Image Quality Assessment Guided Collaborative Learning of Image Enhancement and Classification for Diabetic Retinopathy Grading

Qingshan Hou, Peng Cao, Liyu Jia, Leqi Chen, Jinzhu Yang, and Osmar R. Zaiane

***Abstract*—Diabetic retinopathy (DR) is one of the most serious complications of diabetes and is a prominent cause of permanent blindness. However, the low-quality fundus images increase the uncertainty of clinical diagnosis, resulting in a significant decrease on the grading performance of the fundus images. Therefore, enhancing the image quality is essential for predicting the grade level in DR diagnosis. In essence, we are faced with three challenges: (I) How to appropriately evaluate the quality of fundus images; (II) How to effectively enhance low-quality fundus images for providing reliable fundus images to ophthalmologists or automated analysis systems; (III) How to jointly train the quality assessment and enhancement for improving the DR grading performance. Considering the importance of image quality assessment and enhancement for DR grading, we propose a collaborative learning framework to jointly train the subnetworks of the image quality assessment as well as enhancement, and DR disease grading in a unified framework. The key contribution of the proposed framework lies in modelling the potential correlation of these tasks and the joint training of these subnetworks, which significantly improves the fundus image quality and DR grading performance. Our framework is a general learning model, which may be useful in other medical images with low-quality data. Extensive experimental results have shown that our method outperforms state-of-the-art DR grading methods by a considerable 73.6% ACC/71.2% Kappa and 88.5% ACC/86.3% Kappa on Messidor and EyeQ benchmark datasets, respectively. In addition, our method significantly enhances the low-quality fundus images while preserving fundus structure features and lesion information. To make the framework more general, we also evaluate the enhancement results in more downstream tasks, such as vessel segmentation.**

***Index Terms*—Diabetic retinopathy, grading, quality assessment, image enhancement, joint learning.**

Manuscript received 24 July 2022; revised 11 November 2022; accepted 19 December 2022. Date of publication 23 December 2022; date of current version 7 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62076059 and in part by the Science Project of Liaoning province under Grant 2021-MS-105. *(Corresponding author: Peng Cao.)*

Qingshan Hou, Peng Cao, Liyu Jia, Leqi Chen, and Jinzhu Yang are with the School of Computer Science and Engineering, Northeastern University, Shenyang 110000, China, and also with the Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang 110000, China (e-mail: qingshan_hou@126.com; caopeng@cse.neu.edu.cn; 20194693@stu.neu.edu.cn; 20195188@stu.neu.edu.cn; yangjinzhu@cse.neu.edu.cn).

Osmar R. Zaiane is with the Amii, University of Alberta, Edmonton, Canada (e-mail: zaiane@ualberta.ca).

Digital Object Identifier 10.1109/JBHI.2022.3231276

I. Introduction

DIABETIC retinopathy (DR) is one of the most serious complications of diabetes and is currently the leading cause of blindness in adults [1], and has been identified by the World Health Organization as the second most serious eye disease after cataract. Owing to the safety and cost-effectiveness of acquiring fundus images, they are widely used for early screening and diagnosis of DR [2], [3]. However, due to the limitations of the acquisition equipment and the operation procedure, the fundus images often present significant differences with respect to the image quality as shown in Fig. 1. Automatic identification of lesions, such as microaneurysms (MAs) and hard exudates (EXs) are crucial to the diagnostic assessment of DR. The lower quality leads to the failure of the identification of the suspicious lesions, which decreases the diagnosis performance. Therefore, it is desirable to enhance the image quality for accurately capturing the lesions related to the severity grading.

Disease grading and image quality enhancement are two main fundamental tasks in this area. The image quality enhancement is required to be guided by the image quality assessment (IQA), the aim of which is to measure and control the quality of images. However, IQA is a subjective task depending on the experience of the ophthalmologists. The current solution turns it into learning, data-driven approaches based on neural networks. Through the thorough analysis of disease grading, image quality enhancement and quality assessment in Section IV, we believe that a joint framework incorporating the image quality assessment, image quality enhancement, and disease grading is feasible and significant, but to the best of our knowledge, no such work has been studied in this field. The challenges mainly lie in: (I) how to appropriately evaluate the quality of fundus images; (II) how to effectively enhance the low-quality fundus images, and (III) how to develop an end-to-end collaborative learning framework by integrating image quality enhancement subnetwork, image quality assessment subnetwork and DR subnetwork.

To solve these issues, we propose an image **Q**uality **A**ssessment guided **C**ollaborative **L**earning framework for both image quality **E**nhancing and DR grading, called CLEAQ-DR. The framework takes into account the image quality assessment and image quality enhancement during the DR grading. The underlying assumption is that under the guidance and help of the quality assessment and enhancement, the lesion's identification capability and DR grading performance can be improved. To better explore the potential relationships among the components of the quality assessment, the quality enhancement and the grading in the fundus images, we propose a collaborative learning framework to explore the potential correlation among these tasks and jointly train these subnetworks in a unified deep model for improving the individual performance. To this end, our collaborative learning framework incorporates three subnetworks: a DR disease grading subnetwork for predicting the DR level, a two-branch image quality enhancement (IQE) subnetwork for improving the image quality while preserving the fundus structure, and a two-branch image quality assessment (IQA) subnetwork for capturing the inherent low-quality indicators and predicting the quality level. Specifically, the IQE

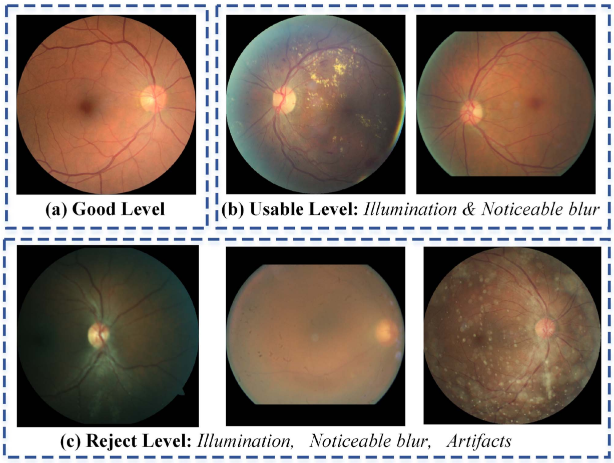

(a) Good Level
(b) Usable Level: Illumination & Noticeable blur
(c) Reject Level: Illumination, Noticeable blur, Artifacts

Fig. 1. Some examples of retinal images with different quality levels. (a) The fundus image of ‘Good’ level(high-quality) provides clear fundus structures and lesions location. (b) The ‘usable’ level (usable-quality) fundus images maintain the major fundus structures and lesions, but some diagnostic interferences are present in the images. (c) The ’Reject’ level(low-quality) fundus images are influenced by the low-quality indicators from uneven illumination, noticeable blurring and artifacts. The poor quality images make the DR diagnostic assessment task challenging. *The examples are from the EyeQ dataset [14].*

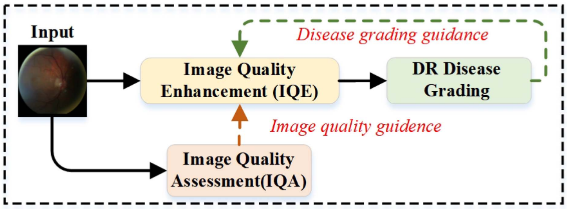


Fig. 2. The relationship among the IQA, IQE and DR Grading subnetworks in our study. The aim of IQE is to improve the quality under the guidance of the IQA and Grading subnetworks, while both DR Grading and IQA provide different quality criteria to guide the image quality enhancement process. By optimizing the three subnetworks jointly, we can achieve DR prediction and image quality enhancement in a unified deep model.

subnetwork consists of two encoder-decoder modules, where an image quality enhancement (U-IQE) module aims to learn the mapping relationship of low-quality images to high-quality images, and a retina vessel structure segmentation (Seg-IQE) module that aims to model the vessel structure to guarantee the preservation of the main fundus structure during the enhancement procedure. Moreover, the IQA subnetwork involves a classification (C-IQA) module for producing a reliable quality level, and an encoder-decoder (LQI-IQA) module for capturing the critical low-quality indicators by reconstructing the input images into the low-quality images. The image quality assessment, image quality enhancement and disease grading tasks are optimized in an end-to-end manner.

There are three notable characteristics for the CLEAQ-DR on the fundus retinal images.

1) *Modeling task relationship:* There exist inherent relationships among image quality assessment, image quality enhancement and DR grading tasks. Fig. 2 illustrates the inherent relationships among the three subnetworks in our framework. Appropriately modeling the task relationship allows to improve the performance of each task.

2) *Exploiting the image quality from different aspects:* To comprehensively guide IQE to improve the image quality, the DR and IQA tasks focus on the image quality from the lesion level and the global image level.

3) *Collaborative Learning:* To better reinforce each task, it is necessary to jointly train the DR grading, IQA and IQE tasks within a unified framework. With such an end-to-end trainable framework, our study establishes the association among the tasks of image quality assessment, enhancement and DR grading by collaborating the three subnetworks for better recovering the image quality and facilitating the precise localization of lesions.

In summary, our contributions can be summarized as follows.

- A major limitation of most current automatic DR grading models is that they ignore the effect of the image quality on the grading performance. To the best of our knowledge, our work is the first attempt to simultaneously perform multiple tasks: image quality assessment, image quality enhancement and disease diagnosis through an end-to-end collaborative learning framework. Considering that image quality assessment is essential for DR grading, our study establishes the association among the quality assessment, the quality enhancement and DR disease grading.
- We propose a two-branch encoder-decoder image enhancement subnetwork for improving low-quality images while preserving major retinal structures for avoiding the distortion occurrence, which helps to improve the DR diagnosis performance. Moreover, we propose a two-branch image quality assessment subnetwork for assessing the quality and guiding the enhancement process. The module can learn the inherent low-quality indicators for enhancing the assessment performance. Both subnetworks can be easily extended to other tasks related to the low-quality medical images.
- Experiment results on two benchmark datasets (Messidor and EyeQ) demonstrate that our approach leads to a significant performance boost over existing networks for DR grading and quality enhancement, notably on the EyeQ dataset that contains a large number of low-quality images. Moreover, we thoroughly analyze the potential correlation among the tasks of image quality assessment, image quality enhancement and DR grading through a series of experiments, demonstrating that these three tasks can benefit from each other. The proposed joint learning framework CLEAQ-DR can be broadly applied to other tasks of medical images with low-quality in general. To make the framework more general, we also further evaluate the enhanced results of low-quality images in more downstream tasks, such as vessel segmentation.

II. Related Work

Our work relates to three research areas: (a) DR disease grading, (b) image quality assessment, and (c) image quality enhancement. We discuss closely related work for each part.

DR disease Grading: In recent years, deep learning approaches have achieved immense success on DR diagnosis and screening [2], [3]. Compared to the shallow models such as kernel machines, deep neural networks have the potential to learn hierarchical representations of the fundus images. Existing deep learning-based DR grading methods can be divided into two categories [19], [21], [25]. The first category is to train a DR grading model for distinguishing the disease severity with the image-level grading label. For example, Zhou et al. [22] proposed the prediction of DR severity by both classification and regression methods based on the relationship between multi-stage images. Wang et al. [25] proposed a hierarchical multi-task learning framework, which accomplished the high-level DR grading by the low-level task of image super-resolution reconstruction and the middle-level task of lesion segmentation. In contrast to the image-level grading approaches that consider the entire fundus image as input, another category is to determine DR grading by identifying the location information of the DR-related lesions, e.g., microaneurysms, hemorrhage. For instance, Wang et al. [21] designed Zoom-in-Net which mimics the magnification process of ophthalmologists to examine fundus images, and generates attention maps highlighting suspicious lesion regions for DR grading. Huang et al. [23] suggested a contrastive learning approach based on the lesion patches to learn highly discriminative representations for

DR grading. In clinical conditions, there are unavoidably interference indicators on many fundus images that affect DR grading, such as uneven illumination, noticeable blur, and artifacts. However, none of the above studies considered how to deal with the interference indicators present in fundus images.

Image Quality Assessment: Image quality assessment (IQA) methods mainly include structure-based methods [4] and feature-based methods [5], [6], [7]. The structure-based approaches utilize segmented structures to determine the quality of fundus images. For example, Köhler et al. [4] recognized the quality of fundus images by using vascular structures. However, the performance of the structure-based approaches is heavily dependent on the segmentation of the fundus structures and cannot directly capture the potential visual features of the fundus images. In contrast to the structure-based approaches, the feature-based approaches evaluate the quality of fundus images by extracting feature representations directly from the images.

Image Quality Enhancement: For image quality enhancement, it mainly consists of traditional machine learning methods [32], [34], [36], [38] and deep learning methods [26], [35]. As we know, the traditional approaches perform image enhancement based on image contrast normalization and contrast limited adaptive histogram equalization techniques [31], [37]. More recently, image reconstruction methods based on deep learning have been soon developed, such as low-light image enhancement [8], image deraining [9] and deblurring [10]. Eilertsen et al. [35] attempted to learn the mapping operator between high-quality images and low-quality images in an end-to-end manner based on the convolutional neural networks. However, these methods focus on generating globally realistic images, but the loacal lesion regions that are critical for clinical decisions are ignored.

From the above analysis and comparison of related work, further exploration of fundus image quality evaluation and enhancement is necessary and feasible for supporting the DR diagnostic grading, but there is no method considering these aspects in the fundus disease grading task.

III. Methodology

Our main goal is to develop a collaborative learning framework that can provide more accurate disease grading performance and improve the quality of retinal fundus images at the same time. In this section, we first describe the formulation of our tasks and introduce an overview of the proposed architecture. Then, we introduce the details of each subnetwork. Finally, we provide the objective function of the CLEAQ-DR framework.

A. Formulation

Conceptually, as shown in the Fig. 3, the input image set $\boldsymbol{X}$ commonly contains three quality levels, including a low-quality image set $\boldsymbol{X}^L$, a usable-quality image set $\boldsymbol{X}^U$ and a high-quality image set $\boldsymbol{X}^H$. Given original fundus images $\boldsymbol{X} = \{\boldsymbol{X}^H, \boldsymbol{X}^U, \boldsymbol{X}^L\}$, pseudo structure masks $\boldsymbol{X_P} = \{\boldsymbol{X_P^H}, \boldsymbol{X_P^U}, \boldsymbol{X_P^L}\}$, the associated DR grading labels $\boldsymbol{Y} = \{\boldsymbol{Y}^H, \boldsymbol{Y}^U, \boldsymbol{Y}^L\}$ and quality labels $\widetilde{\boldsymbol{Y}} = \{\widetilde{\boldsymbol{Y}}^H, \widetilde{\boldsymbol{Y}}^U, \widetilde{\boldsymbol{Y}}^L\}$, the training procedure of the CLEAQ-DR framework in our study can be formulated as follows.

At first, besides the original images $\boldsymbol{X}$, the corresponding low-quality images $\widetilde{\boldsymbol{X}} = \{\boldsymbol{QT_{op2}}(\boldsymbol{X^H}), \boldsymbol{QT_{op1}}(\boldsymbol{X^U}), \boldsymbol{X}^L\}$ are obtained by the image quality transforming module $\boldsymbol{QT}(\cdot)$. To pre-train the IQE subnetwork $\boldsymbol{QE}(\cdot)$ in stage1, a pre-trained learning strategy is designed to achieve quality enhancement according to (1).

$$\min_{\theta_E} \sum_{n=1}^{N_H} L_{IQE}\left(QE\left(QT\left(x_n^H\right)\right), x_n^H, x_{pn}^H\right) \tag{1}$$

where $x_n^H \in \boldsymbol{X}^H$, $x_{pn}^H \in \boldsymbol{X_P^H}$, N_H denote the total number of the high-quality images, θ_E denotes the learnable parameters of the IQE subnetwork, and $L_{IQE}(\cdot,\cdot,\cdot)$ denotes the loss of IQE given the inputs of enhanced images, the original high-quality images and the pseudo fundus structure masks.

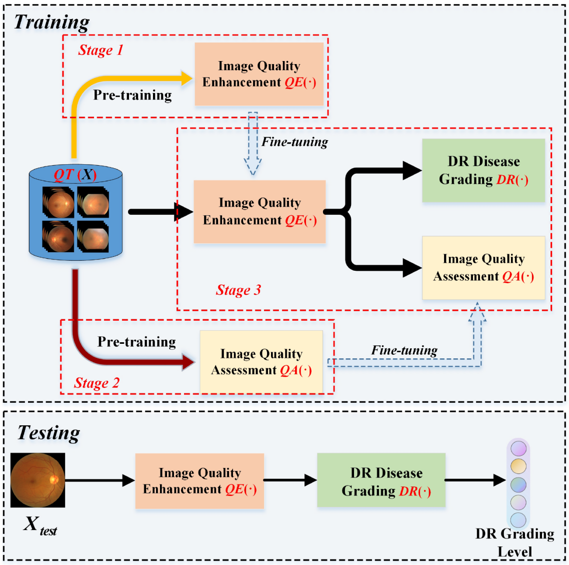


Fig. 3. The diagram of the collaborative learning framework.

Similarly, the pre-training of the IQA subnetwork $\boldsymbol{QA}(\cdot)$ in stage2 can be formulated as:

$$\min_{\theta_A} \sum_{n=1}^{N} L_{IQA}\left(QA\left(x_n\right), QT\left(x_n\right), \widetilde{y}_n\right) \tag{2}$$

where $x_n \in \boldsymbol{X}, \widetilde{y}_n \in \widetilde{\boldsymbol{Y}}$ denotes the IQA label, containing three classes in total, θ_A denotes the learnable parameters of IQA, and $L_{IQA}(\cdot,\cdot,\cdot)$ denotes the loss of IQA given the inputs of the original images, transformed low-quality images and quality labels.

The DR grading subnetwork $\boldsymbol{DR}(\cdot)$ is defined as:

$$\min_{\theta_{DR}} \sum_{n=1}^{N} L_{DR}\left(DR\left(QE\left(QT\left(x_n\right)\right)\right), y_n\right) \tag{3}$$

where $x_n \in \boldsymbol{X}$, and $y_n \in \boldsymbol{Y}$ is the DR grading label of image x_n with one of the five class labels. The input images are labeled as one of five classes [DR-0,...,DR-4] depending on the severity of the disease. θ_{DR} denotes the learnable parameters of $\boldsymbol{DR}(\cdot)$, and $L_{DR}(\cdot,\cdot)$ denotes the loss of $\boldsymbol{DR}(\cdot)$ given the inputs of the enhanced images and DR grading labels. The objective function and the procedure of joint training (Stage3) are described in ***Part F*** of this section.

B. Overview of the CLEAQ-DR Framework

To produce a more accurate DR diagnosis prediction, we propose a collaborative learning framework, called CLEAQ-DR, which is capable of transforming the low-quality fundus images into the usable-quality or high-quality level fundus images, and generating more accurate and reliable disease diagnosis predictions. The overall pipeline of the CLEAQ-DR framework is illustrated in Fig. 4. The training scheme for our CLEAQ-DR framework consists of three stages: the individual pre-training of the IQA subnetwork and the IQE subnetwork, and the joint learning of the three subnetworks.

In the pre-training phase of the IQA subnetwork, it is beneficial to better simulate the disturbances suffered in clinical scenarios, which facilitates the pre-training of the IQA subnetwork. There are three types of common quality degradation interferences, including inadequate illumination, noticeable blur and artifact. Three corresponding filters are defined in [26]. Following it, we design two different image quality degradation operations $\boldsymbol{op1}$ and $\boldsymbol{op2}$ for transforming the image quality. The degradation operation of $\boldsymbol{op1}$ is to randomly select one or two types of degradation interferences to perform quality degradation on the fundus images $\boldsymbol{X}^U$. The degradation operation of $\boldsymbol{op2}$ is to

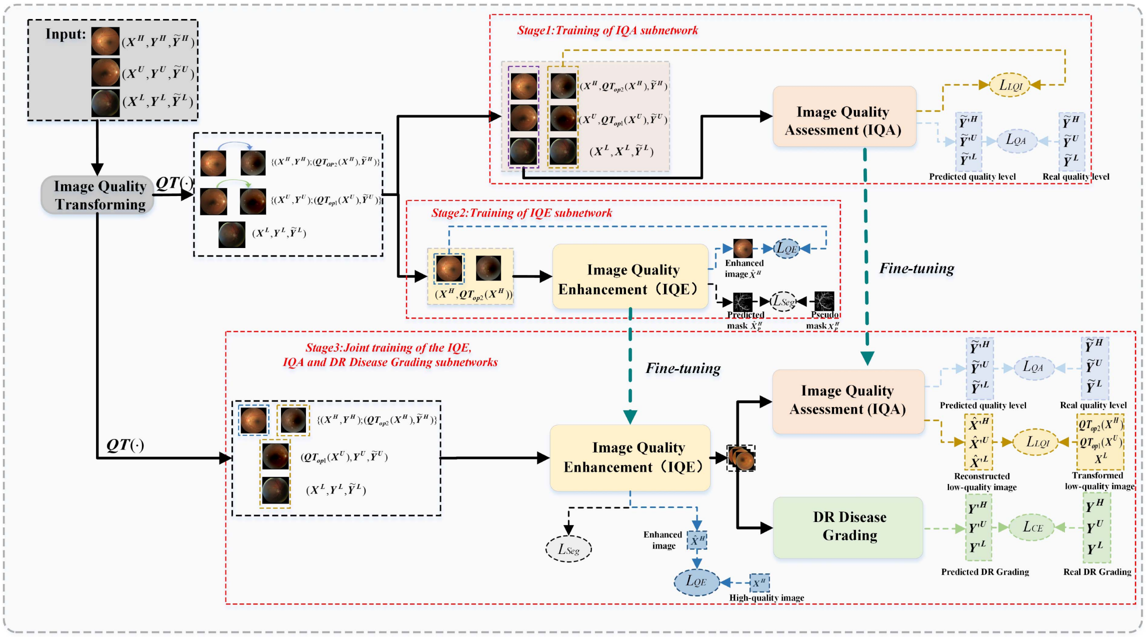

Input:
Image Quality Transforming
QT(·)
Stage1:Training of IQA subnetwork
Image Quality Assessment (IQA)
Predicted quality level
Real quality level
Stage2:Training of IQE subnetwork
Image Quality Enhancement（IQE）
Enhanced image
Predicted mask
Pseudo mask
Fine-tuning
Stage3:Joint training of the IQE, IQA and DR Disease Grading subnetworks
Image Quality Assessment (IQA)
Image Quality Enhancement（IQE）
DR Disease Grading
Reconstructed low-quality image
Transformed low-quality image
Predicted DR Grading
Real DR Grading
Enhanced image
High-quality image

Fig. 4. Illustration of the collaborative learning framework. Note that we need the IQA subnetwork for cooperating with IQE and DR Grading subnetworks as guidance during the joint training. The input images are divided into three groups according to the image quality labels: $\boldsymbol{X}^H$, $\boldsymbol{X}^U$, and $\boldsymbol{X}^L$. All images except $\boldsymbol{X}^L$ are first fed into the image quality transformation module to obtain the corresponding low-quality images. **Stage 1:** With the low-quality fundus images $\tilde{\boldsymbol{X}}$ by image quality transformation, we pre-train the IQA sub-network with the training data triplet $(\boldsymbol{X}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$. **Stage 2:** Following the same pre-training procedure as the IQA subnetwork, we pre-train the IQE subnetwork with the fundus structure pseudo mask $\boldsymbol{X}_P^H$ and the image pairs (the transformed low-quality images $\tilde{\boldsymbol{X}}^{H\to L}$ by the $\boldsymbol{QT_{op2}}(\cdot)$ and the corresponding high-quality images $\boldsymbol{X}^H$). **Stage 3:** Once stage1 and stage2 are independently trained, we collaboratively train the entire framework in an end-to-end manner. When predicting the severity level of the unseen samples, only the IQE and DR grading subnetworks are jointly utilized to enhance the image quality of the input images and produce the DR severity prediction.

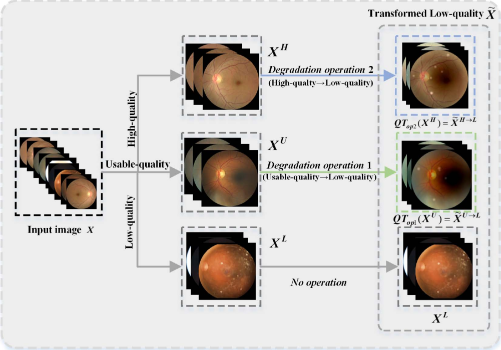

Transformed Low-quality $\tilde{X}$
Input image X
High-quality
Usable-quality
Low-quality
X^H
Degradation operation 2
(High-qualty→Low-quality)
$QT_{op2}(X^H)=\tilde{X}^{H\to L}$
X^U
Degradation operation 1
(Usable-quality→Low-quality)
$QT_{op1}(X^U)=\tilde{X}^{U\to L}$
X^L
No operation
X^L

Fig. 5. Image quality transforming for end-to-end training CLEAQ-DR framework.

perform quality degradation on the fundus images $\boldsymbol{X}^H$ with three types of degradation interferences.

Concretely, for both the transformed original high-quality images $\boldsymbol{X}^H$ and usable-quality $\boldsymbol{X}^U$, we perform the associated quality transformation on them and generate the corresponding low-quality images $\widetilde{\boldsymbol{X}}^{H\to L}$ and $\widetilde{\boldsymbol{X}}^{U\to L}$ with different degrees of image degradation operation as shown in Fig. 5. For low-quality fundus images, we do not perform any image degradation operation. The image quality transforming $\boldsymbol{QT}(\cdot)$ can be formulated as:

$$QT_{op2}\left(\boldsymbol{X}^H\right) = \widetilde{\boldsymbol{X}}^{H\to L}, QT_{op1}\left(\boldsymbol{X}^U\right) = \widetilde{\boldsymbol{X}}^{U\to L} \qquad (4)$$

C. Fundus Images Quality Enhancement Subnetwork, IQE

Medical fundus images acquired from different types of equipment have significant variations in quality. The low-quality fundus images with noticeable blur, low contrast and insufficient illumination lead to the inaccurate diagnosis by ophthalmologists or automated clinical diagnostic systems. In addition, the structure of the high-quality fundus images transformed by the traditional natural image enhancement algorithms is seriously distorted due to the fine-grained characteristics in the fundus images. Therefore, it is crucial to improve the quality of the fundus images while preserving the fundus structure for disease diagnosis and analysis.

Specifically, we take the transformed low-quality fundus images $\widetilde{\boldsymbol{X}}$ as inputs to train IQE in a fully supervised manner. For the quality enhancement module(U-IQE) in Fig. 6, the encoding phase is used to encode input images $\widetilde{\boldsymbol{X}}$ in a lower dimensionality. For each encoder layer, we use a residual block followed by a max pooling layer. Finally, the symmetric decoding phase is designed to the inverse process of encoding, which enables to generate the enhanced fundus image $\hat{\boldsymbol{X}}$ corresponding to the low-quality fundus images $\widetilde{\boldsymbol{X}}$. In addition, it is also used to integrate the structural information from the fundus structure segmentation module (Seg-IQE) into the output feature maps of the first three decoding blocks by concat operation. For each decoder layer, we employ a transposed convolutional layer followed by a residual block with a symmetric structure.

To preserve features of the fundus structure and guide the enhancement procedure, extracting multi-scale fundus structural information is essential for capturing complex scale variations in medical imaging enhancement. Hence, we introduce the Seg-IQE module to assist the U-IQE module. As illustrated in Fig. 6, we choose ResNet-34 as the

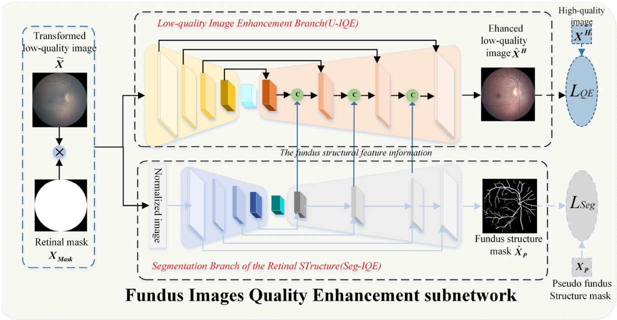

Transformed low-quality image $\widetilde{X}$
Retinal mask X_{Mask}
Low-quality Image Enhancement Branch(U-IQE)
Ehanced low-quality image $\hat{X}^H$
High-quality image X^H
L_{QE}
The fundus structural feature information
Normalized image
Fundus structure mask $\hat{X}_P$
L_{Seg}
X_P
Pseudo fundus Structure mask
Segmentation Branch of the Retinal STructure(Seg-IQE)
Fundus Images Quality Enhancement subnetwork

Fig. 6. Structure of the fundus image quality enhancement subnetwork in our collaborative learning framework.

feature encoder in the Seg-IQE module, and remove the average pooling layer and fully connected layer. In the decoding stage, we adopt the transposed convolution to restore the fundus structure features. More specifically, the decoder mainly consists of a 1×1 convolution, a 3×3 transposed convolution and a 1×1 convolution. With a series of skip connections and decoder blocks, the feature decoder blocks finally produce the fundus structure mask of the same size as the original input. In this way, we can obtain the multi-scale fundus structure information from different depths of the Seg-IQE module, which enables to assist the U-IQE module to enhance the quality of the low-quality fundus images $\widetilde{\boldsymbol{X}}$ while maintaining the major fundus structures.

In medical application, the pixel-level fundus structure annotations are usually not easily accessible in practical situations. To address this, many existing work [22], [24] commonly obtain pseudo-labeling by the pre-trained networks based on auxiliary datasets. To preserve the fundus structure of the enhanced images, we obtain the pixel-level pseudo structure mask $\boldsymbol{X}_P$ of the original image $\boldsymbol{X}$ via the CE-Net of a two-stage training scheme. More specifically, we first pre-train the CE-Net based on the original DRIVE dataset [16]. Then, considering the difference in quality between the low-quality datasets and the DRIVE dataset, we perform the same quality transformation operations on the DRIVE dataset to obtain the corresponding low-quality and usable-quality fundus images. The pre-trained CE-Net is further fine-tuned on the transformed DRIVE dataset to enhance the reliability of the pseudo fundus structure masks. Finally, we apply the trained CE-Net to obtain the pseudo fundus structure masks of as the supervision of the IQE subnetwork. The segmentation loss of fundus structure allows our model to focus on the regions of fundus structure in the input images $\boldsymbol{X}$. Hence, for the Seg-IQE module, we adopt the Dice coefficient loss as our fundus structure segmentation loss $\mathcal{L}_{Seg}$.

$$\mathcal{L}_{Seg} = 1 - \frac{2\left\|\hat{\boldsymbol{X}}_P \circ \boldsymbol{X}_P\right\|_1}{\left\|\hat{\boldsymbol{X}}_P\right\|_1 + \left\|\boldsymbol{X}_P\right\|_1} \tag{5}$$

where $\hat{\boldsymbol{X}}_{\boldsymbol{P}}$ is the fundus segmentation result of the Seg-IQE module, $\boldsymbol{X}_{\boldsymbol{P}}$ indicates the obtained pseudo fundus structure masks by the CE-Net of a two-stage training scheme, and $\circ$ denotes the hadamard product.

For the U-IQE module of the IQE subnetwork, we adopt the widely-used L2 loss as the loss $\mathcal{L}_{QE}$ of the U-IQE module, and the loss function can be formulated as:

$$\mathcal{L}_{QE} = \left\|\boldsymbol{X}^H \otimes \boldsymbol{X}_{Mask} - \Psi\left(\widetilde{\boldsymbol{X}}, W_\Psi\right) \otimes \boldsymbol{X}_{Mask}\right\|_2^2 \tag{6}$$

where $\Psi(\cdot)$ represents the U-IQE module of the IQE subnetwork, W_Ψ denotes the learnable parameters of the U-IQE module, $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{X}^H$ denote the low-quality input images and the corresponding high-quality reference images, $\boldsymbol{X}_{Mask}$ denotes the retinal mask of the input images $\widetilde{\boldsymbol{X}}$, and $\otimes$ indicates an element-wise multiplication.

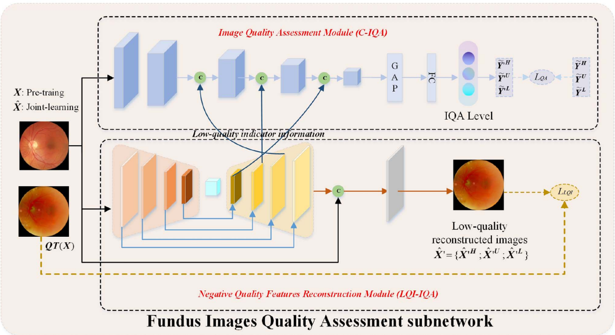


Image Quality Assessment Module (C-IQA)
X: Pre-traing
$\hat{X}$: Joint-learning
G A P
FC
$\widetilde{Y}'^H$ $\widetilde{Y}'^U$ $\widetilde{Y}'^L$
L_{QA}
$\widetilde{Y}^H$ $\widetilde{Y}^U$ $\widetilde{Y}^L$
IQA Level
Low-quality indicator information
$QT(X)$
Low-quality reconstructed images
$\hat{X}' = \{\hat{X}'^H; \hat{X}'^U; \hat{X}'^L\}$
L_{LQI}
Negative Quality Features Reconstruction Module (LQI-IQA)
Fundus Images Quality Assessment subnetwork

Fig. 7. Structure of the two-branch IQA subnetwork.

Once the IQE subnetwork is trained, an enhanced fundus image $\hat{\boldsymbol{X}}$('Usable' or 'Good' level) and a fundus structure mask $\hat{\boldsymbol{X}}_P$ are obtained.

D. Fundus Images Quality Assessment Subnetwork, IQA

Artifacts and noticeable blur in the low-quality input images are removed by the IQE subnetwork. To evaluate the enhancement performance of the IQE subnetwork, we propose a fundus image quality assessment subnetwork to evaluate the quality of enhanced images. With the joint training of both the IQE and IQA subnetworks, the assessment results of the enhanced images are fed back to the IQE subnetwork forcing IQE to remove as many low-quality indicators as possible from the low-quality input images.

A single quality classification in the IQA subnetwork can not well capture the potential quality indicators of the images with different quality levels during the training stage. Hence, we propose a classification module (C-IQA) for producing a reliable quality level and an encoder-decoder module (LQI-IQA) for capturing the critical low-quality indicators in a unified deep model as our IQA subnetwork. The proposed subnetwork is illustrated in Fig. 7. Specifically, we take the original fundus images $\boldsymbol{X}$ (pre-training stage) or enhanced fundus images $\hat{\boldsymbol{X}}$ (joint learning stage) by the IQE subnetwork as inputs to train the IQA subnetwork. We choose ResNet-50 as the classifier in the C-IQA module. We adopt the multi-class cross-entropy, which is expressed as:

$$\mathcal{L}_{QA} = -\frac{1}{N_L}\sum_{i=0}^{N_L}\sum_{k_q=0}^{K_q}\left(\widetilde{y}_{i,k_q} * \log\left(\widetilde{y}'_{i,k_q}\right)\right) \tag{7}$$

where N_L and K_q denote the number of the inputs images and the number of types of the image quality levels, $\widetilde{y}'_{i,k_q}$ and $\widetilde{y}_{i,k_q}$ indicate the predicted image quality probability and true image quality label of the k_q-th quality level.

The structure of the LQI-IQA module is similar to the U-IQE module. Specifically, the module first encodes the input images $\boldsymbol{X}$ or $\hat{\boldsymbol{X}}$ with 4 encoder blocks, and the decoder blocks map the latent representations back to the input for producing the low-quality representations when performing the image reconstruction task. The aim of the LQI-IQA is to capture the low-quality indicators for introducing the prior knowledge associated with the input image quality. The learned low-quality indicator representations are added to the feature maps of the C-IQA module. Through transmitting the multi-scale low-quality indicator representations from LQI-IQA to assist C-IQA, IQA is capable of appropriately evaluating the image quality.

This is formulated as:

$$\mathcal{Y}^\ell = \mathcal{T}_{\text{IQA}}\left(\text{Concat}\left[\mathcal{F}_{\text{C}}(\hat{\boldsymbol{X}}), \mathcal{F}_{\text{LQI}}(\hat{\boldsymbol{X}})\right]\right) \tag{8}$$

where $\mathcal{T}_{\text{IQA}}$ represents a learnable non-linear transformation filter, $\mathcal{F}_{\text{C}}$ and $\mathcal{F}_{\text{LQI}}$ indicate the feature extracted from the C-IQA module and the LQI-module of the IQA subnetwork, respectively.

To be specific, when the input $\boldsymbol{X}$ or $\hat{\boldsymbol{X}}$ is the high-quality fundus image, the decoder contains plenty of the potential low-quality feature information due to the reconstructed image being low-quality. By contrast, when the input $\boldsymbol{X}$ or $\hat{\boldsymbol{X}}$ is the low-quality fundus image, the decoder contains little low-quality feature information. Therefore, the decoder features containing potential low-quality indicators are incorporated into the C-IQA module to improve the quality assessment performance. Furthermore, the low-quality indicators naturally contain patterns of different scales that may be unknown in advance, we therefore propose a multi-scale strategy to sufficiently capture them. Meanwhile, we develop a low-quality reconstruction L_{LQI} to better capture the low-quality indicators associated with the input image quality in the pixel space. As shown in (9), the loss function $\mathcal{L}_{LQI}$ is applied to compute the mean squared error (MSE) between the output of the LQI-IQA module and the low-quality images $\widetilde{X}$. By tightly integrating the encoder-decoder structure and the classification module, the potential low-quality factors aware features can be captured.

$$\mathcal{L}_{LQI} = \left\| \widetilde{\boldsymbol{X}} \otimes \boldsymbol{X}_{Mask} - \Phi\left(\Psi\left(\widetilde{\boldsymbol{X}}, W_{\Psi}\right), W_{\Phi}\right) \otimes \boldsymbol{X}_{Mask} \right\|_2^2 \quad (9)$$

where $\Phi(\cdot)$ and W_{Φ} represent the LQI-IQA module and its learnable parameters. $\boldsymbol{X}_{Mask}$ denotes the retinal mask of the input images.

E. DR Disease Grading Subnetwork

The goal of DR grading is to predict the classification label for the disease severity. ResNet-50 is chosen as the backbone in the DR disease grading subnetwork. Given the enhanced low-quality fundus images $\hat{\boldsymbol{X}}$ obtained by the IQE subnetwork as the inputs, the DR grading subnetwork is trained to produce the severity level. We employ the multi-class cross-entropy loss as the DR grading loss $\mathcal{L}_{DR}$, which is formulated as:

$$\mathcal{L}_{DR} = -\frac{1}{N} \sum_{i=0}^{N} \sum_{l_g=0}^{L_g} \left(y_{i,l_g} * \log \left(y'_{i,k_g} \right) \right) \quad (10)$$

where N and L_g denote the number of the fundus images and the number of the DR levels, y_{i,l_g} and y'_{i,l_g} indicate the real label and the predicted probability of the l_g-th level.

F. The Overall Loss Function.

The learning objective of training our CLEAQ-DR consists of: a) a reconstruction loss for encouraging the output $\hat{\boldsymbol{X}}^H$ of the IQE subnetwork close to the high-quality reference images $\boldsymbol{X}^H$, b) a fundus structures segmentation loss for restoring realistic fundus structure textures, c) a DR grading loss for retaining the features of the lesion areas in the enhanced fundus images and optimizing the DR grading diagnosis subnetwork, d) the C-IQA module loss in the IQA subnetwork for evaluating the performance of the IQE subnetwork as well as constraining the IQE subnetwork for removing more low-quality indicators, and e) the low-quality indicators reconstruction loss for further enhancing the evaluation performance of the IQA subnetwork. Given all the loss functions, the overall loss function for our collaborative learning framework can be defined as:

$$\begin{aligned} \mathcal{L}_{IQE} &= \mathcal{L}_{QE} + \lambda_{Seg}\mathcal{L}_{Seg} \\ \mathcal{L}_{IQA} &= \mathcal{L}_{QA} + \lambda_{LQI}\mathcal{L}_{LQI} \\ \mathcal{L}_{\text{CLEAQ-DR}} &= \mathcal{L}_{DR} + \lambda_{IQE}\mathcal{L}_{IQE} + \lambda_{IQA}\mathcal{L}_{IQA} \end{aligned} \quad (11)$$

where $\mathcal{L}_{IQE}$, $\mathcal{L}_{IQA}$ and $\mathcal{L}_{DR}$ denote the individual loss of the IQE subnetwork, the IQA subnetwork and the DR subnetwork, and λ_{Seg}, λ_{LQI}, λ_{IQE} as well as λ_{IQA} are the regularization weights that balance the losses of different components.

TABLE I
SUMMARY OF THE EYEQ AND MESSIDOR DATASETS

	EyeQ Training set					
	DR-0	DR-1	DR-2	DR-3	DR-4	All
high-quality	6,342	699	1,100	167	39	8,347
usable-quality	1,353	103	283	79	58	1,876
low-quality	1,544	109	426	87	154	2,320
Total	9,239	911	1,809	333	251	**12,543**
	EyeQ Testing set					
	DR-0	DR-1	DR-2	DR-3	DR-4	All
high-quality	5,966	886	1,354	199	65	8,470
usable-quality	3,201	359	721	145	133	4,559
low-quality	2,195	153	569	104	199	3,220
Total	11,362	1,398	2,644	448	397	**16,249**
	Messidor Training/Testing set					
	DR-0	DR-1	DR-2	DR-3		All
high-quality	152 / 22	45 / 18	77 / 12	58 / 16		332 / 68
usable-quality	148 / 18	50 / 13	66 / 21	63 / 21		327 / 73
low-quality	161 / 19	41 / 12	64 / 12	75 / 16		341 / 59
Total	461 / 59	136 / 43	207 / 45	196 / 53		**1,000 / 200**

IV. EXPERIMENT RESULTS

A. Datasets and Performance Metrics

In our experiments, we evaluate the effectiveness of our method by comparing it against existing works on the Messidor dataset and the Eye-Quality (EyeQ) dataset. The information of DR both datasets are shown in Table I.

Eye-Quality (EyeQ) dataset [14]: The EyeQ dataset from the Eye-sPACs dataset [15] is a large-scale public benchmark for fundus image quality assessment and DR grading, which consists of 28,792 fundus images with their IQA labels and DR grading labels.

Messidor dataset [11]: The dataset collects 1200 fundus posterior polar color digital images from three ophthalmology departments. For each image in the dataset, medical experts provide its DR grading annotation, which is used to measure the grade of diabetic retinopathy. DR is divided into four levels based on the severity scale according to different criterions. Similarly, considering the common quality metrics such as blur, uneven illumination, low contrast and artifacts, we divided the Messidor dataset into three quality levels based on image quality transformation.

There are various performance evaluation metrics used with the purpose of quantitatively evaluating the performance of the different task subnetworks in the CLEAQ-DR framework. First, for the DR diagnosis grading subnetwork as well as the IQA subnetwork, we introduced the quadratic weighted Kappa metric [3] in addition to the normal classification accuracy. Second, for the IQE subnetwork, the structural similarity index (SSIM) [18] and the peak signal-to-noise ratio (PSNR) are used as performance measures.

B. Implementation Details

Considering the diverse and large-sized fundus images in the EyeQ and Messidor datasets, we normalize and resize them into 512 × 512 resolution to accelerate the model convergence. Besides, we detect the retinal mask of each input image using the Hough Circle Transform, and then crop the mask regions to reduce the effect of the black background. On the other hand, we employ different data augmentation strategies for each class to alleviate the class imbalance distribution during training. That is, the number of data augmentation depends on the sample number of each class. There are three stages for training the CLEAQ-DR framework. During the first stage, the DR grading subnetwork is pre-trained with DR severity labels, and a variety of data augmentation strategies including random rotation, horizontal flipping and vertical flipping are conducted on the input images. In the second stage, the IQA subnetwork is pre-trained with the quality labels and the transformed low-quality images. In stage3, the DR grading, IQA, and IQE subnetworks are simultaneously fine-tuned in an end-to-end manner.

In the joint training stage, the performance of the CLEAQ-DR framework is highly dependent on the appropriate choice of weights among the losses for all tasks. The different tasks need to be properly balanced, so that the CLEAQ-DR framework can converge to the state which is optimal for all the tasks. For the loss regularization weight values in (11), a naive approach is to assign each individual task with an equal weight. It is not appropriate because the multiple tasks to be optimized have different difficulty levels. The challenge is to find the best regularization weight value for each task at each training step that balances the contribution of each task. We consider it as a multi-task learning paradigm and assign different weights for different tasks. Based on the work of Liu et al. [12], we introduce a dynamic task weighting scheme into the optimization process of the CLEAQ-DR, which enables the entire framework to achieve balanced training automatically for multiple tasks by dynamically tuning gradient magnitudes. The weight of each task changes in every batch. Therefore, for each task in a batch, the optimization considers the loss ratio between the current loss and the initial loss, which measures how well the model has trained for that task. Nevertheless, it is worth noting that the aim of the IQE subnetwork is to maximize the loss $\mathcal{L}_{IQA}$, whereas the aim of the IQA subnetwork is to minimize the loss $\mathcal{L}_{IQA}$. To fit this goal, the IQE subnetwork performs the gradient reversal operation [13] given the gradient from the IQA network and passes it to the preceding layer during the backpropagation.

The whole training process involves 150 epochs in total. The learning rates of all stages are initialized to 1×10^{-3} in the first 60 epochs, and then automatically adjusted in the following epochs for each stage according to the epoch number. The learning rate is multiplied by $(1-\frac{epoch_i}{epoch_{max}})^{\mathcal{T}}$ with $\mathcal{T}$=0.9. The Adam optimizer is used to update the parameters for all stages, and the batch size is set to 32. The CLEAQ-DR is trained using PyTorch with 4 NVIDIA Quadro RTX 6000 GPUs.

C. Comparisons With State-of-the-Art Methods

In this section, we conduct relevant experiments to evaluate the proposed CLEAQ-DR framework on the EyeQ and Messidor datasets. The purpose of our experiments is to investigate the following research questions:

$Q1$. How does CLEAQ-DR's disease grading and image enhancement performance compare to the most advanced methods?

$Q2$. How does the proposed joint learning framework help with the individual task?

$Q3$. Is the Seg-IQE module in the IQE subnetwork beneficial to the improvement of low-quality fundus images?

$Q4$. Does the LQI-IQA module contribute to boosting the performance of quality assessment of the fundus images?

1) The Comparison on the DR Grading: To make our method more convincing, we empirically demonstrate CLEAQ-DR's effectiveness on two benchmark datasets and compare it with state-of-the-art DR grading methods. We evaluate the DR grading performance of our proposed CLEAQ-DR framework with three types of comparable methods: covering the popular networks (e.g. ResNet-50, Inception-v3 and DenseNet-121), the top three places of Kaggle challenge (Min-pooling [19], o_O [19] and Reformed Gamblers [19]) and the current state-of-the-art DR grading models: MMCNN [22], Zoom-in-Net [21], Lesion-base CL [23] and DeepMT-DR [25]. For the grading results of the method [15], the implementation details and source codes are not published. Therefore, the ACC value is absent in the Table. Experimental results are reported in Table II where the best results are boldfaced.

As shown in Table II, CLEAQ-DR consistently achieves the best results across both datasets with respect to both the ACC and Kappa metrics. The improvement demonstrates that CLEAQ-DR presents a notably better DR grading performance than the state-of-the-art methods due to the incorporation of the image quality assessment and enhancement, which are beneficial for the DR grading subnetworks by removing the artifacts, unbalanced illumination, and other diagnostic interferences while preserving lesion characteristics for the low-/usable quality images.

TABLE II
THE COMPARISON BETWEEN OUR METHOD WITH THE STATE-OF-THE-ART METHODS FOR DR GRADING ON EYEQ AND MESSIDOR DATASETS

Methods	EyeQ		Messidor	
	Acc	Kappa	Acc	Kappa
ResNet-50	0.804	0.783	0.672	0.653
Inception-v3	0.798	0.776	0.664	0.649
DenseNet-121	0.813	0.794	0.681	0.657
Reformed Gamblers [19]	/	0.839	/	/
Min-pooling [19]	/	0.849	/	/
o_O [19]	/	0.844	/	/
MMCNN [22]	0.862	0.841	0.692	0.673
Zoom-in-Net [21]	0.873	0.854	0.714	0.694
Lesion-base CL [23]	0.848	0.832	0.687	0.662
DeepMT-DR [25]	0.857	0.839	0.694	0.671
CLEAQ-DR	**0.885**	**0.863**	**0.736**	**0.712**

TABLE III
THE COMPARISON BETWEEN OUR METHOD WITH THE SOTA METHODS FOR LOW-QUALITY FUNDUS IMAGES ENHANCEMENT ON EYEQ DATASETS

Methods	EyeQ		Methods	EyeQ	
	PSNR	SSIM		PSNR	SSIM
Tian et al. [33]	14.71	0.664	Li et al. [38]	9.47	0.547
LIME [31]	13.54	0.868	Fu et al. [37]	14.66	0.882
Fu et al. [32]	9.76	0.564	He et al. [36]	15.56	0.749
Cheng et al. [34]	15.02	0.845	cofe-Net [26]	20.51	0.885
Eilertsen et al. [35]	18.40	0.841	CycleGAN [28]	21.65	0.843
I-SECRET [30]	27.36	0.908	CutGAN [29]	22.76	0.872
cGAN [27]	26.35	0.894	CLEAQ-DR	**28.53**	**0.917**

2) The Comparison on the Low-Quality Image Enhancement: To make our method more convincing, we also empirically demonstrate the effectiveness of the CLEAQ-DR framework in enhancing low-quality fundus images on two benchmark datasets and compare it with the state-of-the-art image enhancement methods. The comparison includes a series of deep learning methods: Eilertsen et al. [35], cGAN [27], CutGAN [29], CycleGAN [28], I-SECRET [30], and cofe-Net [26]. Moreover, we compared the traditional image correction approaches: LIME [31], distribution fitting algorithm [32], Tian et al. [33], variational frameworks [38], latent structure-driven methods [34], [36], Fu et al. [37]. Experimental results are reported in Table III where the best results are boldfaced.

From Table III, we can clearly observe that our method has a remarkable advantage compared with the traditional image correction approaches and deep learning methods in terms of PSNR metrics. By collaborating with the IQA subnetwork, it is able to guide the IQE subnetwork to remove as many diagnostic interferences as possible from the low-quality fundus images. In other words, the IQE subnetwork is viewed as a generator for generating the enhanced images whereas the IQA subnetwork can be considered as a discriminator which is used to evaluate and guide the enhancement optimization of the low-quality images to achieve a progressive refinement. Moreover, Fig. 8 shows the results generated by our CLEAQ-DR and several compared deep learning methods. It can be observed that our method remarkably enhances the low-quality fundus images, while preserving the complete fundus structure information.

To intuitively understand the enhancement effect of the CLEAQ-DR framework on low-quality fundus images, we applied Grad-CAM to visualize the vascular and pathological regions on the low-quality images and the corresponding enhanced images. Fig. 9 shows the original low-quality images and the corresponding higher-quality images generated by our CLEAQ-DR. As shown in Fig. 9, our method achieves image enhancement while preserving the major structural features in the fundus. This result highlights that the quality is obviously improved with the help of the IQA and DR grading subnetworks.

D. Ablation Study

To more comprehensively evaluate our model, we conduct ablation studies to analyze the correlation between different subnetworks.

1) The Ablation Study on the Low-Quality Image Enhancement: The CLEAQ-DR framework mainly involves three subnetworks: IQE subnetwork, IQA subnetwork, DR Disease grading subnet-

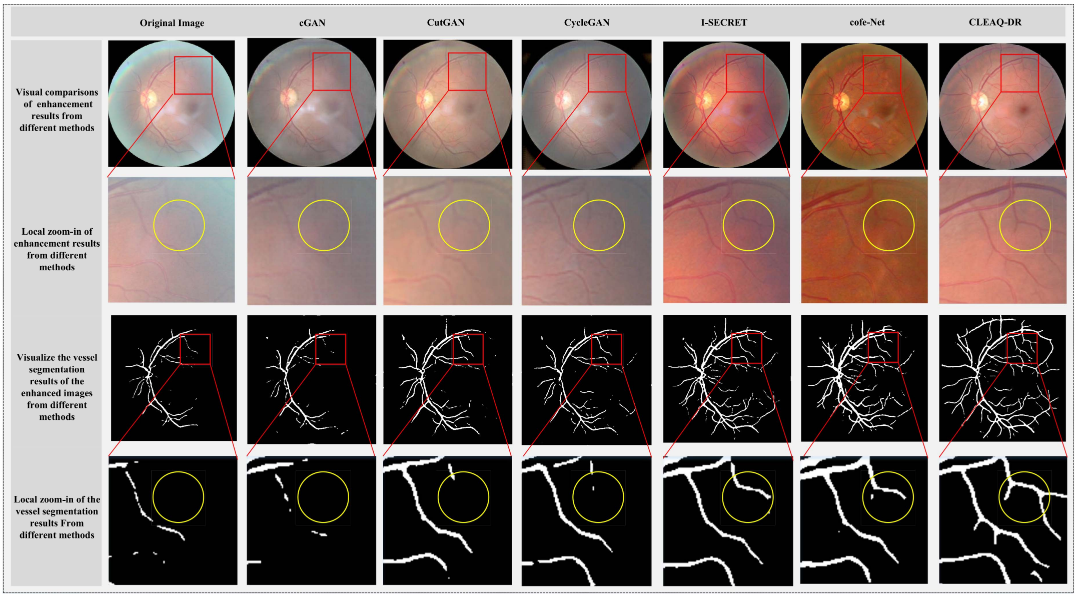
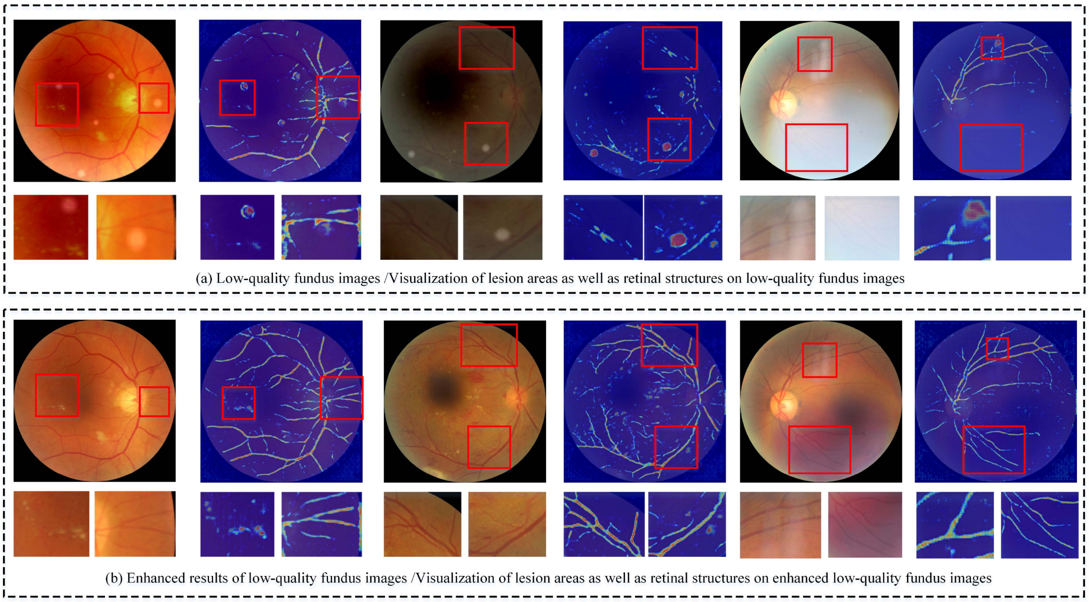

Original Image
cGAN
CutGAN
CycleGAN
I-SECRET
cofe-Net
CLEAQ-DR
Visual comparisons of enhancement results from different methods
Local zoom-in of enhancement results from different methods
Visualize the vessel segmentation results of the enhanced images from different methods
Local zoom-in of the vessel segmentation results From different methods

Fig. 8. Visual comparisons on the low-quality image enhancement between the CLEAQ-DR and other deep learning methods. In addition, we also visualize the vessel segmentation results of the enhanced images obtained by the comparable methods. The proposed CLEAQ-DR framework preserves more structural features of the fundus images, which can more effectively enhance the quality of low-quality fundus images.


(a) Low-quality fundus images /Visualization of lesion areas as well as retinal structures on low-quality fundus images

(b) Enhanced results of low-quality fundus images /Visualization of lesion areas as well as retinal structures on enhanced low-quality fundus images

Fig. 9. Visualization of retinal structures and lesion regions on low-quality images and corresponding enhanced images. The proposed method can significantly improve the low-quality fundus images, and the enhanced low-quality fundus images can provide richer information hidden in the retinal structures and lesions for DR grading.

work. To investigate the effectiveness of the components in the CLEAQ-DR framework for the IQE subnetwork, we compare CLEAQ-DR with its several variants, respectively.

IQE: The IQE sub-network is independently trained for enhancement of the low-quality fundus images.

IQE w/o Seg-IQE module: The IQE subnetwork is trained independently without the Seg-IQE module.

CLEAQ-DR w/o DR, called CLEAQ: The collaborative learning framework without the DR grading subnetwork.

CLEAQ-DR w/o IQA, called CLEQ-DR: The collaborative learning framework without the IQA subnetwork.

The results are summarized in Table IV. We find that CLEAQ-DR outperforms the contender methods in terms of PSNR and SSIM. These results reveal several interesting points:

1) IQE w/o Seg-IQE shows the worst performance among the methods for almost all datasets and metrics. Our results suggest that imposing the fundus structural segmentation module during the quality enhancement training is crucial for improving quality enhancement.
2) CLEAQ performs worse than CLEAQ-DR. The reason is that collaborated with DR grading subnetwork, and the

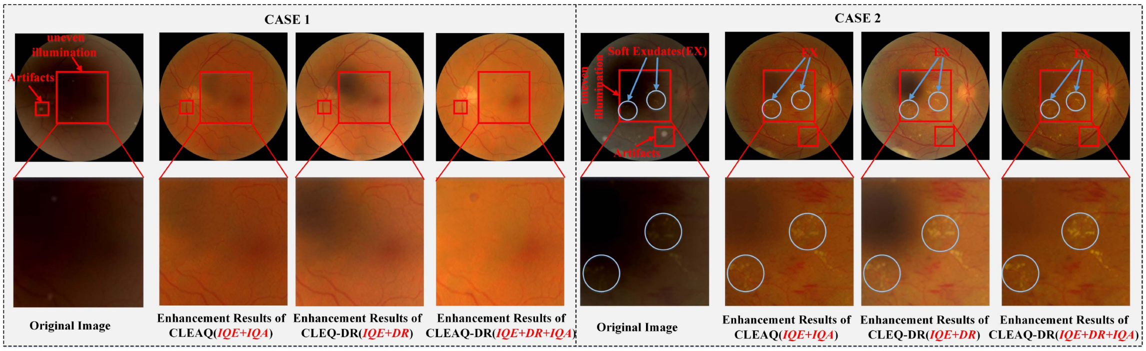

CASE 1
uneven illumination
Artifacts
Original Image
Enhancement Results of CLEAQ(IQE+IQA)
Enhancement Results of CLEQ-DR(IQE+DR)
Enhancement Results of CLEAQ-DR(IQE+DR+IQA)
CASE 2
Soft Exudates(EX)
uneven illumination
Artifacts
EX
EX
EX
Original Image
Enhancement Results of CLEAQ(IQE+IQA)
Enhancement Results of CLEQ-DR(IQE+DR)
Enhancement Results of CLEAQ-DR(IQE+DR+IQA)

Fig. 10. Quality enhancement results of our CLEAQ-DR and its variants. We crop and zoom-in the uneven illumination regions to compare the performance of related methods. The proposed CLEAQ-DR framework enhances low-quality images containing more details.

TABLE IV
THE ABLATION EXPERIMENT RESULTS OF IQE SUBNETWORK

Methods	Messidor		EyeQ	
	PSNR	SSIM	PSNR	SSIM
IQE	17.85	0.758	22.36	0.813
IQE w/o Seg-IQE	16.23	0.732	21.25	0.805
CLEAQ	19.36	0.805	25.43	0.872
CLEQ-DR	18.42	0.783	23.32	0.834
CLEAQ-DR	**21.97**	**0.852**	**28.53**	**0.917**

TABLE V
THE ABLATION EXPERIMENT RESULTS OF DR GRADING SUBNETWORK

Dataset	Methods	Acc	Kappa	Precision	F1-Score
Messidor	DR grading	0.672	0.653	0.678	0.661
	CLEQ-DR	0.712	0.687	0.718	0.694
	CLEAQ-DR	**0.736**	**0.712**	**0.739**	**0.718**
EyeQ	DR grading	0.804	0.783	0.812	0.791
	CLEQ-DR	0.853	0.842	0.861	0.849
	CLEAQ-DR	**0.885**	**0.863**	**0.887**	**0.872**

IQE subnetwork can be guided to discover low-quality indicators. Lack of this collaboration hinders the identification of lesions. The DR grading subnetwork helps the IQE subnetwork to focus on relevant class-specific regions in the images. The performance also decreases IQA guidance, which indicates that the image quality assessment is an important task to provide quality criteria for guiding the image quality enhancement process. In addition, CLEAQ performs worse than CLEA-DR, which demonstrates that the IQA subnetwork has a more significant contribution to the IQE subnetwork compared to the DR grading subnetwork.

3) CLEAQ-DR achieves the best PSNR and SSIM, which verifies the significance of the collaborative learning again.

As shown Fig. 10, we also demonstrate the quality enhancement results of the CLEAQ-DR framework and its two main variants (CLEAQ and CLEQ-DR) for low-quality fundus images. From Fig. 10, it can be observed that the tasks of DR and IQA focus on improving image quality from different aspects. The task of DR diagnosis aims to improve the local image quality from the lesion level, e.g. EXs, under the DR grading supervision, whereas the task of IQA aims to enable that IQE generates higher quality images from the global image level, e.g. uneven illumination, noticeable blurring and artifacts. The image quality enhancement of the CLEAQ-DR can be greatly improved when jointly training the different tasks.

2) The Ablation Study on the DR Disease Grading: In the case of ablation experiments of the DR disease grading, we compare the single DR grading subnetwork and CLEQ-DR on the EyeQ and Messidor datasets.

As shown in Table V, it is obvious that our proposed method shows improvements upon just the DR grading subnetwork, which further confirms our hypothesis that the performance of automated diagnostic systems is highly dependent on image quality. Furthermore, the DR grading performance of CLEAQ-DR greatly degrades when the IQA subnetwork is removed. These results verify the significance of quality assessment in the DR grading on the dataset with a large number of the low-quality images, even on the Messidor dataset where the majority of images are high-quality. The reason is that our collaborative learning framework encourages to further improve the quality under the guidance of the classification.

We further calculate the confusion matrices of CLEAQ-DR, CLEQ-DR and DR grading. As can be seen from Fig. 11, CLEAQ-DR shows the best DR grading performance among the methods. Compared with the DR grading method, CLEAQ-DR and CLEQ-DR exhibit better grading performance, which is probably attributed to the fact that a large number of low-quality fundus images are enhanced by the IQE subnetwork, thus improving the classification performance of the DR grading subnetwork. Comparing the classification performance of CLEAQ-DR and CLEQ-DR, we validate that the quality assessment subnetwork plays an important role in guiding the quality enhancement, which can further help CLEAQ-DR to improve on the DR classification performance.

3) The DR Grading Ablation Study for Different Quality Fundus Images: To evaluate the performance of the DR disease grading on the different quality fundus images, we compare the single DR grading subnetwork and CLEQ-DR on high-quality, usable-quality and low-quality fundus images, respectively. In addition, we also compare it with the recently proposed multi-task learning framework, DeepMT-DR. The comparisons on different quality fundus images are illustrated in Fig. 12, which reveals the following several conclusions:

1) All the comparable methods perform best on high-quality fundus images, and show performance degradation on both usable-quality images as well as low-quality images. Especially on low-quality fundus images, all the methods except our methods obviously suffer from the decreased image quality and obtain the worse performance, which implies that the quality of the fundus images significantly influence the performance of DR grading.
2) Compared with other methods, the performance of the CLEAQ-DR framework is not drastically affected by the low-quality fundus images, which demonstrates the advantage of the proposed method. In comparison with the high-quality fundus images, the Kappa metrics decrease

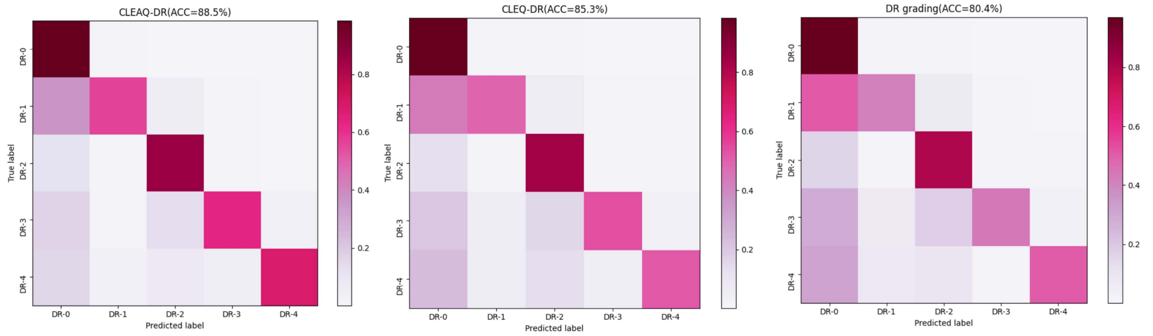


Fig. 11. The classification confusion matrices of the proposed CLEAQ-DR, CLEQ-DR, and DR grading. The horizontal axis represents the predicted categories and the vertical axis represents the true categories.

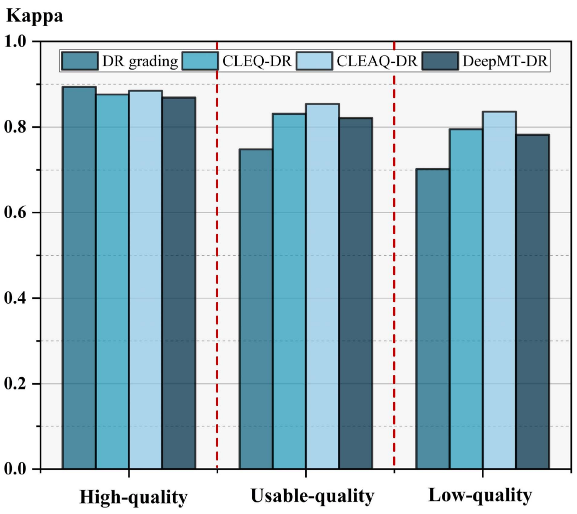


Fig. 12. The ablation experiment results of DR grading subnetwork based on fundus images of different level qualities.

by only 0.031 and 0.049 for CLEAQ-DR on usable-/high-quality fundus images, respectively.

3) When the inputs are high-quality images, the performance of the DR grading subnetwork is the best, outperforming our proposed CLEAQ-DR. The reason is that the aim of the IQE subnetwork is to improve the image quality for lower quality images, leading to inevitable negative influence when the quality enhancement is attempted on the high-quality fundus images.
4) Our proposed method performs better than DeepMT-DR on the different quality levels, especially on low-quality images. Although the DeepMT-DR leverages multi-task learning with consideration of lesion segmentation for improving DR diagnostic performance, it ignores the influence of low quality on the classification task.
5) The proposed CLEAQ-DR framework focuses on the DR grading performance improvement for low-quality and usable-quality fundus images. Hence, the collaborative learning architecture benefits the final results with a gain of 0.106 and 0.134 for Kappa metrics compared with baseline DR grading on the usable-quality images and low-quality images, respectively.

4) The Ablation Study on the Image Quality Assessment: In the ablation experiments for the image quality assessment, we compared our CLEAQ-DR with only the IQA subnetwork, the IQA subnetwork w/o LQI-IQA, and CLEAQ on the EyeQ and Messidor datasets. The quality assessment results of the ablation experiments are shown in Table VI. We can draw the following conclusions:

TABLE VI
THE ABLATION EXPERIMENT RESULTS OF IQA SUBNETWORK

Dataset	Methods	Acc	Kappa	Precision	F1-score
Messidor	IQA	**0.942**	**0.916**	**0.951**	**0.925**
	IQA w/o LQI-IQA	0.867	0.843	0.873	0.851
	CLEAQ	0.662	0.638	0.672	0.643
	CLEAQ-DR	0.641	0.617	0.649	0.624
EyeQ	IQA	**0.865**	**0.841**	**0.872**	**0.849**
	IQA w/o LQI-IQA	0.793	0.774	0.804	0.782
	CLEAQ	0.587	0.569	0.594	0.578
	CLEAQ-DR	0.573	0.554	0.582	0.561

1) CLEAQ-DR expectedly shows the worst performance among the methods, which indicates that the IQE and DR subnetworks have no positive influence on the image quality assessment task. The main reason for the performance drop is due to the fact that the IQE subnetwork improves the quality of Low-/Usable-quality fundus images, which leads the IQA subnetwork to incorrectly classify Low-/Usable-quality images as high-quality images. Although quality label of training data is not changed, the potential quality of training data becomes higher as IQE subnetwork training. That is, the input of the IQA subnetwork is the enhanced fundus images, which do not reflect the inherent quality level of the original images. Therefore, the fine-tuning of the IQA becomes difficult due to inconsistency between the original label and the potential label. In other words, the IQA and IQE subnetworks act as discriminators and generators in adversarial learning, respectively.
2) Comparing the results of IQA and IQA w/o LQI-IQA, an important conclusion is that the low-quality indicators reconstruction can provide more potential critical information associated with the input image quality and help the network to focus on relevant low-quality-specific regions in the image.

5) Collaborative Learning Analysis With Different Epochs: In this part, we thoroughly analyze the correlation among disease diagnostic grading, quality assessment and quality enhancement with different epochs. From Fig. 13, we can draw the following conclusions:

1) It can be observed that the different tasks can mutually reinforce each other by the collaborative learning manner,

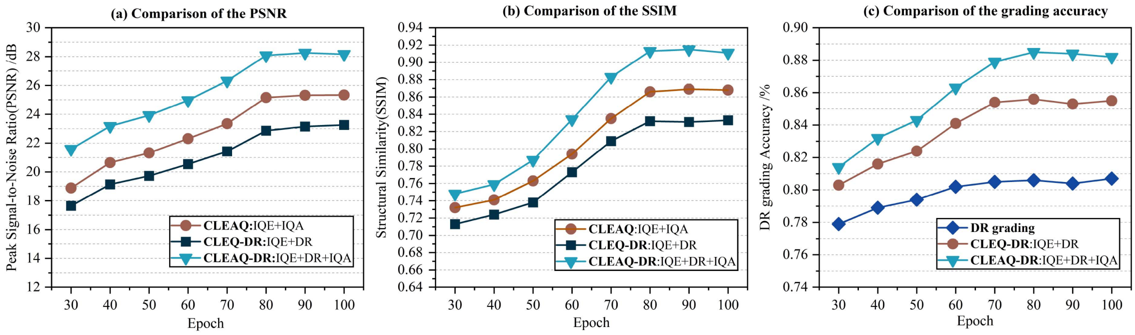


Fig. 13. Task correlation analysis in the procedure of joint learning. (a) The image quality enhancement performance through joint learning of the multiple sub-networks with epoch increasing in terms of PSNR. (b) The image quality enhancement performance through jointly training the multiple sub-networks with epoch increasing in terms of SSIM. (c) The DR grading performance through jointly training the multiple sub-networks with epoch increasing in terms of accuracy.

producing better performance for image enhancement and DR grading.

2) As can be seen in Fig. 13(a) and (b), training the tasks of DR grading and IQA with more epochs enables the IQE to obtain better image quality. Moreover, in comparison with its two variants, CLEAQ-DR is able to obtain higher quality with more guidance.
3) Fig. 13(c) shows that higher image quality leads to more accurate grading diagnostic performance in terms of accuracy. The image quality is enhanced through IQE under the collaboration with the tasks of DR diagnosis and IQA, thereby DR grading becomes easier to be trained, producing better performance.
4) These tasks are required to be jointly optimized in a unified framework. Once DR or IQA is learnt and fixed, there is no feedback from DR or IQA to boost the performance of IQE. One-trial learning for DR grading obtains limited improvement on the non-optimal quality images.
5) The performance of CLEAQ-DR tends to become stable around the 85-th epoch, and starts to gradually and slightly decrease after training further.

V. Discussion

We show some failure cases of low-/usable-quality fundus image enhancement in Fig. 14. In case 1, some low-quality regions (as shown by the green boxes of case 1) with sharp boundaries, appear as vascular-like structures (as shown by the yellow boxes of case 1) after image quality enhancement. It may mislead the ophthalmologists to diagnose it as neovascularisation. This inspire us to incorporate more constraints to distinguish between the real fundus structure and the low-quality regional boundaries. In case 2, although the proposed method is able to remove artifacts (as shown by the blue boxes of case 2) from low-quality fundus images, the fundus structures in the enhanced low-quality images are not sufficiently prominent. Such enhanced images may be detrimental to the ophthalmologist's diagnosis. We need to further improve the IQE subnetwork to enable to capture future of lesions and fundus structures while enhancing image quality. In the last case, the optic disc and part of the blood vessels are lost (as shown by the purple box of case 3) in the enhanced image. The main reasons are 1) extremely uneven illumination (as shown by the red box of case 3) in the usable-quality image, and 2) some of the pseudo-fundus structure masks generated by CE-Net are incomplete, which are required by IQE. Therefore, we will consider introducing structural consistency constraints to alleviate the problem of the partial absence of fundus structures.

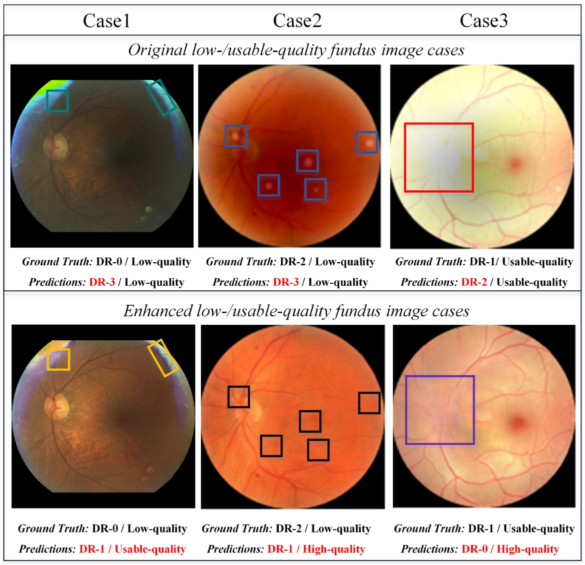

Case1
Case2
Case3
Original low-/usable-quality fundus image cases
Ground Truth: DR-0 / Low-quality
Predictions: DR-3 / Low-quality
Ground Truth: DR-2 / Low-quality
Predictions: DR-3 / Low-quality
Ground Truth: DR-1/ Usable-quality
Predictions: DR-2 / Usable-quality
Enhanced low-/usable-quality fundus image cases
Ground Truth: DR-0 / Low-quality
Predictions: DR-1 / Usable-quality
Ground Truth: DR-2 / Low-quality
Predictions: DR-1 / High-quality
Ground Truth: DR-1 / Usable-quality
Predictions: DR-0 / High-quality

Fig. 14. Some negative cases of low-quality fundus image enhancement. The low-/usable-quality fundus images (top row) and the corresponding enhanced results (bottom row). Case1: the boundaries of low-quality regions appear as vascular-like structures after image quality enhancement. Case2: the fundus structures are not prominent in the enhanced low-quality image. Case3: the fundus structures are partially missing in the enhanced usable-quality image.

We also further discuss some limitations of the proposed method and the future research directions, which mainly include the following aspects:

1) A large fundus image dataset usually comes from multiple institutions. Therefore, it often involves multiple image styles, resulting in performance degradation of the DR grading task. Therefore, how to unify the image styles in the dataset and reduce the inter-domain differences will be the focus of future research.
2) In our CLEAQ-DR framework, the IQE subnetwork is guided by both the IQA and the DR subnetworks from different aspects. The relationship between IQA and DR is implicit through collaboratively guiding IQE, and an explicit relationship needs to be further exploited. How to directly establish the correlations among IQA, DR, and IQE tasks in a unified framework is also a future research direction.

3) The fundus structure annotation is required in our framework. Although the pseudo annotation is obtained by the trained CE-Net, the annotation requirement allows our model to be dependent on the external resource. Therefore, how to collaboratively learn the multiple tasks without any local annotations (fundus structures or lesions) is another future research.

VI. Conclusion

The quality of fundus images is crucial for ensuring the diagnostic reliability of the ophthalmologist or automated medical system. To enhance the DR grading performance on the low-quality fundus images, we propose an end-to-end quality assessment guided collaborative learning framework that (1) improves the disease grading performance given a large number of low-quality images, (2) achieves fundus image quality enhancement, and (3) trains an image quality assessment model. The experimental results demonstrate that our method significantly improves the latest results of DR grading on benchmark fundus datasets, and the low-quality fundus images also gain remarkable enhancement.

References

[1] K. Ogurtsova et al., "IDF diabetes atlas: Global estimates of undiagnosed diabetes in adults for 2021," *Diabetes Res. Clin. Pract.*, vol. 183, 2022, Art. no. 109118.

[2] V. Mayya, S. Kamath, and U. Kulkarni, "Automated microaneurysms detection for early diagnosis of diabetic retinopathy: A comprehensive review," *Comput. Methods Prog. Biomed. Update*, vol. 1, 2021, Art. no. 100013.

[3] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao, "A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 818–828, Mar. 2021.

[4] T. Köhler, A. Budai, M. F. Kraus, J. Odstrčilik, G. Michelson, and J. Hornegger, "Automatic no-reference quality assessment for retinal fundus images using vessel segmentation," in *Proc. 26th IEEE Int. Symp. Comput.-Based Med. Syst.*, 2013, pp. 95–100.

[5] S. Wang, K. Jin, H. Lu, C. Cheng, J. Ye, and D. Qian, "Human visual system-based fundus image quality assessment of portable fundus camera photographs," *IEEE Trans. Med. Imag.*, val. 35, no. 4, pp. 1046–1055, Apr. 2016.

[6] F. Yu, J. Sun, A. Li, J. Cheng, C. Wan, and J. Liu, "Image quality classification for DR screening using deep learning," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2017, pp. 664–667.

[7] G. T. Zago, R. V. Andreão, B. Dorizzi, and E. O. T. Salles, "Retinal image quality assessment using deep learning," *Comput. Biol. Med.*, vol. 103, pp. 64–70, 2018.

[8] C. Lei and Q. Chen, "Robust reflection removal with reflection-free flash-only cues," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14811–14820.

[9] Y. Ye, Y. Chang, H. Zhou, and L. Yan, "Closing the loop: Joint rain generation and removal via disentangled image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2053–2062.

[10] S. J. Cho, S. W. Ji, J. P. Hong, S. W. Jung, and S. J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4641–4650.

[11] E. Decencière et al., "Feedback on a publicly distributed image database: The messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.

[12] S. Liu, Y. Liang, and A. Gitter, "Loss-balanced task weighting to reduce negative transfer in multi-task learning," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9977–9978.

[13] Y. Ganin and L. Victor, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[14] H. Fu et al., "Evaluation of retinal image quality assessment networks in different color-spaces," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2019, pp. 48–56.

[15] "Kaggle diabetic retinopathy detection competition." [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection

[16] J. Staal, M. D. Abramoffff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[17] Z. Gu et al., "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[19] Kaggle diabetic retinopathy detection competition. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection/leaderboard

[20] X. Li, X. Hu, L. Yu, L. Zhu, C. W. Fu, and P. A. Heng, "CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1483–1493, May 2020.

[21] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-in-Net: Deep mining lesions for diabetic retinopathy detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2017, pp. 267–275.

[22] K. Zhou et al., "Multicell multi-task convolutional neural networks for diabetic retinopathy grading," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 2724–2727.

[23] Y. Huang, L. Lin, P. Cheng, J. Lyu, and X. Tang, "Lesion-based contrastive learning for diabetic retinopathy grading from fundus images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2021, pp. 113–123.

[24] Y. Zhou et al., "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2079–2088.

[25] X. Wang, M. Xu, J. Zhang, L. Jiang, and L. Li, "Deep multi-task learning for diabetic retinopathy grading in fundus images," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2826–2834.

[26] Z. Shen, H. Fu, J. Shen, and L. Shao, "Modeling and enhancing low-quality retinal fundus images," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 996–1006, Mar. 2021.

[27] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[28] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[29] T. Park, A. A. Efros, R. Zhang, and J. Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 319–345.

[30] P. Cheng, L. Lin, Y. Huang, J. Lyu, and X. Tang, "I-secret: Importance-guided fundus image enhancement via semi-supervised contrastive constraining," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021. pp. 87–96.

[31] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.

[32] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X. P. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 4572–4576.

[33] Q. C. Tian and L. D. Cohen, "Global and local contrast adaptive enhancement for non-uniform illumination color images," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 3023–3030.

[34] J. Cheng, Z. Li, Z. Gu, H. Fu, D. W. K. Wong, and J. Liu, "Structure-preserving guided retinal image filtering and its application for optic disk analysis," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2536–2546, Nov. 2018.

[35] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–15, 2017.

[36] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[37] X. Fu, D. Zeng, Y. Huang, X. P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2782–2790.

[38] C. Y. Li, J. C. Guo, R. M. Cong, Y. W. Pang, and B. Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5664–5677, Dec. 2016.