

Zero-Reference Low-Light Enhancement via Physical Quadruple Priors

Wenjing Wang
Peking University

Huan Yang
01.AI

Jianlong Fu
Microsoft Research Asia

Jiaying Liu *
Peking University

Abstract

Understanding illumination and reducing the need for supervision pose a significant challenge in low-light enhancement. Current approaches are highly sensitive to data usage during training and illumination-specific hyperparameters, limiting their ability to handle unseen scenarios. In this paper, we propose a new zero-reference low-light enhancement framework trainable solely with normal light images. To accomplish this, we devise an illumination-invariant prior inspired by the theory of physical light transfer. This prior serves as the bridge between normal and low-light images. Then, we develop a prior-to-image framework trained without low-light data. During testing, this framework is able to restore our illumination-invariant prior back to images, automatically achieving low-light enhancement. Within this framework, we leverage a pretrained generative diffusion model for model ability, introduce a bypass decoder to handle detail distortion, as well as offer a lightweight version for practicality. Extensive experiments demonstrate our framework’s superiority in various scenarios as well as good interpretability, robustness, and efficiency. Code is available on our [project homepage](#).

1. Introduction

Restoring images in low-light conditions is an important and challenging task in computer vision. The goal is to unveil concealed details in poorly lit areas, ultimately elevating the overall image quality. Over time, a large number of algorithms have been developed to address this challenge. However, current methods exhibit limitations due to their dependence on supervisory information and their adaptability to unseen domains. In the following, we review recent advancements and then present our primary contributions.

Supervised Methods. Deep learning has significantly influenced the advancement of low-light enhancement. In

*Corresponding author. This work was supported in part by the National Natural Science Foundation of China under Grant 62332010, and in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

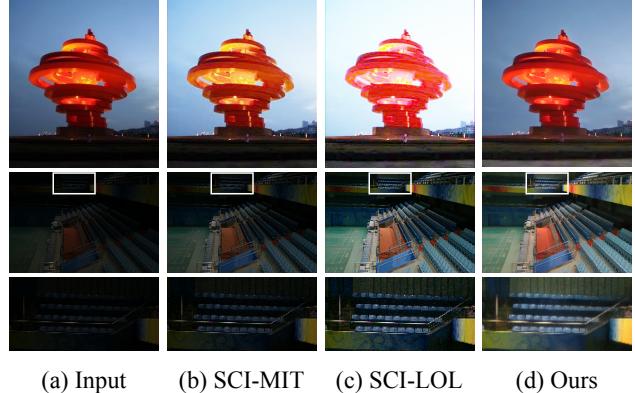


Figure 1. Comparison with a SOTA zero-reference method: SCI [34]. The SCI model, trained on varied datasets like LOL [48] and MIT [2], yields diverse enhancement results. Nevertheless, none effectively maintains a consistent lighting effect across both dark and moderately dark images. In contrast, our model demonstrates greater robustness across various scenarios.

2017, Li *et al.* [31] introduced the initial deep-based low-light enhancement model using a straightforward auto-encoder. Subsequently, a series of studies improved the network design by incorporating concepts from the Retinex theory [48, 49, 63], Fourier transform [16], image processing systems [15], semantics [50], and adopting innovative architectures like Flow-based generative models [47], vision transformers [3] and diffusion models [57, 65]. In addition to RGB images, a body of research is dedicated to RAW data [4, 5, 17], videos [59], and multi-modalities [51, 52]. While these models have achieved notable success, they rely on paired data for training, which can be inflexible and less robust in unforeseen scenarios. Recent studies explore how to reduce this dependency on supervision.

Unsupervised Methods. Some work transitions from the strict requirement of pairwise pairings to only necessitating unpaired normal-low light data for training. EnlightenGAN [19], FlexiCurve [25], and NeRCo [54] employ adversarial learning, where discriminators are constructed to guide low-light enhancement models, *i.e.*, the generators. CLIP-LIT [27] learns prompts from images under varied illuminations. PairLIE [7] instead learns adaptive priors from

paired low-light instances of the same scene. However, these methods retain specific training data requirements, limiting their ability to generalize to unknown scenarios.

Zero-Reference Methods. “Zero-reference” [10] refers to a special unsupervised setting where neither paired nor unpaired data is available for training. This setting is more challenging, but offers greater flexibility in practical applications. Traditional non-deep low-light enhancement algorithms [11, 37] can also be classified as zero-reference. These methods primarily rely on manually designed strategies, such as histogram equalization [37] or Retinex decomposition [11, 26]. As for deep models, Zero-DCE [10] employs a neural network to predict the parameters of a pre-defined curve function and applies it to the input low-light image. A suite of non-reference loss functions is employed to guide the enhancement process. Subsequent works improved speed and curve forms [23, 24]. As an alternative approach, RUAS [30] introduces a neural architecture search strategy based on the Retinex rule. It also implements several reference-free losses. Furthermore, SCI [34] streamlines the iterative process in RUAS into a single step.

Despite the success of zero-reference methods, they often require careful parameter-tuning and can be sensitive to the distribution of training data. For instance, in SCI [34], varying training data leads to distinct enhanced appearances, causing over-exposure or under-exposure as shown in Fig. 1. The crux of the matter is that these zero-reference models lack a genuine concept of illumination. Learning lighting knowledge without reference and depending on artificially set parameters presents an challenging and unsolved problem.

Our Contributions. In this paper, we propose a new zero-reference low-light enhancement framework to address the aforementioned challenges. Our central idea is to develop an illumination-invariant prior and employ it as an intermediary between low-light and normal light images. We devise a unique illumination prior, named the **physical quadruple prior**, originating from the Kubelka–Munk theory of light transfer. Next, we construct a prior-to-image mapping framework **solely using typical normal-light images**, easily obtainable from the Internet or existing open-source visual datasets. During this stage, the model learns the authentic concept of bright lighting from the image distribution. Finally, when tested on low-light images, our physical quadruple prior automatically extracts illumination-invariant features, and prior-to-image mapping framework transfers these features to normal light images. Throughout this process, low-light enhancement can be achieved without requiring any low-light data or illumination-relevant hyper-parameters.

The challenge lies in that, as our prior discards a considerable amount of illumination-relevant information, restoring it back to images is not a straightforward task. To ad-

dress this challenge, we capitalize on the exceptional capabilities of a pretrained large-scale generative model, Stable Diffusion (SD) [42], and construct the prior-to-image mapping by integrating priors as conditions to control the SD model. Unlike typical generative tasks that require high-quality data [13, 35, 43, 66, 67], our framework is insensitive to data quality and is trained on one of the most readily available datasets: COCO [28]. Since SD is originally designed for tasks other than restoration, it faces challenges in detail preservation. Hence, we propose an bypass decoder to address the distortion issue, which also proves useful for other image processing tasks. Finally, considering practical applications, our framework can be used to create a lighter zero-reference model on specific data. By employing a CNN-transformer mixed model, we distill the complex multi-step optimization of a large diffusion model into a single forward pass within a lightweight network. This lightweight version maintains comparable performance while significantly improving inference speed and computational efficiency.

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:

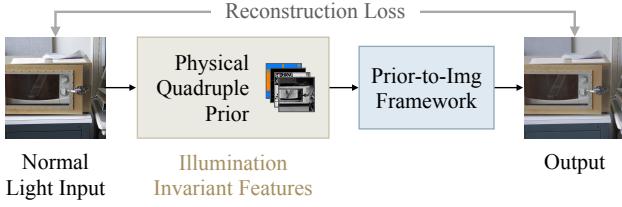
- We present a zero-reference low-light enhancement model that utilizes an illumination-invariant prior as the intermediary between different illumination. Our model exhibits superior performance in various under-lit scenarios without relying on any specific low-light data.
- We establish the physical quadruple prior, a novel learnable illumination-invariant prior derived from a light transfer theory. This prior captures the essence of imaging under diverse lighting conditions, freeing low-light enhancement from dependence on reference samples or artificially set hyper-parameters.
- We develop an effective prior-to-image mapping system by incorporating the prior as a condition to control a pretrained large-scale generative diffusion model. We introduce a bypass decoder to address the distortion issue, and show that our model can be distilled into a lightweight version for practical application.

2. Physical Prior-based Image Restoration

2.1. Motivation

Developing illumination invariant features has a long history in the domain of image restoration. One of the most representative invariants for low-light enhancement is the Retinex model [40]. This model posits that an image X can be decomposed into illumination I and illumination-invariant reflectance R , expressed as $X = I \odot R$, where \odot

Training on *normal light* images



Inference on *low-light* images

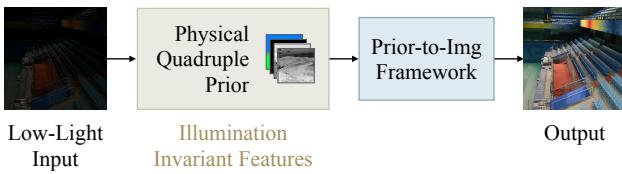


Figure 2. The overall methodology of our zero-reference low-light enhancement approach. Our model is trained to reconstruct images from an illumination-invariant prior (the physical quadruple prior) in the normal light domain. During testing, the model extracts illumination-invariant priors from low-light images and reconstructs them into normal light images.

denotes element-wise multiplication. However, solving this decomposition is challenging. Existing approaches either lean on human-crafted policies [11] or rely on paired data with varying brightness levels [48, 63], which lack robustness when faced with unknown scenarios.

Instead of decomposing images into illumination-invariant and illumination-related information, we suggest to solely extract illumination-invariant features from the input image and subsequently generate illumination-related information to reconstruct the image. In contrast to traditional Retinex-based deep models [48, 63], our framework ensures that our deep model can comprehend lighting without the need for paired data.

The overall training and inference pipeline of our method is illustrated in Fig. 2. We introduce a physical quadruple prior derived from the Kubelka-Munk theory [9] to extract illumination-invariant features. While training on normal light images, the model simultaneously learns how to reconstruct images from priors and learns the illumination distribution of natural images. Acknowledging the complexity of simultaneously addressing these two tasks, we make use of a pre-trained diffusion generative model. During testing, low-light images are initially mapped to the physical quadruple prior and subsequently reconstructed. Given the illumination-invariant nature of our prior, it extracts comparable features from low-light images as it does from normal light images. Combined with the fact that our prior-to-image model is specifically trained to reconstruct priors into normal light images, our model achieves low-light enhancement without the need for low-light data. In the following, we elaborate on the details of each component.

2.2. Learnable Illumination-Invariant Prior

Physical Quadruple Priors. We start from the Kubelka-Munk theory [9] of light transfer. Given wavelength λ , the energy of the incoming spectrum at spatial location \mathbf{x} on the image plane is modeled as

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

where $e(\lambda, \mathbf{x})$ denotes the spectrum of the light source, $i(\mathbf{x})$ the specular reflection, and $R_\infty(\lambda, \mathbf{x})$ the material reflectivity. Note that when the object is matte, *i.e.*, $i(\mathbf{x}) \approx 0$, Eq. (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. It means that the Retinex theory is a special case of Eq. (1).

First of all, we denote some variables for simplicity

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

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

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

Following [8], through simplifying assumptions, we can obtain a series of invariants from Eq. (1). The primary idea is to eliminate i and e , retaining solely R_∞ . As R_∞ is about material property and is independent of illumination, the derived variable will exhibit illumination invariance.

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

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

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

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

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

$$H = \arctan(E^\lambda/E^{\lambda\lambda}). \quad (7)$$

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

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

Similarly, we derive another illumination invariant,

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

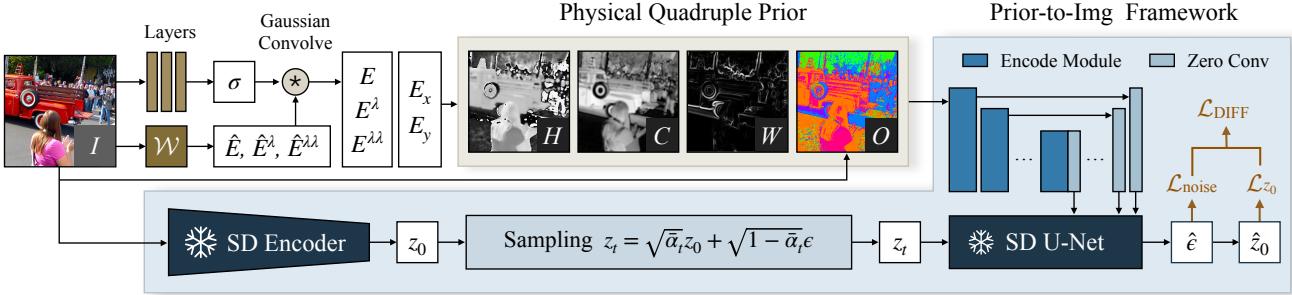


Figure 3. Our illumination-invariant prior and the training process for our prior-to-image model framework. We start by predicting the physical quadruple prior from the input image I . During the training phase, the model dynamically learns the linear mapping \mathcal{W} and the layers for predicting the scale σ . In the process of reconstructing priors into images, a static SD encoder extracts 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 ϵ and z_0 .

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

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

Similarly, we derive our third illumination invariant,

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

The Kubelka-Munk theory [9] is effective for grayscale images but falls short in describing colors. The three aforementioned illumination invariants loss part of the color information, so we add some additional color information. We adopted a straightforward one: the relative relationship between pixel values of the three RGB channels.

- Assuming that illumination maintains the order of colors, we propose the order of the RGB three channels as a fundamental illumination-invariant feature, denoted as O .

Learning through Neural Networks. We follow Gaussian color models [9] and CICConv [22] to obtain priors from RGB images. First, we estimate the observed energy \hat{E} along with its derivatives \hat{E}^λ and $\hat{E}^{\lambda\lambda}$ via linear mapping:

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

where x and y denote positions in the image, and \mathcal{W} is a 3×3 matrix. In [9, 22], \mathcal{W} is manually defined. We instead learn it from the distribution of natural images through our prior-to-image framework.

The spatial derivative $\partial E / \partial \mathbf{x}$ in Eq. (11) is computed in both the x- and y-direction, denoted as $\partial E / \partial \mathbf{x} = (E_x, E_y)$,

with its magnitude given by $|\partial E / \partial \mathbf{x}| = \sqrt{E_x^2 + E_y^2}$. Finally, E , E_x , and E_y are estimated by convolving \hat{E} with Gaussian color smoothing and derivative filters of scale σ . σ is predicted from the input image. Similarly, E^λ is obtained from \hat{E}^λ , and $E^{\lambda\lambda}$ is obtained from $\hat{E}^{\lambda\lambda}$. Now we can compute H , C , and W from the input image.

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

$$O(x, y) = [O_R(x, y), O_G(x, y), O_B(x, y)], \quad (13)$$

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. Firstly, the mathematical form indicates that W represents the intensity-normalized spatial derivatives of the spectral intensity. As for H , according to [9], it is associated with the hue, *i.e.*, $\arctan(\lambda_{\max})$ of the material. As for C , within a color circle based on spectral wavelengths, hue represents the angle while chroma is the distance from the center. Additionally, when converting a Cartesian coordinate (a, b) to polar, the angle becomes $\arctan(b/a)$, and the radius becomes $\sqrt{(a)^2 + (b)^2}$. Connecting this with Eq. (7) and Eq. (9), we find C associates with chroma. A visualization can be found in Fig. 3. More analysis and comparisons will be presented in Sec. 3.3.

2.3. Prior-to-Image via Diffusion Models

Ideally, we want to retain all illumination-invariant information while discarding lighting-relevant information. However, achieving this decomposition is challenging and remains a unsolved problem in image modeling. Despite our physical quadruple prior, *i.e.*, H , C , W , and O , capturing illumination-independent information from various perspectives, some information is still lost. Consequently, reconstructing images from the prior is non-trivial.



Figure 4. Image restoration effect of the SD decoder and ours. (a) Input image I , from which we extract latent z_0 . (b) z_0 decoded by the SD decoder. (c) The distorted version of I . (d) z_0 decoded by our decoder using the encoder features from \tilde{I} .

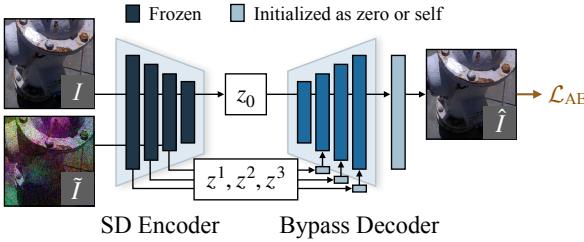


Figure 5. The training strategy of our bypass decoder. We distort the input image I into \tilde{I} , and allow the decoder to reconstruct I using encoder features from the distorted \tilde{I} .

Instead of focusing on improving the illumination-invariant prior, we propose leveraging the capabilities of a large-scale generative model to directly complete the missing information. We employ Stable Diffusion v1-5 [42] and convert it into a conditional generative mode using the ControlNet [60] framework. Our physical quadruple prior serves as the condition to control the SD model.

The overall framework is illustrated in Fig. 3. During training, a frozen SD encoder is employed to map the image I into a compressed latent representation z_0 . We then sample z_t at a random time step $t \in \{1, \dots, T\}$ using

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

where $\{\bar{\alpha}_t\}$ is a sequence of pre-defined parameters [14]. The training objective is to predict ϵ from z_t and our prior. Originally, SD utilizes a U-Net to predict ϵ from z_t , and this U-Net is now frozen. A set of encode modules are added to extract features from our quadruple prior. These features are then incorporated into the SD U-Net. The zero convolution strategy [60] is adopted to ensure that, at the beginning of training, new layers do not influence the original SD. During testing, given an input image I , we extract the physical quadruple prior and use it as the condition to predict z_0 in the reverse diffusion process. Subsequently, z_0 is projected back to the image space through a decoder.

While ControlNet has demonstrated success in various applications, applying it directly suffers from issues includ-

ing slow convergence, detail degradation, and dependence on text prompts. To address these challenges and make it more suitable for our image restoration task, we implement the following improvements.

- In typical diffusion models, the training objective is to predict the Gaussian noise term:

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

We additionally minimize the difference in terms of z_0 , which can accelerate convergence. Combining Eq. (14), we derive

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

We simply combine these two losses as the final objective

$$\mathcal{L}_{\text{DIFF}} = \mathcal{L}_{z_0} + \mathcal{L}_{\text{noise}}. \quad (17)$$

- SD employs an auto-encoder (AE) to compress the image I into a latent representation z_0 , reducing computational cost. However, the auto-encoder introduces severe detail distortion. As shown in Fig. 4(b), the face of the mounted police is completely distorted. To mitigate this issue, we support the decoder with features from the encoder and devise an effective fine-tuning strategy. As shown in Fig. 5, during training, we distort the input image I with random illumination jittering and noise, resulting in \tilde{I} . The decoder then restores z_0 , combining the features z^1, z^2, z^3 extracted from \tilde{I} . This progress guides the decoder to capture details from \tilde{I} while preserving the illumination characteristics of I . We introduce several convolutional layers for feature fusion and a residual block for post-processing. These additional layers are initialized as zero or self, ensuring they have minimal impact on the original decoding process at the beginning of training. The new decoder is named bypass decoder. As shown in Fig. 4(d), noticeable detail restoration is achieved through our bypass decoder. During testing, features from the input image assist in the latent decoding process, as illustrated in Fig. 6. Our bypass decoder utilizes z^1, z^2 , and z^3 , extracted from the input image, to reconstruct details and maintain the enhanced illumination within \hat{z}_0 .
- Stable Diffusion is originally designed as a text-to-image model. However, requiring users to provide text for low-light enhancement is inconvenient. To address this, we set the text input to always be an empty string.

Denoising. Noise poses a significant challenge in low-light enhancement. Although our prior isn't designed for denoising purposes, we implement a simple strategy to suppress noise. During training, we apply random Gaussian-Poisson compound noise to the input image I while extracting the physical quadruple prior. This approach guides the model

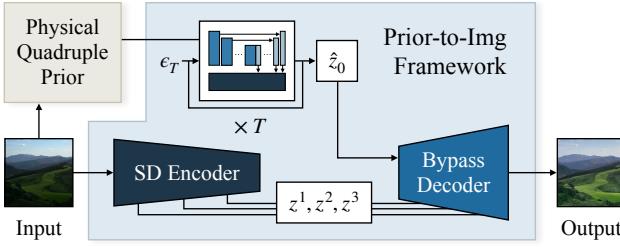


Figure 6. The inference pipeline of our overall framework. Given a low-light image, we extract its physical quadruple prior. Then, this prior serves as the condition for predicting the latent representation \hat{z}_0 from pure noise e_T . Lastly, the bypass decoder utilizes features extracted by the encoder from the low-light image to map the predicted \hat{z}_0 back into images.

to disregard high-frequency details and concentrate solely on low-frequency illumination-invariant information.

Distillation for Efficiency. Diffusion models require multi-step optimization in inference. Even with DPM-Solver++ [32], 10 steps are still cumbersome. In the pursuit of practicality, our framework can create a more lightweight version. In short, we construct a lightweight U-net consists of residual blocks and integrate transformer blocks from Restormer [58] at the bottleneck. Transformers have proven to be effective in low-level vision [29, 38, 39, 53]. Then, we make 1.7k samples using our framework to teach the lightweight model. The training objective is solely the L1 loss. More details are provided in the supplementary.

3. Experiments

3.1. Implementation Details

Framework Development. Our framework is trained on the COCO-2017 [28] train and unlabeled set. We employ a minibatch size of 8, conducting training over 140k steps, approximately 5 epochs, with a learning rate set at 1e-4 and the ADAM optimizer [20]. To accommodate this sizable model within limited GPU memory, we implement float16 precision and DeepSpeed [41]. More implementation details can be found in the supplementary.

Compared Methods. Our model is compared with nine unsupervised low-light image enhancement methods. Among these, EnlightenGAN [19], PairLIE [7], NeRCo [54], and CLIP-LIT [27] utilize unpaired low-light-related data. The remaining five methods, ExCNet [61], ZeroDCE [10], ZeroDCE++ [23], RUAS [30], and SCI [34], are zero-reference. Additionally, we present the results of six supervised methods to show the upper bound of our task.

Benchmark Settings. We report performance on three sets of widely-used low-light datasets. The first two sets are LOL and MIT-Adobe FiveK [2]. For LOL, we adopt the official test sets of LOL v1 [48] and LOL v2 [55], resulting

in 115 low-/normal-light image pairs. For MIT, we follow Retinexformer [3] to split 500 pairs for testing. Additionally, we gather low-light images from LIME [11], NPE [46], MEF [33], DICM [21], and VV [45] that have no ground truth. This set is simply called the “unpaired set”. On LOL and MIT, we report PSNR, SSIM, LPIPS [62], and LOE [46]. On the unpaired set, we report BRISQUE [36] and noise level (NL) estimated by [6].

3.2. Benchmarking Results

Tab. 1 demonstrates that our model surpasses the majority of unsupervised techniques and notably reduces the performance gap compared to supervised methods. Subjective results can be found in Fig. 7 and Fig. 1. Our model is able to more effectively suppress noise and prevent overexposure or excessive darkness. More showcases can be found in the supplementary. SCI [34] demonstrates good performance on training-related datasets but significantly degrades in unseen scenarios. In contrast, our model exhibits robustness across both LOL and MIT datasets simultaneously. This resilience stems from our model’s ability to learn comprehensive illumination knowledge from the physical quadruple prior and normal light images. Consequently, our model proves to be less sensitive to specific datasets.

In Tab. 1, it’s noticeable that supervised methods often tend to overfit to their training sets, exhibiting limited generalization to unseen domains. For instance, Retinexformer [3] and DiffLL [18], trained with LOL, achieve lower performance than our model on MIT, and vice versa. This experiment underscores our model’s superior adaptability to previously unseen scenarios, outperforming even supervised methods.

3.3. Ablation Studies

Prior Design. We first analyze the effect of each element, H , C , W , and O , in our illumination-invariant prior. The evaluation is conducted on LOL, following the same setting as Tab. 1. Removing any element reduces the performance. It is because deleting any element would remove a corresponding part of information. Showing that only the four prior combinations together can reconstruct the original image without extracting the light-related features.

Fig. 8 provides a visual comparison. The absence of H or C leads to color bias or a washed-out white appearance. As previously discussed in Sec. 2.2, H and C are associated with hue and chroma. But on the right of Fig. 8, we can also see that H and C are not exactly hue and chroma. This shows that prior basically conforms to the physical explanation we have derived, but can go on to learn more advanced features. Recall that W represents intensity-normalized spatial derivatives of spectral intensity, capturing local illumination changes. When W is omitted, the alterations in light and shadow are completely lost (as observed in the

Table 1. Benchmarking results for low-light enhancement. Among unsupervised methods, we highlight the top-ranking scores in **red** and the second in **blue**. Additionally, we denote the training set used by each model. “LOL+” indicates a fusion of LOL and other datasets.

Datasets	Train Set	LOL [48, 55]				MIT-Adobe FiveK [2]				Unpaired Sets		
Metrics		PSNR↑	SSIM↑	LPIPS↓	LOE↓	PSNR↑	SSIM↑	LPIPS↓	LOE↓	BRISQUE↓	NL↓	
Supervised	Retinex-Net [48]	LOL	16.19	0.403	0.534	0.346	12.30	0.687	0.258	0.244	27.10	3.254
	KinD [63]	LOL	20.21	0.814	0.147	0.245	14.71	0.756	0.176	0.174	26.89	0.700
	KinD++ [64]	LOL	16.64	0.662	0.410	0.288	15.76	0.650	0.319	0.176	26.16	0.431
	URetinex-Net [49]	LOL	20.93	0.854	0.104	0.245	14.10	0.734	0.182	0.187	23.80	1.319
	Retinexformer [3]	LOL	28.48	0.877	0.117	0.256	13.87	0.692	0.222	0.224	14.77	1.064
	Retinexformer [3]	MIT	13.02	0.426	0.365	0.280	24.93	0.907	0.063	0.162	24.13	0.684
Unsupervised	DiffLL [18]	LOL+	28.54	0.870	0.102	0.253	15.81	0.719	0.244	0.213	14.96	0.888
	ExCNet [61]	test images	16.29	0.455	0.380	0.295	14.21	0.719	0.197	0.197	19.03	1.563
	EnlightenGAN [19]	own data	18.57	0.700	0.302	0.291	13.28	0.738	0.203	0.199	20.65	0.779
	PairLIE [7]	LOL+	19.70	0.774	0.235	0.278	10.55	0.642	0.273	0.225	29.84	1.471
	NeRCo [54]	LSRW [12]	19.67	0.720	0.266	0.310	17.33	0.767	0.208	0.213	22.81	0.603
	CLIP-LIT [27]	own data	14.82	0.524	0.371	0.320	17.00	0.781	0.159	0.194	23.44	1.962
	ZeroDCE [10]	own data	17.64	0.572	0.316	0.296	13.53	0.725	0.201	0.191	21.76	1.569
	ZeroDCE++ [23]	own data	17.03	0.445	0.314	0.391	12.33	0.408	0.280	0.417	19.34	1.150
	RUAS [30]	MIT	13.62	0.462	0.346	0.292	9.53	0.610	0.301	0.272	29.91	2.091
	RUAS [30]	LOL	15.47	0.490	0.305	0.330	5.15	0.373	0.669	0.399	44.70	3.312
	RUAS [30]	FACE [56]	15.05	0.456	0.371	0.292	5.00	0.366	0.685	0.398	46.21	3.633
	SCI [34]	MIT	11.67	0.395	0.361	0.286	16.29	0.795	0.143	0.165	16.73	0.853
	SCI [34]	LOL+	16.97	0.532	0.312	0.289	7.83	0.573	0.360	0.187	24.46	1.893
	SCI [34]	FACE [56]	16.80	0.543	0.322	0.297	10.95	0.684	0.272	0.205	18.33	1.335
Ours		COCO [28]	20.31	0.808	0.202	0.281	18.51	0.785	0.163	0.188	14.64	0.423

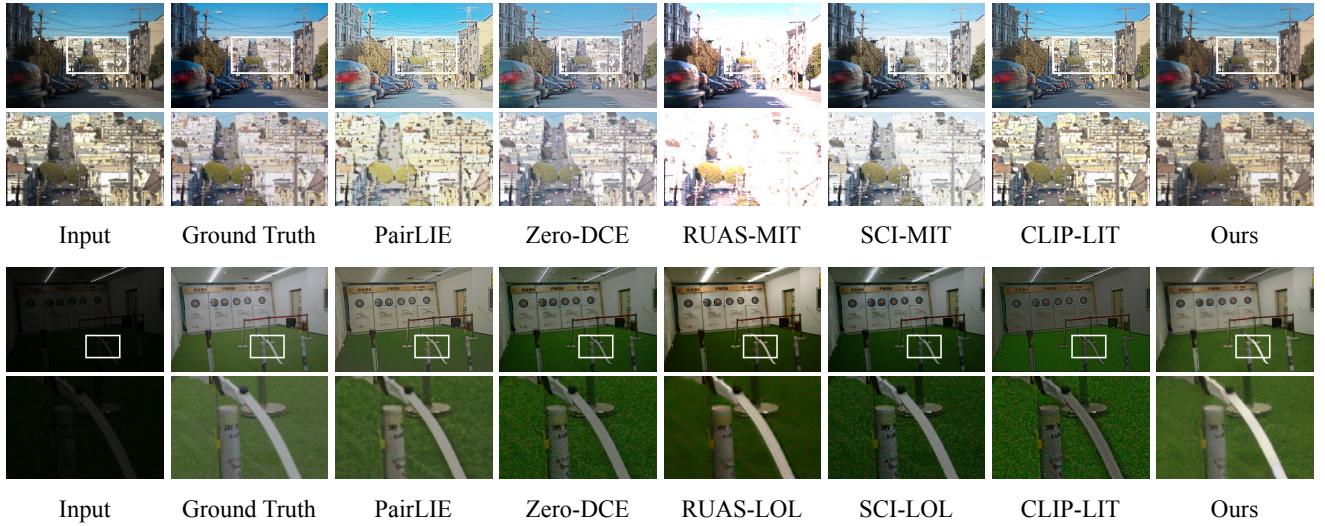


Figure 7. Example low-light enhancement results on the MIT-Adobe FiveK (top row) and LOL datasets (bottom row).

second row of Fig. 8). Additionally, without O , blue was mistakenly enhanced to orange. Ultimately, our full version showcases the most refined details and superior contrast.

We further explore the implications of replacing our prior with alternative representations. We consider three representative ones: (1) Naive HS channels in the HSV color space. (2) CICConv [22], a trainable prior similar to our W . (3) The reflectance estimated by Retinex-based PairLIE [7] trained on LOL. In Tab. 2, the HS channels sacrifice significant content information, particularly impacting SSIM and LPIPS negatively. While CICConv exhibits illumination invariance in high-level vision tasks, it suffers from excessive color loss in the image, leading to perfor-

mance degradation.⁴ In comparison with our prior learned from COCO, the reflectance derived by PairLIE originates from low-light data pairs in LOL. Despite being trained on the target domain data, it still performs worse than our prior.

Prior-to-Image Framework. We discuss the importance of using a pre-trained generative model to build the prior-to-image mapping. We explore an alternative diffusion-based backbone, SR3 [44], recognized for its efficacy in super resolution. However, when used to replace our prior-to-image framework, as illustrated in Fig. 9, SR3 exhibits noticeable color bias and noise issues. It is because, in order to strip away light-related features, our prior discarded some of the image information, requiring the prior-to-image model to

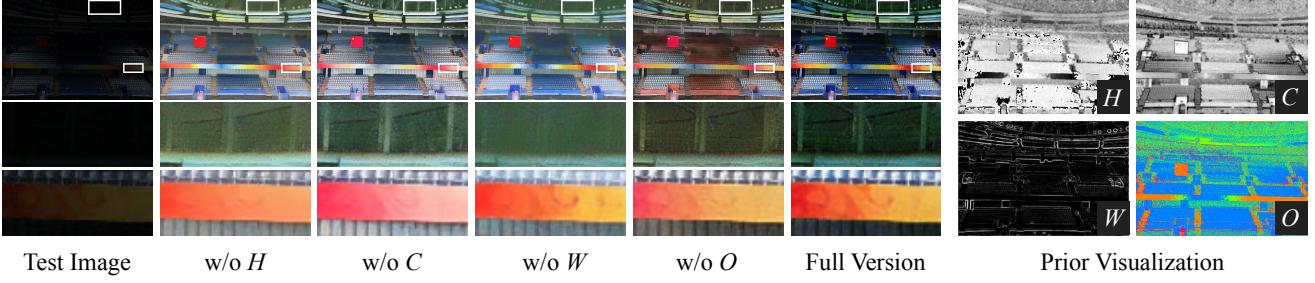


Figure 8. Low-light enhancement effects for different prior designs (left), and the visualization of our physical quadruple prior (right).

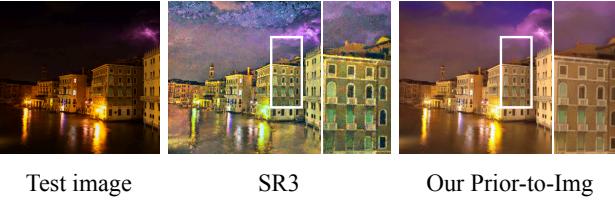


Figure 9. Effects of using different prior-to-image frameworks.

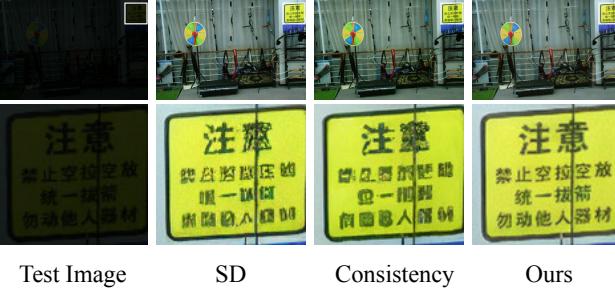


Figure 10. Effects of different decoders in our framework.

Table 2. Ablation studies on the effect of our method designs.

Datasets		LOL [48]			
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
AE	HS channels in HSV	18.04	0.562	0.498	0.410
	CICConv	17.02	0.455	0.551	0.421
	Reflectance by PairLIE [7]	20.16	0.790	0.287	0.296
Ours Final Version		20.25	0.807	0.199	0.278

fill in the gaps and restore the complete details.

Auto-Encoder. We showcase the impact of our bypass decoder, comparing it with the original decoder used in SD. Additionally, we evaluate the Consistency Decoder from DALL-E 3 [1], a recent diffusion-based decoder enhancing SD VAEs’ decoding capabilities. As shown in Fig. 10, both the original decoder in SD and the Consistency Decoder fail to preserve the text. In contrast, our decoder uti-

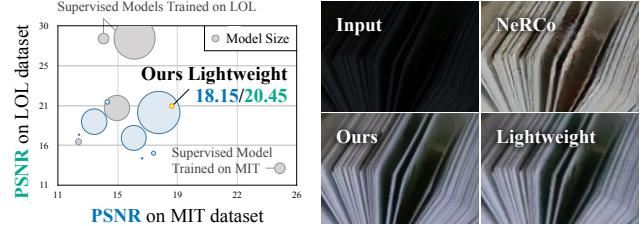


Figure 11. Left: Model size and PSNR comparison between our lightweight model and the SOTA. Detailed scores are in the supplementary. Right: Visual comparison of our full & lightweight versions and NeRCo [54]. Both our full & lightweight versions similarly show improved contrast and reduced color bias.

lizes illumination-irrelevant details from input images, producing clear and undistorted text. Furthermore, we evaluate the impact of the decoders in another image restoration task based on SD: colorization. Due to space limit, please refer to the supplementary for corresponding results.

Framework Distillation. Compared with the full model, our lightweight version reduces the running time to 500x faster. It can process a 1024×1024 image in 0.03 seconds on a Tesla M40. Additionally, the number of parameters is reduced from 1.3G to 327.36k, even smaller than SOTAs with comparable performance, as shown in Fig. 11. The performance of our lightweight model on LOL/MIT datasets for PSNR, SSIM, LPIPS, and LOE is 20.45/18.15, 0.798/0.770, 0.290/0.175, and 0.273/0.174, respectively. Compared with our full model, the lightweight one achieves comparable performance and even marginally improves the PSNR and LOE. This difference might be due to the randomness inherent in the generative diffusion model. During the fitting of the larger full version model, randomness is averaged, further reducing noise and rectifying errors.

4. Conclusion

We introduce a new zero-reference low-light enhancement framework, developed without low-light data. At its core lies a physical quadruple prior derived from the light transfer theory, and an efficient prior-to-image framework based on generative diffusion models. Experimental results show our superior performance across diverse scenarios.

References

- [1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra† Prafulla Dhariwal, Casey Chu† Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. 2023. 8
- [2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR*, 2011. 1, 6, 7
- [3] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, 2023. 1, 6, 7
- [4] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 1
- [5] Chen Chen, Qifeng Chen, Minh N. Do, and Vladlen Koltun. Seeing motion in the dark. In *ICCV*, 2019. 1
- [6] Guangyong Chen, Fengyuan Zhu, and Pheng-Ann Heng. An efficient statistical method for image noise level estimation. In *ICCV*, 2015. 6
- [7] Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *CVPR*, 2023. 1, 6, 7, 8
- [8] Jan-Mark Geusebroek, Rein van den Boomgaard, Arnold W. M. Smeulders, and Hugo Geerts. Color invariance. *IEEE TPAMI*, 23(12):1338–1350, 2001. 3
- [9] Theo Gevers, Arjan Gijsenij, Joost Van de Weijer, and Jan-Mark Geusebroek. *Color in computer vision: Fundamentals and applications*. John Wiley & Sons, 2012. 3, 4
- [10] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, 2020. 2, 6, 7
- [11] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2017. 2, 3, 6
- [12] Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2RNet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation*, 90: 103712, 2023. 7
- [13] Huiguo He, Tianfu Wang, Huan Yang, Jianlong Fu, Nicholas Jing Yuan, Jian Yin, Hongyang Chao, and Qi Zhang. Learning profitable NFT image diffusions via multiple visual-policy guided reinforcement learning. In *ACM MM*, 2023. 2
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 5
- [15] Haofeng Huang, Wenhan Yang, Yueyu Hu, Jiaying Liu, and Ling-Yu Duan. Towards low light enhancement with RAW images. *IEEE TIP*, 31:1391–1405, 2022. 1
- [16] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *ECCV*, 2022. 1
- [17] Haiyang Jiang and Yinjiang Zheng. Learning to see moving objects in the dark. In *ICCV*, 2019. 1
- [18] Hai Jiang, Ao Luo, Songchen Han, Haoqiang Fan, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. In *Siggraph Asia*, 2023. 6, 7
- [19] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE TIP*, 30:2340–2349, 2021. 1, 6, 7
- [20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 6
- [21] Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2D histograms. *IEEE TIP*, 22(12):5372–5384, 2013. 6
- [22] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C. van Gemert. Zero-shot domain adaptation with a physics prior. In *ICCV*, 2021. 4, 7
- [23] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE TPAMI*, 44(8):4225–4238, 2021. 2, 6, 7
- [24] Chongyi Li, Chunle Guo, Ruicheng Feng, Shangchen Zhou, and Chen Change Loy. Cudi: Curve distillation for efficient and controllable exposure adjustment. *arXiv*, 2022. 2
- [25] Chongyi Li, Chunle Guo, Shangchen Zhou, Qiming Ai, Ruicheng Feng, and Chen Change Loy. Flexicurve: Flexible piecewise curves estimation for photo retouching. In *CVPRW*, 2023. 1
- [26] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE TIP*, 27(6):2828–2841, 2018. 2
- [27] Zhixin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *ICCV*, 2023. 1, 6, 7
- [28] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2, 6, 7
- [29] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *CVPR*, 2022. 6
- [30] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, 2021. 2, 6, 7
- [31] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *PR*, 61:650–662, 2017. 1
- [32] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv*, 2022. 6
- [33] Kede Ma, Kai Zeng, and Zhou Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE TIP*, 24(11):3345–3356, 2015. 6

- [34] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 2022. 1, 2, 6, 7
- [35] Yiyang Ma, Huan Yang, Bei Liu, Jianlong Fu, and Jiaying Liu. AI illustrator: Translating raw descriptions into images by prompt-based cross-modal generation. In *ACM MM*, 2022. 2
- [36] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12):4695–4708, 2012. 6
- [37] 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 *VBC*, 1990. 2
- [38] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In *ECCV*, 2022. 6
- [39] Zhongwei Qiu, Huan Yang, Jianlong Fu, Daochang Liu, Chang Xu, and Dongmei Fu. Learning degradation-robust spatiotemporal frequency-transformer for video super-resolution. *IEEE TPAMI*, 45(12):14888–14904, 2023. 6
- [40] Zia-ur Rahman, Daniel J Jobson, and Glenn A Woodell. Retinex processing for automatic image enhancement. *Journal of Electronic Imaging*, 2004. 2
- [41] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*, 2020. 6
- [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 5, 8
- [43] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023. 2
- [44] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 45(4):4713–4726, 2023. 7
- [45] Vassilios Vonikakis, Rigas Kouskouridas, and Antonios Gasteratos. On the evaluation of illumination compensation algorithms. *Multimedia Tools and Applications*, 77:9211–9231, 2018. 6
- [46] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE TIP*, 22(9):3538–3548, 2013. 6
- [47] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex C. Kot. Low-light image enhancement with normalizing flow. In *AAAI*, 2022. 1
- [48] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 1, 3, 6, 7, 8
- [49] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *CVPR*, 2022. 1, 7
- [50] Yuhui Wu, Chen Pan, Guoqing Wang, Yang Yang, Jiwei Wei, Chongyi Li, and Heng Tao Shen. Learning semantic-aware knowledge guidance for low-light image enhancement. In *CVPR*, 2023. 1
- [51] Zhihao Xia, Michael Gharbi, Federico Perazzi, Kalyan Sunkavalli, and Ayan Chakrabarti. Deep denoising of flash and no-flash pairs for photography in low-light environments. In *CVPR*, 2021. 1
- [52] Jinhui Xiong, Jian Wang, Wolfgang Heidrich, and Shree K. Nayar. Seeing in extra darkness using a deep-red flash. In *CVPR*, 2021. 1
- [53] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 6
- [54] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *ICCV*, 2023. 1, 6, 7, 8
- [55] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*, 2020. 6, 7
- [56] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE TIP*, 29:5737–5752, 2020. 7
- [57] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with A generative diffusion model. In *ICCV*, 2023. 1
- [58] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 6
- [59] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *CVPR*, 2021. 1
- [60] Lvmian Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 5
- [61] Lin Zhang, Lijun Zhang, Xinyu Liu, Ying Shen, Shaoming Zhang, and Shengjie Zhao. Zero-shot restoration of back-lit images using deep internal learning. In *ACM MM*, 2019. 6, 7
- [62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6
- [63] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, 2019. 1, 3, 7
- [64] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *IJCV*, 129(4):1013–1037, 2021. 7
- [65] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. In *IJCAI*, 2023. 1
- [66] Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. Moviefactory: Automatic movie creation from text using large generative models for language and images. In *ACM MM*, 2023. 2

- [67] Junchen Zhu, Huan Yang, Wenjing Wang, Huiguo He, Zixi Tuo, Yongsheng Yu, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, Jianlong Fu, and Jiebo Luo. Mobilevidfactory: Automatic diffusion-based social media video generation for mobile devices from text. In *ACM MM*, 2023. 2