# QuadPrior++: Multi-Dimension Augmented Physical Prior for Zero-Reference Illumination Enhancement

Haofeng Huang, *Student Member, IEEE*, Yifan Li, Wenjing Wang, *Member, IEEE*, Wenhan Yang, *Member, IEEE*, Ling-Yu Duan, *Member, IEEE*, and Jiaying Liu, *Fellow, IEEE*

***Abstract*—Existing low-light enhancement methods typically rely on fitting data mappings (pixel-wise mappings through fully supervised methods or distribution-wise mappings through weakly supervised or self-supervised methods). However, their performance is heavily dependent on specific scenes and fails to adequately model the intrinsic prior of natural images, resulting in poor generalization. To tackle this challenge, we leverage the strengths of powerful generative diffusion models, conditioned on a thoughtfully designed prior, and propose a novel zero-reference low-light enhancement framework that gets rid of dependence on the distribution of low-light images. In detail, we address the most fundamental core by proposing an illumination-invariant prior derived from the theory of physical light transfer, bridging the gap between normal and low-light domains, and enabling zero-shot enhancement without the need for low-light-specific training. A prior-to-image restoration framework is built upon generative diffusion models, pre-trained on normal-light data. During inference, the framework extracts the illumination-invariant prior from low-light inputs and maps them back to high-quality images, naturally for low-light enhancement. Additionally, such intrinsic properties of illumination-invariant prior open up opportunities for distilling diffusion models into compact CNN-based networks. We propose a novel prior-injected distillation paradigm incorporating intensity, frequency, and gradient domain-augmented regularization comprehensively. This distillation framework not only reduces computational costs but also maintains high fidelity and perceptual quality in enhanced outputs, making it more efficient and practical for real-world applications. The approach further extends seamlessly to handle over-exposure scenarios, demonstrating its versatility in addressing complex lighting conditions. Extensive experiments demonstrate the superiority of our framework in various scenarios, as well as its strong interpretability, robustness, and efficiency.**

***Index Terms*—Low-light, zero-reference illumination enhancement, physical prior, model distillation.**

Received 7 March 2025; revised 28 August 2025; accepted 16 November 2025. Date of publication 26 November 2025; date of current version 4 February 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62332010, in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology), in part by the Interdisciplinary Frontier Research Project of PCL under Grant PCL2025QYB013, and in part by the Major Key Project of PCL under Grant PCL2025A03. Recommended for acceptance by J. Gu. *(Corresponding author: Jiaying Liu.)*

Haofeng Huang, Yifan Li, Wenjing Wang, and Jiaying Liu are with Peking University, Beijing 100871, China (e-mail: hhf@pku.edu.cn; 2100012520@stu.pku.edu.cn; daooshee@pku.edu.cn; liujiaying@pku.edu.cn).

Wenhan Yang is with Peking University, Beijing 100871, China, and also with Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: yangwenhan@pku.edu.cn).

Ling-Yu Duan is with the National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China, and also with Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: lingyu@pcl.ac.cn).

The code is available on https://github.com/lyf1212/QuadPrior-plus.

Digital Object Identifier 10.1109/TPAMI.2025.3637277

## I. Introduction

RESTORING images under low-light conditions poses a significant and complex challenge within the field of computer vision. The primary objective is to enhance the overall illumination and reveal obscured details in regions that are not adequately illuminated, thereby improving the overall quality of the image to meet the user demand. However, due to severe and complex noise, unnatural illumination distribution, biased color, *etc*, modeling the distortion and restoring satisfactory results remains a challenging problem.

As a long-history topic, a large number of low-light algorithms have been proposed. In the beginning, researchers develop handcrafted methods based on histogram [4], [5], statistics [6], and Retinex theory [7]. These methods analyze the signal distribution of low light images, model the target, and based on the observation or assumption propose the processing function to amplify the intensity and restore details from severe noisy data. The emergence of the Big Data era has prompted researchers to adopt data-driven methodologies for problem-solving, including the low light enhancement task. The pioneering attempt [8] uses a simple CNN (Convolution Neural Network) trained on synthesized low/normal-light data pairs to establish the low-to-normal mapping. With the following numerous improvements on frameworks [9], [10], models [9], [11] and datasets [2], [12], [13], the end-to-end supervision scheme significantly elevates the performance upper bound of such methods. To get rid of the annoying collection of well-aligned paired datasets, unsupervised [14] and zero-reference methods [1], [15], [16] are further explored. This new branch does not need pixel-to-pixel training, which is expected to benefit the generalization capacity of models. However, despite the existence of abundant methodologies and data collections, there still exist persistent challenges that remain unresolved. For supervised methods, the overfitting of training data results in unsatisfactory performance on unseen data distribution. For handcrafted, unsupervised, or current zero-reference branches, they heavily rely on the user-defined assumption, models, and hyper-parameters [1], [15], [16]. It severely increases the difficulty of application to diverse real-world data.

Recently, generative models have achieved significant breakthroughs with the newly proposed impressive diffusion models [17] capable of generating high-quality images with natural

and reasonable texture as well as satisfactory lighting distribution. Researchers have also attempted to leverage its powerful generative capabilities for the field of image restoration [18], [19]. However, although the generation quality is remarkable, applying them to low-light enhancement tasks still poses challenges. To pursue a diversity of generated results, the pretrained generative model usually introduces texture inconsistency compared with the condition input, which is unacceptable for image restoration. Implementing the generative models into a degraded image domain requires a specific dataset and supervised training as well [14], [19], [20], which inevitably introduces bias due to overfitting. For real-world applications, the long inference Markov chain is time-consuming. To respond aforementioned challenges, it is essential to overcome a critical contradiction: The intrinsic nature of the data prior is highly compact, but to accommodate the sampling randomness of the data distribution, the model becomes extremely complex.

Therefore, we propose a novel prior-based zero-reference low-light enhancement framework. Our primary idea is the development of a set of prior, which serve as an illumination-invariant feature between low-light and normal-light images. Drawing inspiration from physical imaging modeling principles, we propose a distinctive illumination prior, termed the *physical quadruple prior*, derived from the Kubelka–Munk theory of light transfer. Our designed prior efficiently captures the intrinsic content of the signals, thereby freeing the remarkable capacity of the pre-trained large-scale generative model, namely, Stable Diffusion (SD) [21]. This enables the development of an efficient enhancement method that does not rely on low-light data. A prior-to-image mapping framework is proposed *with only normal-light images for training*, incorporated with easily accessible data from the Internet or existing open-source visual datasets. The model learns intrinsic features of well-lit scenes, enabling our prior to extract illumination-invariant representations and map them to high-quality normal-light images—achieving low-light enhancement without low-light data or illumination-specific parameters. Our physical quadruple prior serves as a compact and effective condition, capturing illumination-invariant features by learning intrinsic characteristics of well-lit scenes. This enables high-quality normal-light image reconstruction, achieving low-light enhancement without low-light data or illumination-specific parameters. In pursuit of practical applications, we further propose an efficient distillation paradigm, leveraging the physical quadruple prior to build a prior-aware lightweight model with a CNN-transformer hybrid architecture. By condensing denoising iterations into a single forward pass and introducing multi-dimension constraints from the perspectives of the histogram, gradient, and pyramid structures, our model surpasses the teacher in fidelity and efficiency while dramatically boosting inference speed.

In summary, thanks to our physical quadruple prior, prior-to-image framework, the lightweight version, our approach combines *interpretability*, *robustness*, and *efficiency*. Experimental results demonstrate that our model attains favorable subjective and objective performance across diverse datasets. Our main contributions are concluded as:

- We derive a set of physics-based priors from natural images as the illumination-invariant feature across different lighting conditions. With illumination-invariant prior as the essence of imaging under varying illumination, we are capable of building a zero-reference low-light enhancement model requiring no low-light training data.
- A novel prior-to-image mapping system is proposed that leverages the generative capacity of a pre-trained large-scale diffusion model for the restoration of high-quality results with satisfactory illumination. With illumination-invariant prior and a generative model, our model well handles diverse light conditions, including both over-exposed and under-exposed scenes.
- For enhanced practical applicability, we propose a novel lightweight distillation scheme to further pursue efficiency, fidelity, and perceptual quality. It employs a multi-dimension augmentation constraint from perspectives of *intensity*, *frequency* and *gradient*, and a prior injection technique. Experimental results demonstrate the model's superiority in complexity, generalization capacity, and flexible prioritization between perceptual and quantitative quality according to application needs.

This work is an extension of our previous paper, Quad-Prior [22], published in CVPR'24. Compared with the conference version, we make substantial contributions to method improvement and comprehensive content addition as follows:

(1) *Prior-projection distillation:* Different from the original data distillation scheme, we propose to further inject the extracted prior into the distilled model as additional informative guidance.

(2) *Multi-dimension augmentation:* In the intensity domain, we resample the pixel of the illumination-distorted input with histogram matching to provide unbiased texture guidance. In the frequency domain, we leverage a pyramid decomposition-based training strategy to make full use of enhanced illumination provided by the generative prior-to-image model and precise high-frequency details provided by the low-light inputs. In the gradient domain, we additionally incorporate a gradient amplification technique to suppress noise and maintain clearer details.

(3) *Handling more applications:* To further reveal the potential capability of illumination-invariant prior, we develop a two-stage exposure correction strategy that is still unsupervised but even outperforms some supervised methods.

(4) *More experimental analysis:* We conduct more experiments, design analysis, and ablation studies to demonstrate the effectiveness of our designs on both low-light and over-exposure scenarios and show the superior performance and generalizability of our method over existing works.

The rest of our paper is organized as follows. Section II reviews recent advancements. Section III provides a detailed description of the primary motivation and specific methods. Section IV illustrates sufficient experimental results, demonstrating the effectiveness of our design. Section V poses a conclusion of our work.

## II. Related Work

### A. Supervised Methods

Deep learning has profoundly boosted the development of low-light image enhancement (LLIE). Li et al. [8] pioneered the

first deep learning-based model for LLIE, employing a simple auto-encoder architecture. Following this, a multitude of studies have refined network designs by integrating principles from the Retinex theory [2], [11], [23], Fourier transforms [24], [25], image processing systems [26], and semantic information [27], as well as adopting innovative frameworks such as flow-based generative models [20], vision transformers [9] and diffusion models [19], [28], [29]. Beyond RGB images, a significant trend of research has focused on RAW data [12], [13], [30], videos [31], and multi-modal approaches [32], [33]. Despite their notable achievements, these models predominantly rely on paired data for training, which often limits their flexibility and robustness in unfamiliar scenarios. Recent efforts have sought to reduce such reliance on supervised learning, exploring alternative strategies to enhance adaptability and generalization.

### B. Unsupervised Methods

Some unsupervised methods alleviate the need for paired data, instead utilizing unpaired normal-light and low-light data for training, making efforts on learning a distribution-wise transformation between domains with different illumination conditions. EnlightenGAN [14], FlexiCurve [34], and NeRCo [10] adopt adversarial learning, where discriminators are constructed to regularize the generators to perform low-light enhancement. CLIP-LIT [35] utilizes prompt learning to restore backlit images with complex illuminations. PairLIE [36] learns an adaptive prior from paired low-light instances of the same scene. Based on PairLIE, LightenDiffusion [37] performs content-illumination decomposition in the latent space and retrains a diffusion model. However, these approaches are still limited by the prerequisites of training data, exhibiting insufficient generalizability across diverse real-world scenarios.

### C. Zero-Reference Methods

"Zero-reference" [15] represents a unique unsupervised setting which depends on neither paired nor unpaired data for training. Due to the absence of any paired images or inter-domain relationships, this paradigm is more challenging. However, it also liberates itself from data requirements and overfitting traps, allowing for more flexible application to variable scenarios. On one hand, traditional non-deep low-light enhancement algorithms [4], [5], [38] can also be categorized as zero-reference. These methods predominantly depend on manually crafted priors, such as histogram equalization [4], [5] or Retinex decomposition [38], [39]. On the other hand, as for deep methods, Zero-DCE [15] proposes some novel non-reference loss functions, and a set of learnable curves which are applied to low-light images several times, achieving reasonable illumination, color and details with iterative pixel-wise curve-based transformation. Subsequent works boost inference speed with improved curve formulations [40], [41]. As an alternative approach, RUAS [16] employs a neural architecture search strategy based on the Retinex scheme and reference-free losses. SCI [1] simplifies the iterative process in RUAS into a single step. CoLIE [42] leverages neural implicit representation with aforementioned non-reference losses, for more precise control during the restoration process. Despite notable achievements of zero-reference methods, they are over-sensitive to hyper-parameters and training data distribution, leading to unstable and inconsistent performance. As shown in Fig. 1, treating the same low-light input, SCI [1] produces significantly different over-exposure or under-exposure results with only varied training data. The fundamental challenge lies in the fact that these zero-reference models lack a genuine understanding of illumination. Acquiring lighting knowledge without reference data and relying on artificially defined parameters remains a complex and unresolved problem.

## III. Multi-Dimension Augumented Physical Prior-to-Image Framework

In this section, we introduce our proposed Multi-dimension Augmented Physical Prior-to-image Framework. This framework contains three core components: Physical Quadruple Prior, Prior-to-Img module, and Prior-injected Distillation. The training pipeline of the framework is three-staged:

- First, end-to-end train Physical Quadruple Prior extractor and Prior-to-Img framework. Details are introduced in Sections III-B and III-C.
- Then, refine the original auto-encoder of the pretrained diffusion generative model as illustrated in Section III-C.
- Finally, distill the multi-step enhancement model into a one-step transformer for efficiency with the method proposed in Section III-D.

A simplified outline of our proposed methods is provided in Fig. 3.

### A. Motivation

The challenges of developing low light enhancement methods are threefold:

- To construct the mapping from low-light to normal-light distribution, supervised learning requires well-aligned paired data collection and unsupervised learning introduces bias due to specific low light training datasets and human-defined parameters.
- Chasing high-quality restoration results, recent works [14], [19], [20] choose to embed generative models into the enhancement frameworks. However, the intrinsic diversity of generated results conflicts with the need for signal fidelity for enhancement.
- As modeling and generation tools become increasingly powerful, the computational complexity is unacceptable for real-world scenes, especially for the Markov-based diffusion models.

To mitigate such a problem, 1) we are motivated by illumination-invaiant-based physical principles to solely extract *illumination-invariant priors* and subsequently generate illumination-related information to reconstruct the image, avoiding cumbersome decomposition of both illumination-invariant features and illumination-related counterparts; 2) With the wealth of such illumination-invariant priors, we build a *prior-to-image framework* to adapt a pre-trained diffusion model for restoration of high-quality results with satisfactory illumination;

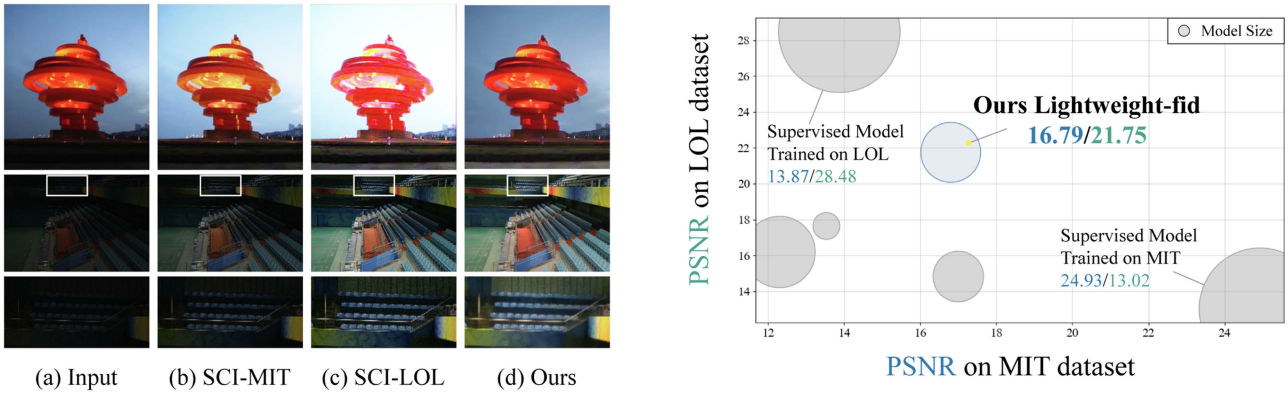


Fig. 1. (Left) A cross-datasets comparison between SCI [1] and the proposed method. SCI is a SOTA method in a zero-shot manner, which is trained on multiple datasets i.e. LOL [2] and MIT [3]. However, it yields inconsistent results in terms of illumination, which indicates its fragility to different data distributions, e.g. dark and moderately dark. In contrast, our method demonstrates greater robustness. (Right) By virtue of our intrinsic prior and insightful distillation scheme, the proposed model indicates better robustness and efficiency.

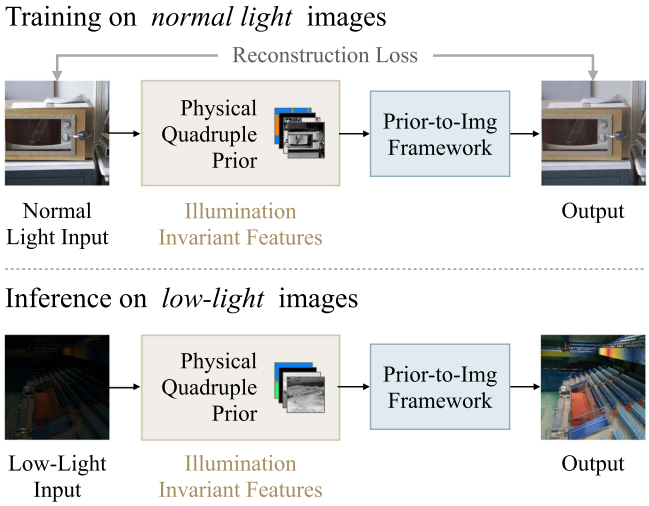


Fig. 2. The primary motivation of our zero-reference low-light enhancement approach. Our model learns the illumination-invariant physical quadruple prior as an effective bridge. The model is constrained to map the physical quadruple prior (whether extracted from low-light or normal-light images) to the normal-light version during training. Therefore, during inference, the model can reconstruct corresponding images with satisfactory illumination even with the priors extracted from the low-light images.

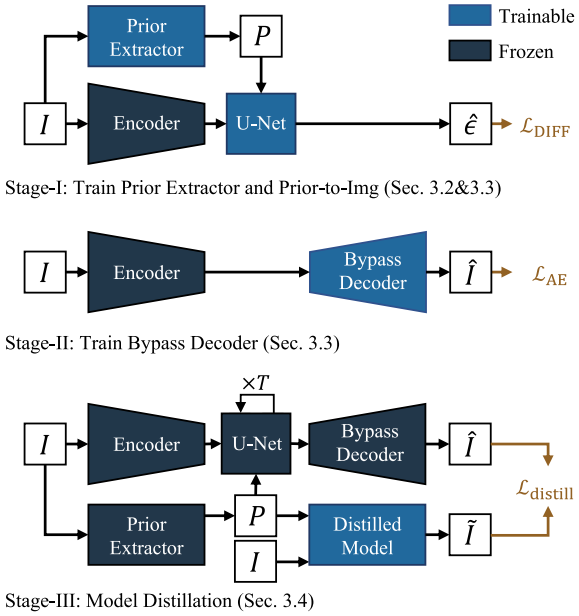


Fig. 3. A simplified outline of our proposed methods which includes a three-stage training. $I$ denotes the input normal-light image, $P$ denotes the learned physical priors and $\epsilon$ denotes predicted diffusion noise. Details can be found in related subsections.

3) we design a *multi-dimension augmented distillation* to avoid the Markov inference chain and further enhance the quality. Fig. 2 depicts the general training and inference pipeline of our method, and the overall distillation pipeline is shown in Fig. 9. We will elaborate on the details of each component as follows.

### B. Learnable Illumination-Invariant Prior

*Physical Quadruple prior:* We start with a brief introduction to the Kubelka-Munk theory [43] of light propagation. As for each spatial location $\mathbf{x}$ on the image plane, the energy it receives can be modeled as

$$E(\lambda, \mathbf{x}) = e(\lambda, \mathbf{x}) \left( (1 - i(\mathbf{x}))^2 R_\infty(\lambda, \mathbf{x}) + i(\mathbf{x}) \right), \quad (1)$$

where $\lambda$ denotes wavelength, $e(\lambda, \mathbf{x})$ denotes the spectrum of the light source, $i(\mathbf{x})$ the specular reflection, and $R_\infty(\lambda, \mathbf{x})$ the material reflectivity. In fact, the Retinex theory is a special case of (1). When we assume the object surface is matte, i.e., $i(\mathbf{x}) \approx 0$, (1) can be reduced to

$$E(\lambda, \mathbf{x}) = e(\lambda, \mathbf{x}) R_\infty(\lambda, \mathbf{x}), \quad (2)$$

which is the same as the Retinex model.

First of all, we denote some variables for simplicity

$$E^\lambda = \frac{\partial E(\lambda, \mathbf{x})}{\partial \lambda}, \; R_\infty^\lambda = \frac{\partial R_\infty(\lambda, \mathbf{x})}{\partial \lambda}, \quad (3)$$

$$E^{\lambda\lambda} = \frac{\partial^2 E(\lambda, \mathbf{x})}{\partial \lambda^2}, \; R_\infty^{\lambda\lambda} = \frac{\partial^2 R_\infty(\lambda, \mathbf{x})}{\partial \lambda^2}. \tag{4}$$

Intuitively, $E$ represents spectral intensity, $E^\lambda$ signifies spectral slope, and $E^{\lambda\lambda}$ denotes spectral curvature.

Following [44], we can derive a set of invariants from (1) based on a series of simplified assumptions. Our goal is to eliminate $i$ and $e$, retaining solely $R_\infty$. As $R_\infty$ is only about material property and is independent of illumination, the derived variable will exhibit illumination invariance. To this end, we make several assumptions as follows:

- Assuming *equal energy* illumination, i.e., $e(\lambda, \mathbf{x})$ is simplified to $\lambda$-independent $\tilde{e}(\mathbf{x})$, and (1) is reduced to

$$E(\lambda, \mathbf{x}) = \tilde{e}(\mathbf{x}) \left( (1 - i(\mathbf{x}))^2 R_\infty(\lambda, \mathbf{x}) + i(\mathbf{x}) \right), \tag{5}$$

  Substituting (5) into $E^\lambda / E^{\lambda\lambda}$ gives

$$\frac{E^\lambda}{E^{\lambda\lambda}} = \frac{\tilde{e}(\mathbf{x})(1 - i(\mathbf{x}))^2 R_\infty^\lambda}{\tilde{e}(\mathbf{x})(1 - i(\mathbf{x}))^2 R_\infty^{\lambda\lambda}} = \frac{R_\infty^\lambda}{R_\infty^{\lambda\lambda}}, \tag{6}$$

  where illumination properties $i$ and $e$ are eliminated. As the material property $R_\infty$ is independent of illumination, it establishes the illumination-invariance of $E^\lambda / E^{\lambda\lambda}$. Now we derive our first illumination invariant,

$$H = \arctan\left( E^\lambda / E^{\lambda\lambda} \right). \tag{7}$$

- *Further* assuming that the surface is *matte*, i.e., $i(\mathbf{x}) \approx 0$, then (1) is reduced to

$$E(\lambda, \mathbf{x}) = \tilde{e}(\mathbf{x}) R_\infty(\lambda, \mathbf{x}), \tag{8}$$

  Similarly, we derive another illumination invariant,

$$\begin{aligned} C &= \log\left( \frac{(E^\lambda)^2 + (E^{\lambda\lambda})^2}{E(\lambda, \mathbf{x})^2} \right) \\ &= \log\left( \frac{(R_\infty^\lambda)^2 + (R_\infty^{\lambda\lambda})^2}{R_\infty(\lambda, \mathbf{x})^2} \right). \end{aligned} \tag{9}$$

- *Further* assuming *uniform* illumination, i.e., $\tilde{e}(\mathbf{x})$ is reduced to a spatial-independent constant $\bar{e}$, and (1) is reduced to

$$E(\lambda, \mathbf{x}) = \bar{e} R_\infty(\lambda, \mathbf{x}), \tag{10}$$

  Similarly, we derive our third illumination invariant,

$$\begin{aligned} W &= \tan\left( \left| \frac{\partial E(\lambda, \mathbf{x})}{\partial \mathbf{x}} \frac{1}{E(\lambda, \mathbf{x})} \right| \right) \\ &= \tan\left( \left| \frac{\partial R_\infty(\lambda, \mathbf{x})}{\partial \mathbf{x}} \frac{1}{R_\infty(\lambda, \mathbf{x})} \right| \right). \end{aligned} \tag{11}$$

Although effective, the Kubelka-Munk theory [43] is typically applied to grayscale images, which is limited to represent colors. Considering the three aforementioned illumination invariants do not fully encompass color information, we directly provide some compact color compensation with inter-channel relativity between RGB components.

- Assuming that chromatic order maintains approximate invariance under varying photometric conditions, we leverage the relative order of RGB channels as an essential illumination-invariant prior, denoted as $O$.

*Learning through Neural Networks:* Drawing inspiration from Gaussian color models [43] and CIConv [45], we first estimate the observed energy $\hat{E}$ along with its derivatives $\hat{E}^\lambda$ and $\hat{E}^{\lambda\lambda}$ via linear mapping:

$$\begin{bmatrix} \hat{E}(x, y) \\ \hat{E}^\lambda(x, y) \\ \hat{E}^{\lambda\lambda}(x, y) \end{bmatrix} = \mathcal{W} \begin{bmatrix} R(x, y) \\ G(x, y) \\ B(x, y) \end{bmatrix}, \tag{12}$$

where $x$ and $y$ denote positions in the image, and $\mathcal{W}$ is a $3 \times 3$ matrix. In our method, the weight matrix $\mathcal{W}$, conventionally heuristically specified in prior works [43], [45], is instead derived through data-driven optimization of natural image distributions via our prior-image framework. It means that the learning of the Physical Quadruple Prior is not directly supervised by any ground truth (which is inaccessible) but optimized via the subsequent Prior-to-Image module and diffusion loss, which is introduced in Section III-C. It is initialized with the predefined $\mathcal{W}$ as an informative prior and tuned for the generative model.

The spatial derivative $\partial E / \partial \mathbf{x}$ in (11) is derived along both the x- and y-axis, denoted as $\partial E / \partial \mathbf{x} = (E_x, E_y)$, with its magnitude computed as $|\partial E / \partial \mathbf{x}| = \sqrt{E_x^2 + E_y^2}$. Ultimately, $E$, $E_x$, and $E_y$ are computed by convolution operations using Gaussian smoothing kernels and derivative filters with a spatially adaptive scale parameter $\sigma$, which is predicted from the input image. Similarly, $E^\lambda$ is obtained from $\hat{E}^\lambda$, and $E^{\lambda\lambda}$ is obtained from $\hat{E}^{\lambda\lambda}$. Now we have estimated $H$, $C$, and $W$ from the input image.

Our last illumination invariant, the order of RGB channels, is defined as three channels as follows,

$$O(x, y) = \left[ O_R(x, y), O_G(x, y), O_B(x, y) \right], \tag{13}$$

where $O_R$ represents the order of the $R$ channel in RGB, normalized to [-1,1]. $O_G$ and $O_B$ are treated similarly.

Finally, $H$, $C$, $W$, and $O$ are concatenated in the channel dimension to form our physical quadruple prior.

*Physical Explanation:* First, the gradient magnitude $W$ is derived from intensity-normalized spatial derivatives of spectral intensity, which is established by the mathematical formulation in 11. As for $H$, according to [43], it indicates the hue, i.e., $\arctan(\lambda_{\max})$ of the material. As for $C$, within the chromaticity plane parametrized by spectral wavelengths, chroma can be represented by radial distance to the origin, while hue refers to the angular position. Moreover, to establish an intuitive geometric interpretation, we can transform the Cartesian coordinate $(a, b)$ to polar system, where the angle is given by $\arctan(b/a)$, and the radius is computed as $\sqrt{(a)^2 + (b)^2}$. Acknowledging the definition in (7) and (9), it becomes evident that $C$ is associated with chroma. Fig. 4 provides a visualization of our physical quadruple prior. Further analysis and comparative evaluations will be detailed in Section IV-C.

### C. Prior-to-Image via Diffusion Models

The ultimate goal is to preserve illumination-invariant information while completely removing lighting-dependent components. However, achieving this precise decomposition poses

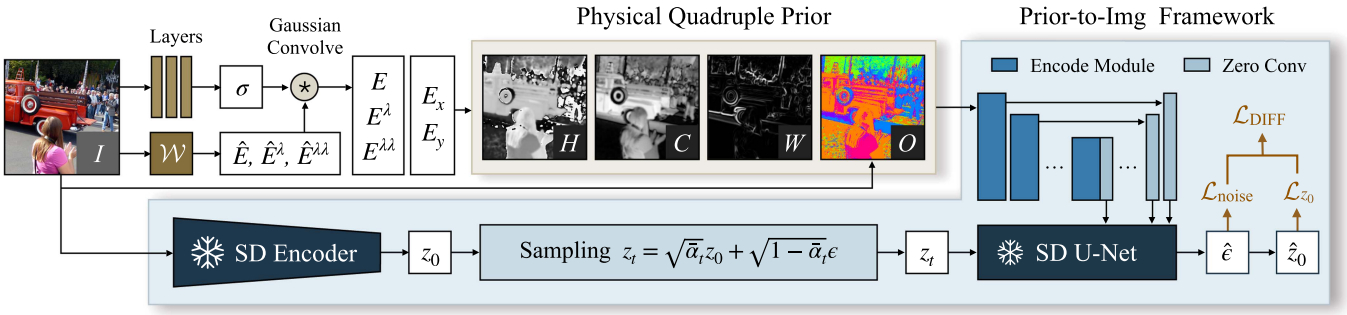


Fig. 4. The training process of our prior-to-image framework and illustration of our physical quadruple prior. We start by estimating the illumination-invatiant prior from the input image $I$. During the training phase, the model dynamically learns the weight matrix $\mathcal{W}$ and the layers for predicting the scale $\sigma$. Subsequently, a static SD encoder is leveraged to extract the latent representation $z_0$ from the input image $I$. Following this, we sample noisy latent $z_t$ based on $z_0$. Finally, the physical quadruple prior is encoded by convolutional and transformer modules, and is then merged with a frozen SD U-net to predict both noise $\epsilon$ and $z_0$.

significant challenges and continues to be an open problem in image modeling. While our proposed physical quadruple prior - $H, C, W$ and $O$ - effectively capture illumination-independent characteristics from multiple perspectives, certain information loss is unavoidable. Consequently, image reconstruction from this prior representation remains a complex and non-trivial undertaking.

Considering the difficulty of improving the representative ability of prior while maintaining its illumination-invariant, we utilize a large-scale pretrained diffusion model with powerful generative abilities to predict the incomplete components from the extracted prior. Stable Diffusion v1-5 [21] is adopted and converted into the conditional mode by integrating the ControlNet [46] architecture. In this setup, the physical quadruple prior serves as the conditioning input to guide the Stable Diffusion model.

The overall pipeline is depicted in Fig. 4. In the training phase, the frozen pretrained SD encoder first compresses the image $I$ into a latent representation $z_0$. $z_t$ is sampled at a random noise schedule $t \in \{1, \ldots, T\}$ using

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{14}$$

where $\{\bar{\alpha}_t\}$ is a set of hyperparameters [17]. The model is optimized to predict $\epsilon$ from $z_t$ using our prior as the conditioning input. Our prior-to-image model consists of a frozen pre-trained U-Net, a trainable duplicate of this U-Net that extracts features from the quadruple prior, and zero convolutions introduced by [46], which help stabilize the initial training phase. Essentially, the model learns to predict $\epsilon$ from $z_t$, effectively performing the denoising process. During testing, the illimination-invariant prior is estimated from input image $I$, which is employed as the condition for the diffusion-based model to iteratively predict $z_0$ in the reverse diffusion process. Finally, $z_0$ is projected back into the pixel domain using a pre-trained VAE decoder.

- The standard training objective for diffusion models is based on the distance between actual Gaussian noises $\epsilon$ and the predicted ones:

$$\mathcal{L}_{\text{noise}} = ||\epsilon - \hat{\epsilon}||_2^2. \tag{15}$$

(a) $I$ (b) SD Decoded (c) $\tilde{I}$ (d) Ours

Fig. 5. Structural details preservation with our bypass decoder. (a) Input image $I$, which results in latent $z_0$. (b) $z_0$ decoded by the SD decoder. (c) The synthesized distortion of $I$. (d) $z_0$ decoded by our decoder with encoder features from $\tilde{I}$ as compensation.

To speed up convergence, we incorporate additional regularization in the context of $z_0$. Using (14), we derive:

$$\mathcal{L}_{z_0} = ||z_0 - \hat{z}_0||_2^2 = \left|\left| z_0 - \frac{z_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right|\right|_2^2. \tag{16}$$

We combine these two losses to form the final objective:

$$\mathcal{L}_{\text{DIFF}} = \mathcal{L}_{z_0} + \mathcal{L}_{\text{noise}}. \tag{17}$$

- Stable Diffusion (SD) uses an auto-encoder (AE) to compress the image $I$ into a latent representation $z_0$, reducing computational costs. However, the auto-encoder introduces significant structural distortion. For instance, as shown in Fig. 5(b), the face of the police on horseback is severely compromised. To address this, we improve the decoder by incorporating intermediate features of the encoder with an effective fine-tuning scheme. As illustrated in Fig. 6, during training, we synthesize degradations with random illumination jittering and noise, producing distorted images $\tilde{I}$ based on clean ones $I$. Note that the degradations are based on predefined hyper-parameters and do not introduce any dataset or distribution. It does not conflict with our zero-shot setting. The decoder then reconstructs $z_0$ by combining features $z^1, z^2$, and $z^3$ extracted from $\tilde{I}$. This process encourages the decoder to capture informative details from $\tilde{I}$ while preserving the illumination characteristics of $I$. We utilize convolutional blocks for feature fusion and a

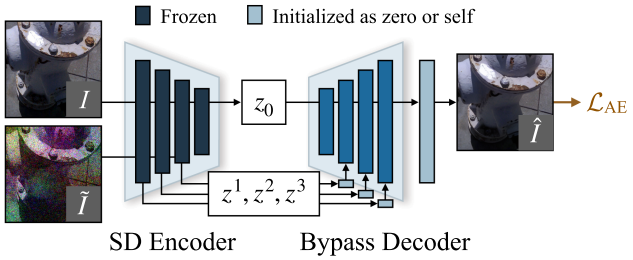


Fig. 6. The design of our bypass decoder. We augment the input data $I$ for the corresponding degraded counterparts $\tilde{I}$, and fine-tune the pretrained decoder network to reconstruct $I$ with distorted encoder features from $\tilde{I}$.

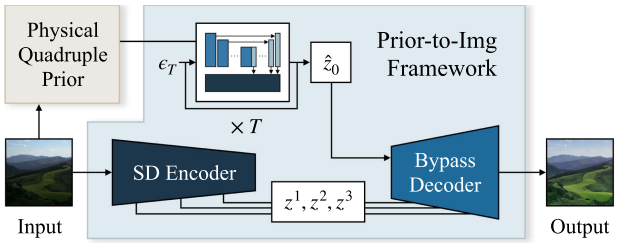


Fig. 7. The inference process of our approach. To start with, we extract a physical quadruple prior of a low-light image and use it as the condition of prior-to-image framework. Subsequently, the denoising model iteratively performs a reverse diffusion process starting from Gaussian noise $\epsilon_T$, yielding the latent representation $\hat{z}_0$. Ultimately, our bypass decoder project $\hat{z}_0$ back into the pixel domain with the guidance of intermediate features extracted from illumination-distorted input.

residual module for post-processing. To minimize harmless disruption on the original decoding process at the start of training, we initialize these layers to zero or identity, and adopt a combination of L1 loss and Perceptual Loss [47] as the $\mathcal{L}_{AE}$:

$$\mathcal{L}_{\text{AE}}(I, \hat{I}) = \mathcal{L}_1 \left( \text{Dec}_{\text{bypass}}(\text{Enc}(\tilde{I})), \hat{I} \right) \tag{18}$$

$$+ \mathcal{L}_{\text{vgg}} \left( \text{Dec}_{\text{bypass}}(\text{Enc}(\tilde{I})), \hat{I} \right), \tag{19}$$

where Enc is the original encoder with extra intermediate features as output, and $\text{Dec}_{\text{bypass}}$ is the finetuned decoder with features injection which is named as the *bypass decoder*. As demonstrated in Fig. 5(d), significant detail preservation is achieved using our bypass decoder. During testing, intermediate features extracted from the encoder help recover finer details in the decoding process, as shown in Fig. 7. Our bypass decoder effectively reconstructs details while maintaining the enhanced illumination in $\hat{z}_0$.

- Noise poses a significant challenge in low-light image enhancement. While our prior is not specifically designed for denoising, we devise a straightforward strategy to suppress noise. During training, we distort the input image $I$ with random Gaussian-Poisson compound noise while extracting the physical quadruple prior. This approach enhances the model's robustness to high-frequency disturbances and makes it concentrate exclusively on essential illumination-invariant clues.

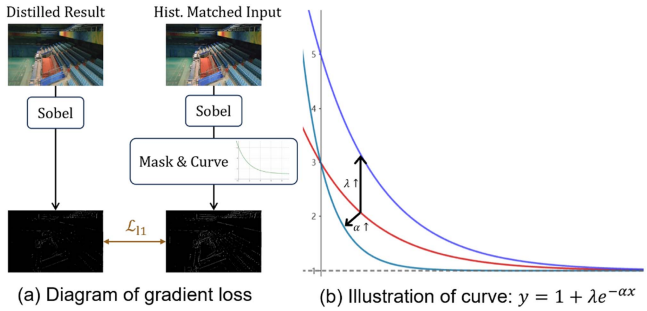


Fig. 8. Illustration of gradient amplification. (a) We leverage the Sobel operator to extract gradients of enhanced results by the distilled model and histogram matched results, and close the distance between the former with amplified ones. (b) We design a flexible curve-based amplification method, effectively balances the trade-off between fidelity and sharpness.

### *D. Prior-Injected Distillation With Multi-Dimension Augmentation*

Although some enhanced sampler like DPM-Solver++ [48] accelerates inference of diffusion models, 10 iterations of the denoising process are still time-consuming and impractical. In our previous work [22], we employ a simple distillation mechanism with only images enhanced by our main prior-to-image model. However, such a vanilla distillation approach only depends on data-level alignment between teacher and student, which fails to fully leverage the robust and powerful illumination-invariant prior that integrates physical principles and general knowledge of pre-trained large models. Additionally, it is constrained by potential noise, color shifts, and detail incompleteness of original model results, leading to suboptimal performance. Considering the above drawbacks, we conduct a thorough exploration of low-light inputs from both the feature and data perspective. We propose a novel prior-aware distillation paradigm that receives illumination-invariant prior as conditions. We have observed that SD is inherently constrained by the structure degradation within encoder-decoder architecture, leading to an imbalance between noise and detail preservation in the results. Despite we have refined the original VAE of SD via Bypass Decoder design, the encoder module is still unchanged because it determines the latent space for the pretrained diffusion model, which still affects the signal fidelity. Therefore, we propose a pyramid decomposition-based regularization, involving detail-rich histogram matched results as the compensation for informative high frequency components. For better perceptuality, we embrace a gradient amplification technique, effectively enhances sharpness and overall visual quality. Fig. 9 shows the pipeline of our distillation scheme. We shall introduce more details as follows.

*Prior-aware Feature Projection:* We utilize to pretrained illumination-invariant extractor used in our diffusion-based prior-to-image model, achieving a robust physical quadruple prior. We regard it as general conditions with rich and impact information, and insert them through SFT layers [49] into the distilled model. SFT layers modulate the features in an adaptive manner, which can be formulated as (20), where $f$ and $\hat{f}$ denote original features and the ones after projection. $F_{\text{scale}}$ and $F_{\text{shift}}$ denote learnable transformation module. We initialize $F_{\text{sihft}}$ with

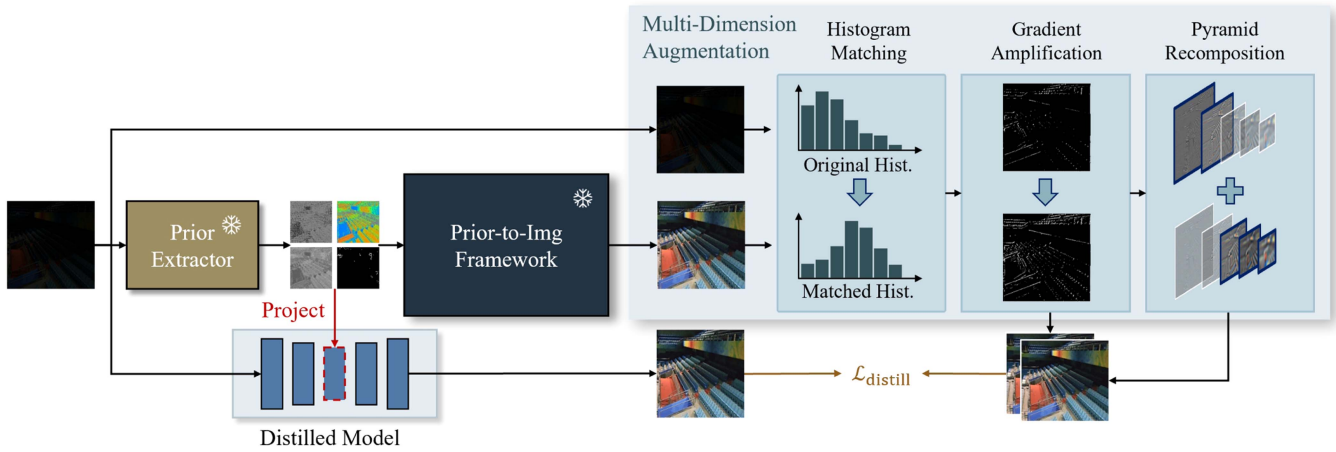


Fig. 9. Our distillation pipeline, includes *prior-projection lightweight model* and *multi-dimension augmentation paradigm*, which includes (a) intensity matching, (b) pyramid-based frequency recompisition, and (c) gradient amplification. We make full use of the illumination-distorted inputs to construct more informative pseudo-targets for distillation, achieving results with both high-quality illumination and clear details.

zeros to stabilize the training process.

$$\hat{f} = F_{\text{scale}}(f) \cdot f + F_{\text{shift}}(f) \tag{20}$$

*Intensity Matching and Frequency Decoupling:* Despite our prior-to-image system achieving high-quality results with plausible low-frequency information such as illumination and colors, the resolution is inherently limited by pretrained diffusion models, hampering the reconstruction of high-frequency components. On the other side, although the low-light images are corrupted with severe noise, they preserve sharp details to some extent. Therefore, to fully unleash the potential structural information that lies in the low-light images, we employ an augmentation strategy on both intensity and frequency domains with histogram matching and pyramid decomposition. For the intensity domain, we want to fully utilize the clue from the single input data while the range of input data is distinct from the wanted target data. Benefiting from the well-trained teacher model which provides a reasonable distribution assumption of target illumination, we globally augmented the intensity of the input data via histogram matching:

$$I_{\text{match}} = \text{HM}(I, I_{\text{enhance}}), \tag{21}$$

where HM denotes the histogram matching algorithm. For the frequency domain, we believe that as for low-frequency components, enhanced results $I_{\text{enhance}}$ by our full model will be served as the target with high-quality illumination and colors, and as for high-frequency components, our histogram matching images $I_{\text{match}}$ provides good assumption. To decouple different parts of frequencies, we utilize Laplacian Pyramid $\mathcal{P}$ to decompose $I_{\text{enhance}}$ and $I_{\text{match}}$, merging high-frequency details from $I_{\text{match}}$ and low-frequency information from $I_{\text{enhance}}$ with a reweighted MSE loss:

$$\begin{aligned}\mathcal{L}_{\text{pyramid}} = & \left\|\mathcal{P}^{\text{low}}(I_{\text{res}}), \mathcal{P}^{\text{low}}(I_{\text{enhance}})\right\|_2^2 \\ & + \lambda_{\text{details}} \cdot \left\|\mathcal{P}^{\text{high}}(I_{\text{res}}), \mathcal{P}^{\text{high}}(I_{\text{match}})\right\|_2^2,\end{aligned} \tag{22}$$

where $I_{\text{res}}$ is generated by distilled models, $\mathcal{P}^{\text{low}}, \mathcal{P}^{\text{high}}$ represents low- and high-frequency parts of pyramid. $\lambda_{\text{details}}$ is hyperparameter and is set to 5.

*Gradient Amplification for Better Perceptual Quality:* Although details information has been compensated through pyramid decomposition, the neural network will inadvertently tend to over-smooth and details incompleteness, degrading visual perceptuality. To mitigate such an issue, we propose a gradient amplification technique, which over-optimizes gradients of target images, achieving more effective regularization on the distilled model, as shown in Fig. 8

In particular, we extract the gradient of the target and the model's outputs with the Sobel operator, denoted as $\nabla I_{\text{gt}}$ and $\nabla I_{\text{res}}$, respectively. Note that we omit the x- and y-axis for simplicity. Subsequently, we filter out small gradients in $\nabla I_{\text{gt}}$, which can be regarded as noise, and amplify the remaining parts with an exponential curve, which can be expressed as:

$$\mathcal{T}(\nabla I_{\text{gt}}) = \begin{cases} 0, & \nabla I_{\text{gt}} < p_\gamma, \\ \nabla I_{\text{gt}}, & \nabla I_{\text{gt}} \geq p_\gamma \end{cases}, \tag{23}$$

$$\nabla \hat{I}_{\text{gt}} = 1 + \lambda \cdot e^{-\alpha \cdot \mathcal{T}(\nabla I_{\text{gt}})}. \tag{24}$$

where $p_\gamma$ denotes $p-$quantile of gradient values. We set $\lambda = 5$. We use the amplified gradient as a pseudo-label, regularize the distilled model with L1 loss:

$$\mathcal{L}_{\text{grad}} = \left\|\nabla I_{\text{res}} - \nabla \hat{I}_{\text{gt}}\right\|_1. \tag{25}$$

$\alpha$, $\lambda$, and $\gamma$ are hyper-parameters controlling the slope, intensity, and range of amplificatio,n respectively. As illustrated by Fig. 8, larger $\alpha$ promotes slope and pulls the curve to the origin $O$, while large $\gamma$ tugs the intersection point of the curve and the y-axis, enhancing overall amplified intensity. Moreover, such a gradient amplification method simultaneously enhances both large and medium gradients and eliminates the disturbance of small noise, effectively improving the sharpness of the results.

The overall objective of distillation is:

$$\mathcal{L}_{\text{distill}} = \lambda_{\text{pm}} \cdot \mathcal{L}_{\text{pyramid}} + \lambda_{\text{grad}} \cdot \mathcal{L}_{\text{grad}} + \lambda_{\text{lpips}} \cdot \mathcal{L}_{\text{vgg}}, \quad (26)$$

where $\mathcal{L}_{\text{vgg}}$ indicates perceptual loss based on VGG [47] network and we experimnetally set $\lambda_{\text{lpips}} = 2$ and $\lambda_{\text{pm}} = \lambda_{\text{grad}} = 1$. With different settings of $\alpha$ and $\lambda$ in $\mathcal{L}_{\text{grad}}$, we achieve two distilled models that perform better on fidelity or percetualit,y respectively.

### E. Generalization for Zero-Shot Exposure Correction

Intuitively, our proposed physical quadruple prior essentially concentrates on the reflectance of object surfaces, and can be generalized to diverse illumination conditions, contributing to the reconstruction of high-quality normal-light images. Motivated by such robust properties of the illumination-invariant prior and prior-to-image model, we extend our applications to more general and complex illuminations, i.e., exposure correction in a zero-shot manner.

However, the matte surface assumption in (8) limits proper illumination modeling under over-exposed scenes. To this end, we first normalize the illumination distribution of the overexposed image through histogram equalization, which serves as the input of our prior-to-image model. It is noteworthy that we applied a **mask** to filter out severely overexposed regions, mitigating the impact of excessively large overexposed areas on the global color distribution during histogram equalization. Subsequently, we calculated the cumulative density function of the histogram with the unmasked region, resulting in the final transformation of the light distribution to be applied globally.

For better sharpness and applicable efficiency, we conduct the aforementioned distillation paradigm on over-exposed scenarios, achieving two versions of the model, catering to diverse requirements of realism and sharpness. We will show more results as follows.

## IV. Experiments

### A. Implementation Details

*Framework Development:* We leverage COCO-2017 [56] train and unlabeled set as the training dataset. As for the distilled applications on over-exposed scenes, we adopt the dataset collected by [57] as the training dataset. We train our framework over 140 k iterations with the ADAM optimizer [58], learning rate as 1e-4 and batch size as 8. We adopt FP16 mixed precision and DeepSpeed [59] to save GPU memory. All evaluations and training of distillation models are conducted on RTX 4090 GPUs. We distill two versions of our lightweight model with different hyper-parameter settings of gradient amplification, namely $\mathcal{F}_{\text{fide}}$ and $\mathcal{F}_{\text{per}}$, focusing on fidelity and perceptuality, respectively. Refering to (24), (26), for the former one, we set $\alpha = 1$ and $\gamma = 0.7$, while for the later one, we set $\alpha = 2$ and $\gamma = 0.8$.

*Evaluation Protocal:* As for low-light enhancement, we conduct a comparison on LOL [2], [50], LSRW [51] and FiveK [3] datasets. For LOL, we adopt the official test sets of LOL v1 [2] and LOL v2 [50], resulting in 115 low-/normal-light image pairs. LSRW provides 50 pairs of testing images with rich details. For MIT, we follow Retinexformer [9] to split 500 pairs for testing. As for the exposure correction task, we randomly select 203 over-exposed images from SICE dataset [54]. We adopt PSNR, SSIM, LPIPS [60], and LOE [61] as Full-Reference (FR) metrics. We additionally leverage BRISQUE [62], NIQE [63], CLIPIQA [64] and NIMA [65] as Non-Reference (NR) metrics for better alignment with human perception. All metrics are computed after bilinear interpolation to align diverse output sizes with ground truth. Besides, we present the computational complexity (multiple-accumulate operations, MACs), network parameters, and processing time for input images of resolution $512 \times 512 \times 3$ in Table II.

*Compared Methods:* Our model is compared with twelve unsupervised low-light image enhancement methods. Among these, EnlightenGAN [14], PairLIE [36], NeRCo [10], CLIP-LIT [35] and LightenDiffusion [37] utilize unpaired low-light-related data. The remaining sevel methods, ExCNet [53], ZeroDCE [15], ZeroDCE++[40], RUAS [16], SCI [1], CoLIE [42] and FourierDiff [25], are zero-reference. Additionally, we benchmark seven supervised methods which indicate a quantitative upper bound of our task. As for exposure correction, we compared our method with traditional histogram-based algorithm [4], [5], supervised model trained on low-light dataset [19] and multi-exposure dataset [57], and some representative unsupervised low-light enhancement method [10], [15], [23], [25], [35], [37].

### B. Benchmarking Results

*Low-Light Enhancement:* Tables I and II represent quantitative results with Full-Reference (FR) and Non-Reference (NR) metrics, respectively. Our framework surpasses the majority of unsupervised approaches and substantially reduces the performance disparity with supervised methods. Although supervised methods achieve superior performance, they may overfit to specific training data. For instance, Retinexformer [9], trained on the LOL dataset, exhibits significant performance degradation on MIT-Adobe FiveK dataset, even worse than our method. Some unsupervised methods trained with unpaired data also overfit to relevant training data, such as NerCO [10] and SCI [1]. Meanwhile, non-data-driven unsupervised methods like ExCNet [53], CoLIE [42], and FourierDiff [25] present slightly inferior performance on both FR and NR scores. On the contrary, our model demonstrates consistent robustness across the LOL, LSRW, and MIT datasets simultaneously, without requiring dataset-specific adjustments. This adaptability arises from our model's capacity to learn comprehensive illumination insights from the physical quadruple prior and normal light images. As a result, our model is less prone to overfitting and demonstrates superior effectiveness in diverse and unfamiliar scenarios. Although LightenDiffusion [37] achieves commendable performance, it requires paired pixel-aligned images of diverse brightness levels for training, increasing the cost of data collection. Additionally, its performance on NR metrics indicates that it does not align well with human perception.

TABLE I

BENCHMARKING RESULTS FOR LOW-LIGHT ENHANCEMENT WITH FR METRICS. WE CATEGORIZE THESE METHODS INTO **S**UPERVISED AND **US**UPERVISED. WE COMPARE OUR METHODS WITH UNSUPERVISED ONES, AND HIGHLIGHT THE TOP-RANKING SCORE IN **RED**, THE SECOND IN **BLUE** AND THE THIRD IN **GREEN**.

| Datasets | | Train Set | LOL [2], [50] | | | | LSRW [51] | | | | MIT-Adobe FiveK [3] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | | | PSNR↑ | SSIM↑ | LPIPS↓ | LOE↓ | PSNR↑ | SSIM↑ | LPIPS↓ | LOE↓ | PSNR↑ | SSIM↑ | LPIPS↓ | LOE↓ |
| S | Retinex-Net [2] | LOL | 16.19 | 0.403 | 0.534 | 0.346 | 15.61 | 0.414 | 0.454 | 0.336 | 12.30 | 0.687 | 0.258 | 0.244 |
| | KinD [23] | LOL | 20.21 | 0.814 | 0.147 | 0.245 | 16.41 | 0.484 | 0.337 | 0.374 | 14.71 | 0.756 | 0.176 | 0.174 |
| | KinD++ [52] | LOL | 16.64 | 0.662 | 0.410 | 0.288 | 16.08 | 0.402 | 0.371 | 0.394 | 15.76 | 0.650 | 0.319 | 0.176 |
| | URetinex-Net [11] | LOL | 20.93 | 0.854 | 0.104 | 0.245 | 18.27 | 0.518 | 0.419 | 0.352 | 14.10 | 0.734 | 0.182 | 0.187 |
| | Retinexformer [9] | LOL | 28.48 | 0.877 | 0.117 | 0.256 | 14.97 | 0.342 | 0.359 | 0.389 | 13.87 | 0.692 | 0.222 | 0.224 |
| | Retinexformer [9] | MIT | 13.02 | 0.426 | 0.365 | 0.280 | 15.03 | 0.399 | 0.327 | 0.418 | 24.93 | 0.907 | 0.063 | 0.162 |
| | DiffLL [29] | LOL | 28.54 | 0.870 | 0.102 | 0.253 | 17.26 | 0.398 | 0.437 | 0.397 | 15.81 | 0.719 | 0.244 | 0.213 |
| | Diff-Retinex [19] | LOL | 26.20 | 0.897 | 0.082 | 0.218 | 15.94 | 0.487 | 0.339 | 0.399 | 16.81 | 0.728 | 0.241 | 0.244 |
| US | ExCNet [53] | test images | 16.25 | 0.454 | 0.382 | 0.277 | 15.74 | 0.408 | 0.339 | 0.388 | 14.21 | 0.719 | 0.197 | 0.197 |
| | EnlightenGAN [14] | own data | 18.57 | 0.700 | 0.302 | 0.291 | 17.08 | 0.470 | 0.327 | 0.387 | 13.28 | 0.738 | 0.203 | 0.199 |
| | PairLIE [36] | LOL+SICE [54] | 19.70 | 0.774 | 0.235 | 0.278 | 17.60 | 0.512 | 0.329 | 0.390 | 10.55 | 0.642 | 0.273 | 0.225 |
| | NeRCo [10] | LSRW | 19.67 | 0.720 | 0.266 | 0.310 | 19.45 | 0.549 | 0.288 | 0.372 | 17.33 | 0.767 | 0.208 | 0.213 |
| | CLIP-LIT [35] | own data | 14.82 | 0.524 | 0.371 | 0.320 | 13.48 | 0.405 | 0.353 | 0.417 | 17.00 | 0.781 | 0.159 | 0.194 |
| | ZeroDCE [15] | SICE [54] | 17.64 | 0.572 | 0.316 | 0.296 | 15.85 | 0.453 | 0.317 | 0.393 | 13.53 | 0.725 | 0.201 | 0.191 |
| | ZeroDCE++ [40] | own data | 17.03 | 0.445 | 0.314 | 0.391 | 16.25 | 0.462 | 0.329 | 0.389 | 12.33 | 0.408 | 0.280 | 0.417 |
| | RUAS [16] | MIT | 13.62 | 0.462 | 0.346 | 0.292 | 13.02 | 0.359 | 0.379 | 0.372 | 9.53 | 0.610 | 0.301 | 0.272 |
| | RUAS [16] | LOL | 15.47 | 0.490 | 0.305 | 0.330 | 14.27 | 0.470 | 0.465 | 0.376 | 5.15 | 0.373 | 0.669 | 0.399 |
| | RUAS [16] | FACE [55] | 15.05 | 0.456 | 0.371 | 0.292 | 14.03 | 0.403 | 0.384 | 0.392 | 5.00 | 0.366 | 0.685 | 0.398 |
| | SCI [1] | MIT | 11.67 | 0.395 | 0.361 | 0.286 | 11.79 | 0.317 | 0.401 | 0.380 | 16.29 | 0.795 | 0.143 | 0.165 |
| | SCI [1] | LOL+LSRW | 16.97 | 0.532 | 0.312 | 0.289 | 15.24 | 0.424 | 0.322 | 0.380 | 7.83 | 0.573 | 0.360 | 0.187 |
| | SCI [1] | FACE [55] | 16.80 | 0.543 | 0.322 | 0.297 | 15.16 | 0.408 | 0.326 | 0.393 | 10.95 | 0.684 | 0.272 | 0.205 |
| | CoLIE [42] | test images | 14.90 | 0.499 | 0.327 | 0.273 | 14.00 | 0.408 | 0.351 | 0.384 | 18.50 | 0.789 | 0.167 | 0.199 |
| | FourierDiff [25] | test images | 16.95 | 0.604 | 0.293 | 0.257 | 15.64 | 0.466 | 0.327 | 0.377 | 17.81 | 0.793 | 0.168 | 0.203 |
| | LightenDiffusion [37] | LOL+LSRW | 20.44 | 0.801 | 0.202 | 0.261 | 18.44 | 0.534 | 0.320 | 0.375 | 21.23 | 0.797 | 0.172 | 0.208 |
| | **Ours** | COCO [56] | 20.31 | 0.808 | 0.202 | 0.281 | 16.96 | 0.564 | 0.408 | 0.367 | 18.51 | 0.785 | 0.163 | 0.188 |
| | **Ours Lightweight-fid** | LOL+LSRW | 21.75 | 0.825 | 0.196 | 0.244 | 18.82 | 0.555 | 0.305 | 0.369 | 16.79 | 0.764 | 0.173 | 0.187 |
| | **Ours Lightweight-per** | LOL+LSRW | 20.92 | 0.811 | 0.195 | 0.254 | 18.64 | 0.549 | 0.311 | 0.371 | 12.56 | 0.717 | 0.201 | 0.228 |

TABLE II

BENCHMARKING RESULTS FOR LOW-LIGHT ENHANCEMENT WITH NR METRICS AND COMPUTATIONAL COMPLEXITY. WE CATEGORIZE THESE METHODS INTO **S**UPERVISED AND **US**UPERVISED. WE COMPARE OUR METHODS WITH UNSUPERVISED ONES, AND HIGHLIGHT THE TOP-RANKING SCORE IN **RED**, THE SECOND IN **BLUE** AND THE THIRD IN **GREEN**.

| Datasets | | Train Set | LOL [2], [50] | | LSRW [51] | | MIT-Adobe FiveK [3] | | Computaional Complexity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | | | BRISQUE↓ | NIMA↑ | BRISQUE↓ | NIMA↑ | BRISQUE↓ | NIMA↑ | MACs | Params | Times (s) |
| S | Retinex-Net [2] | LOL | 25.496 | 5.017 | 29.373 | 4.673 | 22.781 | 4.631 | 87.28 G | 555.21 k | 0.0332 |
| | KinD [23] | LOL | 26.882 | 4.805 | 24.220 | 4.760 | 19.918 | 4.702 | 159.56 G | 8.02 M | 0.356 |
| | KinD++ [52] | LOL | 30.677 | 5.004 | 28.763 | 4.766 | 20.204 | 4.759 | 258.62 G | 8.27 M | 0.3592 |
| | URetinex-Net [11] | LOL | 27.642 | 4.975 | 30.020 | 4.597 | 22.014 | 4.628 | 229.00 G | 340.11 K | 0.0959 |
| | Retinexformer [9] | LOL | 21.939 | 4.478 | 28.114 | 4.403 | 19.194 | 4.891 | 68.39 G | 1.61 M | 0.1327 |
| | Retinexformer [9] | MIT | 27.180 | 4.383 | 29.424 | 4.507 | 16.107 | 4.921 | 68.39 G | 1.61 M | 0.1327 |
| | DiffLL [29] | LOL | 19.152 | 4.393 | 20.423 | 4.514 | 19.472 | 4.174 | 21.88 G | 22.08 M | 0.3399 |
| | Diff-Retinex [19] | LOL | 29.724 | 4.745 | 26.061 | 4.647 | 26.694 | 4.708 | 6.08 T | 47.70 M | 1.0257 |
| US | ExCNet [53] | test images | 40.628 | 4.474 | 28.997 | 4.477 | 17.809 | 4.558 | / | 8.27 M | 12.973 |
| | EnlightenGAN [14] | own data | 17.515 | 4.466 | 19.079 | 4.617 | 16.184 | 4.433 | 65.88 G | 8.64 M | 0.014 |
| | PairLIE [36] | LOL+ | 23.128 | 4.107 | 28.975 | 4.191 | 30.651 | 3.809 | 89.99 G | 341.77 k | 0.029 |
| | NeRCo [10] | LSRW | 14.265 | 4.848 | 14.710 | 4.820 | 22.700 | 4.562 | 822.8 G | 23.30 M | 0.538 |
| | CLIP-LIT [35] | own data | 37.167 | 4.100 | 25.881 | 4.205 | 15.518 | 4.359 | 73.04 G | 278.79 k | 0.018 |
| | ZeroDCE [15] | SICE [54] | 34.342 | 4.334 | 24.531 | 4.366 | 24.239 | 4.166 | 20.92 G | 79.42 k | 0.010 |
| | ZeroDCE++ [40] | own data | 33.185 | 4.394 | 22.881 | 4.529 | 20.382 | 4.183 | 20.37 M | 10.56 k | 0.002 |
| | RUAS [16] | MIT | 25.736 | 4.470 | 24.820 | 4.621 | 17.586 | 4.525 | 950.8 M | 3.438 k | 0.009 |
| | RUAS [16] | LOL | 35.158 | 4.364 | 37.882 | 4.329 | 16.013 | 4.411 | 950.8 M | 3.438 k | 0.009 |
| | RUAS [16] | FACE [55] | 41.506 | 4.489 | 43.342 | 4.549 | 34.598 | 4.572 | 950.8 M | 3.438 k | 0.009 |
| | SCI [1] | MIT | 18.578 | 4.395 | 20.327 | 4.603 | 19.466 | 4.503 | 95.16 M | 0.258 k | 0.001 |
| | SCI [1] | LOL+LSRW | 31.613 | 4.480 | 26.100 | 4.606 | 17.914 | 4.348 | 95.16 M | 0.258 k | 0.001 |
| | SCI [1] | FACE [55] | 31.816 | 4.442 | 22.234 | 4.577 | 18.160 | 4.095 | 95.16 M | 0.258 k | 0.001 |
| | CoLIE [42] | test images | 31.182 | 4.463 | 24.112 | 4.584 | 20.756 | 4.404 | 865.7 G | 133.07 k | 1.787 |
| | FourierDiff [25] | test images | 25.782 | 4.392 | 25.899 | 4.623 | 22.183 | 4.363 | 110.84 T | 547.41 M | 15.440 |
| | LightenDiffusion [37] | LOL+LSRW | 14.769 | 4.372 | 18.093 | 4.588 | 25.522 | 4.219 | 10.7 G | 26.99 M | 0.423 |
| | **Ours** | COCO [56] | 11.577 | 4.043 | 12.897 | 4.615 | 23.237 | 4.980 | / | 1.313 B | 2.896 |
| | **Ours Lightweight-fid** | LOL+LSRW | 10.281 | 4.297 | 18.145 | 4.625 | 17.369 | 4.599 | 10.57 G | 389.81 k | 0.0125 |
| | **Ours Lightweight-per** | LOL+LSRW | 10.412 | 4.486 | 17.490 | 4.643 | 15.212 | 4.535 | 10.57 G | 389.81 k | 0.0125 |

Notably, by virtue of our novel distillation paradigm which incorporates illumination-invariant prior injection and pyramid decomposition-based regularization, our lightweight model also achieves leading FR and NR metrics, even surpassing the teacher model. Our proposed gradient-amplification strategy further improves the sharpness of enhanced results that better align with human perception, as verified by better NR scores in Table II. Qualitative comparisons are presented in Figs. 1, 10, and 11. Compared with the full model, our lightweight version boosts the inference speed up to 230x faster and reduces model parameters by 1200 times.

*Exposure Correction:* Previous methods for exposure correction predominantly rely on supervised learning with paired datasets, which may lead to overfitting and limited generalizability. Our physical prior-based model can comprehensively learn illumination distribution from normal-light images, thus

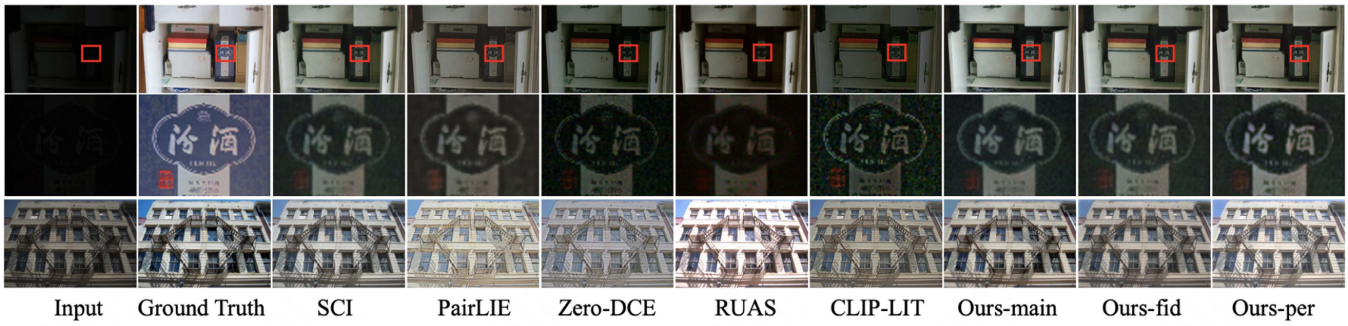


Fig. 10. Example low-light enhancement results on the LOL datasets (top two rows) and MIT-Adobe FiveK dataset (bottom row). Ours-main denotes the proposed diffusion-based enhancement model. Ours-fid and Ours-per denote distilled models optimized for fidelity and perceptual quality.


Fig. 11. More low-light enhancement competing results on the LOL datasets with our main and distilled model.

TABLE III
BENCHMARKING RESULTS FOR EXPOSURE CORRECTION, INCLUDING TRADITIONAL METHODS (TRAD.), SUPERVISED METHOD TRAINED ON OVER-EXPOSED DATA (SUP. ON EXP.), SUPERVISED METHODS ONLY TRAINED ON LOW-LIGHT DATA (SUP. ON LOW), UNSUPERVISED METHODS TRAINED ON UNPAIRED LOW-/NORMAL-LIGHT DATA (UNSUP.) AND ZERO-SHOT METHODS WHICH HAVE NEVER BEEN TRAINED ON OVER-/UNDER-EXPOSED DATA. WE COMPARE OUR METHODS WITH UNSUPERVISED ONES, AND HIGHLIGHT THE TOP-RANKING SCORE IN **RED**, THE SECOND IN **BLUE** AND THE THIRD IN **GREEN**. HE MEANS "HISTOGRAM EQUALIZATION" AND $*$ DENOTES THE EXPOSURE CORRECTION VERSION OF OUR MODELS.

| Metrics | | Full-Reference | | | | Non-Reference | | | Computational Complexity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Category | PSNR↑ | SSIM↑ | LPIPS↓ | LOE↓ | BRISQUE↓ | NIQE↓ | CLIPIQA↑ | MACs | Params | Times (s) |
| HE [4] | Trad. | 16.57 | 0.786 | 0.223 | 0.159 | 25.615 | 3.683 | 0.513 | / | / | 0.1783 |
| CLAHE [5] | Trad. | 16.18 | 0.813 | 0.201 | 0.157 | 22.952 | 3.562 | 0.574 | / | / | 0.2154 |
| MSEC [57] | sup. on exp. | 17.01 | 0.790 | 0.191 | 0.204 | 20.640 | 2.965 | 0.567 | 13.02 G | 433.57 k | 0.0734 |
| KinD [23] | sup. on low | 12.87 | 0.806 | 0.216 | 0.165 | 22.192 | 3.410 | 0.499 | 159.56 G | 8.02 M | 0.0356 |
| Diff-Retinex [19] | sup. on low | 13.24 | 0.699 | 0.350 | 0.273 | 27.392 | 3.920 | 0.502 | 6.08 T | 47.70 M | 1.0257 |
| CLIP-LIT [35] | unsup. | 12.56 | 0.796 | 0.208 | 0.189 | 23.246 | 3.631 | 0.564 | 73.04 G | 278.79 k | 0.0179 |
| NerCO-LOL [10] | unsup. | 17.48 | 0.758 | 0.331 | 0.239 | 31.508 | 4.000 | 0.315 | 822.8 G | 23.30 M | 0.538 |
| NerCO-LSRW [10] | unsup. | 17.31 | 0.788 | 0.290 | 0.219 | 21.182 | 3.504 | 0.347 | 822.8 G | 23.30 M | 0.538 |
| ZeroDCE [15] | unsup. | 11.37 | 0.733 | 0.281 | 0.219 | 26.929 | 3.451 | 0.471 | 20.92 G | 79.42 k | 0.0099 |
| LightenDiffusion [37] | unsup. | 15.31 | 0.811 | 0.201 | 0.210 | 22.811 | 3.628 | 0.499 | 10.7 G | 26.99 M | 0.423 |
| FourierDiff [25] | zero-shot | 9.81 | 0.653 | 0.413 | 0.292 | 42.877 | 4.949 | 0.442 | 110.84 T | 547.41 M | 15.440 |
| **Ours*** | zero-shot | 19.06 | 0.832 | 0.210 | 0.198 | 10.564 | 3.569 | 0.492 | / | 1.313 B | 4.6157 |
| **Ours*-lightweight-fid** | zero-shot | 19.73 | 0.859 | 0.148 | 0.199 | 16.168 | 4.068 | 0.487 | 10.57 G | 389.81 k | 0.0125 |
| **Ours*-lightweight-per** | zero-shot | 16.06 | 0.790 | 0.201 | 0.186 | 14.457 | 3.183 | 0.571 | 10.57 G | 389.81 k | 0.0125 |

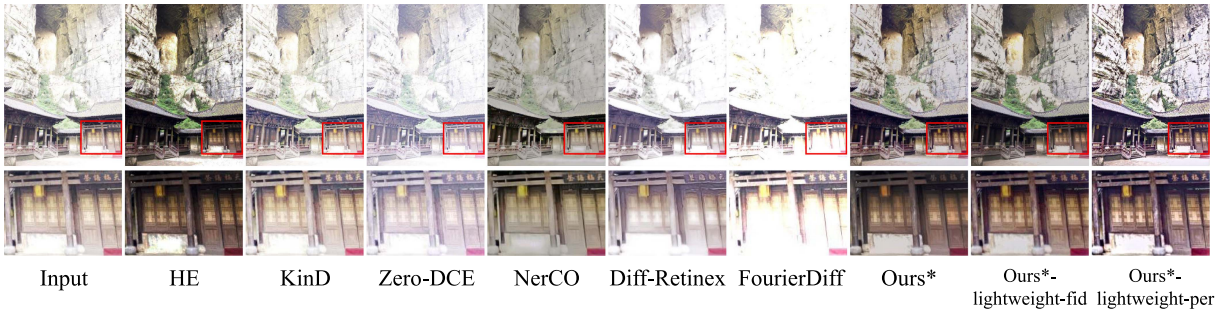


Fig. 12. Example exposure correction comparisons on the SICE dataset. $*$ denotes the sexposure correction version of our models.

effortlessly extending its application to exposure correction, particularly demonstrating robust capabilities in overexposure suppression. Table III and Fig. 12 show quantitative and qualitative results on over-exposed scenarios. With our fruitful distillation paradigm, our lightweight model performs better than the diffusion-based teacher model and also enhances sharpness and contrast with the gradient amplification strategy, as indicated by better NR scores. Traditional methods such as histogram equalization [4] suffer from color artifacts and hue shifts, such as the yellowish color spots on the white wall inside the red bounding box. As for low-light enhancement methods, while supervised ones perform well on specific under-exposed datasets, they fail to suppress over-exposed regions, such as KinD [23] and Diff-Retinex [19]. Some semi-supervised methods [15], utilize a subset of over-exposed images from the SICE dataset [54] for training, but are still unsatisfactory to dynamically correct the exposure. FourierDiff [25] suggests a novel zero-shot training-free method, however is limited to under-lit scenes, further exacerbating the overexposed regions. Although LightenDiffusion [35], [37] makes remarkable progress on low-light enhancement within an unsupervised paradigm, they struggle to generalize to more complex illumination conditions. NerCO [10] demonstrates its potential generalizability on overexposed inputs, but exhibits obvious color shifts. Notably, our method surpasses the supervised MSEC [57] method trained on exposure correction datasets in terms of Full-Reference (FR) metrics, with clearer details and perception-friendly illumination. In summary, our method leverages illumination-invariant prior to learn normal-light distribution and the knowledge of high-quality images. Our model paradigm achieves robust and superior performance, in a zero-shot manner that requires no additional training or fine-tuning. Our productive distillation mechanism also demonstrates its effectiveness and generalization on both low-light and over-exposed scenes.

### C. Ablation Studies

*Prior Design:* We begin by examining the impact of each component, $H$, $C$, $W$, and $O$, within our illumination-invariant prior. Each part plays a crucial role, and their combination forms illumination-invariant features that achieve the best performance. As shown in Fig. 13, the removal of $H$ or $C$ causes color bias or faded white appearance. As previously analyzed in Section III-B, $H$ and $C$ are associated with hue and chromacity. However, as depicted at the right side of Fig. 13, $H$ and $C$ do not strictly correspond to typical hue and chroma. This indicates that while the prior aligns with the physical explanation we have established, it can also learn more sophisticated representations.

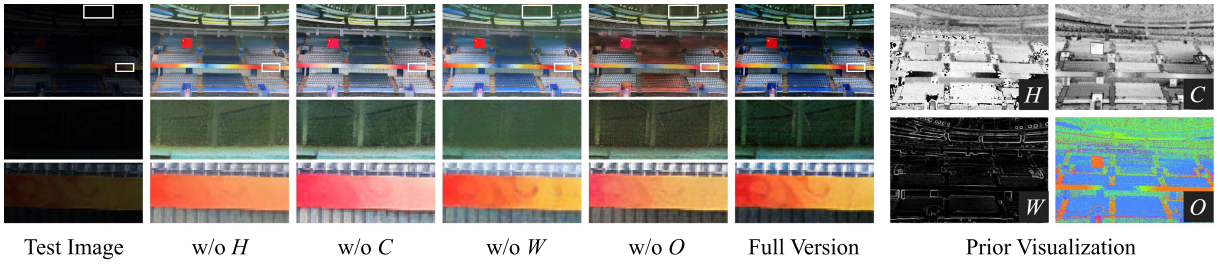


Fig. 13. (Left) Ablation study of different prior designs. (Right) The visualization of our physical quadruple prior.

TABLE IV
ABLATION STUDIES ON THE EFFECT OF OUR METHOD DESIGNS

| Datasets | | LOL [2] | | | |
|---|---|---|---|---|---|
| Metrics | | PSNR↑ | SSIM↑ | LPIPS↓ | LOE↓ |
| Prior | Ours w/o $H$ | 17.60 | 0.756 | 0.262 | 0.314 |
| | Ours w/o $C$ | 17.60 | 0.762 | 0.262 | 0.313 |
| | Ours w/o $W$ | 17.77 | 0.749 | 0.291 | 0.313 |
| | Ours w/o $O$ | 18.63 | 0.764 | 0.285 | 0.315 |
| | HS channels in HSV | 18.04 | 0.562 | 0.498 | 0.410 |
| | CIConv | 17.02 | 0.455 | 0.551 | 0.421 |
| | Reflectance by PairLIE [36] | 20.16 | 0.790 | 0.287 | 0.296 |
| AE | SD Decoder [21] | 19.26 | 0.665 | 0.243 | 0.353 |
| | Consistency Decoder [67] | 19.35 | 0.686 | 0.235 | 0.350 |
| | Ours w/o noise aug. | 19.54 | 0.714 | 0.265 | 0.348 |
| | Ours w/o illum. aug. | 19.85 | 0.783 | 0.213 | 0.246 |
| Ours Final Version | | 20.31 | 0.808 | 0.202 | 0.281 |


Fig. 14. Ablation Study of different prior-to-image architectures.


Fig. 15. Ablation study of different decoders in our framework.

As previously explained, $W$ refers to intensity-normalized spatial derivatives of spectral intensity, which encode local illumination variations. When $W$ is excluded, the model confuses the boundary between light and shadow, as evident in the second row of Fig. 13. Furthermore, the lack of $O$ introduces significant color distortions, such as the erroneous conversion of blue tones to orange. Ultimately, the full version of our prior demonstrates the most precise details, proper colors, and enhanced contrast, highlighting the importance of integrating all components for optimal performance.

We further demonstrate the unique effectiveness of our physical quadruple prior by substituting it with alternative features. Specifically, we examine three representative ones: (1) Naive HS channels in the HSV color space. (2) CIConv [45], a trainable prior similar to our $W$. (3) The reflectance estimated by Retinex-based PairLIE [36] trained on LOL. In Table IV, the HS channels result in a substantial loss of content information, hampering PSNR and SSIM scores significantly. Although CIConv achieves better performance on high-level vision tasks with illumination invariance, it is limited by color fidelity due to information sacrifices, resulting in a notable decline in overall performance. In addition, PairLIE still relies on training within a specific low-light image domain to obtain reflectance as illumination prior. Without additional fine-tuning, its performance remains inferior to our method, which exhibits stronger generalizability. *Prior-to-Image Framework:* We conduct an ablative trial to demonstrate the superiority of our ControlNet-based prior-to-image model design. We investigate a substitute diffusion-based architecture, SR3 [66], which performs well in super-resolution tasks. However, when integrated into our prior-to-image framework, as depicted in Fig. 14, SR3 demonstrates significant color bias and noise-related artifacts. This limitation arises because our prior, designed to eliminate light-related features, inherently discards certain image information. Consequently, the prior-to-image model must compensate for these gaps and reconstruct the missing details. Since SR3 directly incorporates prior conditions as the inputs of the denoising network, which compromises the inherent generative capabilities and results in unsatisfactory performance in reconstructing high-quality images compared to our framework.

*Auto-Encoder:* We evaluate the effectiveness of our bypass decoder by comparing it with the original decoder used in SD and the Consistency Decoder from DALL-E 3 [67], a cutting-edge diffusion-based decoder designed for better decoding performance of SD VAEs. As shown in Fig. 15, both the original SD decoder and the Consistency Decoder fail to preserve textual details in the reconstructed images. By comparison, our decoder utilizes illumination-independent features from the input images, achieving sharp and distortion-free reconstructions.

Besides, we also evaluate the necessity of introducing data distortion during bypass decoder training. As shown in Table IV, the data distortion benefits the final performance of our method with better robustness.

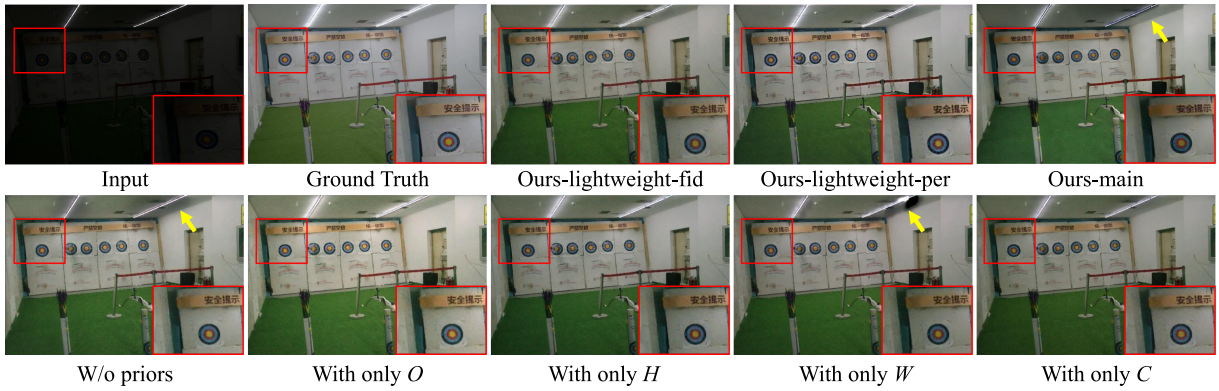


Fig. 16. Ablative examples of prior insertion in our distillation paradigm. "Our-main" means our full model without distillation.

TABLE V
ABLATION STUDY OF OUR DISTILLATION PARADIGM

| Datasets | | LOL [2] | | | |
|---|---|---|---|---|---|
| Metrics | | PSNR↑ | SSIM↑ | LPIPS↓ | LOE↓ |
| Prior | Ours w/o prior insertion | 20.45 | 0.798 | 0.290 | 0.273 |
| | Ours with only $H$ | 21.51 | 0.807 | 0.203 | 0.246 |
| | Ours with only $C$ | 21.73 | 0.816 | 0.199 | 0.246 |
| | Ours with only $W$ | 21.82 | 0.816 | 0.202 | 0.246 |
| | Ours with only $O$ | 20.42 | 0.804 | 0.207 | 0.248 |
| Loss | w/o $\mathcal{L}_{\text{pyramid}}$ | 20.84 | 0.808 | 0.205 | 0.252 |
| | w/o gradient amp. | 21.17 | 0.819 | 0.197 | 0.247 |
| | w/o $\mathcal{L}_{\text{pyramid}}$ & $\mathcal{L}_{\text{grad}}$ | 20.89 | 0.813 | 0.214 | 0.250 |
| | $\mathcal{L}_{\text{pyramid}} \rightarrow \mathcal{L}_2$ on HM | 21.24 | 0.813 | 0.203 | 0.249 |
| Ours lightweight-fidelity | | 21.75 | 0.825 | 0.197 | 0.244 |
| Ours lightweight-perceptuality | | 20.92 | 0.811 | 0.195 | 0.254 |

*Framework Distillation:* We analyze two-fold innovations of our distillation paradigm, namely prior insertion and loss regularization. Example low-light enhancement results on LOL [2] are depicted in Table V, Fig. 16.

On one hand, we abolish some or all prior during distillation, finding worse illumination and details results. As pointed by the yellow arrows in Fig. 16, without $H$, $C$, and $O$ prior which represent hue, chromacity, and RGB channel order, respectively, black shadows and artifacts appeared near the light tubes. Although prior $O$ helps to achieve better brightness, it struggles with color shift, as the white wall becomes greenish. Overall, the combination of all the prior helps to obtain better illumination and colors. Notably, by virtue of gradient amplification, our lightweight model achieves much sharper results. Since the lightweight model is trained with the results from the main model based on all prior, qualitative differences may not be pronounced.

On the other hand, we discard pyramid-based or gradient-based loss, demonstrating that both are indispensable for high-quality reconstruction. Note that leveraging only the gradient amplification without histogram matching results in detail compensation leads to worse results on PSNR and SSIM. This is because the network rigidly learns to enhance contrast without sufficient structural hints of low-light images, resulting in more artifacts. Furthermore, to verify the unique superiority of our frequency domain augmentation, we test on another intuitive loss on histogram matching results, i.e., utilize MSE loss on the downsampled enhanced output of our full models and original MSE loss on the histogram matching results. We denote it as $\mathcal{L}_{pyramid} \rightarrow \mathcal{L}_2$ *on HM*. However, such naive regularization struggles to decouple low-frequency illumination and high-frequency details, resulting in blurry and under-lit results.

We compare our lightweight model with SOTA in terms of model size and PSNR performance in Fig. 1. After distillation, our lightweight versions reduce running time to 231x faster and the parameter size reduces from 1.313B to 389.81 k. Compared to some supervised methods that exhibit obvious overfitting, our lightweight model achieves better generalization, even surpassing some supervised approaches.

## V. CONCLUSION

We present a novel zero-reference low-light enhancement framework, uniquely designed without the need for any low-light training data. Our primary innovation is the physical quadruple prior derived from principles of light transfer theory, and an effective prior-to-image mapping framework based on generative diffusion models and an improved bypass encoder. With the robustness of our illimination-invariant prior, our prior-based system can be easily extended to diverse lighting environment, i.e., over-exposed scenarios. For enhanced practical applications, we propose a novel distillation scheme with prior-aware architecture and a powerful regularization strategy, which has been proved to be effective on both under-lit and over-exposed scenes. Extensive experimental evaluations demonstrate that our approach achieves superior performance across a wide range of scenarios, showcasing its robustness and versatility.

## REFERENCES

[1] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5627–5636.
[2] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proc. Brit. Mach.Vis. Conf.*, 2018, p. 155.
[3] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 97–104.
[4] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Muller, "Contrast-limited adaptive histogram equalization: Speed and effectiveness," in *Proc. 1st Conf. Vis. Biomed. Comput.*, 1990, pp. 337–345.
[5] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *J. Signal Process. Syst.*, vol. 38, pp. 35–44, 2004.

[6] S. Malik and R. Soundararajan, "A low light natural image statistical model for joint contrast enhancement and denoising," *Signal Process. Image Commun.*, vol. 99, 2021, Art. no. 116433.

[7] Z.-U. Rahman, D. J. Jobson, and G. A. Woodell, "Retinex processing for automatic image enhancement," *J. Electron. Imag.*, vol. 13, pp. 100–110, 2004.

[8] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, 2017.

[9] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 12470–12479.

[10] S. Yang, M. Ding, Y. Wu, Z. Li, and J. Zhang, "Implicit neural representation for cooperative low-light image enhancement," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 12872–12881.

[11] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5891–5900.

[12] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5891–5900.

[13] H. Jiang and Y. Zheng, "Learning to see moving objects in the dark," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 7323–7332.

[14] Y. Jiang et al., "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.

[15] C. Guo et al., "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1777–1786.

[16] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10556–10565.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[18] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 23593–23606.

[19] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 12268–12277.

[20] Y. Wang, R. Wan, W. Yang, H. Li, L. Chau, and A. C. Kot, "Low-light image enhancement with normalizing flow," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2604–2612.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10674–10685.

[22] W. Wang, H. Yang, J. Fu, and J. Liu, "Zero-reference low-light enhancement via physical quadruple priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26057–26066.

[23] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 1632–1640.

[24] J. Huang et al., "Deep Fourier-based exposure correction network with spatial-frequency interaction," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 163–180.

[25] X. Lv et al., "Fourier priors-guided diffusion for zero-shot joint low-light enhancement and deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 25378–25388.

[26] H. Huang, W. Yang, Y. Hu, J. Liu, and L. Duan, "Towards low light enhancement with RAW images," *IEEE Trans. Image Process.*, vol. 31, pp. 1391–1405, 2022.

[27] Y. Wu et al., "Learning semantic-aware knowledge guidance for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1662–1671.

[28] D. Zhou, Z. Yang, and Y. Yang, "Pyramid diffusion models for low-light image enhancement," in *Proc. Int. Joint Conf. Artif. Intell*, 2023, pp. 1795–1803.

[29] H. Jiang, A. Luo, S. Han, H. Fan, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Trans. Graph.*, vol. 42, no. 6, pp. 238:1–238:14, 2023.

[30] C. Chen, Q. Chen, M. N. Do, and V. Koltun, "Seeing motion in the dark," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 3184–3193.

[31] F. Zhang, Y. Li, S. You, and Y. Fu, "Learning temporal consistency for low light video enhancement from single images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4965–4974.

[32] J. Xiong, J. Wang, W. Heidrich, and S. K. Nayar, "Seeing in extra darkness using a deep-red flash," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10000–10009.

[33] Z. Xia, M. Gharbi, F. Perazzi, K. Sunkavalli, and A. Chakrabarti, "Deep denoising of flash and no-flash pairs for photography in low-light environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2063–2072.

[34] C. Li, C. Guo, S. Zhou, Q. Ai, R. Feng, and C. C. Loy, "FlexiCurve: Flexible piecewise curves estimation for photo retouching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Workshops*, 2023, pp. 1092–1101.

[35] Z. Liang, C. Li, S. Zhou, R. Feng, and C. C. Loy, "Iterative prompt learning for unsupervised backlit image enhancement," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 8060–8069.

[36] Z. Fu, Y. Yang, X. Tu, Y. Huang, X. Ding, and K. Ma, "Learning a simple low-light image enhancer from paired low-light instances," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22252–22261.

[37] H. Jiang, A. Luo, X. Liu, S. Han, and S. Liu, "LightenDiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 161–179.

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

[39] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2828–2841, Jun. 2018.

[40] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4225–4238, Aug. 2022.

[41] C. Li, C. Guo, R. Feng, S. Zhou, and C. C. Loy, "CuDi: Curve distillation for efficient and controllable exposure adjustment," 2022, *arXiv:2207.14273*.

[42] T. Chobola, Y. Liu, H. Zhang, J. A. Schnabel, and T. Peng, "Fast context-based low-light image enhancement via neural implicit representations," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 413–430.

[43] T. Gevers, A. Gijsenij, J. Van de Weijer, and J.-M. Geusebroek, *Color in Computer Vision: Fundamentals and Applications*. Hoboken, NJ, USA: Wiley, 2012.

[44] J. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1338–1350, Dec. 2001.

[45] A. Lengyel, S. Garg, M. Milford, and J. C. van Gemert, "Zero-shot domain adaptation with a physics prior," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 4379–4389.

[46] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 3813–3824.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[48] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-solver : Fast solver for guided sampling of diffusion probabilistic models," 2022, *arXiv:2211.01095*.

[49] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 606–615.

[50] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3060–3069.

[51] J. Hai et al., "R2RNet: Low-light image enhancement via real-low to real-normal network," *J. Vis. Commun. Image Representation*, vol. 90, 2023, Art. no. 103712.

[52] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1013–1037, 2021.

[53] L. Zhang, L. Zhang, X. Liu, Y. Shen, S. Zhang, and S. Zhao, "Zero-shot restoration of back-lit images using deep internal learning," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 1623–1631.

[54] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.

[55] W. Yang et al., "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Trans. Image Process.*, vol. 29, pp. 5737–5752, 2020.
[56] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[57] M. Afifi, K. G. Derpanis, B. Ommer, and M. S. Brown, "Learning multi-scale photo exposure correction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9157–9167.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[59] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, "Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters," in *Proc. ACM SIGKDD Int. Conf.*, 2020, pp. 3505–3506.
[60] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[61] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, Sep. 2013.
[62] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[63] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blin," image quality analyzer," *IEEE Sign. Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
[64] J. Wang, K. C. Chan, and C. C. Loy, "Exploring CLIP for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intel*, 2023, pp. 2555–2563.
[65] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.
[66] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.
[67] J. Betker et al., "Improving image generation with better captions," 2023, *arXiv: 2006.11807*.

**Haofeng Huang** (Student Member, IEEE) received the BS degree in computer science from Peking University, Beijing, China, in 2021. He is currently working toward the PhD degree with the Wangxuan Institute of Computer Technology. His current research interests include deep-learning-based image/video compression for collaborative intelligence, intelligent visual enhancement, and generative image/video editing.

**Yifan Li** is currently working toward the BS degree in data science from Yuanpei College, Peking University, Beijing, China. He is currently a research intern with the Wangxuan Institute of Computer Technology. His current research interests include image enhancement and generative models for low-level vision.

**Wenjing Wang** (Member, IEEE) received the BS and PhD degrees in computer science from Peking University, Beijing, China, in 2019 and 2024. She has authored more than 20 technical articles in refereed journals and proceedings, and she holds five granted patents. Her current research interests include image synthesis, image enhancement, and deep learning.

**Wenhan Yang** (Member, IEEE) received the BS and PhD degrees (Hons.) in computer science from Peking University, Beijing, China, in 2012 and 2018. He is currently an associate researcher with Pengcheng Laboratory, Shenzhen, Guangdong, China. His current research interests include image/video processing/restoration, bad weather restoration, human-machine collaborative coding. He has authored more than 70 technical articles in refereed journals and proceedings, and holds 9 granted patents. He received the 2023 IEEE Multimedia Rising Star Runner-Up Award, the IEEE ICME-2020 Best Paper Award, the IFTC 2017 Best Paper Award, the IEEE CVPR-2018 UG2 Challenge First Runner-up Award, and the MSA-TC Best Paper Award of ISCAS 2022. He was recognized as one of the World's Top 2% Scientists by Stanford/Elsevier since 2021. He was the Candidate of CSIG Best Doctoral Dissertation Award in 2019. He served as the area chair of IEEE ICME-2021-2025, the session chair of IEEE ICME-2021, and the organizer of IEEE CVPR-2019/2020/2021 UG2+ Challenge and Workshop.

**Ling-Yu Duan** (Member, IEEE) received the PhD degree in information technology from the University of Newcastle, Callaghan, Australia, in 2008. He is a full professor with the National Engineering Laboratory of Video Technology (NELVT), School of Computer Science, Peking University (PKU), China, and has served as the associate director with the Rapid-Rich Object Search Laboratory (ROSE), a joint lab between Nanyang Technological University (NTU), Singapore, and Peking University (PKU), China since 2012. He is also with Peng Cheng Laboratory, Shenzhen, China, since 2019. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, and video analytics, etc. He has published about 200 research papers. He received the IEEE ICME Best Paper Award in 2019/2020, the IEEE VCIP Best Paper Award in 2019, and EURASIP Journal on Image and Video Processing Best Paper Award in 2015, the Ministry of Education Technology Invention Award (First Prize) in 2016, the National Technology Invention Award (Second Prize) in 2017, China Patent Award for Excellence (2017), the National Information Technology Standardization Technical Committee "StandardizationWork Outstanding Person" Award in 2015. He was a co-editor of MPEG Compact Descriptor for Visual Search (CDVS) Standard (ISO/IEC 15938-13) and MPEG Compact Descriptor for Video Analytics (CDVA) standard (ISO/IEC 15938-15). Currently, he is an associate editor of *IEEE Transactions on Multimedia*, *ACM Transactions on Intelligent Systems and Technology* and *ACM Transactions on Multimedia Computing, Communications, and Applications*, and serves as the area chairs of ACM MM and IEEE ICME. He is a member of the MSA Technical Committee in IEEE-CAS Society.

**Jiaying Liu** (Fellow, IEEE) received the PhD degree (Hons.) in computer science from Peking University, Beijing China, 2010. She is currently an associate professor, Boya young fellow with the Wangxuan Institute of Computer Technology, Peking University, China. She has authored more than 100 technical articles in refereed journals and proceedings, and holds 70 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She is a senior member of IEEE/CSIG, and a distinguished member of CCF. She was a visiting scholar with the University of Southern California, Los Angeles, California, from 2007 to 2008. She was a visiting researcher with Microsoft Research Asia, in 2015 supported by the Star Track Young Faculties Award. She has served as a member of Multimedia Systems and Applications Technical Committee (MSA TC), and Visual Signal Processing and Communications Technical Committee (VSPC TC) in IEEE Circuits and Systems Society. She received the IEEE ICME 2020 Best Paper Award and IEEE MMSP 2015 Top10% Paper Award. She has also served as the associate editor of *IEEE Transactions on Image Processing*, *IEEE Transactions on Circuits Systems for Video Technology* and *Journal of Visual Communication and Image Representation*, the technical program chair of ACM MM Asia-2023/IEEE ICME-2021/ACM ICMR-2021/IEEE VCIP-2019, the area chair of CVPR-2021/ECCV-2020/ICCV-2019, ACM ICMR Steering Committee member and the CAS representative with the ICME Steering Committee. She was the APSIPA distinguished lecturer (2016–2017).