The Loop Game: Quality Assessment and Optimization for Low-Light Image Enhancement

Baoliang Chen\textsuperscript{1*}, Lingyu Zhu\textsuperscript{1*}, Hanwei Zhu\textsuperscript{1}, Wenhan Yang\textsuperscript{1}, Fangbo Lu\textsuperscript{2}, Shiqi Wang\textsuperscript{1†}

\textsuperscript{1}Department of Computer Science, City University of Hong Kong
\textsuperscript{2}Kingsoft Cloud, Beijing
blchen6-c@my.cityu.edu.hk, lingyzhu-c@my.cityu.edu.hk

Abstract

There is an increasing consensus that the design and optimization of low light image enhancement methods need to be fully driven by perceptual quality. With numerous approaches proposed to enhance low-light images, much less work has been dedicated to quality assessment and quality optimization of low-light enhancement. In this paper, to close the gap between enhancement and assessment, we propose a loop enhancement framework that produces a clear picture of how the enhancement of low-light images could be optimized towards better visual quality. In particular, we create a large-scale database for QUality assessment Of The Enhanced LOw-Light Image (QUOTE-LOL), which serves as the foundation in studying and developing objective quality assessment measures. The objective quality assessment measure plays a critical bridging role between visual quality and enhancement and is further incorporated in the optimization in learning the enhancement model towards perceptual optimally. Finally, we iteratively perform the enhancement and optimization tasks, enhancing the low-light images continuously. The superiority of the proposed scheme is validated based on various low-light scenes. The database as well as the code will be available.

1. Introduction

Low light image enhancement fills in the gap between the prevalence of image acquisition devices and poor quality photos under low-light conditions, and provides the meaningful surrogate for expensive hardware solutions. Typically, the enhancement algorithms aim to improve the visual quality from multiple perspectives due to a series of visual quality degradations by inadequate and unbalanced lighting, such as low visibility, undesirable color cast, and intensive noise. Pioneering efforts are dedicated to the reconstruction from the low-light one to the normal illumination, including the Histogram Equalization (HE) based approaches [1, 2, 5, 44] and Retinex theory-based approaches [13,17,21,50]. However, only enlightening the image without inducing any other image priors could inevitably lead to amplified noise. Recently, the data-driven approaches based upon deep-learning [27, 39, 48] have achieved substantial breakthroughs. In particular, they rely on the large-scale database with paired low/high-quality images, such that the enhancement mappings can be learned in a fully-supervised manner.

One of the major challenges of fully-supervised based methods lies in the optimization process that is rarely guided by perceptual quality models. Despite the fact that many fidelity measures have been committed to optimizing low-light image enhancement methods [27, 39, 55, 58], there are still underlying limits. When it comes to error vis-
ibility approaches, mean absolute error (MAE) and mean squared error (MSE) are poorly correlated with the human vision system [51]. Though the structural similarity index was motivated by human perception [52], the gray image assumption made it unsuitable for optimization in the color domain. Learning-based methods have recently played an important role in perceptual image optimization [7,61], but those methods are prone to overfitting the training data, and the optimized images may contain extra artifacts [8, 34]. Furthermore, the quality of reference image limits the supervised low-light image enhancement methods. As a result, to develop a method that can generate visually pleasing results beyond the ground truth, some works attempted to learn the low-light enhancement model using unsupervised learning strategy [20, 35, 36]. Though such methods have great potentials in exploiting the scene statistics of natural images, it is still quite questionable that the learned statistics could well reflect human perception without clear guidance. Talebi et al. provided another solution that combined the fidelity loss with a no-reference quality assessment (NR-IQA) model dubbed NIMA [47], resulting in a considerable quality improvement [46]. The fidelity loss (SSIM [52]) was employed to approximate the ground truth, while the quality loss (NIMA [47]) was used to monitor image style. However, when we applied the loss to the low-light image enhancement task, the addition of NIMA [47] has not achieved a perfect result. There several reasons behind this. First, the NIMA quality model lacks enough generalization capability due to the large domain gap between its training set and the enhanced low-light image. Second, the fidelity loss and quality loss may conflict with each other when they are simultaneously optimized, which is verified by their scores changes as shown in Fig 1. As such, the quality measure may not always convey the right information to the enhancement quality, resulting in unexpected distortions (e.g., biased color and residual noise).

In this paper, we address those issues via closing the gap between quality assessment and enhancement in a loop manner. In particular, we first contribute a new dedicated database consists of diverse images and different enhancement models, and assign each image with a pseudo-Mean Opinion Score (pseudo-MOS) by the score-mapping of several reliable FR-IQA models. The database construction aims to develop a low-light specific NR-IQA measure. To solve the dilemma between fidelity measure and quality measure, we disentangle an image into the fidelity and quality by the signal decomposition and construct a fidelity compatible quality measure based on the NR-IQA model. Finally, the loop between quality assessment and enhancement is preformed, through which, the generalization capability of the NR-IQA model can be further improved by continuously updated samples. Extensive experimental results demonstrate this loop-closing framework successfully achieves a large leap in terms of visual quality and guarantees better quality of the resulting images after each loop. Our main contributions are as follows,

- We create a large-scale database termed as QUality Assessment Of The Enhanced LOw-Light Image (QUOTE-LOL), consisting of 290 image contents and 10 enhancement methods. In total, 2,900 images as well as their pseudo-MOSs are created, providing a reliable way for measuring and optimizing the visual quality of the enhanced images.
- We develop an NR-IQA model based on the statistics of deep features, aiming for the image style modulation. The proposed quality measure is then incorporated with our fidelity measure, which could improve the visual quality and preserve image content simultaneously. The high generalization capability is further verified on an unsupervised enhancement framework.
- We propose the first loop solution for low-light image enhancement. The proposed joint loss guides the enhancement network to achieve high-quality results, possessing a potential even superior to the reference, and the enhanced images, in turn, facilitate the quality measure. The game between quality evaluation and quality optimization is performed iteratively, building the closed loop for continuous quality improvement.

2. Related Works

2.1. Image Quality Assessment

Image Quality Assessment Database. The image quality assessment (IQA) database plays a vital role in IQA research. Numerous databases have been created for IQA, including LIVE [42], TID2013 [37], CSIQ [23] and Waterloo Exploration Database [28], which have greatly advanced this field. Comparing with those databases with only synthetically distorted images, quality evaluation for images in the wild has been widely explored in LIVE Challenge [15], KonIQ-10k [18] and SPAQ [11]. The quality of images generated by algorithms for different tasks (e.g., super-resolution, deblurring, denoising, compression) are further evaluated in Berkeley-Adobe Perceptual Patch Similarity (BAPPS) [61] and Perceptual Image-Error Assessment through Pairwise Preference (PieAPP) [38] databases. In particular, rather than labeling images with a subjective opinion score, the database can also be annotated with pseudo-MOSs generated by several reliable full-reference quality assessment (FR-IQA) models [29, 56].

Image Quality Assessment Methods. The earliest FR-IQA originated from mean square error (MSE) or peak signal-to-noise ratio (PSNR). They benefit from clear physical meaning, but have a low correlation with human perception. The well-known structural similarity (SSIM) [52]...
measurement consists of three independent components, luminance, contrast, and structure, providing a prototype for the subsequent FR-IQA models, including feature representation similarity (FSIM) [60], gradient similarity (GSM) [26], and multi-scale analysis (MS-SSIM) [54]. The signal distortion was measured from an information perspective in VIF [41]. Inspired by VIF, Wang et al. [53] developed an information content weighted SSIM measure (IW-SSIM). In AWDS [9], Ding et al. introduced adaptive weighting strategy to predict cross-content-type images. In another scenario, NR-IQA methods rely on natural scene statistics (NSS) to evaluate the destruction of “naturalness”. For example, Mittal et al. [31] investigated NSS features by exploiting the local spatial normalized luminance coefficients. Sophisticated deep learning based NR-IQA methods have also been developed for IQA [4, 45, 47, 62, 64]. The Neural Image Assessment (NIMA) model [47] which tackles the problem of understanding visual aesthetics was trained on the large-scale Aesthetic Visual Analysis (AVA) database [33] to predict the distribution of quality ratings.

2.2. Low-light Image Enhancement

Acquiring images under low-lighting conditions often suffers from poor visibility. To improve the image quality, numerous low-light enhancement methods have been developed to brighten dark regions and simultaneously remove annoying hidden artifacts. Pioneering research works fall into the following two categories: Histogram Equalization (HE) [1, 2, 5, 44] and Retinex theory-based approaches [13, 17, 21, 50]. For example, the Dynamic Histogram Equalization (DHE) [1] was proposed to eliminate undesirable artifacts. Moreover, Guo et al. [17] imposed structure-aware prior to guide the final illumination layer, and Fu et al. [12] developed the fusion strategy to blend different techniques together in an effort to produce an adjusted illumination layer. The Neural Image Assessment (NIMA) model [47] which tackles the problem of understanding visual aesthetics was trained on the large-scale Aesthetic Visual Analysis (AVA) database [33] to predict the distribution of quality ratings.

Table 1. Performance comparisons of classical IQA methods on the tailored TID2013 [37] database.

| Methods   | NIQE [32] | BRISQUE [31] | PSNR | SSIM [52] | VIF [41] | MSSSIM [54] | FSIM [60] | GMSD [23] | MAD [23] | IWSSIM [53] | AWDS [9] | Fused |
|-----------|-----------|---------------|------|-----------|---------|-------------|--------|--------|--------|------------|---------|-------|
| PLCC      | 0.2437    | 0.2195        | 0.6313 | 0.6127    | 0.8052  | 0.8049       | 0.8521 | 0.7908 | 0.7996 | 0.8332     | 0.8182  | 0.8643|
| SRCC      | 0.2600    | 0.1566        | 0.6578 | 0.5919    | 0.7084  | 0.7833       | 0.8013 | 0.7835 | 0.7165 | 0.7119     | 0.7538  | 0.8182|

3. Closing the Loop of Quality Assessment and Optimization for Low-Light Enhancement

3.1. Motivation

Our goal is to enhance the low-light image into a normal-light one with both signal fidelity and perceptual quality. Previous works usually learn an enhancement network supervised by the reference (normal-light) images. However, as indicated in [10], the fidelity measures between enhanced image and its reference do not always reflect the perceptual quality change. Namely, beyond the fidelity measurement, an NR-IQA model should also be introduced for final quality optimization.

As we discussed before, a reliable and valid NR-IQA model should: 1) be generalized to reflect the quality changes and provide useful feedback continuously during the enhancement. 2) be compatible with the fidelity loss, namely the profiting of quality should not be paid at a cost of fidelity. For the first issue, we create a low-light specific database (QUOTE-LOL) for NR-IQA model learning, aiming to capture and model the intrinsic perceptual characteristics especially for low-light images. Besides, by feeding our IQA model with new samples at each loop, its generalization capability can be further improved, leading to continuous quality improvement. For the second issue, inspired by the SSIM and its related works [30, 49, 52], we decompose an image patch $x$ into five components:

$\begin{align*}
    x &= \|x - \mu_x\|_2 \cdot \frac{x - \mu_x}{\|x - \mu_x\|_2} + \mu_x \\
    &= \|\hat{x}\|_2 \cdot \frac{\hat{x}}{\|\hat{x}\|_2} + \mu_x \\
    &= c_x \cdot s_x + \mu_h + \mu_s + \mu_v,
\end{align*}$

where $\mu_x$ denotes the mean value of the patch and we further decompose it into HSV color space i.e., hue $\mu_h$, saturation $\mu_s$ and lightness $\mu_v$. The $c_x$ and $s_x$ represent the
contrast and structure components of $x$, respectively. By such decomposition, the $\mu_h$ and $s_x$ can be regarded as the content (fidelity) information of an image while the $\mu_s$, $\mu_v$ and $c_x$ control the image style (quality).

Following this vein, we propose a fidelity and quality compatible joint loss for low-light image optimization. In particular, the structure and hue information is maintained by our fidelity measure and the style is modulated by our NR-IQA model. In the following subsections, we will introduce our database construction first, then the loop between quality assessment and quality optimization is presented.

### 3.2. Database Construction

We create a dedicated database QUOTE-LOL which contains 2,900 enhanced low-light images and corresponding quality labels. Those images are generated by ten methods including both traditional and deep learning based schemes with various image contents. Based on the developed large-scale database, we demonstrate that a well-behaved IQA algorithm and readily-deployed optimization pipeline can be further developed to ultimately improve the performance of the image enhancement. In particular, we adopt the images in the publicly accessible LOw-Light database (LOL) [55] as the source for database construction.

The LOL database is acquired from realistic and diverse scenes, consisting of 689 low/normal-light training pairs and 100 low/normal-light testing pairs. In the LOL, 247 and 43 low-light images are selected by the contents from the training and testing sets in LOL, respectively. Subsequently, we apply ten classical low-light image enhancement methods, including Camera Response Model (CRM) [59], Joint Enhancement and Denoising Method (JED) [40], Multiple Fusion (MF) [12], Multiscale Retinex (MR) [21], Simultaneous Reflectance and Illumination Estimation (SRIE) [13], EnlightenGAN [20], Deep Retinex Decomposition (DRD) [55], ZeroDCE [16], Self-supervised [63] and Deep Recursive Band Network (DRBN) [58] for enhancement. In particular, EnlightenGAN and ZeroDCE are the pioneering unsupervised methods for low-light image enhancement.

As shown in Fig. 2 and the supplementary materials, these methods inevitably introduce a wide variety of quality degradation in the enhancement process, including blurring, noise contamination, contrast reduction, detail loss, over-exposure, under-exposure and color shift. More importantly, the noise could even be amplified. These distinct artifacts deliberately or unintentionally introduced with different distortion patterns are prone to quality issues, and present a stark contrast compared with the synthetic distortions.

To study the quality of the enhanced images and behaviors of low-light enhancement models, inspired by the database construction in [29, 56], we explore 11 classical IQA models, from which three reliable IQA models are selected for a pseudo-MOS generation. In particular, we first tailor the TID2013 database and collect a subset that only contains the possible distortion types introduced by enhancement, including various types of noise, blur, intensity shift, contrast as well as change of saturation. The performance comparison of 11 classical IQA models is shown in Table 1. Without loss of generalization, we then adopt a four parameters linear mapping function to combine the three selected FR-IQA models (FSIM, IWSSIM, AWDS),

$$Q = \lambda_1 Q_{FS} + \lambda_2 Q_{IW} + \lambda_3 Q_{AW} + \lambda_4,$$

where $\lambda_1, \lambda_2, \lambda_3$ and $\lambda_4$ are the optimized parameters. The mapping result is also presented in Table 1, which achieves the best results on the tailored TID2013 database. By means of the mapping function, the pseudo-MOSs of our QUOTE-LOL database finally can be generated. In Fig. 2, the pseudo-MOS of each enhanced image is annotated, from
which we can observe that our database covers a wide range of quality levels. In addition, very few enhanced images achieve a pleasant visual quality. The proposed database sheds a light on how to develop better enhancement systems.

3.3. Loop Closed Optimization

As shown in Fig. 3, we learn to enhance the low-light images based on a full operation chain that transfers the knowledge between quality assessment and enhancement. In particular, the NR-IQA model is learned from the annotated samples and integrated with a style modulation. Subsequently, we aim at learning to enhance the images driven by our IQA model, and finally the loop of quality optimization and assessment can be achieved with the mapping scores of three FR-IQA models as the pseudo-MOS annotations.

3.3.1 From Enhancement to Assessment

Benefiting from the proposed QUOTE-LOL database, a deep learning based IQA measure dedicated to enhanced low-light images is developed. It has been known that the content and the style of an image can be captured by the deep neural networks (DNNs) and are somehow separable [14, 19, 24]. In particular, the convolutional feature statistics (mean, variance/std, etc.) usually encode the style of an image. Inspired by the style transfer network, we design our NR-IQA model based on the widely used VGG16 [43] network. The VGG16 is pretrained on the ImageNet [6] with its extracted features contain the content information. When it comes to quality optimization, we fix all parameters of VGG16. The image quality is measured and regressed only by the evaluation of feature statistics (mean and std) of different layers. We find such style-sensitive architecture can highly benefit the content preservation of the final enhanced results which has been further verified in our experiments. The detailed architecture of our NR-IQA network is shown in Fig. 3.

Given an image \( I \), it passes through the pretrained VGG16 network for multi-scale feature extraction. To encode the style information, the features in two convolution stages (the third layer \( \text{Conv}3 \) and the fifth layer \( \text{Conv}5 \)) are aggregated with a global average (mean) pooling and standard deviation (std) pooling. As shown in Fig. 3, following those feature statistics, a style modulation subnetwork is introduced. In this subnetwork, for each mean/std feature, we
adopt two fully-connected (FC) layers with the units set as 128 and 64 to reduce their dimensions. Subsequently, the processed mean and std features of each layer are combined by a bilinear operation. Such bilinear fusion aims to capture the pairwise interactions between each channel. Compared with concatenating the mean and std features directly, in our experiments, we find the bilinear operation leads to better quality modeling and achieves more pleasant results. After the mean and std features are combined, three FC layers with the units set as 256, 64 and 1 are utilized for layer-wise multi-scale quality regressions are optimized as follows,

$$L_{reg} = \frac{1}{N} \sum_{i=1}^{N} (\|Q_p - Q_o\|_1 + \|Q_p - Q_{l1}\|_1 + \|Q_p - Q_{l2}\|_1),$$

where $\|\cdot\|$ represents the L1-norm, $i$ is the image index in a batch with the batch size is set to $N$.

3.3.2 From Assessment to Enhancement

The general methodology of low-light enhancement is to recover the images with the Maximum A Posteriori (MAP) framework, based on the simultaneous maintenance of fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3]. Following this vein, a fidelity and alignment of the scene statistics [3].

Regarding the learning of enhancement model $f_E$, we adopt the general architecture EDSR [25] as the backbone of our enhancement model. In particular, we replace the original loss functions used in EDSR with the loss function $L$ defined in Eqn. (6). After the enhancement model is learned, the IQA model can be renewed by model finetuning with the enhanced results and the loop will be performed iteratively until the number of iterations reaches a maximum value. Algorithm 1 summarizes our training procedure.

### Algorithm 1: Loop Closed Optimization for Low-light Enhancement

**Input:** The LOL training set $D_{LOL}$, QUOTE-LOL database $D_{Qua}$, updatable enhanced image set $D_{End}$, enhancement model $f_E$, IQA model $f_Q$.  

**Output:** High-quality enhanced images.

1. Train $f_E$ on $D_{LOL}$ supervised by SSIM loss.
2. $D_{End} ←$ Images in $D_{LOL}$ enhanced by $f_E$.
3. Assign pseudo-labels for $D_{End}$ with Eqn. (2).
4. Train $f_Q$ on $D_{Qua}$ and $D_{End}$ by Eqn. (3).
5. $D_{End} ← \emptyset$.
6. for $t ← 1$ to Max loop times do
   7.   Fine-tune $f_E$ by Eqn. (6) on $D_{LOL}$.
   8.   $D_{End} ←$ Images in $D_{LOL}$ enhanced by $f_E$.
   9.   Assign pseudo-labels for $D_{End}$ with Eqn. (2).
   10. Fine-tune $f_Q$ on $D_{Qua}$ and $D_{End}$.
   11. $D_{End} ← \emptyset$.
7. end
8. Performance testing on LOL database.

To reflect the underlying performance of our framework, we adopt the general architecture EDSR [25] as the backbone of our enhancement model. In particular, we replace the original loss functions used in EDSR with the loss function $L$ defined in Eqn. (6). After the enhancement model is learned, the IQA model can be renewed by model finetuning with the enhanced results and the loop will be performed iteratively until the number of iterations reaches a maximum value. Algorithm 1 summarizes our training procedure.

4. Experiments

4.1. Experimental Settings

Regarding learning of enhancement model $f_E$, we follow the default data splitting strategy in the LOL database [55] for generating the training and testing sets. Regarding the learning of NR-IQA model $f_Q$, we split the constructed QUOTE-LOL into the training set and testing set according to the content, which ensuring no content overlapping with the enhancement model. We augment the images by randomly cropping the images to $256 \times 256$. The batch size is 32 and we adopt Adam optimizer for optimization with the learning rate set to 1e-4. In total, 200 and 100 epochs are allowed during training $f_E$ and $f_Q$. After 200 epochs, we begin to train the closed-loop framework. The learning rate drops by 0.5, when the models are fine-tuned and up to 15 and 50 epochs $f_E$ and $f_Q$ optimized at each loop. The mapping parameters $\lambda_1 \sim \lambda_4$ in Eqn. (2) are set to -1.8041, 2.6152, -0.2776, and 0.2563, respectively. The hyper-parameters $c$ and $\lambda_5$ in Eqn. (4) and Eqn. (6) are
4.2. Enhancement Results

To verify the effectiveness of our proposed method, we compare our enhanced results with existing methods. As shown in Fig. 4, it is easy to find that our result tends to produce normal illumination and correct color deviation, while most of the existing methods fail to reconstruct a pleasant visual result especially in the area inside the red box. The distortions like overexposure, noise amplification and color bias are obviously introduced, while our method can adaptively restore the details according to the contexts, achieving a better balance of light enhancement and contrast modulation. From the Fig. 4, we can also observe that even the reference image shows an unpleasant result while our method exhibits a high potential in outperforming the reference with the quality measure introduced. We provide more comparison results in our supplementary. For a quantitative study of our method, we have not adopted the wide-used fidelity measures like PSNR or SSIM, as they are limited by the quality of reference. Instead, we perform the subjective study on the enhanced results. In particular, the two-alternative forced choice (2AFC) method is adopted. Given two enhanced images in each trial, subjects are forced to select the one that has better visual quality. The 2AFC method has been widely adopted in psychophysical studies due to the discrimination capability, leading to more reliable quality labels. In our subjective study, we invited 10 inexperienced subjects for quality preference collection. In each trial, for a specific scene, our enhanced result and the result achieves the highest pseudo-MOS in QUOTE-LOL are compared and selected by each subject. The 2AFC testing is performed on all scenes in the test set and the average individual preferences are shown in Fig. 5 (a), revealing a huge leap in quality preference our method achieves.

4.3. Ablation Study

Study of Loss Function. To demonstrate the superiority of our joint loss, we compare the optimization results with the original SSIM, which is widely used in the optimization framework. In particular, we deploy the SSIM into the optimization framework on the same baseline model to comprehensively compare the proposed perception optimization protocol. In Fig. 6 (a), we present the comparison result of SSIM optimization and our method. In addition, the subjective (2AFC) experiment is also performed for reliable performance evaluation. The results are provided in Fig. 5 (b) and it can be clearly observed that all the subject preferences are larger than 0.79, revealing the proposed closed-loop enhancement architecture provides a visually pleasing gain in human subjective measurement.

Compare with Perceptual Loss. Since the same architecture (VGG16) our IQA model shared with the well-known perceptual loss [22], we further compare the optimization results of perceptual loss with ours in Fig. 6 (b), from which we can observe a better contrast adjustment our method
holds. More experimental results are provided in our supplementary.

**Bilinear v.s. Concatenation.** In addition, we also verify that the bilinear fusion strategy in the proposed framework is superior to the concatenated mean and standard deviation value directly. From Fig. 6 (c), we can observe that the blur is usually caused by the concatenation, demonstrating the channel-wise interaction plays a critical role in quality optimization. More experimental results are provided in our supplementary.

**Parameter fixed v.s. Trainable VGG16.** Our IQA model is learned based on the parameter fixed VGG16. As shown in Fig. 6 (d), the trainable parameter introduced brings a negative effect on the image fidelity, indicating the parameters maintaining benefits for the compatibility between fidelity and quality optimization. More experimental results are provided in our supplementary.

**Study of Training Steps.** A straightforward way to improve the previous enhanced output is to iterate the enhanced images repeatedly under the proposed perceptual framework. Fig. 7 compares the results of different loop times. It may be observed that the results show the proposed optimization approach is able to successfully yield superior subjective than before and bring continuous quality improvement.

### 4.4. Application to Unsupervised Enhancement.

In order to verify the generalization capability of our NR-IQA model, we further incorporate our IQA model into the latest unsupervised framework ZeroDCE. We compare the enhancement results with/without our quality loss combined. The results shown in Fig. 6 (i) ~ (l) reveal that our quality loss can still supervise the quality and maintain the content in the reference absent scenario.

### 5. Conclusions

In this paper, we propose the closed-loop framework for quality assessment and perceptual optimization of low-light image enhancement. Experimental results show the proposed framework is able to continuously improve the image quality and achieve high perceptual pleasant results. The generalization capability of our IQA measure is further verified in the unsupervised manner. As one of the first attempts on this research topic, the proposed scheme could be further improved by quality labels annotated by humans, and the interaction between human opinion and model learning opens up new innovative space for future exploration.
References

[1] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan, and Oksam Chae. A dynamic histogram equalization for image contrast enhancement. IEEE Transactions on Consumer Electronics, 53(2):593–600, 2007. 1, 3

[2] Tarik Arici, Salih Dikbas, and Yucel Altunbasak. A histogram modification framework and its application for image contrast enhancement. IEEE Transactions on image processing, 18(9):1921–1935, 2009. 1, 3

[3] Mark R Banham and Aggelos K Katsaggelos. Digital image restoration. IEEE signal processing magazine, 14(2):24–41, 1997. 6

[4] Sebastian Boss, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. IEEE Transactions on image processing, 27(1):206–219, 2017. 3

[5] Dinu Coltuc, Philippe Bolon, and J-M Chassery. Exact histogram specification. IEEE Transactions on Image Processing, 15(5):1143–1152, 2006. 1, 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009. 5

[7] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear, 2020. 2

[8] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. International Journal of Computer Vision, 129(4):1258–1281, 2021. 2

[9] Li Ding, Ping Wang, and Hua Huang. Unified quality assessment of natural and screen content images via adaptive weighting on double scales. Signal Processing: Image Communication, 99:16446, 2021. 3

[10] Zhengfang Duanmu, Wentao Liu, Zhongling Wang, and Zhou Wang. Quantifying visual image quality: A bayesian view. Annual Review of Vision Science, 7(1):437–464, 2021. 3

[11] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3677–3686, 2020. 2

[12] Xueyang Fu, Delu Zeng, Yue Huang, Yinghao Liao, Xinghao Ding, and John Paisley. A fusion-based enhancing method for weakly illuminated images. Signal Processing, 129:82–96, 2016. 3, 4

[13] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2782–2790, 2016. 1, 3, 4

[14] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2414–2423, 2016. 5

[15] Deepthi Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. IEEE Transactions on Image Processing, 25(1):372–387, 2015. 2

[16] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1780–1789, 2020. 3, 4

[17] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. IEEE Transactions on image processing, 26(2):982–993, 2016. 1, 3

[18] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing, 29:4041–4056, 2020. 2

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In IEEE International Conference on Computer Vision, pages 1501–1510, 2017. 5

[20] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. EnlightenGAN: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing, 30:2340–2349, 2021. 2, 4

[21] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Transactions on Image Processing, 6(7):965–976, 1997. 1, 3, 4

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision, pages 694–711, 2016. 7

[23] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of electronic imaging, 19(1):011006, 2010. 2, 3

[24] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2479–2486, 2016. 5

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 136–144, 2017. 6

[26] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. IEEE Transactions on Image Processing, 21(4):1500–1512, 2011. 3

[27] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Li-net: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition, 61:650–662, 2017. 1, 3

[28] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo
[29] Kede Ma, Xuelin Liu, Yuming Fang, and Eero P Simoncelli. Blind image quality assessment by learning from multiple annotators. In *IEEE International Conference on Image Processing*, pages 2344–2348, 2019. 2, 4

[30] Kede Ma, Kai Zeng, and Zhou Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 24(11):3345–3356, 2015. 3

[31] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 3

[32] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 3

[33] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA:A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012. 3

[34] Aamir Mustafa, Aliaksei Mikhailiuk, Dan Andrei Iliescu, Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. Learned perceptual illumination images. *IEEE Transactions on Image Processing*, 29(12):4695–4708, 2020. 1, 3

[35] Zhangkai Ni, Wenhan Yang, Shiqi Wang, Lin Ma, and Sam Kwong. Towards unsupervised deep image enhancement with generative adversarial network. *IEEE Transactions on Image Processing*, 29:9140–9151, 2020. 2, 3

[36] Zhangkai Ni, Wenhan Yang, Shiqi Wang, Lin Ma, and Sam Kwong. Unpaired image enhancement with quality-attention generative adversarial network. In *ACM International Conference on Multimedia*, pages 1697–1705, 2020. 2

[37] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015. 2, 3

[38] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. PieAPP: Perceptual image-error assessment through pairwise preference. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 2

[39] Wenhui Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, 28(9):4364–4375, 2019. 1, 3

[40] Xutong Ren, Mading Li, Wen-Huang Cheng, and Jiaying Liu. Joint enhancement and denoising method via sequential decomposition. In *International Symposium on Circuits and Systems*, pages 1–5, 2018. 4

[41] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006. 3

[42] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 2

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, pages 1–14, 2015. 5

[44] J Alex Stark. Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE Transactions on image processing*, 9(5):889–897, 2000. 1, 3

[45] ShaoLin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 3

[46] Hossein Talebi and Peyman Milanfar. Learned perceptual image enhancement. In *IEEE International conference on computational photography*, pages 1–13, 2018. 2

[47] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. 1, 2, 3

[48] Ruixing Wang, Qiang Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6849–6857, 2019. 1

[49] Shiqi Wang, Kede Ma, Hojatollah Yeganeh, Zhou Wang, and Weisi Lin. A patch-structure representation method for quality assessment of contrast changed images. *IEEE Signal Processing Letters*, 22(12):2387–2390, 2015. 3

[50] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, 22(9):3538–3548, 2013. 1, 3

[51] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009. 2

[52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1, 2, 3

[53] Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on image processing*, 20(5):1185–1198, 2010. 3

[54] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, pages 1398–1402, 2003. 3

[55] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiayi Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, pages 1–12, 2018. 1, 4, 6

[56] Jinjin Wu, Jupo Ma, Fuhu Liang, Weisheng Dong, Guangming Shi, and Weisi Lin. End-to-end blind image quality prediction with cascaded deep neural network. *IEEE Transactions on Image Processing*, 29:7414–7426, 2020. 2, 4

[57] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient
perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2013.

[58] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3063–3072, 2020.

[59] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new low-light image enhancement algorithm using camera response model. In *IEEE International Conference on Computer Vision Workshops*, pages 3015–3022, 2017.

[60] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[62] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018.

[63] Yu Zhang, Xiaoguang Di, Bin Zhang, and Chunhui Wang. Self-supervised image enhancement network: Training with low light images only. *CoRR*, abs/2002.11300, 2020.

[64] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14143–14152, 2020.