinal structure, retinal fundus images have been extensively used in ophthalmology clinical diagnosis and treatment. Moreover, computer-aided diagnosis systems have been developed based on the fundus images to automatically diagnose ocular diseases [1], [2]. However, the quality of retinal fundus images suffers a considerable variation in the clinical environment. The clinical fundus images are always collected by people under various lighting conditions using different cameras, leading to the floating in imaging quality. The low quality of fundus images may impact the analysis and diagnosis of ophthalmologists as well as computer-aided diagnosis systems [3].

With the success of deep learning on image processing, studies of how image quality affects CNN models have been increasingly reported. In the field of natural images, Samuel et al. [4], [5] evaluated the performance of CNN models and human visual system under quality distortions. Samil et al. [6] and Klemen et al. [7] investigated the effects of degradation on CNN models in the scenario of face recognition. The effect of image degradations and degradation removal is demonstrated in the experiments on nine kinds of degraded images [8], [9]. To quantify the quality of no-reference images, an model using meta-learning is proposed by Zhu et al. [10] for quality assessment. On the other hand, CNN-based models have been increasingly introduced to medical scenarios [11], especially the automatic diagnosis on fundus images. Efforts have thus been made to study the QD of fundus images. Fu et al. [12] proposed a classification model for image quality to select the input for computer-aided diagnosis. To remove the artifacts in fundus images, Yoo et al. [13] proposed a scheme to control the fundus image quality. As low-quality fundus images lead to misdiagnosis, Shen et al. [14] proposed a degradation model to simulate the inferior-quality interferences and presented an enhancement network to suppress the interferences. Since the interferences caused by cataracts lead to a higher risk of misdiagnosis, Li et al. [15] proposed a restoration model to improve the quality of the cataract fundus image via domain adaptation. However, how QD affects CNN-based diagnosis on fundus images needs further study, since the collection and classification of fundus images are different from the natural ones.

This paper investigates the effects of inferior-quality interferences on diabetic retinopathy (DR) grading systems to extend the study of QD on fundus images. The overview of this study is shown in Fig. 1. Following the degradation

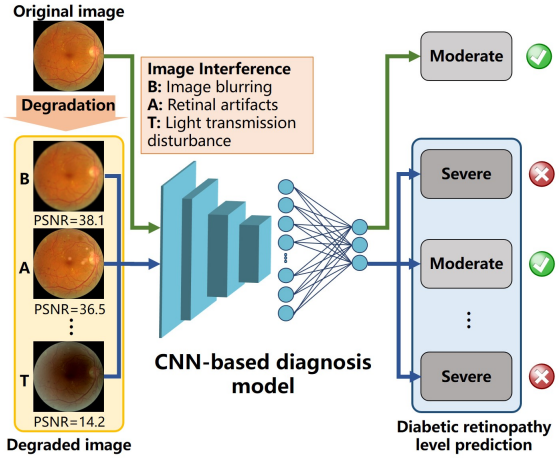

Fig. 1. Imaging interferences degrade the quality of fundus image and increase the risk of misdiagnosis from CNN-based diagnosis models. Considering the interferences present different effects on the CNN-based models, this paper using quantified degradation investigates the impacts from various interferences on the diagnosis for promoting following studies on fundus images with degradation.

* Correspondence: lih3, huy3@sustech.edu.cn
[1] Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[2] Cixi Institute of Biomedical Engineering, Ningbo Institute of Industrial Technology, Chinese Academy of Sciences, Ningbo 315201, China

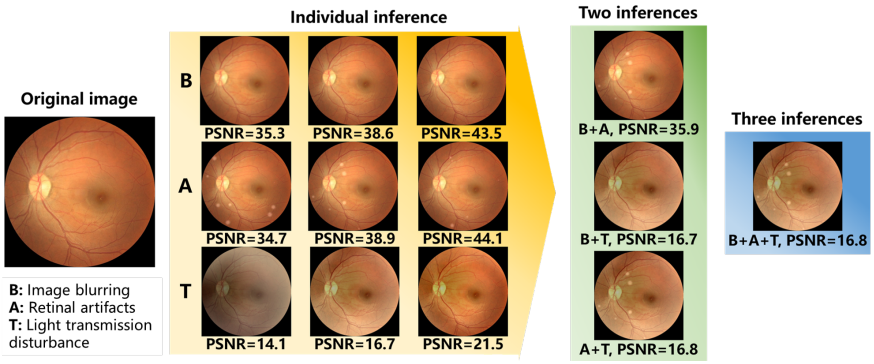

Fig. 2. The fundus images degraded by various combinations of the interferences.

model proposed in [14], three main inferior-quality interferences (e.g., image blurring, retinal artifacts, and light transmission disturbance) are applied on fundus images in our study. Moreover, we introduce the Peak Signal-to-Noise Ratio (PSNR) to quantify the degradation level. The DR grading models are built on the dataset of EyePACS with networks, including Alexnet [16], VGG [17], ResNet [18], and DenseNet [19]. Subsequently, the effects of QD are presented by applying the DR grading models to the quantitatively degraded fundus images. The main contributions of this paper are summarised as follows:

- To investigate the effects of QD on the DR grading models, a protocol is presented to control the QD on fundus images by simulating and quantifying the inferior-quality interferences.
- Using the degradation protocol, state-of-the-art DR grading models are compared to understand the effects of QD on CNN-based systems.
- Based on this study, the effects of QD on DR grading models are presented and analyzed, and inferences are provided to promote further efforts on the low-quality fundus images.

II. METHODOLOGY

To investigate the effects of QD on the CNN-based DR grading models of fundus images, we elaborate on the simulation and quantification of QD, as well as the construction of DR grading models in this section.

A. Fundus image degradation

Retinal fundus images often suffer from QD when collected in a complex clinical environment. By analyzing the optical feed-forward system, three significant interferences of the low-quality fundus images, including image blurring, retinal artifacts, and light transmission disturbance are recognized in [14].

Caused by the movement of the eyeball or a wrong setting of the focal length, image blurring may appear in fundus images. Undesired objects on the lens of the imaging plane, such as dust and grains, leads to retinal artifacts in fundus images. Due to the lighting source mixed with unstable stray light during image collection, fundus images are prone to suffer light transmission disturbance, which consists of uneven global exposure and local uneven illumination fundus images. To understand how QD of fundus images affects CNN-based diagnosis, fundus images with QD is simulated for diagnosis following [14].

B. Quality degradation quantification

To quantitatively analyze the effects of QD, PSNR is introduced to measure the degree of degradation. PSNR is a measurement for the quality of reconstruction of signal and has been widely applied as an evaluation metric to image denoising. For calculating the PSNR of a test image b, the original image a is necessary. And the PSNR is given by:

$$PSNR(a,b) = 10\log_{10}\left(I^2/MSE(a,b)\right), \quad (1)$$

$$MSE(a,b) = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}\|a(i,j) - b(i,j)\|^2. \quad (2)$$

where I represents the maximum possible pixel value of the image (e.g., I is equal to 255 for 8-bit representation), W and H refer to the width and height of the image. According to the definition, the smaller the PSNR, the more severe degradation has been imported into the image.

The interferences are quantitatively applied to the fundus image to control the degradation using the following protocol. Firstly, the interference simulation is repeated 100 times with random parameters and quantified by PSNR. The PSNR of three interferences lays within the range of (31, 49), (29, 48), and (12, 27). Accordingly, for image blurring and retinal artifacts, the degradation levels are constrained within PSNR from 35 to 44, while light transmission disturbance is within 14 to 23. Subsequently, the fundus images at each level are collected from the images degraded by the parameters with the closest PSNR. Additionally, the PSNR range of (-2, 2) around a specific level, respectively, covers 92.34%, 99.76%, and 83.84% of the collected images degraded by image blurring, retinal artifacts, and light transmission disturbance. Further, to comprehensively simulate the clinical environment, multiple interferences are also combined to implement the degradation. An example of the images degraded by individual and combined interferences is shown in Fig. 2.

C. DR grading model and baseline construction

EyePACS is the largest publicly available fundus images dataset collected from diverse models and cameras with real-world settings. And in total, 88,702 fundus images are contained in EyePACS with DR grading labels. Therefore, the diagnosis models are constructed to grade DR using EyePACS. In this study, five state-of-the-art CNNs are utilized to comprehensively construct the models to investigate how low-quality images affect DR grading models. In particular, AlexNet, SqueezeNet, VGG, ResNet, and DenseNet are respectively implemented with PyTorch, and ResNet is implemented with various depths.

The images are loaded with the size of 512×512 and batch size of 12. Random horizontal flips, vertical flips, as well as random rotation are conducted to reduce overfitting. The models are optimized by stochastic gradient descent of cross-entropy loss with an initial learning rate of 0.005 and momentum of 0.9. Fifty epochs are executed, during which the learning rate decays with a factor of 0.8 every three epochs.

Considering EyePACS is collected with real-world settings, the fundus images have a large variation in resolution, intensity, and quality. The DR grading models learned from EyePACS have hence been equipped with robustness to noise and variation. The models are optimized by the 35,126 training images and validated using the 10,906 public test images. The models that have the minimum loss on the validation set are selected as the final DR grading models. The 42,670 test images are employed to evaluate the DR grading models.

TABLE I
BASELINE OF DR GRADING MODELS (%).

Models	Precision	Recall	F1-score	QKappa
AlexNet	76.55	81.27	77.53	69.77
SqueezeNet	82.07	84.58	82.36	77.96
VGG-16	83.60	85.32	83.98	79.82
DenseNet-121	83.98	85.27	84.27	80.41
ResNet-18	82.96	85.16	83.43	79.11
ResNet-34	83.46	85.35	83.54	79.89
ResNet-50	83.84	85.70	84.29	80.56
ResNet-101	83.69	85.53	83.85	80.25
ResNet-152	83.34	85.38	83.57	79.43

The diagnosis baselines summarized in Table I are demarcated by applying the models to grade DR in the test images. The metrics of precision, recall, F1-score, and quadratic weighted kappa (QKappa) [20] are presented to quantify the baselines.

III. EXPERIMENTS

In the experiments, various combinations of the interferences are imported to the test data to understand the clinical QD of fundus images comprehensively. And the DR grading models are then applied to the degraded test images for investigating the effects of QD. Each experiment is conducted three times by random repeating the interference simulation. Comparisons of the F1-score are plotted in the following parts to visualize the impacts of QD on the DR grading models.

A. Degradation from individual interference

In this experiment, the three interferences are individually imported into the test images following the degradation levels of PSNR. The diagnosis results are illuminated in Fig. 3, where (a-c) refer to the test images degraded by image blurring, retinal artifacts, and light transmission disturbance. The horizontal axis represents the PSNR levels, and the vertical axis represents the F1-score.

It can be observed from Fig. 3 (a-c), image blurring has the most significant impact on the diagnosis, followed by light transmission disturbance, and the most negligible impact comes from retinal artifacts. Along with the degradation of image blurring aggravating, a sharp drop presents on the curve of F1-scores. For image blurring, a gap of about 10% in F1-scores is observed between the least and the highest PSNR. And the performance gap is about 4% under light transmission disturbance. While retinal artifacts only lead to a gap of less than 2%.

For the DR grading models, outstanding performance is achieved by DenseNet, ResNet, and VGG, while SqueezeNet and AlexNet perform mediocrely on the grading of DR. Superior robustness has been presented by DenseNet. It achieves the top 2 performance in most degradation conditions. VGG appears especially robust to light transmission disturbance, but sensitive to image blurring. Such that VGG outperforms DenseNet and ResNet in Fig. 3 (c), but drops precipitously in Fig. 3 (a).

Among ResNets, ResNet-101 always performs comparably to the best one in all three interferences. Although ResNet-50 achieves the best diagnosis results in the baselines, it only has superior performances under degradation of retinal artifacts and low-level image blurring. The rest ResNets perform variously under degradation. Using ResNet-101 as a representative of ResNets, it is not as robust as DenseNet to the degradation but outperforms other networks.

B. Degradation from two interferences

Two interferences are randomly combined to degrade the image quality in this experiment. One interference fixes to a specific PSNR based on the median PSNR value under the individual interference, and the other is applied with the ten PSNR levels for the degradation. The condition of interferences will swap when the first round of degradation finishes. Fig. 3 (d-i) exhibits the diagnosis results under two combined interferences. The condition of interferences is marked over each subfigure, where the PSNR of fixed interference refers to the subscript.

According to the comparison of Fig. 3 (a-c) and 3 (d-i), it can be inferred that consistent effects are exhibited by the interferences regardless of presence independently or simultaneously. Specifically, the combined degradation in Fig. 3 (d) (g), demonstrate a similar variation to the results of individual interference in Fig. 3 (a). Furthermore, unanimous phenomenons are observed from the comparisons between

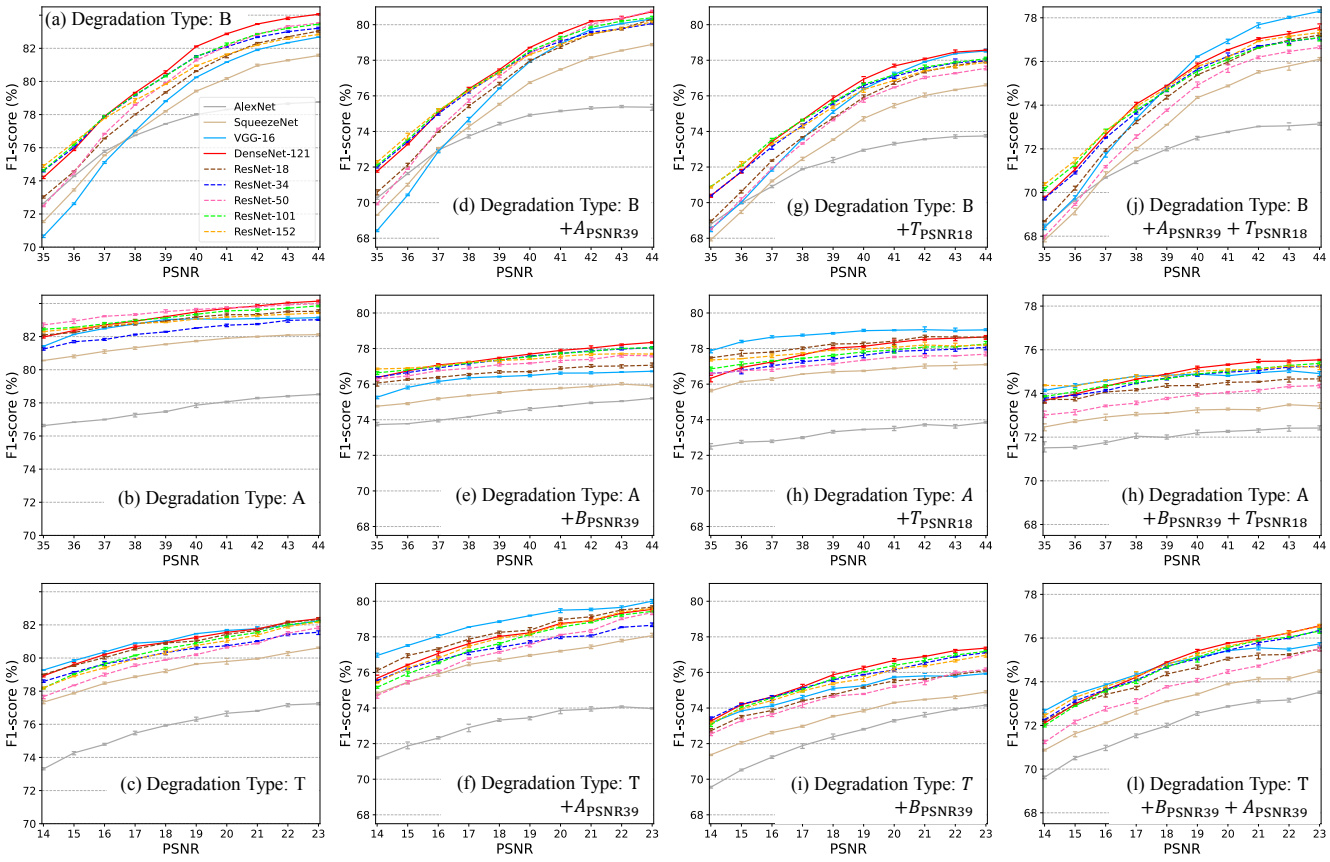

Fig. 3. Diagnosis results of the images degraded by two interferences. $_{\mathrm{PSNR}n}$ denotes limiting the interference to a specific PSNR of n.

Fig. 3 (e) (h) and Fig. 3 (b), as well as Fig. 3 (f) (i) and Fig. 3 (c). Furthermore, comparing to Fig. 3 (a) (c), the impact from retinal artifacts in Fig. 3 (d) (f) is little. However, as a result ofimage blurring and light transmission disturbance, remarkable declines in performance appear in Fig. 3 (e) (i) and Fig. 3 (g) (h).

On the other hand, DenseNet and ResNet-101 exhibit robust to the combination of image blurring and retinal artifacts, as well as image blurring and light transmission disturbance. However, VGG has elegant robustness to the combination of retinal artifacts light transmission disturbance, so that it outperforms DenseNet and ResNet in Fig. 3 (f) (h).

Due to the robustness to light transmission disturbance, ResNet-18 outperforms other ResNets in Fig. 3 (f) (h). And the performance of ResNet-50 has been severely impacted by image blurring and light transmission disturbance.

C. Degradation from three interferences

For investigating the degradation from three interferences, two of the three interferences are fixed, while the rest is performed with ten PSNR levels during the degradation processing. And as illuminated in Fig. 3 (j-l), the interferences are iteratively adjusted and marked over each subfigure.

Under the degradation for all three interferences, DenseNet still performs steadily. Due to the robustness to light transmission disturbance, VGG presents outstanding robustness when image blurring is weak in the combined degradation. The degradation has significantly impacted the performance of ResNet-50. And the robustness grows with the depth increasing for other ResNets. ResNet-152 has the top performance in RestNets.

D. Discussion

The experiments demonstrated that the three interferences, including image blurring, retinal artifacts, and light transmission disturbance, have disparate impacts on the DR grading. As a result of image blurring and light transmission disturbance, global degradation appears on fundus images. The models are thus most prone to misdiagnosis under the degradation of image blurring. And the secondary impact on the diagnosis comes from light transmission disturbance, which does not distort the image information as much as image blurring. The local degradation caused by retinal artifacts impacts the DR grading models slightly among the interferences. Therefore, overcoming global degradation, especially image blurring, is essential to implement a CNN-based diagnosis. And during the collection of fundus images, efforts should be made to address the defocus and unstable stray light.

For the DR grading models, AlexNet and SqueezeNet perform not as well as VGG, DenseNet, and ResNet, since they are constructed by lightweight networks. The commensurate results of VGG, DenseNet, and ResNet in the

baselines, demonstrate that once been fully optimized, the models with strong convolutional layers are competent for DR grading. The robustness of the networks is illuminated in the experiments on controlled degraded images. Because the denser skip-connections between layers propagate low-level features to the classifier, impressive robustness is provided by DenseNet. The large receptive field endows VGG robustness to light transmission disturbance. ResNet-50 achieves the best performance in the absence of degradation but is not robust to the strong degradation. However, ResNet-101 and -152 present remarkable results in most degradation conditions, which illuminates that the depth growth generally enhances the robustness to QD.

Accordingly, ResNet-50 can provide the most accurate diagnosis in the ideal condition. However, to implement in a clinical scenario, DenseNet, ResNet-101, and -152 are recommended to construct the DR grading model. If the image blurring can be constrained, robust performances are presented by VGG.

To further understand the effects of degradation on diagnosis models, other ocular diseases, such as glaucoma age-related macular degeneration, will be tested in future studies.

IV. CONCLUSION

In this paper, we investigated the effects of image degradation on the performance of CNN-based diagnosis models. A protocol is presented to simulate three inferior-quality interferences quantitatively to control the QD of fundus images. And DR grading models based on various CNNs are used to diagnose the degraded images. The experiments indicated that image blurring has the most significant impact on the diagnosis, followed by light transmission disturbance, and the most negligible impact comes from retinal artifacts. And DenseNet and ResNets with deep layers are robust to the QD in fundus images.

ACKNOWLEDGMENT

This work was supported in part by Guangdong Basic and Applied Fundamental Research Fund Committee (2020A1515110286), Guangdong Provincial Department of Education (2020ZDZX3043), Guangdong Provincial Key Laboratory (2020B121201001), and Shenzhen Natural Science Fund (JCYJ20200109140820699 and 20200925174052004).

REFERENCES

[1] 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.

[2] Heng Li, Haofeng Liu, Yan Hu, Huazhu Fu, Yitian Zhao, Hanpei Miao, and Jiang Liu, "An annotation-free restoration network for cataractous fundus images," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2022.

[3] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis, "A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2020, pp. 1–12, ACM.

[4] Samuel Dodge and Lina Karam, "Understanding how image quality affects deep neural networks," in *2016 eighth international conference on quality of multimedia experience (QoMEX)*, 2016, pp. 1–6.

[5] Samuel Dodge and Lina Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," in *2017 26th international conference on computer communication and networks (ICCCN)*, 2017, pp. 1–7.

[6] Samil Karahan, Merve Kilinc Yildirum, Kadir Kirtac, Ferhat Sukru Rende, Gultekin Butun, and Hazim Kemal Ekenel, "How image degradations affect deep cnn-based face recognition?," in *2016 international conference of the biometrics special interest group (BIOSIG)*, 2016, pp. 1–5.

[7] Klemen Grm, Vitomir Štruc, Anais Artiges, Matthieu Caron, and Hazım K Ekenel, "Strengths and weaknesses of deep learning models for face recognition against image degradations," *Iet Biometrics*, vol. 7, no. 1, pp. 81–89, 2017.

[8] Yanting Pei, Yaping Huang, Qi Zou, Yuhang Lu, and Song Wang, "Does haze removal help cnn-based image classification?," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 682–697.

[9] Yanting Pei, Yaping Huang, Qi Zou, Xingyuan Zhang, and Song Wang, "Effects of image degradation and degradation removal to cnn-based image classification," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[10] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020.

[11] Heng Li, Yan Hu, Sanqian Li, Wenjun Lin, Peng Liu, Risa Higashita, and Jiang Liu, "Ct scan synthesis for promoting computer-aided diagnosis capacity of covid-19," in *International Conference on Intelligent Computing*. Springer, 2020, pp. 413–422.

[12] Huazhu Fu, Boyang Wang, Jianbing Shen, Shanshan Cui, Yanwu Xu, Jiang Liu, and Ling Shao, "Evaluation of retinal image quality assessment networks in different color-spaces," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 48–56.

[13] Tae Keun Yoo, Joon Yul Choi, and Hong Kyu Kim, "Cyclegan-based deep learning technique for artifact reduction in fundus photography," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 258, no. 8, pp. 1631–1637, 2020.

[14] Ziyi Shen, Huazhu Fu, Jianbing Shen, and Ling Shao, "Modeling and enhancing low-quality retinal fundus images," *IEEE Transactions on Medical Imaging*, 2020.

[15] Heng Li, Haofeng Liu, Yan Hu, Risa Higashita, Yitian Zhao, Hong Qi, and Jiang Liu, "Restoration of cataract fundus images via unsupervised domain adaptation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 516–520.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[17] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] 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.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[20] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Humaine association conference on affective computing and intelligent interaction*, 2013, pp. 245–251.