SUBJECTIVE QUALITY ASSESSMENT OF ENHANCED RETINAL IMAGES

Guanghui Yue[1]*, Shaoping Zhang*[1]*, Yuan Li*[1]*, Xiaoyan Zhou*[2]*, Tianwei Zhou*[3]*, Wei Zhou*[4]

[1] Health Science Center, Shenzhen University, Shenzhen 518060, China
[2] University of Hong Kong-Shenzhen Hospital, Shenzhen 518040, China
[3] College of Management, Shenzhen University, Shenzhen 518060, China
[4] University of Waterloo, Waterloo N2L 3G1, Canada

ABSTRACT

Many retinal images sometimes suffer from uneven illumination, which influences the analysis and diagnosis of retinal diseases. To improve the image quality of those retinal images, one feasible solution is to utilize low-light image enhancement (LIE) algorithms. However, how to evaluate the perceptual quality of enhanced retinal images (ERIs) generated by different LIE algorithms remains a challenging problem. In this paper, we conduct subjective experiments to investigate the quality assessment of ERIs. First, we collect 250 retinal images with the authentic low-light distortion, and then adopt eight LIE algorithms to produce 2000 ERIs. Second, a subjective experiment is conducted, resulting in the proposed Enhanced Retinal Image Quality Assessment Database (ERIQAD). Finally, we test some well-known no reference image quality assessment (NR IQA) methods on our proposed ERIQAD. Experimental results demonstrate that existing mainstream NR IQA methods merely achieve ordinary performance to predict the perceptual quality of ERIs.

Index Terms— Retinal images, image quality assessment (IQA), subjective assessment, no reference (NR)

1. INTRODUCTION

Nowadays, retinal images are widely used in the screening and diagnosis of retinal diseases such as diabetic retinopathy, glaucoma, and age-related macular degeneration [1, 2]. However, the visual quality of retinal images is various due to the different operating levels of ophthalmologists and the influence of acquisition environment [3, 4]. Generally, retinal images have some distortion interferences especially uneven illumination [5]. Low-quality retinal images not only affect the clinical diagnosis of related diseases, but also are fatal defects for computer-aided diagnosis systems. Thus, it is necessary to enhance and improve the retinal image quality.

In recent years, many low-light image enhancement (LIE) algorithms have been reported to improve image quality in the literature [6–10]. However, existing LIE algorithms are mostly designed for natural scene images (NSIs), which are significantly different from retinal images regarding the characteristics and purposes. Moreover, when processing different kinds of images, the performance of LIE methods varies considerably. Therefore, exploring reliable quality assessment methods for enhanced retinal images (ERIs) is important.

In the literature, the quality assessment of NSIs has been extensively discussed [11–13]. In general, it can be divided into subjective and objective methods. The subjective method requires the experiment to be set up scientifically and strictly, and the experimental task determines the scoring rules, which are the key part of a subjective test. The mainstream scoring rules for NSIs are based on a continuous impairment scale or five-grade rating scale, with key points on the scale corresponding to 'Excellent', 'Good', 'Fair', 'Poor', and 'Bad' quality levels. Specifically, the quality here refers mainly to the visual experience. To date, many quality databases have been constructed based on subjective experiments, which are usually taken as the test platforms for objective methods.

Compared to NSIs, the quality assessment of ERIs has distinctive characteristics and obvious differences. On the one hand, we should simultaneously consider both visual experience and image utility to grade the ultimate quality. Different from NSIs, which focus on a global image quality information, in clinical practice, doctors focus more on whether fundus images can accurately reflect the retinal anatomical structure (such as optic cup/disc, blood vessels) and pathological information (such as exudates, hemorrhages, and macula). On the other hand, ERIs have more complex artifacts caused by LIE algorithms, which makes the quality assessment for ERIs more challenging.

Therefore, we perform an in-depth investigation on the quality assessment of ERIs from the subjective experiment and build a specific database called Enhanced Retinal Image Quality Assessment Database (ERIQAD). Following the clinical practice, we first collect 250 retinal images with authentic low-light distortion. Then, we adopt eight LIE algorithms to process these original images for obtaining 2000 ERIs, which form our database to be scored. Secondly, for each ERI, its quality is reported in the form of mean opinion scores (MOSs) through a strict subjective experiment and data processing procedure. Lastly, as most of the current quality

assessment methods are designed for NSIs, we further investigate whether existing mainstream no reference (NR) objective quality assessment methods are effective in the quality assessment task of ERIs.

2. THE ENHANCED RETINAL IMAGE QUALITY ASSESSMENT DATABASE

2.1. Image Collection and Processing

To establish our ERIQAD, we first collect 250 low-light retinal images from the University of Hong Kong-Shenzhen Hospital. All retinal images are obtained according to standard imaging protocols. To protect privacy, we then remove patient information in each image, such as name, gender, age, treatment time, etc. Moreover, these retinal images are uniformly trimmed into the resolution of 512×512 pixels, as original/raw images. In addition, to meet the clinical practice, these images have sufficient content diversity. That is, there are different retinal pathological features on the images, such as exudate, hemorrhage, macula, etc. Some examples of the selected raw images are given in the first column of Fig. 1.

After obtaining the processed 250 low-light retinal images, we adopt eight representative LIE algorithms to generate ERIs. Specifically, the LIE algorithms include six traditional algorithms (i.e., CLAHE [7], MF [8], NPEA [9], PM [14], LIME [15], NCE [10]) and two deep learning-based algorithms (i.e., ZeroDCE [16], StillGAN [17]), where StillGAN [17] is specially designed for medical images. To be rigorous, we run the official source code to obtain the models, or directly load the pre-trained models provided by the authors for each LIE algorithm. The low-light retinal images are then fed into each model for enhancement processing. Thus, we have a total of 2000 ERIs. Some ERI examples are shown in Fig. 1, where each row represents enhanced results of the same low-light image through eight LIE algorithms.

From Fig. 1, we can also find that the visual effects of these ERIs are dependent on the characteristics of raw images and adopted LIE algorithms. Specifically, some images can be greatly enhanced by the algorithm to achieve a good quality level. However, some enhanced counterparts are worse than the corresponding raw retinal image. This may be because in these cases, the LIE algorithm is not suitable for such distortion scenarios. Additionally, ERIs generated by various LIE algorithms have differences in brightness, contrast, and color. Besides, different from considering only the visual senses, the quality of retinal images could be unsatisfactory when they contain excessive brightness that obscure the physicians' observation of pathological information in small areas. Therefore, a comprehensive subjective study of ERIs is essential to help us better understand the performance of different LIE algorithms.

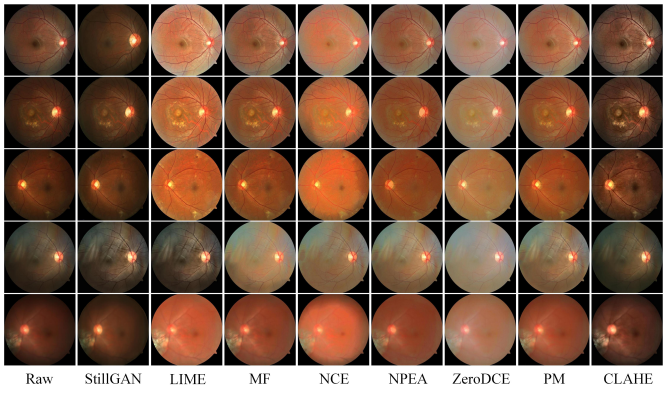

Fig. 1. Raw retinal images and the corresponding enhanced results generated by different LIE algorithms.

2.2. Subjective Experiment

2.2.1. Scoring Rules

To obtain the subjective ratings of the produced ERIs, we first ask the subjects to have a mentally expected score for the raw retinal image as a base score, and this rating should take into account not only the visual perception but also the image utility. Then, the subjects rate the enhanced results, which are added or subtracted on the base score according to the following elements: 1) Whether the important structures of ERIs (e.g. optic cup/disc, blood vessels) are enhanced to be more clearly displayed. If they are enhanced, 1 to 2 points are added to the base score according to the degree of optimization; if there is no significant change, no points need to be added or subtracted from the base score; if the clarity of important structures is diminished, the base score will be reduced by 1 to 2 points depending on the degree. 2) Whether the brightness of the ERIs can promote the observation of retinal anatomy and pathological information. Similarly, according to the judgment, add or subtract 1-2 points on the base score according to the degree, or nothing to do if there is no significant change. 3) Subtract 1-2 points from the base score if ERIs own additional distortions such as blur, artifacts, etc., compared to the original image; or add 1-2 points if some distortions are removed relative to the original image. The rules are summarized in Table 1, and note that all the operations in Table 1 are performed on the base score. Therefore, the final quality scores of ERIs are determined after adding or subtracting points from the base score by considering all the above three elements. In the whole rating process, the subjective scores of ERIs range from 1 to 10.

2.2.2. Subjective Testing Procedure

The subjective testing involves eighteen participants (including 8 males and 10 females with normal or corrected-to-normal vision, 21 to 26 years old) major in biomedical

Table 1. Scoring rules for ERIs quality assessment.

Elements	Descriptions
Structures	If structures are enhanced, add 1-2 points If there are no significant changes, the score remains If structures become less clear, subtract 1-2 points
Brightness	If image is more suitable to observe, add 1-2 points If there are no significant changes, the score remains If the image deteriorates, subtract 1-2 points
Other distortions	If distortions are reduced, add 1-2 points If there are no significant changes, the score remains If the image adds extra distortions, subtract 1-2 points

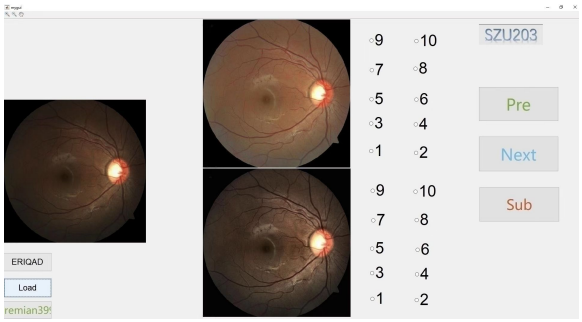

Fig. 2. The graphical user interface of the scoring software.

engineering. Before the experiment, all subjects first receive careful training from two ophthalmologists, and basically master the knowledge of retinal disease screening. They are then asked to sit in front of a screen during the subjective testing, which is conducted in a laboratory environment. It should be noted that the indoor lighting of our experiment is similar to that of the ophthalmologist's office. To conform to clinical practice, we set flexible viewing distances and participants could maintain a comfortable position within the designated viewing area.

In the formal subjective testing, there exist both training and test phases. As for the training phase, subjects are required to be trained to understand the task and manipulation of the experiment. When they fully understand the experiment and can achieve high rating accuracy on the prepared training samples, we then begin the test phase. Note that the training samples are not included in the test phase as well as in the formal database. During the test phase, we simultaneously display the raw low-light image and the two corresponding ERIs at one time, with score selection buttons beside each ERI. The graphical user interface of the scoring software is shown in Fig. 2. The subjects are required to observe the raw image and then rigorously grade each ERI according to the experimental scoring rules. In addition, subjects are encouraged to stop and relax visual fatigue every 10 minutes. All ERIs are randomly presented with the original resolution without repetition on a 27-inch 1920×1080 Philips screen. The scoring software will automatically record all the scoring data of ERIs. To avoid accumulated visual fatigue, the whole subjective experiment is divided into two sessions on three different days, with 800 ERIs scored each time.

2.3. Subjective Data Processing

Generally, there are some differences in each participant's rating data due to their different understanding of the experimental task. Therefore, for the rationality of the experiment, we first clean the collected scoring data. Here, we strictly follow the outlier rejection method recommended by ITU-R BT.500 [18] to remove the scoring data of outliers. By the data analysis and processing, we find that no rating data needed to be removed from the subjects. Then, we utilize the collected subjective data to calculate the MOS value for each ERI. Let $r_{p,q}$ denotes the q-th participant's rating of the p-th image. We calculate the MOS value (i.e., M_p) of the p-th image by

$$M_p = \frac{1}{q}\sum_{q=1}^{Q} r_{p,q}, \tag{1}$$

where Q is the number of raters after the process of data cleaning. Fig. 3 illustrates the MOS distribution of all images in the proposed ERIQAD. As shown in Fig. 3, the subjective quality scores are mainly distributed in the middle-right intervals. The possible reason is that the brightness of the image has a great influence on human perception. Thus, the visual experience of ERIs is basically improved after the processing of LIE algorithms. But it is not enough for the clinically used retinal images. In addition, there are few perfect ERIs.

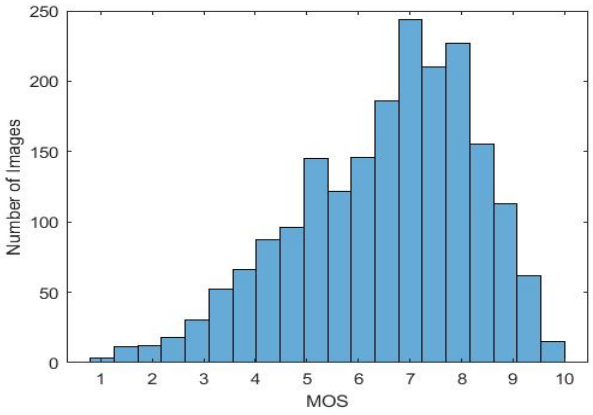

Fig. 3. MOS distribution for all ERIs in the proposed ERIQAD.

3. EXPERIMENTS AND RESULTS

3.1. Experiment Setup

To the best of our knowledge, there are very few quality assessment methods specifically designed for ERIs, thus we select eleven popular NR IQA algorithms to test their feasibility in the ERIs quality assessment task. Specifically, they

can be further classified into three categories depending on their applications. The first category includes one method, named GWH-GLBP [19], which is specifically proposed for evaluating authentically distorted images. The second category is designed for quality evaluation of synthesized distortions, including BRISQUE [20], GM-LOG [21], IL-NIQE [22], CIQA [23] , and NPQI [24]. The last category is designed for quality evaluation of contrast change, including NR-CDIQA [25], NIQMC [26], BIQME [27], MDM [28], and NUIQ [29]. All these methods are implemented using the released codes provided by the authors. To evaluate the performance of these methods, we adopted four evaluation criteria, i.e., Spearman rank correlation coefficient (SRCC), Kendalls rank correlation coefficient (KRCC), Pearson linear correlation coefficient (PLCC), and Root mean-squared error (RMSE). Besides, we adopt a five-parameter nonlinear regression function suggested by video quality expert group [30] before computing PLCC and RMSE. The nonlinear function can be formed as

$$f(s_o) = \kappa_1 \left[\frac{1}{2} - \frac{1}{e^{\kappa_2 (s_o - \kappa_3)} + 1} \right] + \kappa_4 \cdot s_o + \kappa_5, \quad (2)$$

where s_o and $f(s_o)$ are the predicted scores obtained by an objective quality assessment method and the mapped quality score with the same scale of MOS, respectively. κ_i ($i \in \{1, 2, \cdots 5\}$) are the parameters to be fitted using iterative least squares estimation. Generally, a superior NR IQA method should have higher values of SRCC, KRCC, and PLCC, while lower value of RMSE.

3.2. Experimental Results and Analysis

Table 2 lists the performance results of the selected NR IQA methods. According to Table II, first, we can find that the performance of all the methods varies greatly on our proposed ERIQAD. Specifically, CIQA shows the best performance among all 11 methods, with the PLCC, SRCC, KRCC and RMSE of 0.7982, 0.7783, 0.6386, and 0.9798, respectively. In contrast, NIQMC is in the opposite and achieves very mediocre performance. Such low performance indicates that there will be a large number of images being incorrectly rated, which is obviously not applicable to the clinical examination and diagnosis of retinal diseases. Second, it is clear that opinion-aware methods, e.g., BRISQUE, GM-LOG, and BIQME, are easier to obtain satisfactory performance than those opinion-unware methods, such as IL-NIQE, NIQMC, and NPQI. Third, some learning-based contrast-specific methods, e.g., NR-CDIQA and MDM are not competent for the quality assessment task of ERIs.

The possible reasons for above phenomenon are as follows. First, apart from the common texture and color distortions, the subjective evaluation of retinal image quality considers the semantic distortions, e.g., important anatomical structures and pathological information. However, these methods commonly extract low-level features, such as texture, color, etc., which is insufficient to characterize those semantic distortions. Therefore, they only achieve the limited performance on ERIQAD. Second, with supervised learning, opinion-aware methods are easier to build mapping relationship between feature space to quality space than opinion-unaware methods. Third, NR-CDIQA and MDM only consider limited statistical entropy-based features, which are insufficient to represent the complex distortions in ERIs.

Table 2. Performance results of objective quality assessment methods on the proposed ERIQAD.

Metrics	Evaluation Criteria			
	PLCC	SRCC	KRCC	RMSE
GWH-GLBP [19]	0.7796	0.7630	0.6114	1.0141
BRISQUE [20]	0.7681	0.7358	0.5792	1.0383
GM-LOG [21]	0.7796	0.7594	0.6101	1.0145
IL-NIQE [22]	0.5435	0.4452	0.3158	1.3614
CIQA [23]	0.7982	0.7783	0.6386	0.9798
NPQI [24]	0.5566	0.4538	0.3238	1.3525
NR-CDIQA [25]	0.4730	0.5038	0.3544	1.4390
NIQMC [26]	0.2385	0.2058	0.1422	1.5800
BIQME [27]	0.7879	0.7700	0.6233	0.9985
MDM [28]	0.5029	0.5089	0.3626	1.4055
NUIQ [29]	0.7885	0.7708	0.6238	0.9983

4. CONCLUSION AND FUTURE WORK

In clinics, retinal images often suffer from uneven illumination problems. These low-quality retinal images will affect the observation and diagnosis of retinal diseases. Although these years have reported many LIE algorithms to improve image quality, very little work has been devoted to the quality assessment of ERIs processed by LIE algorithms. Therefore, in this study, we construct a subjective quality database called ERIQAD aiming at evaluating the perceptual quality of ERIs generated by various LIE algorithms. Furthermore, we investigate the performance of some well-known NR IQA methods on the ERIQAD. The experimental results show that these methods are not fully qualified for predicting the visual quality of ERIs. In the future, we plan to analyze and quantify the specific distortion properties for ERIs, especially the characteristics of the optic cup/disc, blood vessels, as well as the capture of pathological information.

5. ACKNOWLEDGEMENT

This work was supported in part by the NSF of Shenzhen under Grants JCYJ20190808145011259 and RCBS20200714114920379, in part by the NSFC under Grant 62001302, and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515011348.

6. REFERENCES

[1] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE REV BIOMED ENG*, vol. 3, pp. 169–208, 2010.

[2] G. Yue, Y. Li, T. Zhou, X. Zhou, Y. Liu, and T. Wang, "Attention-driven cascaded network for diabetic retinopathy grading from fundus images," *BIOMED SIGNAL PROCES*, vol. 80, p. 104370, 2023.

[3] A. Raj, A. K. Tiwari, and M. G. Martini, "Fundus image quality assessment: survey, challenges, and future scope," *IET IMAGE PROCESS*, vol. 13, no. 8, pp. 1211–1224, 2019.

[4] M. A. Zapata, D. Royo-Fibla, O. Font, J. I. Vela, I. Marcantonio, E. U. Moya-Sánchez, A. Sánchez-Pérez, D. Garcia-Gasulla, U. Cortés, E. Ayguadé *et al.*, "Artificial intelligence to identify retinal fundus images, quality validation, laterality evaluation, macular degeneration, and suspected glaucoma," *CLIN OPHTHALMOL*, pp. 419–429, 2020.

[5] H. Fu, B. Wang, J. Shen, S. Cui, Y. Xu, J. Liu, and L. Shao, "Evaluation of retinal image quality assessment networks in different color-spaces," in *MICCAI 2019*. Springer, 2019, pp. 48–56.

[6] C. Li, J. Guo, F. Porikli, and Y. Pang, "Lightennet: A convolutional neural network for weakly illuminated image enhancement," *PATTERN RECOGN LETT*, vol. 104, pp. 15–22, 2018.

[7] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Muller, "Contrast-limited adaptive histogram equalization: Speed and effectiveness," in *Proc. 1st Conf. Vis. Biomed. Comput.*, vol. 337, May 1990, p. 1.

[8] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and J. Paisley, "A fusion-based enhancing method for weakly illuminated images," *SIGNAL PROCESS*, vol. 129, pp. 82–96, 2016.

[9] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE T IMAGE PROCESS*, vol. 22, no. 9, pp. 3538–3548, 2013.

[10] L. Tao and V. K. Asari, "Adaptive and integrated neighborhood-dependent approach for nonlinear enhancement of color images," *J ELECTRON IMAGING*, vol. 14, no. 4, pp. 043 006–043 006, 2005.

[11] G. Yue, C. Hou, K. Gu, T. Zhou, and G. Zhai, "Combining local and global measures for dibr-synthesized image quality evaluation," *IEEE T IMAGE PROCESS*, vol. 28, no. 4, pp. 2075–2088, 2019.

[12] G. Yue, C. Hou, K. Gu, T. Zhou, and H. Liu, "No-reference quality evaluator of transparently encrypted images," *IEEE T MULTIMEDIA*, vol. 21, no. 9, pp. 2184–2194, 2019.

[13] G. Yue, C. Hou, and T. Zhou, "Blind quality assessment of tone-mapped images considering colorfulness, naturalness, and structure," *IEEE T IND ELECTRON*, vol. 66, no. 5, pp. 3784–3793, 2019.

[14] Y. Wu, J. Zheng, W. Song, and F. Liu, "Low light image enhancement based on non-uniform illumination prior model," *IET IMAGE PROCESS*, vol. 13, no. 13, pp. 2448–2456, 2019.

[15] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE T IMAGE PROCESS*, vol. 26, no. 2, pp. 982–993, 2017.

[16] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *CVPR*, 2020, pp. 1780–1789.

[17] Y. Ma, J. Liu, Y. Liu, H. Fu, Y. Hu, J. Cheng, H. Qi, Y. Wu, J. Zhang, and Y. Zhao, "Structure and illumination constrained gan for medical image enhancement," *IEEE T MED IMAGING*, vol. 40, no. 12, pp. 3955–3967, 2021.

[18] B. Series, "Methodology for the subjective assessment of the quality of television pictures," *Rec. ITU-R BT*, vol. 500, pp. 500–13, 2012.

[19] Q. Li, W. Lin, and Y. Fang, "No-reference quality assessment for multiply-distorted images in gradient domain," *IEEE SIGNAL PROC LET*, vol. 23, no. 4, pp. 541–545, 2016.

[20] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE T IMAGE PROCESS*, vol. 21, no. 12, pp. 4695–4708, 2012.

[21] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE T IMAGE PROCESS*, vol. 23, no. 11, pp. 4850–4862, 2014.

[22] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE T IMAGE PROCESS*, vol. 24, no. 8, pp. 2579–2591, 2015.

[23] H. Chen, X. Chai, F. Shao, X. Wang, Q. Jiang, M. Chao, and Y.-S. Ho, "Perceptual quality assessment of cartoon images," *IEEE T MULTIMEDIA*, 2021.

[24] Y. Liu, K. Gu, X. Li, and Y. Zhang, "Blind image quality assessment by natural scene statistics and perceptual characteristics," *ACM Trans. Multimedia Comput. Comm. Appl. (TOMM)*, vol. 16, no. 3, pp. 1–91, 2020.

[25] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE SIGNAL PROC LET*, vol. 22, no. 7, pp. 838–842, 2014.

[26] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "No-reference quality metric of contrast-distorted images based on information maximization," *IEEE T CYBERNETICS*, vol. 47, no. 12, pp. 4559–4565, 2016.

[27] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE T NEUR NET LEAR*, vol. 29, no. 4, pp. 1301–1313, 2018.

[28] H. Z. Nafchi and M. Cheriet, "Efficient no-reference quality assessment and classification model for contrast distorted images," *IEEE T BROADCAST*, vol. 64, no. 2, pp. 518–523, 2018.

[29] Q. Jiang, Y. Gu, C. Li, R. Cong, and F. Shao, "Underwater image enhancement quality evaluation: Benchmark dataset and objective metric," *IEEE T CIRC SYST VID*, vol. 32, no. 9, pp. 5959–5974, 2022.

[30] V. Q. E. Group *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii," *2003 VQEG*, 2003.