QUALITATIVE AND QUANTITATIVE QUALITY ASSESSMENT OF LOW-LIGHT ENHANCED IMAGES: A DATASET AND BENCHMARK METRIC

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-light image enhancement (LLIE) improves visibility and restores details in challenging lighting conditions. It is crucial to fairly evaluate LLIE methods to foster the development of more effective models. However, quality assessment of low-light enhanced images proves to be as challenging as the enhancement itself. From a quantitative perspective, full-reference image quality assessment (FR-IQA) metrics (e.g., PSNR and SSIM) are commonly employed to assess the perceptual quality of enhanced images. However, they are not suitable when a pristine reference image is unavailable, which is often the case in real-world applications. From a qualitative perspective, the absence of a standardized and reproducible evaluation pipeline makes it extremely difficult to ensure fair comparisons across different studies. To confront these challenges, we present the Low-light Image Distortions and Quality (LIDQ) dataset, featuring both overall quality scores and distortion distribution annotations collected through formal subjective testing. Leveraging LIDQ, we propose a no-reference Low-light Enhanced Image Quantitative and Qualitative Quality Assessment (LIQ^3A) method that not only estimates perceptual quality without requiring a reference, but also provides qualitative assessments of enhancement-induced distortions. Experiments show that LIQ^3A aligns closely with human perception while accurately identifying distortion patterns. We anticipate that the proposed dataset and metric will facilitate future advances in low-light image enhancement by providing reliable evaluation feedback.

1 INTRODUCTION

Images captured in low-light environments frequently suffer from visual degradations, e.g., poor visibility, low contrast, and severe noise, which can significantly compromise visual perception and hinder the performance of computer vision tasks (Yang et al., 2020). Although advancements in imaging hardware and specialized photographic techniques can partially alleviate these issues, they often fail to completely eliminate noise due to the limited light available to camera sensors. Increasing exposure time may reduce noise but frequently introduces motion blur, further deteriorating image quality (Wang et al., 2023c). As a cost-effective alternative, computational low-light image enhancement (LLIE) methods have gained considerable attention (Guo et al., 2016). These LLIE methods focus on improving visibility, enhancing contrast, and suppressing noise, rendering images with higher perceptual quality, and boosting downstream applications' performance (Guo et al., 2020).

Despite recent advances, LLIE methods still produce artifacts such as amplified noise, color distortions, and over-smoothing that compromise image quality (Wang et al., 2024a). Assessing the perceptual quality of enhanced images is thus critical for evaluating LLIE performance and guiding refinement, typically through quantitative and qualitative evaluations (Chen et al., 2023; Zhai et al., 2021). Quantitative assessment employs image quality assessment (IQA) methods, which assign scalar values to enhancement performance. Depending on reference availability, IQA methods are categorized as full-reference, reduced-reference, or no-reference (blind) (Zhang et al., 2023b). Blind IQA (BIQA) is particularly practical since it does not require pristine references, which are often unavailable in real-world scenarios (Mittal et al., 2012b). Qualitative assessments, by contrast, rely on visual inspection to reveal strengths and weaknesses more intuitively than a single score, but they

Table 1: Summary of the previous IQA datasets for low-light (enhanced) images. 2AFC: Two-alternative forced choice. SS: Single stimulus. DS: Double stimulus. Con.: Contrast. Alg.: Algorithm. ACJ: Adjectival categorical judgement. CQR: Continuous quality rating. QSD: Quality semantic description. DTSR: Distortion types and severity ratings.

Dataset	# Reference images	Enhancement types	# Enhancement methods	# Image	# Subjects	Judgment type
Chen14 (Chen et al., 2014)	100	Alg. outputs	5	500	-	2AFC
CCID2014 (Gu et al., 2015)	15	Con.-enhanced	5	655	22	SS-CQR
NNID (Xiang et al., 2019)	448	Real-captured	-	2240	74	SS-ACJ
LIEQ (Zhai et al., 2021)	100	Alg. outputs	10	1,000	21	SS-CQR
LEISD (Lin et al., 2023)	255	Alg. outputs	8	2,040	20	SS-CQR
EHNQ (Yang et al., 2023b)	100	Alg. outputs	15	1,500	50	DS-ACJ
SQUARE-LOL (Chen et al., 2023)	290	Alg. outputs	10	2,900	30	2AFC
RNTIEQA (Wang et al., 2024b)	200	Alg. outputs	10	2,000	15	2AFC
MLIQ (Wang et al., 2024a)	1,360	Real-captured	-	1,360	26	SS-CQR & QSD
LIDQ (Ours)	253	Alg. outputs	22	5,566	34	SS-ACJ & DTSR

are usually limited to a small sample set, leading to sample bias or the so-called *cherry-picking* issue (Cao et al., 2021). Existing BIQA metrics further lack support for such qualitative comparisons, hindering large-scale dataset-level benchmarking.

In this work, we reformulate qualitative assessment as the estimation of distortion distributions in enhanced images. This approach makes the qualitative assessment process quantifiable, enabling evaluations on full datasets and ensuring the reproducibility of enhanced images across different studies. To this end, we introduce the Low-light Image Distortions and Quality (LIDQ) dataset, comprising (21+1) $\times$ 253 = 5,566 images with the most comprehensive quantitative and qualitative annotations to date. Specifically, we assemble a total of 253 distinct low-light images from existing paired LLIE datasets to serve as reference inputs for enhancement, each accompanied by 1 normal-light ground truth. We then enhance these reference images utilizing over 21 state-of-the-art LLIE methods, resulting in the collection of 5,566 images. A comprehensive subjective quality assessment is conducted to gather human quantitative mean opinion scores (MOSs) of image quality, alongside qualitative annotations of enhancement-induced distortions.

We also leverage the LIDQ dataset to develop the blind Low-light Enhanced Image Quantitative and Qualitative Quality Assessment (LIQ^{3}A) model, a strong baseline that evaluates the quality of low-light enhanced images from both quantitative and qualitative perspectives, without relying on pristine ground-truths. Adopting a multitask learning framework, LIQ^{3}A seamlessly integrates qualitative insights into BIQA learning. The model is trained to simultaneously predict quantitative quality scores and estimate qualitative distortion patterns. Built on a pretrained vision-language model, LIQ^{3}A bridges the two tasks using textual templates. By computing a joint probability based on the cosine similarities between visual and textual embeddings, the model makes predictions for both tasks and optimizes them through carefully designed loss functions.

Overall, our contributions are threefold:

- We establish LIDQ, a comprehensive quality assessment dataset consisting of 5,566 annotated images. Both quantitative quality ratings and qualitative annotations are collected for each image through formal subjective testing.
- Based on LIDQ, we propose LIQ^{3}A, a computational quality metric that assesses low-light enhanced images from both quantitative and qualitative perspectives.
- Extensive experiments on multiple datasets show that LIQ^{3}A achieves closer alignment with human quantitative annotations than other BIQA methods and effectively identifies distortion patterns in enhanced images.

2 Related Works

Low-light Image Enhancement. Classical LLIE methods, such as histogram equalization (Pizer et al., 1987) and Retinex-based techniques (Wei et al., 2018), rely on handcrafted priors and complex optimization, often resulting in limited performance or high computational cost (Li et al., 2015; Ying et al., 2017; Zheng et al., 2022). In contrast, deep neural network (DNN)-based LLIE approaches automatically learn features from data and enhance brightness, contrast, and detail via end-to-end

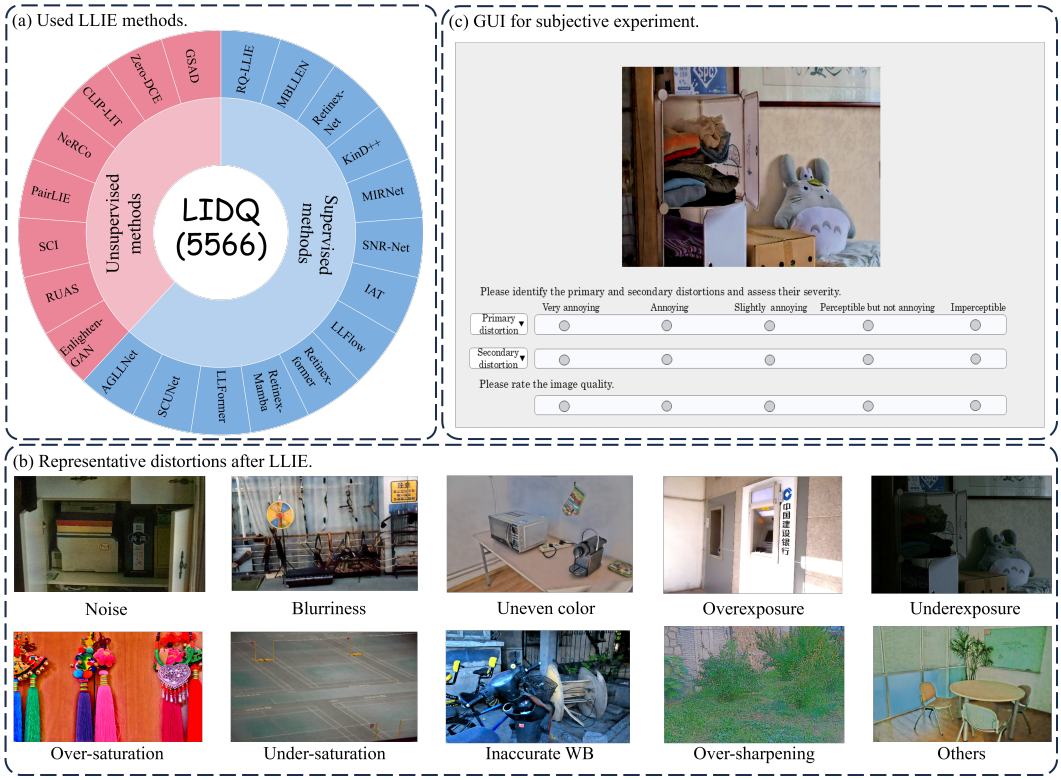
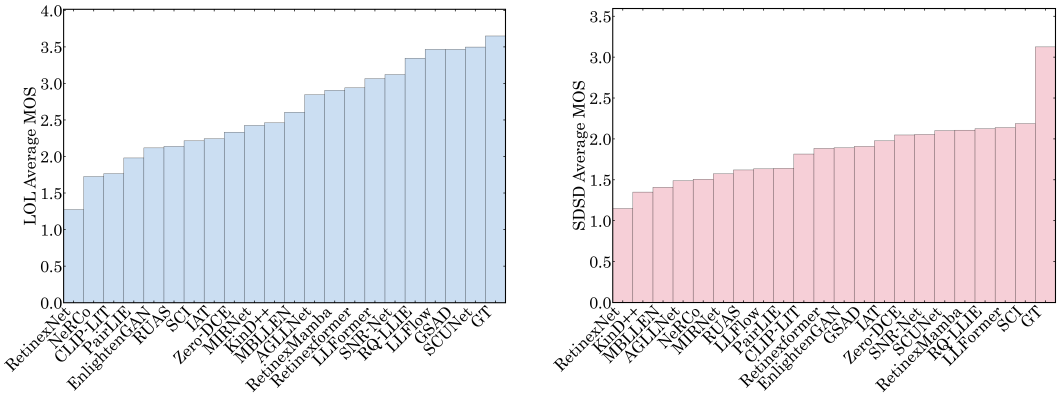


Figure 1: LIDQ comprises enhanced images generated by various LLIE algorithms (see (a)), featuring diverse algorithm-dependent artifacts (see (b)). Both quality and distortion annotations are obtained through formal subjective testing using the graphical interface shown in (c).

training (Lore et al., 2017). These models are typically trained on paired datasets—either synthetic or real—though data collection is often costly (Zheng et al., 2022; Zhang et al., 2019; 2021b). To address this, alternative learning strategies such as unsupervised (Jiang et al., 2021), semi-supervised (Yang et al., 2021a), and zero-shot learning (Zhang et al., 2024) have emerged. Despite their effectiveness, LLIE methods may still introduce distortions like structural artifacts, color shifts, or noise (Zhai et al., 2021), underscoring the need for reliable quality assessment.

Blind Image Quality Assessment. BIQA estimates perceptual image quality without reference images, providing an efficient alternative to subjective testing (Mittal et al., 2012b). Early methods relied on handcrafted features (Mittal et al., 2012a) but lacked robustness across diverse content and distortions. Deep learning greatly improved BIQA by modeling complex content–distortion interactions (Zhang et al., 2021a; Ke et al., 2021), and recent multimodal vision–language approaches further enhance performance (Zhang et al., 2023b; Wu et al., 2024b). For LLIE, however, BIQA faces two challenges: (1) methods tuned for synthetic or natural distortions perform poorly on enhanced low-light images with distinct artifacts (Wang et al., 2024a; 2023d); and (2) they mainly yield scalar scores without detailed distortion analysis (Wu et al., 2024a). Incorporating qualitative insights is thus critical for advancing LLIE evaluation and guiding model development.

Low-light IQA Datasets. LLIE methods distinguish themselves from traditional image IQA datasets by their capability to generate realistic yet artificially enhanced details and textures (Gu et al., 2020). These characteristics pose unique challenges for quantitative assessment using existing BIQA methods optimized for traditional image IQA scenarios (Chen et al., 2023). Consequently, there is a growing interest in developing specialized datasets tailored specifically for evaluating LLIE performance. Table 1 summarizes commonly used IQA datasets for LLIE, which predominantly focus on algorithm-enhanced images, with the exception of NNID (Xiang et al., 2019) and MLIQ (Wang et al., 2024a), which assess real-captured low-light images. While most provide quantitative scores like MOS, these datasets often lack qualitative analysis of distortions introduced by LLIE, such as amplified noise, color deviation, and blurriness. To fill this gap, we propose creating a new dataset dedicated to quantitative and qualitative LLIE assessment.



(a) LOL dataset (b) SDSD dataset

Figure 2: The average MOS scores of enhanced images produced by different LLIE methods on the LOL and SDSD subsets, sorted in ascending order.

3 Proposed Dataset: LIDQ

Reference Image Collection. We compile 253 low-light images for enhancement, including 15 from the LOL-v1 test set (Wei et al., 2018), 100 from the LOL-v2 test set (Yang et al., 2021b), and 138 frames from the SDSD dataset (Wang et al., 2021), where one representative frame is extracted from each video. These datasets are selected because they cover diverse indoor and outdoor scenes and are widely recognized benchmarks in the LLIE literature (Yan et al., 2025), frequently adopted to enable fair and consistent comparisons. Regarding the overlapping in the LOL datasets, we leverage the shared samples between LOL-v1 and LOL-v2 to verify the consistency of subjective ratings, ensuring that no overlapping samples appear simultaneously across the training, validation, or test splits.

Low-Light Enhancement Methods. We employ 21 recent LLIE methods[1] (see Figure 1 (a)) for enhanced image generation. For methods such as Zero-DCE, EnlightenGAN, RUAS, SCI, GDP, PairLIE, NeRCo, and CLIP-LIT, we adopt the official models released by the authors. For the remaining methods, we use publicly available models trained on LOL-v2 or retrain them following the authors' default configurations. This process produces 5,313 enhanced images, and with the corresponding ground truth from the source datasets, yields a total of 5,566 annotated images.

Subjective Testing. We design a graphical user interface (GUI) (see Figure 1(b)) to collect both quantitative and qualitative annotations of low-light enhanced images. Prior to the experiment, subjects are instructed to perform the evaluation on high-resolution monitors and are provided with detailed guidelines, including the definition of "technical image quality," examples of common distortion types, and reference images of varying quality levels. For quantitative assessment, we adopt the standard 5-point ACJ scale (Hosu et al., 2020). For qualitative evaluation, participants identify the two most prominent distortion types from nine categories: *noise, blurriness, uneven color, overexposure, underexposure, over-saturation, under-saturation, inaccurate white balance, over-sharpening, and others* (Wang et al., 2024a; Lin et al., 2023), and then rate the severity of each selected distortion using a 5-point ACJ scale. Compared with LOL-v1/v2 images, SDSD samples typically exhibit more realistic distortions, such as low-light noise and blurriness. To account for these differences, we conduct separate subjective evaluations for enhanced images from LOL-v1/v2 and SDSD, resulting in two distinct subsets.

Table 2: Min, max, median and mean SRCC, and PLCC between two randomized subgroups with equal size across 100 splits.

Criterion	Min	Max	Median	Mean
SRCC ↑	0.792	0.901	0.870	0.868
PLCC ↑	0.794	0.898	0.870	0.869

[1]These methods include Zero-DCE (Guo et al., 2020), EnlightenGAN (Jiang et al., 2021), RUAS (Risheng et al., 2021), SCI (Ma et al., 2022), GSAD (Hou et al., 2023), PairLIE (Fu et al., 2023), NeRCo (Yang et al., 2023a), CLIP-LIT (Liang et al., 2023), AGLLNet (Lv et al., 2021), RQ-LLIE (Liu et al., 2023), MBLLEN (Lv et al., 2018), RetinexNet (Wei et al., 2018), KinD++ (Zhang et al., 2019), MIRNet (Zamir et al., 2020), SNR-Net (Xu et al., 2022), IAT (Cui et al., 2022), LLFlow (Wang et al., 2022), Retinexformer (Cai et al., 2023), RetinexMamba (Bai et al., 2024), LLFormer (Wang et al., 2023b), and SCUNet (Zhang et al., 2023a).

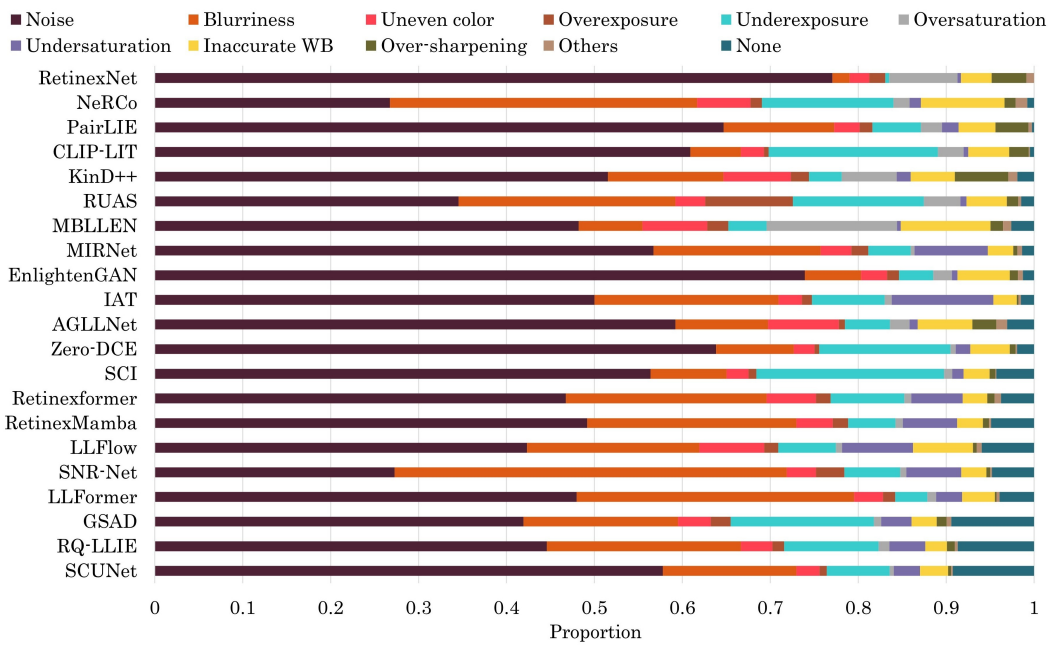


Figure 3: The distortion distributions of all LLIE methods on the LIDQ dataset.

We recruited 34 subjects with normal or corrected-to-normal vision and verified color perception using the Ishihara test (Wang et al., 2023e). Each subject evaluated a subset of images, with every image annotated by at least 15 subjects (Wang et al., 2024a). The evaluation started with qualitative annotations, followed by overall quality ratings (Fang et al., 2020). In total, we collected 89,454 annotations across 5,566 images.

Subjective Data Processing. We apply rigorous outlier and subject filtering based on the ITU-R BT.500-13 methodology (BT.500 ITU-R, 2002) to ensure annotation reliability. Annotations deviating by more than three standard deviations from the mean are marked as outliers, and subjects with outlier rates above 5% are excluded. After filtering, all subjects remain valid, and 2.61% of ratings are discarded as outliers. The mean of the remaining valid ratings is used as the ground-truth mean opinion score (MOS). For qualitative annotations, subjects identify the two most prominent distortion types from ten predefined categories. The final primary distortions are determined by aggregating selections across subjects and choosing the top two. We then convert the selected distortion types and their severity scores into a continuous probability distribution vector over all distortion categories.

Subjective Results and Analysis. To assess annotation reliability, we repeatedly split subjects into two subgroups and computed the Spearman rank correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC) between their mean MOS ratings. Averaged over 100 trials, both metrics exceeded 0.85 (see Table 2), indicating strong rating consistency. Fig. 2 shows the average MOS for each LLIE method, leading to several key observations. First, GT images consistently exhibit higher perceptual quality than all enhanced versions, indicating significant room for improvement. Second, the quality gap between GT and enhanced images is notably larger in the SDSD subset than in LOL, highlighting the greater challenges of enhancing low-light images in the wild. Third, the relative rankings of different LLIE algorithms vary across the two subsets. For example, SCI (Ma et al., 2022) ranks highest among all LLIE algorithms on the SDSD subset, but its ranking significantly drops on the LOL subset. This highlights the limited generalizability of LLIE models and the need for diverse evaluation to ensure reliable performance across varied scenarios.

We show the distortion distributions of enhanced images corresponding to all LLIE methods in Fig. 3. It is evident that noise and blurriness are the most common artifacts in enhanced low-light images, followed by underexposure and white balance (WB) issues, highlighting key challenges faced by current LLIE methods. Fig. 3 also provides an intuitive overview of the relative strengths and weaknesses of various LLIE methods, e.g., NeRCo and SNR-Net show stronger resistance to noise artifacts but are more susceptible to blurriness.

4 PROPOSED METRIC: LIQ^{3}A

Preliminaries. Given an enhanced image $\mathbf{x} \in \mathbb{R}^N$ produced by a specific LLIE algorithm, LIQ^{3}A is designed to estimate the perceptual quality of $\mathbf{x}$ through a mapping $\hat{q} : \mathbb{R}^N \to \mathbb{R}$, and to characterize the distortion profile of $\mathbf{x}$ by converting the qualitative analysis into a distribution over M candidate distortion types via a mapping $\hat{d} : \mathbb{R}^N \to \mathbb{R}^M$. For quality prediction, we adopt a five-level Likert scale $\mathcal{C} = \{1, 2, 3, 4, 5\}$ ({"bad", "poor", "fair", "good", "perfect"}) and define the predicted quality score $\hat{q}$ as

$$\hat{q}(\mathbf{x}) = \sum_{c=1}^{C} \hat{p}(c \mid \mathbf{x}) \times c, \tag{1}$$

where $C = 5$ and $\hat{p}(c \mid \mathbf{x})$ denote the estimated marginal probability of level c. As for distortion analysis, we consider $M = 11$ distortion types as specified in Sec. 3. To build LIQ^{3}A, we leverage the strong representational power of a CLIP-style (Radford et al., 2021) vision-language model, SigLIP-2 (Tschannen et al., 2025), pre-trained on 12 billion image-text pairs. We utilize the built-in language module of SigLIP-2 to bridge the two tasks by constructing a textual template that combines labels from both: "*a photo with {d} artifacts, which is of {c} quality*", yielding $5 \times 11 = 55$ textual descriptions.

Model Specification. We use the NaFlex variant of SigLIP-2, which inherently supports multi-resolution inputs while preserving aspect ratios. Given a patch size and a target sequence length, NaFlex resizes input images to dimensions that are multiples of the patch size, minimizing aspect ratio distortion while ensuring the sequence length remains within bounds. The model comprises a visual encoder $\boldsymbol{f}_{\boldsymbol{\phi}} : \mathbb{R}^N \to \mathbb{R}^K$ and a language encoder $\boldsymbol{g}_{\boldsymbol{\varphi}} : \mathcal{T} \to \mathbb{R}^K$, parameterized by $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$, respectively, where $\mathcal{T}$ denotes the text prompt set. Thanks to the NaFlex mechanism, we can efficiently extract multi-scale visual representations. Each input image $\boldsymbol{x}$ is resized to U different resolutions, from which we derive the visual embedding matrix $\boldsymbol{F}(\boldsymbol{x}) \in \mathbb{R}^{U \times K}$. In parallel, we encode $V = 55$ candidate text prompts to obtain the textual embedding matrix $\boldsymbol{G}(\boldsymbol{x}) \in \mathbb{R}^{V \times K}$.

We then compute the cosine similarity between the visual embedding of the u-th image $\boldsymbol{F}_{u\bullet}$ (as a row vector and for $1 \leq u \leq U$) and the v-th candidate textual embedding $\boldsymbol{G}_{v\bullet}$ (corresponding to a particular set of $\{c, d\}$), averaging across U sub-images to obtain the image-level correspondence score:

$$\text{logit}(c, d|\boldsymbol{x}) = \frac{1}{U} \sum_{u=1}^{U} \frac{\boldsymbol{F}_{u\bullet}(\boldsymbol{x}) \boldsymbol{G}_{v\bullet}^{\intercal}(\boldsymbol{x})}{\|\boldsymbol{F}_{u\bullet}(\boldsymbol{x})\|_2 \|\boldsymbol{G}_{v\bullet}(\boldsymbol{x})\|_2}. \tag{2}$$

After matching the image with all candidate descriptions, we apply a softmax with learnable temperature τ and bias β to compute the joint probability:

$$\hat{p}(c, d \mid \boldsymbol{x}) = \frac{\exp\left(\text{logit}(c, d \mid \boldsymbol{x})/\tau + \beta\right)}{\sum_{c,d} \exp\left(\text{logit}(c, d \mid \boldsymbol{x})/\tau + \beta\right)}. \tag{3}$$

Loss for Quantitative Assessment. We marginalize $\hat{p}(c, d|\boldsymbol{x})$ to compute $\hat{p}(c|\boldsymbol{x})$, from which we obtain the quality estimate $\hat{q}(\boldsymbol{x}) \in \mathbb{R}$ by Eq. equation 1. During training, we sample a mini-batch $\mathcal{B} = \{\boldsymbol{x}_i, q(\boldsymbol{x}_i)\}_{i=1}^{|\mathcal{B}|}$ at each iteration, where $q(\boldsymbol{x}_i)$ is the MOS of $\boldsymbol{x}_i$. We compute a binary label indicating the relative quality ranking of two images within $\mathcal{B}$:

$$p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} 1 & \text{if } q(\boldsymbol{x}_i) \geq q(\boldsymbol{x}_j) \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

Following Thurstone's model (Thurstone, 1927), we estimate the probability that $\boldsymbol{x}_i$ is perceived as better than $\boldsymbol{x}_j$ by:

$$\hat{p}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi\left(\frac{\hat{q}(\boldsymbol{x}_i) - \hat{q}(\boldsymbol{x}_j)}{\sqrt{2}}\right), \tag{5}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. We adopt the fidelity loss (Tsai et al., 2007) as the statistical distance measure:

$$\ell_f(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{\{(\boldsymbol{x}_i, \boldsymbol{x}_j), p\} \in \mathcal{B}} \left(1 - \sqrt{p(\boldsymbol{x}_i, \boldsymbol{x}_j)\hat{p}(\boldsymbol{x}_i, \boldsymbol{x}_j)} - \sqrt{(1 - p(\boldsymbol{x}_i, \boldsymbol{x}_j))(1 - \hat{p}(\boldsymbol{x}_i, \boldsymbol{x}_j))}\right). \tag{6}$$

Table 3: Performance comparison of different IQA algorithms on Subset 1 and Subset 2 (↑ means higher is better, ↓ means lower is better).

Method	Subset 1 (LOL-v1 & LOL-v2)			Subset 2 (SDSD)		
	SRCC (↑)	PLCC (↑)	EMD (↓)	SRCC (↑)	PLCC (↑)	EMD (↓)
MUSIQ (Ke et al., 2021)	0.6871	0.7017	–	0.4643	0.5517	–
CLIPIQA (Wang et al., 2023a)	0.1748	0.2190	–	0.3929	0.4124	–
QualiCLIP+ (Agnolucci et al., 2024a)	0.5788	0.6060	–	0.5952	0.7132	–
UNIQUE (Zhang et al., 2021a)	0.5881	0.6109	–	0.4004	0.5409	–
VisualQuality-R1 (Wu et al., 2025)	0.6897	0.7052	–	0.6346	0.6314	–
DBCNN (Zhang et al., 2018)	0.8072	0.8267	–	0.9284	0.9078	–
HyperIQA (Su et al., 2020)	0.7201	0.7385	–	0.9100	0.9231	–
MANIQA (Yang et al., 2022)	0.8047	0.8260	–	0.9213	0.9354	–
ARNIQA (Agnolucci et al., 2024b)	0.7335	0.7472	–	0.9159	0.9260	–
TOPIQ (Chen et al., 2024)	0.7563	0.7659	–	0.9210	0.9355	–
Q-Align Wu et al. (2024c)	0.8410	0.8512	–	0.9062	0.9047	–
LIQE (Zhang et al., 2023b)	0.8532	0.8657	0.0681	0.9211	0.8888	0.0699
LIQ^{2}A	0.8660	0.8730	0.1742	0.9304	0.8800	0.2853
LIQ^{3}A (Ours)	0.8753	0.8836	0.0740	0.9322	0.9068	0.0664

To improve the precision of quantitative quality prediction, we incorporate an additional loss term based on the PLCC:

$$\ell_p = 1 - \frac{\sum_{i=1}^{|\mathcal{B}|} \left(\hat{q}(\boldsymbol{x}_i) - \overline{\hat{q}}\right)\left(q(\boldsymbol{x}_i) - \overline{q}\right)}{\sqrt{\sum_{i=1}^{|\mathcal{B}|} \left(\hat{q}(\boldsymbol{x}_i) - \overline{\hat{q}}\right)^2} \sqrt{\sum_{i=1}^{|\mathcal{B}|} \left(q(\boldsymbol{x}_i) - \overline{q}\right)^2}}. \tag{7}$$

where $\overline{q} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} q(\boldsymbol{x}_i)$ and $\overline{\hat{q}} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} \hat{q}(\boldsymbol{x}_i)$.

Loss for Qualitative Assessment. Given $\hat{p}(c, d|\boldsymbol{x})$, we marginalize it to obtain $\hat{p}(d|\boldsymbol{x})$. We again use the fidelity loss to measure the distance between the predicted and ground-truth distortion distributions $p(d|\boldsymbol{x}) \in \mathbb{R}^M$:

$$\ell_d(\boldsymbol{x}) = \frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{x} \in \mathcal{B}} \left(1 - \sqrt{p(d|\boldsymbol{x})\, \hat{p}(d|\boldsymbol{x})} - \sqrt{(1 - p(d|\boldsymbol{x}))(1 - \hat{p}(d|\boldsymbol{x}))}\right). \tag{8}$$

Finally, we compute the overall loss as: $\ell = \lambda_1 \ell_f + \lambda_2 \ell_p + \lambda_3 \ell_d$, where λ_1, λ_2, and λ_3 are weighting factors that balance the contribution of each loss term.

5 Experiments and Results

Experimental Setups. We conduct experiments on both subsets of the LIDQ dataset. To ensure meaningful supervision, we perform joint training on both subsets using a pairwise learning-to-rank strategy (Zhang et al., 2021a) restricted to within-subset comparisons. Each of Subset 1 and Subset 2 is partitioned into training, validation, and testing splits in a 7:1:2 ratio, ensuring that visually similar content is confined to a single split to avoid content leakage. Our model is instantiated using the SigLIP-2-base-NaFlex variant (Tschannen et al., 2025), which employs a shared ViT-B/16 architecture (Dosovitskiy et al., 2021) for both visual and language encoders. The visual encoder leverages the NaFlex mechanism to preprocess inputs in an aspect-ratio-preserving manner, based on a preset maximum patch count. We adopt three such settings ($U = 3$): 196, 529, and 1024 patches. The language encoder processes tokenized text truncated to the first 64 tokens, using the Gemma tokenizer (Team et al., 2024). Training is conducted using the AdamW optimizer (Loshchilov & Hutter, 2019) with a weight decay of 10^{-3} and an initial learning rate of 5×10^{-6}, scheduled via cosine annealing (Loshchilov & Hutter, 2017). We train LIQ^{3}A for 8 epochs with a mini-batch size of 16 for each subset. All loss weights, i.e., λ_1, λ_2, and λ_3, are set to 1. We use SRCC and PLCC as prediction monotonicity and precision measures, respectively. Additionally, the Earth Mover's Distance (EMD) (Levina & Bickel, 2001) is utilized to measure the closeness between the predicted and ground-truth distortion distributions.

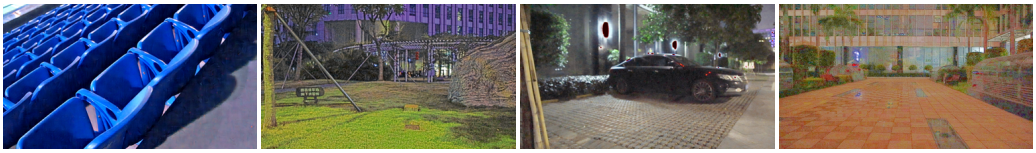
Comparison Methods. We compare the performance of the proposed LIQ^{3}A with eleven BIQA methods, including five pre-trained models—MUSIQ (Ke et al., 2021), CLIPIQA (Wang et al., 2023a), QualiCLIP+ (Agnolucci et al., 2024a), UNIQUE (Zhang et al., 2021a), and VisualQuality-

(a) CLIP-LIT
Quality: 1.438
Underexposure: 0.385
Noise: 0.268
Blurriness: 0.112

(b) LLFlow
Quality: 4.955
None: 0.529
Blurriness: 0.196
Underexposure: 0.106

(c) MIRNet
Quality: 1.478
Under-saturation: 0.421
Noise: 0.218
Blurriness: 0.153

(d) RUAS
Quality: 2.793
Overexposure: 0.453
None: 0.126
Blurriness: 0.111



(e) MBLLEN
Quality: 2.101
Over-saturation: 0.399
None: 0.144
Uneven color: 0.121

(f) KinD++
Quality: 1.177
Noise: 0.425
Over-sharpening: 0.307
Blurriness: 0.117

(g) Retinextformer
Quality: 1.662
Noise: 0.474
Blurriness: 0.278
Others: 0.101

(h) RetinexNet
Quality: 1.000
Noise: 0.522
Inaccurate WB: 0.156
Uneven color: 0.088

Figure 4: Quality scores and probabilities of primary distortions predicted by LIQ^{3}A for enhanced images produced by different LLIE methods.

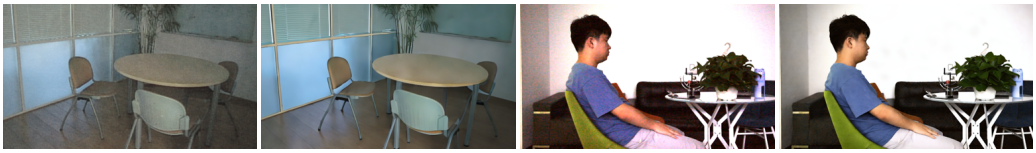
R1 (Wu et al., 2025)—as well as seven models re-trained on LIDQ[2]: DBCNN (Zhang et al., 2018), HyperIQA (Su et al., 2020), MANIQA (Yang et al., 2022), ARNIQA (Agnolucci et al., 2024b), TOPIQ (Chen et al., 2024), LIQE (Zhang et al., 2023b), and Q-Align Wu et al. (2024c). We also evaluate a degenerated variant of our model, referred to as LIQ^{2}A, which is trained solely using MOSs. Among all competing methods, only LIQE and the proposed LIQ^{3}A are capable of performing both quantitative quality prediction and qualitative distortion distribution estimation.

Quantitative Results. We list the quantitative results in Table 3, from which we have several insightful observations. First, by leveraging strong prior knowledge from an advanced multi-modal large language model (MLLM), VisualQuality-R1 outperforms other pre-trained methods across both subsets. Second, all re-trained BIQA models outperform pre-trained ones, highlighting a clear domain shift between low-light enhanced images and standard IQA datasets, and underscoring the need for task-specific fine-tuning. Third, sharing a similar design, LIQ^{3}A outperforms LIQE, validating the superiority of the SigLIP-2 backbone over CLIP and the effectiveness of the NaFlex mechanism in enabling multi-scale representations that better capture quality-aware image features. Fourth, compared to LIQ^{2}A, LIQ^{3}A enables both qualitative distortion pattern estimation and slightly improved quantitative performance, suggesting effective knowledge transfer between the two tasks.

Qualitative Results. We present qualitative examples in Fig. 4 to intuitively demonstrate LIQ^{3}A's ability to perform both quantitative and qualitative evaluations of low-light enhanced images. Thanks to our pairwise learning-to-rank training scheme across both subsets, LIQ^{3}A effectively learns a common perceptual space in which images from Subset 1 (Fig.4(a)–(e)) and Subset 2 (Fig.4(f)–(h)) are well aligned, even though their MOSs are not directly comparable. In addition, LIQ^{3}A's predictions clearly indicate that Noise and Blurriness are the most dominant distortion types, which are consistent with the overall distortion distribution illustrated in Fig. 3.

Generalizability Testing. To evaluate the generalizability of LIQ^{3}A, we perform cross-dataset testing on LIEQ Zhai et al. (2021) and LEISD Lin et al. (2023), corresponding to diverse contents and LLIE algorithms. In addition, we select five low-light images from each of DICM (Lee et al., 2013), MEF (Ma et al., 2015), LIME (Guo et al., 2016), NPE (Wang et al., 2013), and VV (Vonikakis et al., 2018), and enhance them with the same 21 algorithms used in LIDQ, resulting in 550 images. We perform formal subjective testing to annotate their perceptual quality. The resulting dataset, termed Hybrid-LLIE, is used to evaluate the cross-scene generalizability of IQA methods. The results are reported in Table 4, from which we have two primary observations. First, VisualQuality-R1 and

[2]These models are re-trained and evaluated using the same data splits as LIQ^{3}A.



(a) MIRNet (b) MAP-refined (c) RUAS (d) MAP-refined

(e) RetinexNet (f) MAP-refined (g) EnlightenGAN (h) MAP-refined

Figure 5: Visual examples of MAP-refined images from outputs generated by different LLIE methods, where LIQ^{3}A is employed to guide the perceptual optimization.

Table 4: SRCC and PLCC results on three datasets under the cross-dataset setup. The top two performances are highlighted in **bold**.

Method	LIEQ		LEISD		Hybrid-LLIE	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
MUSIQ (Ke et al., 2021)	0.7434	0.7411	0.4512	0.4346	0.4689	0.4566
CLIPIQA (Wang et al., 2023a)	0.4500	0.4664	0.1864	0.1942	0.4306	0.4486
QualiCLIP+ (Agnolucci et al., 2024a)	0.7348	0.7273	0.4849	0.4554	0.5984	0.5789
UNIQUE (Zhang et al., 2021a)	0.7809	0.7770	0.5346	0.5472	0.5371	0.5415
VisualQuality-R1 (Wu et al., 2025)	**0.8487**	**0.8479**	0.7005	0.7100	**0.7120**	**0.7095**
DBCNN (Zhang et al., 2018)	0.6233	0.6298	0.6136	0.6580	0.4747	0.4948
HyperIQA (Su et al., 2020)	0.5485	0.5570	0.5594	0.6247	0.4664	0.4743
MANIQA (Yang et al., 2022)	0.7247	0.7308	0.6333	0.6779	0.6400	0.6424
ARNIQA (Agnolucci et al., 2024b)	0.4563	0.4829	0.5569	0.6148	0.4744	0.4786
TOPIQ (Chen et al., 2024)	0.7130	0.7237	0.6606	0.6967	0.5052	0.5117
Q-Align Wu et al. (2024c)	0.8133	0.8007	0.7329	0.7460	0.6864	0.6898
LIQE (Zhang et al., 2023b)	0.7549	0.7680	**0.7506**	**0.7919**	0.6231	0.6406
LIQ^{3}A (Ours)	**0.8165**	**0.8121**	**0.7729**	**0.7979**	**0.7470**	**0.7398**

Q-Align demonstrate strong performance across all three datasets, highlighting the generalizability of MLLM-based IQA models. Second, with far fewer parameters, LIQ^{3}A attains comparable or superior cross-dataset performance to MLLM-based methods, validating the soundness of our design.

Perceptual Optimization. Beyond serving as a performance measure for LLIE, it is also highly beneficial to explore the use of a quality metric for perceptual optimization. We plug LIQ^{3}A into the maximum a posteriori (MAP) estimation within the diffusion latents framework (Zhang et al., 2025) to perform post-enhancement on the outputs generated by LLIE methods. To evaluate this, we show in Fig. 5 image pairs, consisting of the outputs from different LLIE methods and our MAP-refined results. It is evident that LIQ^{3}A performs effectively within the MAP estimation framework, removing distortions introduced by various LLIE methods and showcasing its reliable understanding of degradation factors through an analysis-by-synthesis manner (Grenander & Miller, 2007).

6 Conclusion

In this work, we conduct a comprehensive quality assessment study for low-light enhanced images. We first present LIDQ, a new dataset containing 5,566 enhanced images produced by 21 modern LLIE methods. Both quantitative and qualitative quality annotations, in the form of MOSs and distortion distributions, respectively, are obtained via well-controlled subjective testing. Building on LIDQ, we develop LIQ^{3}A, a quality metric that effectively quantifies overall image quality and estimates distortion distributions in low-light enhanced images. We believe our dataset and metric will drive progress in both the development of LLIE methods and the advancement of LLIE evaluation metrics. A key limitation of this study is the small volume of annotated data, highlighting the need for larger and more diverse datasets. Methodologically, integrating MLLMs to enhance both quantitative and qualitative LLIE assessments remains an open direction.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Quality-aware image-text alignment for real-world image quality assessment. *arXiv preprint arXiv:2403.11176*, 2024a.

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. ARNIQA: Learning distortion manifold for image quality assessment. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 189–198, 2024b.

Jiesong Bai, Yuhao Yin, and Qiyuan He. Retinexmamba: Retinex-based Mamba for low-light image enhancement. *arXiv preprint arXiv:2405.03349*, 2024.

BT.500 ITU-R. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union*, 2002. URL `https://www.itu.int/rec/R-REC-BT.500`.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *IEEE International Conference on Computer Vision*, pp. 12504–12513, 2023.

Peibei Cao, Zhangyang Wang, and Kede Ma. Debiased subjective assessment of real-world image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 711–721, 2021.

Baoliang Chen, Lingyu Zhu, Hanwei Zhu, Wenhan Yang, Linqi Song, and Shiqi Wang. Gap-closing matters: Perceptual quality evaluation and optimization of low-light image enhancement. *IEEE Transactions on Multimedia*, 2023.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024.

Zhengying Chen, Tingting Jiang, and Yonghong Tian. Quality assessment for comparing image enhancement algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3003–3010, 2014.

Ziteng Cui, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, ZhengKai Jiang, Yu Qiao, and Tatsuya Harada. You only need 90K parameters to adapt light: A light weight transformer for image enhancement and exposure correction. In *British Machine Vision Conference*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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*, pp. 3677–3686, 2020.

Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 22252–22261, 2023.

Ulf Grenander and Michael I. Miller. *Pattern Theory: From Representation to Inference*. Oxford University Press, 2007.

Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pp. 633–651, 2020.

Ke Gu, Guangtao Zhai, Weisi Lin, and Min Liu. The analysis of image contrast: From quality assessment to automatic enhancement. *IEEE Transactions on Cybernetics*, 46(1):284–297, 2015.

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*, pp. 1780–1789, June 2020.

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.

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.

Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 2023.

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.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE International Conference on Computer Vision*, pp. 5148–5157, 2021.

Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2d histograms. *IEEE Transactions on Image Processing*, 22(12):5372–5384, 2013.

Elizaveta Levina and Peter Bickel. The earth mover's distance is the mallows distance: Some insights from statistics. In *IEEE International Conference on Computer Vision*, pp. 251–256, 2001.

Lin Li, Ronggang Wang, Wenmin Wang, and Wen Gao. A low-light image enhancement method for both denoising and contrast enlarging. In *IEEE International Conference on Image Processing*, pp. 3730–3734, 2015.

Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *IEEE International Conference on Computer Vision*, pp. 8094–8103, 2023.

Weitao Lin, Yuxuan Wu, Lishi Xu, Weiling Chen, Tiesong Zhao, and Hongan Wei. No-reference quality assessment for low-light image enhancement: Subjective and objective methods. *Displays*, 78:102432, 2023.

Yunlong Liu, Tao Huang, Weisheng Dong, Fangfang Wu, Xin Li, and Guangming Shi. Low-light image enhancement with multi-stage residue quantization and brightness-aware attention. In *IEEE International Conference on Computer Vision*, pp. 12140–12149, 2023.

Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. MBLLEN: Low-light image/video enhancement using cnns. In *British Machine Vision Conference*, pp. 220, 2018.

Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *International Journal of Computer Vision*, 129(7):2175–2193, 2021.

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.

Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5637–5646, 2022.

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, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012b.

Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Liu Risheng, Ma Long, Zhang Jiaao, Fan Xin, and Luo Zhongxuan. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10561–10661, 2021.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3667–3676, 2020.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, Jul. 1927.

Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: A ranking method with fidelity loss. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 383–390, 2007.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLiP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Vassilios Vonikakis, Rigas Kouskouridas, and Antonios Gasteratos. On the evaluation of illumination compensation algorithms. *Multimedia Tools and Applications*, 77:9211–9231, 2018.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023a.

Miaohui Wang, Zhuowei Xu, Mai Xu, and Weisi Lin. Blind multimodal quality assessment of low-light images. *International Journal of Computer Vision*, pp. 1–24, 2024a.

Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *IEEE International Conference on Computer Vision*, pp. 9700–9709, 2021.

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.

Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 2654–2662, 2023b.

Xuejin Wang, Leilei Huang, Hangwei Chen, Qiuping Jiang, Shaowei Weng, and Feng Shao. Benchmark dataset and pair-wise ranking method for quality evaluation of night-time image enhancement. *IEEE Transactions on Multimedia*, 2024b.

Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *AAAI Conference on Artificial Intelligence*, pp. 2604–2612, 2022.

Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Exposurediffusion: Learning to expose for low-light image enhancement. In *IEEE International Conference on Computer Vision*, pp. 12438–12448, 2023c.

Zhihua Wang, Zhi-Ri Tang, Jianguo Zhang, and Yuming Fang. Toward a blind image quality evaluator in the wild by learning beyond human opinion scores. *Pattern Recognition*, 137:109296, 2023d.

Zhihua Wang, Keshuo Xu, Yang Yang, Jianlei Dong, Shuhang Gu, Lihao Xu, Yuming Fang, and Kede Ma. Measuring perceptual color differences of smartphone photographs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10114–10128, 2023e.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations*, 2024a.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 25490–25500, 2024b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-ALIGN: teaching lmms for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, pp. 54015–54029, 2024c.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*, 2025.

Tao Xiang, Ying Yang, and Shangwei Guo. Blind night-time image quality assessment: Subjective and objective approaches. *IEEE Transactions on Multimedia*, 22(5):1259–1272, 2019.

Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. SNR-aware low-light image enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 17714–17724, 2022.

Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. HVI: A new color space for low-light image enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5678–5687, 2025.

Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *IEEE International Conference on Computer Vision*, pp. 12918–12927, 2023a.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognitio Workshops*, pp. 1191–1200, 2022.

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*, pp. 3063–3072, 2020.

Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality. *IEEE Transactions on Image Processing*, 30:3461–3473, 2021a.

Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021b.

Ying Yang, Tao Xiang, Shangwei Guo, Xiao Lv, Hantao Liu, and Xiaofeng Liao. EHNQ: Subjective and objective quality evaluation of enhanced night-time images. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4645–4659, 2023b.

Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new image contrast enhancement algorithm using exposure fusion framework. In *Computer Analysis of Images and Patterns*, volume 10425, pp. 36–46. 2017.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European Conference on Computer Vision*, pp. 492–511, 2020.

Guangtao Zhai, Wei Sun, Xiongkuo Min, and Jiantao Zhou. Perceptual quality assessment of low-light image enhancement. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(4):1–24, 2021.

Fengqi Zhang, Zhigang Tu, Weifeng Hao, Yihao Chen, Fei Li, and Mao Ye. Zero-shot parameter learning network for low-light image enhancement in permanently shadowed regions. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023a.

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.

Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021a.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081, 2023b.

Weixia Zhang, Dingquan Li, Guangtao Zhai, Xiaokang Yang, and Kede Ma. When no-reference image quality models meet map estimation in diffusion latents. *arXiv preprint arXiv:2403.06406*, 2025.

Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM International Conference on Multimedia*, pp. 1632–1640, 2019.

Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *International Journal of Computer Vision*, 129:1013–1037, 2021b.

Shen Zheng, Yiling Ma, Jinqian Pan, Changjie Lu, and Gaurav Gupta. Low-light image and video enhancement: A comprehensive survey and beyond. *arXiv preprint arXiv:2212.10772*, 2022.