Zero-Shot Low-Light Image Enhancement via Latent Diffusion Models

Yan Huang[1], **Xiaoshan Liao**[1], **Jinxiu Liang**[2,3#], **Yuhui Quan**[1,4], **Boxin Shi**[2,3], **Yong Xu**[1,4]

[1]School of Computer Science and Engineering, South China University of Technology
[2]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[4]Pazhou Lab
{aihuangy, csyhquan, yxu}@scut.edu.cn, {csxsliao}@mail.scut.edu.cn,
{cssherryliang, shiboxin}@pku.edu.cn

Abstract

Low-light image enhancement (LLIE) aims to improve visibility and signal-to-noise ratio in images captured under poor lighting conditions. While deep learning has shown promise in this domain, current approaches require extensive paired training data, limiting their practical utility. We present a novel framework that reformulates low-light image enhancement as a zero-shot inference problem using pre-trained latent diffusion models (LDMs), eliminating the need for task-specific training data. Our key insight is that the rich natural image priors encoded in LDMs can be leveraged to recover well-lit images through a carefully designed optimization process. To address the ill-posed nature of low-light degradation and the complexity of latent space optimization, our framework introduces an exposure-aware degradation module that adaptively models illumination variations and a principled latent regularization scheme with adaptive guidance that ensures both enhancement quality and natural image statistics. Experimental results demonstrate that our framework outperforms existing zero-shot methods across diverse real-world scenarios.

Code — https://github.com/Eileen000/LLIEDiff

Extended version —
https://github.com/Eileen000/LLIEDiff/raw/main/paper.pdf

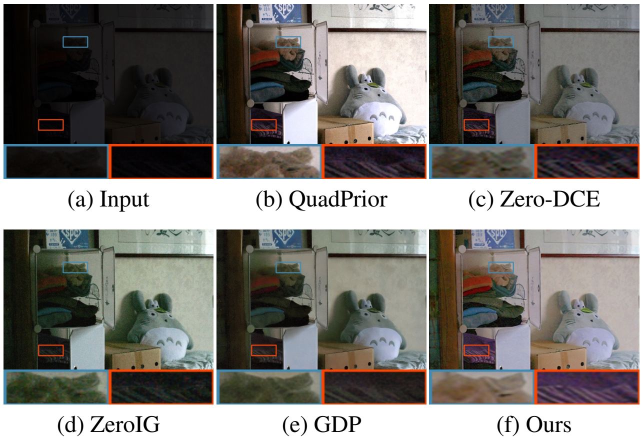

(a) Input (b) QuadPrior (c) Zero-DCE
(d) ZeroIG (e) GDP (f) Ours

Figure 1: Comparison of LLIE results between state-of-the-art unsupervised learning methods (b) QuadPrior (Wang et al. 2024), (c) ZeroDEC (Guo et al. 2020), and zero-shot learning methods (d) ZeroIG (Shi et al. 2024), (e) GDP (Fei et al. 2023), and (f) the proposed one based on pre-trained LDMs (Rombach et al. 2022).

Introduction

Low-light image enhancement (LLIE) stands as a critical challenge in computer vision, fundamentally impacting applications from autonomous navigation to medical diagnostics (Li et al. 2022). The complexity of this problem arises from the non-linear degradation processes in low-light conditions, which introduce multiple interrelated challenges: severe noise contamination, compromised contrast, and significant color distortion. These factors not only degrade visual quality but also pose risks to downstream vision tasks.

Traditional LLIE approaches have primarily relied on analytical models and hand-crafted priors (Fu et al. 2016; Guo, Li, and Ling 2017; Wei et al. 2018; Zhang, Zhang, and Guo 2019; Liu et al. 2021b). While theoretically grounded, these methods struggle to capture the intricate statistics of natural images, resulting in limited effectiveness across diverse real-world scenarios The advent of deep learning has dramatically advanced the field (Chen et al. 2018; Wang et al. 2019; Xu et al. 2020; Yang et al. 2020; Liu et al. 2021b; Liang et al. 2023a), yet fundamental challenges persist: *i)* The requirement for extensive paired training data presents a significant bottleneck, as acquiring such data is both costly and often impractical in real-world settings. Although unsupervised methods (Jiang et al. 2021; Guo et al. 2020) attempt to address this limitation, their dependence on large unpaired datasets still constrains practical deployment. *ii)* The challenge of achieving robust generalization across diverse lighting conditions and scene types remains particularly critical in high-stakes applications like forensic analysis and medical imaging (Drozdowski et al. 2020; Abbasi-Sureshjani et al. 2020; Liang et al. 2022). This limitation fundamentally stems from the difficulty of learning effective image priors from quite limited training data.

Recent breakthroughs in latent diffusion models (LDMs) (Rombach et al. 2022) offer a promising new direction. These models learn a compressed latent space

[#]Corresponding author

that captures the manifold of natural images through a diffusion process, enabling powerful generative capabilities without pixel-space computation overhead. Their success in solving various inverse problems without task-specific training (Chung et al. 2023; Rout et al. 2023) suggests particular promise for LLIE, where they could leverage rich learned priors to handle diverse lighting conditions and provide principled uncertainty estimates.

This observation motivates our novel approach: leveraging pre-trained LDMs for zero-shot LLIE. This strategy offers several compelling advantages: First, it eliminates task-specific training by utilizing rich priors learned from massive general-domain datasets, effectively transferring knowledge from foundation models. Second, it capitalizes on substantial existing computational investments in model pre-training. Third, the probabilistic nature of generative models enables principled sampling from the posterior distribution, naturally mitigating the regression-to-mean artifacts common in deterministic approaches.

However, adapting LDMs for zero-shot LLIE presents unique technical challenges. The degradation process in low-light imaging involves complex, spatially-varying interactions between scene radiance, sensor characteristics, and noise sources. These relationships become even more ambiguous in the latent space of pre-trained LDMs. Furthermore, balancing the preservation of input image content with the generation of naturally-lit outputs requires careful consideration - too strong a guidance towards the input may retain unwanted low-light characteristics, while too weak a guidance risks content modification or detail hallucination.

We address these challenges through the following strategies: *i)* An exposure-aware degradation modeling module that incorporates the bright channel as a representation of illumination to model the degradation process with an image-adaptive exposure factor. *ii)* A principled latent space regularization scheme with adaptive guidance, which penalizes latents whose decoded images fall outside the manifold of natural images. Our framework demonstrates significant advantages: it eliminates the need for paired or unpaired training data and generalizes robustly across diverse scenarios. As illustrated in Figure 1, our method outperforms existing zero-shot and zero-reference approaches.

To summarize, this paper proposes the first zero-shot latent diffusion-based LLIE framework using LDMs with the following contributions:

- We develop an exposure-adaptive bright channel-based degradation modeling module, adapting dynamically to varying illumination conditions.
- We introduce a principled latent regularization term with adaptive guidance that simultaneously optimizes enhancement quality and maintains natural image statistics.

Related Works

The field of LLIE has witnessed remarkable progress, transitioning from conventional hand-crafted methods to sophisticated data-driven deep learning approaches (Li et al. 2022; Liu et al. 2021a). In the following, we delve into a focused review of the most relevant deep learning-based techniques.

Regression LLIE methods Deep learning-based regression methods learn a mapping between low-light and normal-light images by leveraging existing architectures or incorporating problem-related information (Guo et al. 2020; Yang et al. 2020; Liang et al. 2021; Zhou et al. 2023). Examples include HWMNet (Fan, Liu, and Liu 2022), which integrates half wavelet attention with CNN, and IAT (Cui et al. 2022), a lightweight illumination adaptive Transformer under different light conditions. Retinex-based deep learning methods, such as RUAS (Liu et al. 2021b), KinD (Zhang, Zhang, and Guo 2019), and KinD++ (Zhang et al. 2021), have also been proposed. While excelling in structure recovery and distortion metrics, they tend to produce over-smoothed images lacking high-frequency details, which harms perceptual realism.

Generative LLIE methods Generative methods are known for their exceptional perceptual quality and ability to produce high-frequency details (Jiang et al. 2021; Zhou, Yang, and Yang 2023; Jiang et al. 2023). They differ in their underlying generative models and learning principles. EnlightenGAN integrates attention mechanisms with image-related regularization (Jiang et al. 2021). Normalizing flow models, such as LLFlow (Wang et al. 2022), have also been utilized. However, GANs face challenges like training instability and artifact introduction, while normalizing flows have limitations in their expressive capacity.

Diffusion-based LLIE methods Diffusion models (DMs) have revolutionized image generation, with main formulations: Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020), Stochastic Differential Equations (SDE) (Song et al. 2021), and Noise Conditional Score Networks (NCSN) (Song and Ermon 2020). DDPMs consist of a noise-added diffusion process and a noise removal-based reverse process, while NCSN models focus on score-based generative modeling for denoising and enhancement. SDE-based models generalize these concepts through forward and reverse SDEs (Huang, Lim, and Courville 2021). DMs have been applied to various tasks, such as image restoration in adverse weather conditions (Özdenizci and Legenstein 2023), image shadow removal (Guo et al. 2023), and low-resolution latent space diffusion (Luo et al. 2023).

Diffusion-based LLIE approaches employ a noise estimation network for the reverse process (Yin et al. 2023; Jiang et al. 2023; Zhou, Yang, and Yang 2023). Basic DDPMs (Ho, Jain, and Abbeel 2020) lack spatial adaptation and may fail to preserve fine details in complex textures. PyDiff (Zhou, Yang, and Yang 2023) enhances low-light images by progressively increasing resolution and globally correcting degradation, while DiffLL (Jiang et al. 2023) uses wavelet transformation. A diffusion-based post-processing framework has also been proposed (Panagiotou and Bosman 2023). LLDiffusion (Wang et al. 2023a) integrates image degradation and priors. CLEDiff offers enhancement and region-specific controllability (Yin et al. 2023). There is also a zero-reference method (Wang et al. 2024) that leverages pre-trained DMs, whose weights are fine-tuned with simulated data.

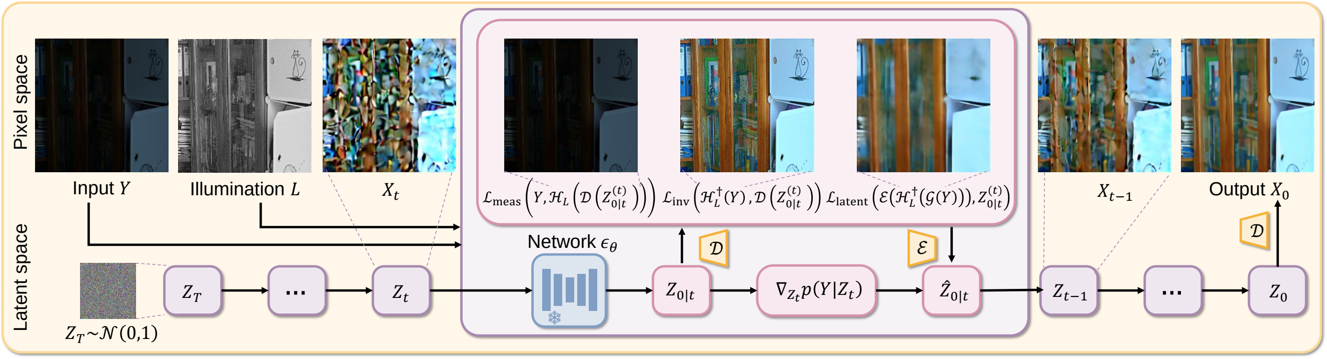

Pixel space
Latent space
Input Y
Illumination L
X_t
$\mathcal{L}_{\text{meas}}\left(Y, \mathcal{H}_L\left(\mathcal{D}\left(Z_{0|t}^{(t)}\right)\right)\right)$
$\mathcal{L}_{\text{inv}}\left(\mathcal{H}_L^{\dagger}(Y), \mathcal{D}\left(Z_{0|t}^{(t)}\right)\right)$
$\mathcal{L}_{\text{latent}}\left(\mathcal{E}(\mathcal{H}_L^{\dagger}(\mathcal{G}(Y))), Z_{0|t}^{(t)}\right)$
X_{t-1}
Output X_0
Network ϵ_θ
$\mathcal{D}$
$\mathcal{E}$
$\mathcal{D}$
Z_T
...
Z_t
$Z_{0|t}$
$\nabla_{Z_t} p(Y|Z_t)$
$\hat{Z}_{0|t}$
Z_{t-1}
...
Z_0
$Z_T \sim \mathcal{N}(0,1)$

Figure 2: Illustration of the proposed zero-shot LLIE framework, which leverages the power of pre-trained LDMs, *i.e.*, Stable Diffusion (Rombach et al. 2022), for superior image priors. The latent and pixel spaces are connected via the encoder $\mathcal{E}$ and decoder $\mathcal{D}$: $X = \mathcal{D}(Z), Z = \mathcal{E}(X)$. Starting with pure Gaussian noise $Z_T \sim \mathcal{N}(0, I)$, we iteratively estimate the latent representation $Z_{0|t}$ of the desired result using the pre-trained neural network ϵ_θ, ensuring the generation adheres to natural image characteristics. Given a low-light input image Y and its estimated illumination map L (defining the degradation model $\mathcal{H}$ and its inverse $\mathcal{H}^\dagger$), we ensure fidelity to the input via minimization of objectives derived from the low-light image likelihood $p(Y|Z_t)$ at each timestep t. These objectives include regularizations $\mathcal{L}_{\text{meas}}$, $\mathcal{L}_{\text{inv}}$, and $\mathcal{L}_{\text{latent}}$ defined in Eqs. (18) to (20), respectively, where a blurring operation $\mathcal{G}$ is incorporated to emphasize the low-frequency information. Finally, at timestep $t = 0$, the enhanced image is obtained via $X_0 = \mathcal{D}(Z_0)$.

Methodology

In this section, we present our framework for zero-shot latent diffusion-based LLIE, leveraging LDMs pre-trained on massive data. The overall pipeline is shown in Figure 2. Our approach builds upon the recent advancements in diffusion-based generative models, particularly the Stable Diffusion (Rombach et al. 2022), which has demonstrated remarkable performance in various image generation tasks. We extend this approach to the problem of LLIE by introducing several novel components that address the specific challenges associated with this task.

Degradation Model

Given a low-light image $Y \in \mathbb{R}^{H\times W\times 3}$, our goal is to estimate the corresponding normal-light image $X \in \mathbb{R}^{H\times W\times 3}$. We assume that the degradation process $\mathcal{H}$ can be modeled as an element-wise multiplication between the normal-light image and an illumination map $L \in \mathbb{R}^{H\times W}$:

$$Y = \mathcal{H}(X) = X \odot L^{\beta} + N, \tag{1}$$

where $\odot$ denotes the element-wise multiplication, β is an exposure factor, and N models the additive noise. This degradation model is motivated by the physical properties of light and has been widely adopted in the literature on LLIE (Guo, Li, and Ling 2017; Li et al. 2018), which is built upon the Retinex model. Here, L represents the illumination map. We assume that for color images, three channels share the same illumination map. Based on the Maximum a Posterior (MAP) framework,

$$\begin{aligned} X^* &= \text{argmax}_X \, p(X|Y, L) \\ &= \text{argmax}_X \, p(Y, L|X)p(X), \end{aligned} \tag{2}$$

where $p(Y, L|X)$ is the likelihood corresponding to the data fidelity, and the priors $p(X)$ model the latent normal-light image.

Estimating the illumination map is a crucial step in our framework, as it provides a representation of the low-light conditions in the input image. To achieve this, we employ the bright channel prior (Guo, Li, and Ling 2017), which is effective in capturing illumination information. The bright channel is defined as the maximum value among the three color channels for each pixel:

$$L_0(i, j) = \max_{c\in\{R,G,B\}} Y^c(i, j), \tag{3}$$

where Y^c represents the c-th color channel of the low-light image Y, and (i, j) denotes the pixel location. To obtain a smooth and spatially consistent illumination map, we apply a Gaussian filter $\mathcal{G}$ to the bright channel:

$$L = \mathcal{G}(L_0). \tag{4}$$

It helps to remove high-frequency noise and artifacts that may be present in the bright channel, resulting in a more reliable estimation of the illumination map.

Assuming there is no noise, we can recover the normal-light image X by simply performing an inverse operation $\mathcal{H}^\dagger$ of the forward operation in Eq. (1):

$$\widehat{X} = \mathcal{H}^\dagger(Y) = \frac{Y}{L^{\beta}}. \tag{5}$$

Posterior Based on Latent Diffusion

In LDMs such as Stable Diffusion (Rombach et al. 2022), the diffusion occurs in the latent space. The latent and pixel spaces are connected via the encoder $\mathcal{E}(\cdot) : \mathbb{R}^{H\times W} \to \mathbb{R}^k$ and decoder $\mathcal{D}(\cdot) : \mathbb{R}^k \to \mathbb{R}^{H\times W}$: $X = \mathcal{D}(Z), Z = \mathcal{E}(X)$. We consider how to construct the posterior distribution in Eq. (2) given the prior distribution $p(X)$ of natural images through their latent $p(Z)$ modeled by LDMs.

In the score-based perspective, the forward diffusion process of the data $Z_t, t \in [0, T]$ is defined with a linear SDE

$$dZ = -f(Z, t)dt + g(t)dw, \tag{6}$$

where w is the standard Brownian motion. During sampling, starting with pure Gaussian noise $Z_T \sim \mathcal{N}(0, I)$, the reverse diffusion process is run and then a normal-light image X_0 can be obtained by passing $Z_0|Z_T$ through the decoder $\mathcal{D}$. Then, the corresponding reverse SDE is given by

$$dZ = [-f(Z,t) - g^2(t)\nabla_{Z_t} \log p_t(Z_t)]dt + g(t)dw, \qquad (7)$$

where $\nabla_{Z_t} \log p_t(Z_t)$ is the time-dependent score function, typically approximated with denoising score matching (Song et al. 2021)

$$\min_{\theta} \mathbb{E}_{t,Z_t,Z_0}[||\epsilon_\theta(Z_t,t) - \nabla_{Z_t} \log p(Z_t|Z_0), ||_2^2], \qquad (8)$$

where neural network $\epsilon_\theta(Z,t)$ is trained to predict the score $\nabla_Z \log p(Z)$ of a diffusion process. Once θ^* is acquired by training, one can use the approximation $\nabla_{Z_t} \log p_t(Z_t) \approx \epsilon_{\theta^*}(Z_t, t)$ as a plug-in estimate to replace the score function in Eq. (7), and solve by discretization (*e.g.*, ancestral sampling of Ho, Jain, and Abbeel (2020)), effectively sampling from the prior distribution $p(Z_0)$.

Given low-light image Y modeled on the latent Z_0 of normal-light image X_0 by Eq. (1), the posterior distribution $p(Z_0|Y)$ can be sampled by running a modified Reverse SDE that depends on the unconditional score $\nabla_{Z_t} \log p(Z_t)$ and the term $\nabla_{Z_t} \log p(Y|Z_t)$:

$$\begin{aligned} dZ =& [-f(Z,t) - g^2(t)(\nabla_{Z_t} \log p(Z_t) \\ &+ \nabla_{Z_t} \log p(Y|Z_t))]dt + g(t)dw, \end{aligned} \qquad (9)$$

where we have used the fact that

$$\nabla_{Z_t} \log p_t(Z_t|Y) = \nabla_{Z_t} \log p(Z_t) + \nabla_{Z_t} \log p(Y|Z_t), \quad (10)$$

based on Bayes' rule. To compute the former term involving $p(Z_t)$, we can simply use the pre-trained score function ϵ_{θ^*}. The latter term captures how much the current iterate explains the observed low-light image Y, it is hard to acquire in closed-form due to the dependence on the time t, as there only exists explicit dependence between low-light image Y and estimated normal-light image $\mathcal{D}(Z_0)$. The term $p(Y|Z_t)$ can be factorized as

$$\begin{aligned} p(Y|Z_t) &= \int p(Y|Z_0, Z_t)p(Z_0|Z_t)dZ_0 \\ &= \int p(Y|Z_0)p(Z_0|Z_t)dZ_0. \end{aligned} \qquad (11)$$

Noted that for the case of DMs such as variance preserving (VP) SDE or DDPM, the forward diffusion can be simply represented by

$$Z_t = \sqrt{\bar{\alpha}_t} Z_0 + \sqrt{1 - \bar{\alpha}_t} z, \quad z \sim \mathcal{N}(0, I), \qquad (12)$$

so that we can obtain the specialized representation of the posterior mean through Tweedie's approach as follows (Chung et al. 2023):

$$\begin{aligned} Z_{0|t} &= \mathbb{E}[Z_0|Z_t] \\ &= \frac{1}{\sqrt{\bar{\alpha}_t}}(Z_t + (1 - \bar{\alpha}_t)\nabla_{Z_t} \log p_t(Z_t)), \end{aligned} \qquad (13)$$

where the last term $\nabla_{Z_t} \log p_t(Z_t)$ can be replaces by $\epsilon_{\theta^*}(Z_t)$. Given the posterior mean $\bar{Z}_0$ that can be efficiently computed at the intermediate steps, we propose to provide a tractable approximation for $p(Y|Z_t)$ such that one can use the surrogate function to maximize the low-light image *likelihood* $p(Y|\mathcal{D}(Z_{0|t}))$ between the low-light image Y and estimated normal-light image $\mathcal{D}(Z_{0|t})$-yielding approximate posterior sampling. Specifically, given the interpretation $p(Y|Z_t) = \mathbb{E}_{Z_0 \sim p(Z_0|Z_t)}[p(Y|\mathcal{D}(Z_0))]$ we use the following approximation in Chung et al. (2023):

$$\begin{aligned} p(Y|Z_t) &\approx p(Y|Z_{0|t} := \mathcal{D}(\mathbb{E}[Z_0|Z_t])) \\ &= p(Y|\mathcal{D}(Z_{0|t})\odot L^\beta). \end{aligned} \qquad (14)$$

Essentially, such an approximation substitutes the unknown normal-light image Z_0 with its conditional expectation given the noisy input, $\mathbb{E}[Z_0|Z_t]$. Under this approximation, the term $p(Y|Z_t)$ becomes tractable.

Low-light Image Likelihood

In a probabilistic setting, where the random noise N is assumed to follow Gaussian distributions with standard deviation σ, the maximization of the likelihood in (14) of observing the low-light image Y, given an estimated normal-light image $\mathcal{D}(Z_{0|t})$ and illumination map L can be expressed as minimizing

$$\mathcal{L} = \frac{1}{\sigma^2}||Y - \mathcal{D}(Z_{0|t})\odot L^\beta||_2^2. \qquad (15)$$

Plugging in the score function into Eq. (10), we can obtain

$$\nabla_{Z_t} \log p_t(Z_t|Y) = \epsilon_{\theta^*}(Z_t, t) - s\nabla_{Z_t}\mathcal{L}, \qquad (16)$$

where s denotes the scale of guidance from the given Y.

To impose constraints in the latent space of pre-trained LDMs, we propose a latent regularization term to guide the diffusion process towards latent that explains measurements and remains fixed points of the decoder-encoder composition, ensuring generated samples stay on the data manifold. The overall objective $\mathcal{L}$ to maximize the likelihood is defined as

$$\mathcal{L} = \frac{1}{\sigma_{\text{p}}^2}\mathcal{L}_{\text{meas}} + \frac{1}{\sigma_{\text{i}}^2}\mathcal{L}_{\text{inv}} + \frac{1}{\sigma_{\text{l}}^2}\mathcal{L}_{\text{latent}} + \frac{1}{\sigma_{\text{c}}^2}\mathcal{L}_{\text{col}}, \qquad (17)$$

where σs are scaler for balancing different terms,

$$\mathcal{L}_{\text{meas}} = ||Y - \mathcal{D}(Z_{0|t})\odot L^\beta||_2^2, \qquad (18)$$

$$\mathcal{L}_{\text{inv}} = ||\frac{Y}{L^\beta} - \mathcal{D}(Z_{0|t})||_2^2, \qquad (19)$$

and

$$\mathcal{L}_{\text{latent}} = ||\mathcal{E}(\mathcal{G}(\frac{Y}{L^\beta})) - Z_{0|t}||_2^2, \qquad (20)$$

respectively. These terms serve as the regularization in different aspects, which is particularly important in the context of LLIE, where the increased ambiguity in the latent space can lead to artifacts and unnatural-looking images if not properly constrained. By incorporating this term into the optimization objective, we effectively regularize the solution space and improve the overall quality of the enhanced images. Furthermore, to maintain the inter-relationship between image color channels during the sampling process and to avoid color shifts, we have also incorporated a channel consistency loss:

$$\mathcal{L}_{\text{col}} = \sum_{\forall(p,q)\in\Omega} \left(\mathcal{D}(Z_{0|t})_p - \mathcal{D}(Z_{0|t})_q\right)^2, \qquad (21)$$

where (p, q) denotes a pair of channels sampled from $\Omega = \{(R, G), (R, B), (G, B)\}$ and $D_p(Z_{0|t})$ represents the pixel values of the p-th channel of $D(Z_{0|t})$.

Table 1: Comparisons of state-of-the-art methods on the real-world dataset proposed by Wei et al. (2018). Yellow, orange, and red highlights indicate the best-performing method among the dataset-based supervised (SL) and unsupervised (UL) learning methods, and the zero-shot (ZS) methods learned from the input test image only. ↑ (↓) indicates that higher (lower) values are better.

Type	Method	PSNR↑	SSIM↑	LPIPS↓
SL	KinD (Zhang, Zhang, and Guo 2019)	17.65	0.775	0.171
	DRBN (Yang et al. 2020)	16.29	0.551	0.260
	KinD++ (Zhang et al. 2021)	14.71	0.799	0.207
	RetinexNet (Yang et al. 2021)	17.65	0.648	0.379
	URetinexNet (Wu et al. 2022)	19.80	0.826	0.128
	SNR-Aware (Xu et al. 2022)	24.61	0.840	0.151
	DiffLL (Jiang et al. 2023)	21.84	0.871	0.202
	GSAD (Hou et al. 2024)	22.96	0.917	0.104
UL	EnlightenGAN (Jiang et al. 2021)	17.48	0.652	0.322
	PairLIE (Fu et al. 2023)	19.15	0.736	0.248
	NeRCo (Yang et al. 2023)	19.74	0.832	0.234
	CLIP-LIT (Liang et al. 2023b)	12.39	0.663	0.382
	ZeroDCE (Guo et al. 2020)	14.58	0.736	0.401
	RUAS (Liu et al. 2021b)	16.40	0.771	0.270
	SCI (Ma et al. 2022)	14.78	0.710	0.339
	QuadPrior (Wang et al. 2024)	18.34	0.859	0.213
ZS	ExCNet (Zhang et al. 2019)	13.88	0.648	0.370
	GDP (Fei et al. 2023)	15.83	0.688	0.338
	ZeroIG (Shi et al. 2024)	17.63	0.632	0.390
	Ours	19.82	0.841	0.242

Adaptive Guidance Scale

During the sampling, it is observed that the optimal guidance scale in Eq. (16) may vary depending on the specific image content and the current state of the optimization. Therefore, another key component of our framework is the adaptive guidance scale, which dynamically adapts and controls the strength of the guidance from the low-light image during the denoising process.

To adaptively adjust the fidelity introduced by the guidance from the given low-light image Y, we propose to adapt the guidance scale based on the distance between the current estimation and the previous one in the latent space. We define the adaptive guidance scale as:

$$s_t = \frac{\|Z_t - Z_{t-1}\|_2^2}{(\nabla_{Z_t}\mathcal{L} - \nabla_{Z_{t-1}}\mathcal{L})(Z_t - Z_{t-1})}. \tag{22}$$

The adaptive guidance scale plays a crucial role in balancing the influence of the low-light image and the prior knowledge captured by the diffusion model. By dynamically adjusting the guidance strength, our framework can effectively handle a wide range of low-light conditions and image contents, leading to more robust and visually appealing results.

Experiments

Experimental Settings

Datasets We conduct experiments on two widely used datasets with paired low-light and normal-light images, LOL-v1 (Wei et al. 2018) and LOL-v2 (Yang et al. 2020). We also provide visual comparisons on datasets without ground truth (DICM (Lee, Lee, and Kim 2013), MEF (Ma, Zeng, and Wang 2015), NPE (Wang et al. 2013), and LIME (Guo, Li, and Ling 2017)) in supplementary materials.

Table 2: Evaluations of low-light enhancement performance on the real-world dataset proposed by (Yang et al. 2020).

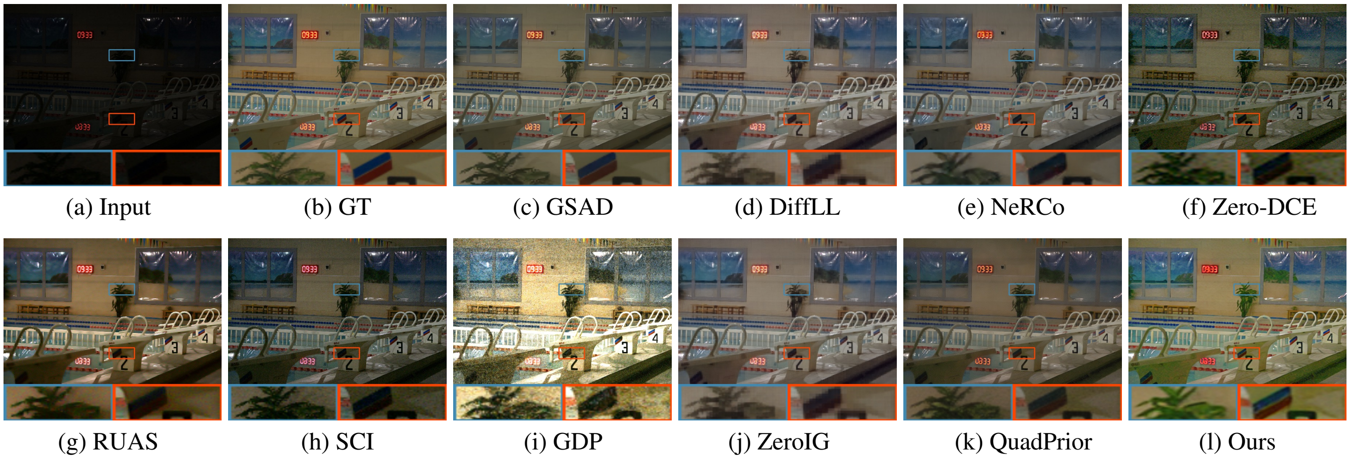
Type	Method	PSNR↑	SSIM↑	LPIPS↓
SL	KinD (Zhang, Zhang, and Guo 2019)	14.74	0.641	0.447
	DRBN (Yang et al. 2020)	20.13	0.830	0.147
	RetinexNet (Yang et al. 2021)	18.33	0.723	0.365
	URetinexNet (Wu et al. 2022)	21.16	0.840	0.196
	SNR-Aware (Xu et al. 2022)	21.48	0.849	0.193
	DiffLL (Jiang et al. 2023)	23.47	0.882	0.189
	GSAD (Hou et al. 2024)	20.16	0.890	0.112
UL	EnlightenGAN (Jiang et al. 2021)	18.23	0.617	0.308
	PairLIE (Fu et al. 2023)	19.89	0.778	0.291
	NeRCo (Yang et al. 2023)	19.67	0.777	0.270
	CLIP-LIT (Liang et al. 2023b)	15.18	0.697	0.368
	ZeroDCE (Guo et al. 2020)	18.06	0.744	0.312
	RUAS (Liu et al. 2021b)	15.33	0.745	0.309
	SCI (Ma et al. 2022)	17.30	0.721	0.307
	QuadPrior (Wang et al. 2024)	20.24	0.831	0.232
ZS	ExCNet (Zhang et al. 2019)	15.50	0.574	0.410
	GDP (Fei et al. 2023)	14.36	0.630	0.364
	ZeroIG (Shi et al. 2024)	15.66	0.607	0.408
	Ours	19.95	0.781	0.272

Table 3: Ablation studies on the regularization term. Red highlights indicate the best-performing settings.

$\mathcal{L}_{\text{meas}}$	$\mathcal{L}_{\text{inv}}$	$\mathcal{L}_{\text{latent}}$	$\mathcal{G}$	PSNR↑	SSIM↑	LPIPS↓
✗	✓	✗	✓	12.47	0.514	0.781
✓	✗	✗	✓	16.59	0.607	0.720
✗	✗	✓	✓	18.69	0.735	0.535
✓	✗	✓	✓	18.88	0.743	0.541
✓	✓	✗	✓	10.57	0.453	0.858
✗	✓	✓	✓	19.18	0.766	0.461
✓	✓	✓	✗	18.91	0.703	0.411
✓	✓	✓	✓	19.15	0.767	0.452

Metrics We evaluate the performance of each method using various metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

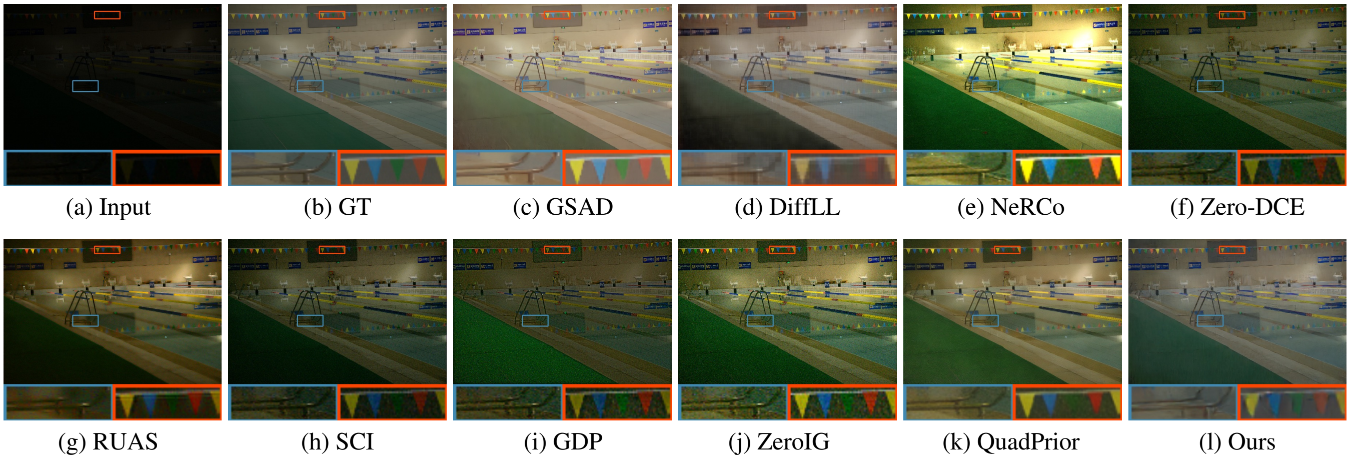
Implementation Details Our method is implemented using PyTorch and run on a single NVIDIA GTX 3090 Ti GPU. In all experiments, we use DDIM sampling (Song, Meng, and Ermon 2020) with $T = 1000$ steps. To further enhance local contrast in the degradation model, we apply pixel-wise exposure $\beta(i) = L(i)^{\exp{(L(i)-\phi)}}$ adaptive to the illumination map L at each pixel i, where ϕ is a hyperparameter set to 0.3. The Gaussian blur filter $\mathcal{G}$ applied to the bright channel in Eq. (4) uses a kernel size of 5×5 with standard deviations 1.5. To address input with resolutions different from the pre-trained model's training resolution, we employ the method proposed in Wang et al. (2023b).



(a) Input (b) GT (c) GSAD (d) DiffLL (e) NeRCo (f) Zero-DCE

(g) RUAS (h) SCI (i) GDP (j) ZeroIG (k) QuadPrior (l) Ours

Figure 3: Comparison results on real data in the LOL-v1 dataset (Wei et al. 2018).



(a) Input (b) GT (c) GSAD (d) DiffLL (e) NeRCo (f) Zero-DCE

(g) RUAS (h) SCI (i) GDP (j) ZeroIG (k) QuadPrior (l) Ours

Figure 4: Comparison results on real data in the LOL-v2 dataset (Yang et al. 2020).

Comparison with State-of-the-Art Methods

We compare our method with several state-of-the-art LLIE methods on the LOL-v1 and LOL-v2 datasets, including eight supervised methods (KinD (Zhang, Zhang, and Guo 2019), DRBN (Yang et al. 2020), KinD++ (Zhang et al. 2021), RetinexNet (Yang et al. 2021), URetinexNet (Wu et al. 2022), SNR-Aware (Xu et al. 2022), DiffLL (Jiang et al. 2023), and GSAD (Hou et al. 2024)), four unpaired learning methods (EnlightenGAN (Jiang et al. 2021), PairLIE (Fu et al. 2023), NerCo (Yang et al. 2023) and CLIP-LIT (Liang et al. 2023b)), four zero-reference methods (Zero-DCE (Guo et al. 2020), RUAS (Liu et al. 2021b), SCI (Ma et al. 2022) and QuadPrior (Wang et al. 2024)), and two zero-shot methods (ExCNet (Zhang et al. 2019), GDP (Fei et al. 2023) and ZeroIG (Shi et al. 2024)).

The quantitative results are presented in Table 1 and Table 2, showing that our method achieves superior performance on both datasets. As shown in Table 1, our method significantly outperforms existing zero-shot approaches on the LOL-v1 test set in terms of PSNR, SSIM, and LPIPS. Compared to unpaired training and zero-reference methods, ours achieves higher PSNR and well-performing SSIM and LPIPS, reaching the performance levels of some supervised methods. The reason we cannot surpass the current best-supervised methods in SSIM and LPIPS is that supervised methods typically learn the data distribution bias through paired training sets. As shown in Table 2, our method continues to significantly outperform zero-shot techniques in terms of PSNR, SSIM, and LPIPS on the LOL-v2 test set, with SSIM even exceeding that of unpaired learning methods. This indicates that our approach is capable of generating high-quality images and effectively addressing real-world scenarios.

We present visual comparisons of our method and competitive methods on the paired datasets in Figure 3 and Figure 4. It can be seen that previous methods result in noise amplification, under or overexposure, or color distortion. However, our method removes noise, reconstructs texture details, and produces well-illuminated images. For example, in Figure 3 our method not only brightens the trademarks on objects within the red box but also removes noise and restores the texture details. In contrast, zero-reference and zero-shot methods fail

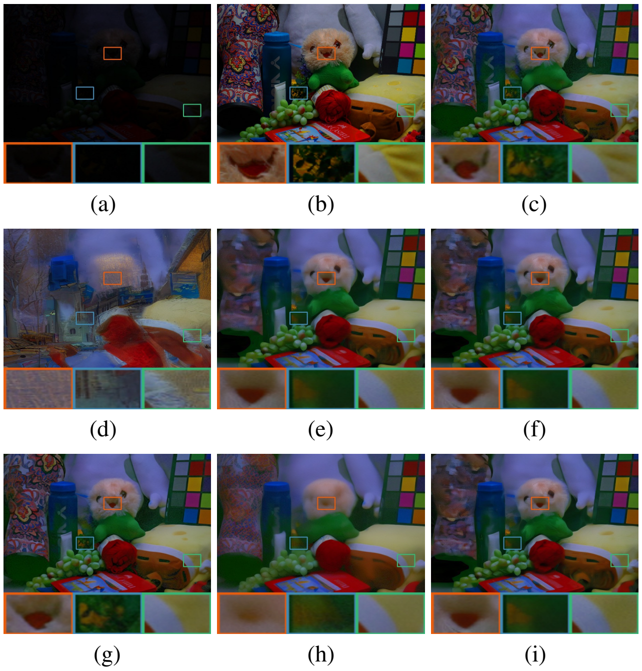


Figure 5: Comparison of results generated by using different settings of regularization terms. (a) Input. (b) Ground truth. (c) Ours. (d) W/ $\mathcal{L}_{\text{meas}}$. (e) W/ $\mathcal{L}_{\text{inv}}$. (f) W/ $\mathcal{L}_{\text{latent}}$. (g) W/o $\mathcal{L}_{\text{meas}}$. (h) W/o $\mathcal{L}_{\text{inv}}$. (i) W/o $\mathcal{L}_{\text{latent}}$.

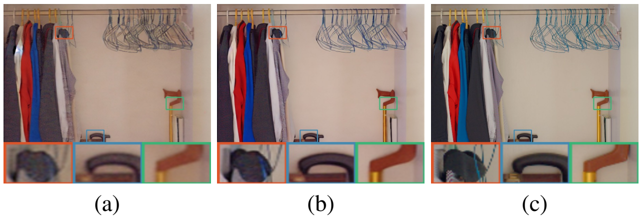


Figure 6: Comparison of results generated with (a) constant ($s_t = 0.7$), (b) adaptive guidance scale, and (c) ground truth.

to handle this area correctly. Additionally, the blue box in Figure 4 generated by our method is closer to the Ground Truth compared to GSAD (Hou et al. 2024), indicating our effectiveness in addressing color shifts. This demonstrates that our method significantly enhances visibility in the dark region, achieving results close to those of supervised methods.

Ablation Study

We conduct ablation studies on the LOL-v1 dataset to investigate the effectiveness of our main contributions. All experiments in this section were conducted with inputs of 256×256 resolution to accelerate the sampling process.

Regularization term in the latent space By comparing our full model with a variant that does not include the regularization term, we study their effect maintaining the naturalness of the results. It is noted that $\mathcal{L}_{\text{col}}$ is not included in these experiments. The experimental results in Table 3 indicate that the regularization term helps in preserving the naturalness and avoiding artifacts in the enhanced images. As illustrated in Figure 5, $\mathcal{L}_{\text{meas}}$ alone can convert low-light images to non-low-light ones but at the cost of missing image content. Both the $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{latent}}$ can enhance texture details, as shown in the red box. However, the combinations of these terms in pairs do not yield optimal results. When all three terms are combined, the image details are further enriched, as seen in the green box with correctly generated wrinkles. This highlights the significance of enforcing a prior distribution in the latent space for generating visually pleasing results.

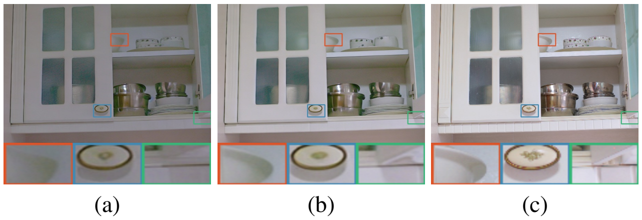


Figure 7: Comparison of results generated by using degradation model defined by (a) gamma correction with $\gamma = 0.7$, (b) the proposed degradation model defined by bright channel, and (c) ground truth.

As shown in the last two rows in Table 3, using the low-frequency extraction operation $\mathcal{G}$ in $\mathcal{L}_{\text{latent}}$ leads to significant improvements. This validates it can effectively remove image noise while appropriately preserving texture.

Adaptive guidance scale We investigate the impact of the adaptive guidance scale, which balances the effects of guided low-light images during the denoising process. The visual results in Figure 6 show clearer edges and colors that are closer to the Ground Truth, demonstrating that the adaptive guidance scale enables our method to effectively handle different levels of noise and distortions in low-light images.

Bright channel for degradation modeling We evaluate the impact of using the bright channel to model the degradation process in low-light images. As illustrated in Figure 7, after applying the bright channel prior, the edges of the bowl in the red box and the corner of the cabinet in the green box become clearer, indicating that the method can effectively handle local regions. In contrast, using a constant illumination value leads to a loss of contrast information. This demonstrates the importance of explicitly modeling the degradation process in the brightness domain for effective LLIE.

Conclusion

We presented the first zero-shot framework that leverages pre-trained latent diffusion models as powerful image priors, eliminating the need for task-specific training data or architectural modifications. The key technical innovations - the exposure-adaptive bright channel prior and principled latent space regularization with adaptive guidance - address core challenges in utilizing generative priors for enhancement tasks. Through extensive experiments on challenging real-world datasets, we demonstrated that our method not only outperforms existing zero-shot approaches but also exhibits superior generalization across diverse lighting conditions.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (No. 2024YFE0105400), National Natural Science Foundation of China (Nos. 62302019, 62088102, 62072188, 62472179, 62372186, 62136001), National Foreign Expert Project of the Ministry of Science and Technology of China (No. G2023163015L), Science and Technology Plan Project of Guangzhou (No. 2023A04J1681), and the China Postdoctoral Science Foundation (No. 2022M720236).

References

Abbasi-Sureshjani, S.; Raumanns, R.; Michels, B. E.; Schouten, G.; and Cheplygina, V. 2020. Risk of Training Diagnostic Algorithms on Data with Demographic Bias. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, 183–192.

Chen, C.; Chen, Q.; Xu, J.; and Koltun, V. 2018. Learning to See in the Dark. In *Proc. of Computer Vision and Pattern Recognition*, 3291–3300.

Chung, H.; Ryu, D.; McCann, M. T.; Klasky, M. L.; and Ye, J. C. 2023. Solving 3D Inverse Problems Using Pre-trained 2D Diffusion Models. In *Proc. of Computer Vision and Pattern Recognition*, 22542–22551.

Cui, Z.; Li, K.; Gu, L.; Su, S.; Gao, P.; Jiang, Z.; Qiao, Y.; and Harada, T. 2022. You Only Need 90K Parameters to Adapt Light: a Light Weight Transformer for Image Enhancement and Exposure Correction. In *Proc. of British Machine Vision Conference*, 238–255.

Drozdowski, P.; Rathgeb, C.; Dantcheva, A.; Damer, N.; and Busch, C. 2020. Demographic Bias in Biometrics: A Survey on an Emerging Challenge. *IEEE Transactions on Technology and Society*, 1(2): 89–103.

Fan, C.-M.; Liu, T.-J.; and Liu, K.-H. 2022. Half Wavelet Attention on M-Net+ for Low-Light Image enhancement. In *Proc. of International Conference on Image Processing*, 3878–3882.

Fei, B.; Lyu, Z.; Pan, L.; Zhang, J.; Yang, W.; Luo, T.; Zhang, B.; and Dai, B. 2023. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 9935–9946.

Fu, X.; Zeng, D.; Huang, Y.; Liao, Y.; Ding, X.; and Paisley, J. 2016. A Fusion-Based Enhancing Method for Weakly Illuminated Images. *Signal Processing*, 129: 82–96.

Fu, Z.; Yang, Y.; Tu, X.; Huang, Y.; Ding, X.; and Ma, K.-K. 2023. Learning a Simple Low-Light Image Enhancer from Paired Low-Light Instances. In *Proc. of Computer Vision and Pattern Recognition*, 22252–22261.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 1780–1789.

Guo, L.; Wang, C.; Yang, W.; Huang, S.; Wang, Y.; Pfister, H.; and Wen, B. 2023. Shadowdiffusion: When Degradation Prior Meets Diffusion Model for Shadow Removal. In *Proc. of Computer Vision and Pattern Recognition*, 14049–14058.

Guo, X.; Li, Y.; and Ling, H. 2017. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Transactions on Image Processing*, 26(2): 982–993.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Adv. of Neural Information Processing Systems*, volume 33, 6840–6851.

Hou, J.; Zhu, Z.; Hou, J.; Liu, H.; Zeng, H.; and Yuan, H. 2024. Global Structure-Aware Diffusion Process for Low-Light Image Enhancement. In *Adv. of Neural Information Processing Systems*, volume 36.

Huang, C.-W.; Lim, J. H.; and Courville, A. C. 2021. A Variational Perspective on Diffusion-Based Generative Models and Score Matching. In *Adv. of Neural Information Processing Systems*, volume 34, 22863–22876.

Jiang, H.; Luo, A.; Han, S.; Fan, H.; and Liu, S. 2023. Low-light Image Enhancement with Wavelet-Based Diffusion Models. *ACM Transactions on Graphics*, 42(6): 1–15.

Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021. EnlightenGAN: Deep Light Enhancement Without Paired Supervision. *IEEE Transactions on Image Processing*, 30: 2340–2349.

Lee, C.; Lee, C.; and Kim, C.-S. 2013. Contrast Enhancement Based on Layered Difference Representation of 2D Histograms. *IEEE Transactions on Image Processing*, 22(12): 5372–5384.

Li, C.; Guo, C.; Han, L.; Jiang, J.; Cheng, M.-M.; Gu, J.; and Loy, C. C. 2022. Low-Light Image and Video Enhancement Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9396–9416.

Li, M.; Liu, J.; Yang, W.; Sun, X.; and Guo, Z. 2018. Structure-Revealing Low-Light Image Enhancement Via Robust Retinex Model. *IEEE Transactions on Image Processing*, 27(6): 2828–2841.

Liang, J.; Wang, J.; Quan, Y.; Chen, T.; Liu, J.; Ling, H.; and Xu, Y. 2021. Recurrent Exposure Generation for Low-Light Face Detection. *IEEE Transactions on Multimedia*, 24: 1609–1621.

Liang, J.; Xu, Y.; Quan, Y.; Shi, B.; and Ji, H. 2022. Self-supervised low-light image enhancement using discrepant untrained network priors. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7332–7345.

Liang, J.; Yang, Y.; Li, B.; Duan, P.; Xu, Y.; and Shi, B. 2023a. Coherent Event Guided Low-Light Video Enhancement. In *Proc. of International Conference on Computer Vision*, 10615–10625.

Liang, Z.; Li, C.; Zhou, S.; Feng, R.; and Loy, C. C. 2023b. Iterative Prompt Learning for Unsupervised Backlit Image Enhancement. In *Proc. of International Conference on Computer Vision*, 8094–8103.

Liu, J.; Xu, D.; Yang, W.; Fan, M.; and Huang, H. 2021a. Benchmarking Low-Light Image Enhancement and Beyond. *International Journal of Computer Vision*, 129: 1153–1184.

Liu, R.; Ma, L.; Zhang, J.; Fan, X.; and Luo, Z. 2021b. Retinex-Inspired Unrolling With Cooperative Prior Architecture Search for Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 10561–10570.

Luo, Z.; Gustafsson, F. K.; Zhao, Z.; Sjölund, J.; and Schön, T. B. 2023. Refusion: Enabling Large-Size Realistic Image Restoration with Latent-space Diffusion Models. In *Proc. of Computer Vision and Pattern Recognition*, 1680–1691.

Ma, K.; Zeng, K.; and Wang, Z. 2015. Perceptual Quality Assessment for Multi-Exposure Image Fusion. *IEEE Transactions on Image Processing*, 24(11): 3345–3356.

Ma, L.; Ma, T.; Liu, R.; Fan, X.; and Luo, Z. 2022. Toward Fast, Flexible, and Robust Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 5637–5646.

Özdenizci, O.; and Legenstein, R. 2023. Restoring Vision in Adverse Weather Conditions with Patch-Based Denoising Diffusion Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10346–10357.

Panagiotou, S.; and Bosman, A. S. 2023. Denoising diffusion post-processing for low-light image enhancement. *Pattern Recognition*, 156: 110799.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proc. of Computer Vision and Pattern Recognition*, 10684–10695.

Rout, L.; Raoof, N.; Daras, G.; Caramanis, C.; Dimakis, A.; and Shakkottai, S. 2023. Solving Linear Inverse Problems Provably via Posterior Sampling with Latent Diffusion Models. In *Adv. of Neural Information Processing Systems*.

Shi, Y.; Liu, D.; Zhang, L.; Tian, Y.; Xia, X.; and Fu, X. 2024. ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images. In *Proc. of Computer Vision and Pattern Recognition*, 3015–3024.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *Proc. of International Conference on Learning Representations*, 1–20.

Song, Y.; and Ermon, S. 2020. Generative Modeling by Estimating Gradients of the Data Distribution. In *Adv. of Neural Information Processing Systems*, volume 32, 1–13.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *Proc. of International Conference on Learning Representations*, 1–12.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019. Underexposed Photo Enhancement Using Deep Illumination Estimation. In *Proc. of Computer Vision and Pattern Recognition*, 6849–6857.

Wang, S.; Zheng, J.; Hu, H.-M.; and Li, B. 2013. Naturalness Preserved Enhancement Algorithm for Non-Uniform Illumination Images. *IEEE Transactions on Image Processing*, 22(9): 3538–3548.

Wang, T.; Zhang, K.; Shao, Z.; Luo, W.; Stenger, B.; Kim, T.-K.; Liu, W.; and Li, H. 2023a. LLDiffusion: Learning Degradation Representations in Diffusion Models for Low-Light Image Enhancement. *2023, arXiv:2307.14659*, 1–16.

Wang, W.; Yang, H.; Fu, J.; and Liu, J. 2024. Zero-Reference Low-Light Enhancement via Physical Quadruple Priors. In *Proc. of Computer Vision and Pattern Recognition*, 26057–26066.

Wang, Y.; Wan, R.; Yang, W.; Li, H.; Chau, L.-P.; and Kot, A. 2022. Low-Light Image Enhancement with Normalizing Flow. In *Proc. of AAAI Conference on Artificial Intelligence*, volume 36, 2604–2612.

Wang, Y.; Yu, J.; Yu, R.; and Zhang, J. 2023b. Unlimited-Size Diffusion Restoration. In *Proc. of Computer Vision and Pattern Recognition*, 1160–1167.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *Proc. of British Machine Vision Conference*, 103948.

Wu, W.; Weng, J.; Zhang, P.; Wang, X.; Yang, W.; and Jiang, J. 2022. Uretinex-Net: Retinex-Based Deep Unfolding Network for Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 5901–5910.

Xu, K.; Yang, X.; Yin, B.; and Lau, R. W. H. 2020. Learning to Restore Low-Light Images via Decomposition-and-Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 2281–2290.

Xu, X.; Wang, R.; Fu, C.-W.; and Jia, J. 2022. SNR-Aware Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 17714–17724.

Yang, S.; Ding, M.; Wu, Y.; Li, Z.; and Zhang, J. 2023. Implicit Neural Representation for Cooperative Low-light Image Enhancement. In *Proc. of International Conference on Computer Vision*, 12918–12927.

Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; and Liu, J. 2020. From Fidelity to Perceptual Quality: A Semi-Supervised Approach for Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 3063–3072.

Yang, W.; Wang, W.; Huang, H.; Wang, S.; and Liu, J. 2021. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30: 2072–2086.

Yin, Y.; Xu, D.; Tan, C.; Liu, P.; Zhao, Y.; and Wei, Y. 2023. CLE Diffusion: Controllable Light Enhancement Diffusion Model. In *Proc. of ACM International Conference on Multimedia*, 8145–8156.

Zhang, L.; Zhang, L.; Liu, X.; Shen, Y.; Zhang, S.; and Zhao, S. 2019. Zero-shot Restoration of Back-Lit Images Using Deep Internal Learning. In *Proc. of ACM International Conference on Multimedia*, 1623–1631.

Zhang, Y.; Guo, X.; Ma, J.; Liu, W.; and Zhang, J. 2021. Beyond Brightening Low-Light images. *International Journal of Computer Vision*, 129: 1013–1037.

Zhang, Y.; Zhang, J.; and Guo, X. 2019. Kindling the Darkness: A Practical Low-Light Image Enhancer. In *Proc. of ACM International Conference on Multimedia*, 1632–1640.

Zhou, C.; Teng, M.; Han, J.; Liang, J.; Xu, C.; Cao, G.; and Shi, B. 2023. Deblurring Low-Light Images with Events. *International Journal of Computer Vision*, 131(5): 1284–1298.

Zhou, D.; Yang, Z.; and Yang, Y. 2023. Pyramid Diffusion Models for Low-Light Image Enhancement. In *Proc. of International Joint Conference on Artificial Intelligence*, 1795–1803.

Zero-Shot Low-Light Image Enhancement via Latent Diffusion Models Supplementary Material

Yan Huang[1], **Xiaoshan Liao**[1], **Jinxiu Liang**[2,3#], **Yuhui Quan**[1,4], **Boxin Shi**[2,3], **Yong Xu**[1,4]

[1]School of Computer Science and Engineering, South China University of Technology
[2]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[4]Pazhou Lab

{aihuangy, csyhquan, yxu}@scut.edu.cn, {csxsliao}@mail.scut.edu.cn,
{cssherryliang, shiboxin}@pku.edu.cn

Algorithm 1: Zero-Shot Latent Diffusion-based LLIE.

```
Require : $\epsilon_\theta, \mathcal{E}, \mathcal{D}, T, \{\gamma_t\}_{t=1}^T, \{\sigma_t\}_{t=1}^T$ from the pre-trained diffusion models, $\mathcal{G}$, degradation model parameter $\phi$, regularization parameters $\sigma_\mathrm{p}, \sigma_\mathrm{i}, \sigma_\mathrm{l}, \sigma_\mathrm{c}$.
Input   : Low-light input image $Y$.
Output  : Enhanced normal-light image $X_0$.
$L \leftarrow \mathcal{G}(\max_{c\in\{R,G,B\}} Y^c(i,j))$    ▷ Eqs. (3), (4)
$\beta(i) = L(i)^{\exp{(L(i)-\phi)}}, \forall \text{pixel } i$
$\widehat{X} \leftarrow \frac{Y}{L^\beta}$    ▷ Eq. (5)
$Z_T \sim N(0, I)$
$\bar\alpha_0 = 1$
for $t = 1$ to $T$ do
    $\alpha_t \leftarrow 1 - \gamma_t$.
    $\bar\alpha_t \leftarrow \alpha_t \bar\alpha_{t-1}$
end for
for $t = T$ to $1$ do
    $Z_{0|t} \leftarrow \frac{1}{\sqrt{\bar\alpha_t}}(Z_t + (1-\bar\alpha_t)\epsilon_{\theta^*}(Z_t, t))$    ▷ Eq. (13)
    $X_{0|t} \leftarrow \mathcal{D}(Z_{0|t})$
    $z \sim N(0, I)$
    $Z'_{t-1} \leftarrow \frac{\sqrt{\alpha_i(1-\bar\alpha_{t-1})}}{1-\bar\alpha_t} Z_t + \frac{\sqrt{\bar\alpha_{t-1}\gamma_t}}{1-\bar\alpha_t} Z_{0|t} + \tilde\sigma_t z$
    $\mathcal{L}_\text{meas} \leftarrow \|Y - X_{0|t} \odot L^\beta\|_2^2$    ▷ Eqs. (1), (18)
    $\mathcal{L}_\text{inv} \leftarrow \|\widehat{X} - X_{0|t}\|_2^2$    ▷ Eq. (19)
    $\mathcal{L}_\text{latent} \leftarrow \|\mathcal{E}(\mathcal{G}(\widehat{X})) - Z_{0|t}\|_2^2$    ▷ Eq. (20)
    $\mathcal{L}_\text{col} \leftarrow \sum_{\text{pixel } i} \sum_{\text{colors } p,q} \left(X_{0|t,p}(i) - X_{0|t,q}(i)\right)^2$    ▷ Eq. (21)
    $\nabla_{Z_t}\mathcal{L} \leftarrow \frac{1}{\sigma_\mathrm{p}^2}\nabla_{Z_t}\mathcal{L}_\text{meas} + \frac{1}{\sigma_\mathrm{i}^2}\nabla_{Z_t}\mathcal{L}_\text{inv} + \frac{1}{\sigma_\mathrm{l}^2}\nabla_{Z_t}\mathcal{L}_\text{latent} + \frac{1}{\sigma_\mathrm{c}^2}\nabla_{Z_t}\mathcal{L}_\text{col}$    ▷ Eq. (17)
    $s_t \leftarrow \frac{\|Z_t - Z_{t-1}\|_2^2}{(\nabla_{Z_t}\mathcal{L} - \nabla_{Z_{t-1}}\mathcal{L})(Z_t - Z_{t-1})}$.    ▷ Eq. (22)
    $Z_{t-1} \leftarrow Z'_{t-1} - s_t \nabla_{Z_t}\mathcal{L}$
end for
$X_0 \leftarrow \mathcal{D}(Z_0)$
return $X_0$
```

More Implementation Details

Pseudo-codes The pseudo-codes of the proposed zero-shot LLIE framework are shown in Algorithm 1.

Pre-trained LDMs We utilize the Stable Diffusion 1.5 model (Rombach et al. 2022) pre-trained on the LAION-2B dataset. Specifically, we employ the Exponential Moving Average (EMA) weights of the $\epsilon_\theta, \mathcal{E}, \mathcal{D}$. During testing, the weights of the pre-trained model remain frozen. The noise schedule follows a linear rule, with γ values linearly spaced between $\gamma_\text{min} = 0.0001$ and $\gamma_\text{max} = 0.02$. The classifier-free guidance scale is set to 0.0.

Regularization term The regularization term incorporates parameters $\{\sigma_\mathrm{p}^2, \sigma_\mathrm{i}^2, \sigma_\mathrm{l}^2, \sigma_\mathrm{c}^2\}$, which are weighted to adapt to the characteristics of different datasets. σ_p controls the fidelity to the given low-light image. However, as the contrast of the low-light image is very low, the error tends to be small, and can not well regularize the fidelity. σ_i emphasizes the error in darker regions, benefiting retain details from the input, however, it also emphasizes the noise. σ_l ensures the injectivity of the estimated latent between pixel space and latent space, however, it is very sensitive to the noise. σ_c adjusts the weight of the color constancy constraint. These adjustments balance the influence of the different regularization terms and prevent excessive noise or blurriness in the enhanced images. The default values are set to $\{1.0, 1.0, 1.0, 0.1\}$. The LOL-v2 dataset exhibits lower low-light severity and noise levels compared to LOL-v1. The adjusted parameter set is: $\{\sigma_\mathrm{p}^2, \sigma_\mathrm{i}^2, \sigma_\mathrm{l}^2, \sigma_\mathrm{c}^2\} = \{10.0, 2.0, 1.0, 0.1\}$. For real-world unpaired test sets characterized by uneven lighting, we increase the parameters to mitigate overexposure. The adjusted parameter set is: $\{\sigma_\mathrm{p}^2, \sigma_\mathrm{i}^2, \sigma_\mathrm{l}^2, \sigma_\mathrm{c}^2\} = \{20.0, 15.0, 10.0, 0.1\}$. The low-frequency extraction operation $\mathcal{G}$ within the $\mathcal{L}_\text{latent}$ regularization term in Eq. (20) uses a kernel size of 5×5 with standard deviations 1.5.

Limitations

While our proposed method achieves significant improvements for zero-shot LLIE, there are some limitations.

Handling extreme low-light regions Our method faces

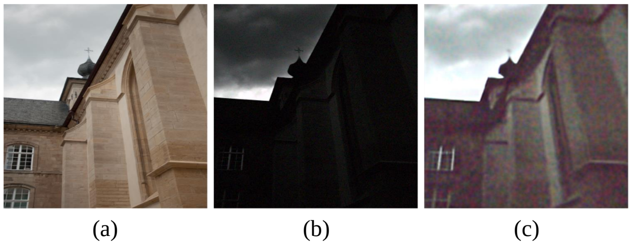


Figure 8: A failure case on the LOL-v2-sync dataset. (a) Ground truth, (b) Input, (c) Our result. While avoiding overexposure in bright regions, our method struggles to enhance extremely dark areas or completely suppress noise patterns.

Table 4: Computational complexity of different methods in terms of FLOPs (G) and number of parameters (M).

Type	Method	PSNR↑	FLOPs (G)	#Param. (M)
SL	DiffLL	21.84	102.60	702.15
SL	GSAD	22.96	86.91	37.45
UL	EnlightenGAN	17.48	114.35	67.80
UL	CLIP-LIT	12.39	110.98	428.00
ZS	GDP	15.83	209.17	763.00
ZS	Ours	19.82	106.93	859.00

challenges in extreme low-light regions where the image contains very little visible information. In such cases, the enhanced images may still suffer from low contrast and residual noise. Figure 8 shows a failure case. Future work can explore incorporating additional priors or leveraging more source signals to handle extreme low-light scenarios.

Computational overhead While the computational overhead of pre-trained LDMs is justified by their superior enhancement quality, it presents practical constraints for real-time applications. Table 4 shows comparisons of FLOPs (G), and the number of parameters (M) of different methods on images of 256×256 resolution. It is noted that compared to GDP, another method based on pre-trained models, our GFLOPs are lower, and we achieve better qualitative and quantitative results. While our method introduces some computational overhead due to the iterative sampling process, it also provides significant flexibility and adaptability in a zero-shot setting. It can accommodate alternative diffusion models of varying complexity—whether lightweight or high-performance models—based on application needs, thus enabling scalability without compromising zero-shot capabilities. For example, lower-parameter models can be substituted for resource-sensitive tasks, while higher-parameter models can be applied where higher-quality output is required. Future research could explore model compression techniques or efficient sampling strategies to reduce computational requirements while maintaining enhancement performance.

More Experimental Results

Effect of adaptive guidance scale To validate the effectiveness of the proposed adaptive guidance scale, we conducted a quantitative comparison between with and without it on the LOL-v1 dataset, as shown in Table 5. Incorporating the adaptive guidance scale shows improvements in PSNR, SSIM, and LPIPS, supporting its contribution to performance. While the quantitative improvements are moderate, the qualitative benefits are more evident as shown in Figure 6 of the main text. It enhances local details, such as the sharper edges in the red and green boxes, illustrating the effectiveness of managing fine details within local regions. We also compare the performance of the proposed adaptive guidance scale and those with constant guidance scales of 1.0 and 10.0. The quantitative results of intermediate normal-light image estimations, denoted as $\mathcal{D}(Z_{0|t})$, are evaluated at different timesteps from $t = 1000$ to $t = 0$. Figure 9 shows that the adaptive guidance scale exhibits faster convergence compared to the constant guidance scales. This dynamic adjustment based on image content leads to better denoising performance and cleaner, enhanced images at timestep $t = 0$.

Table 5: Analysis of the adaptive guidance scale (AGS).

Setting	PSNR↑	SSIM↑	LPIPS↓
w/o AGS	19.01	0.762	0.489
w/ AGS	19.15	0.767	0.452

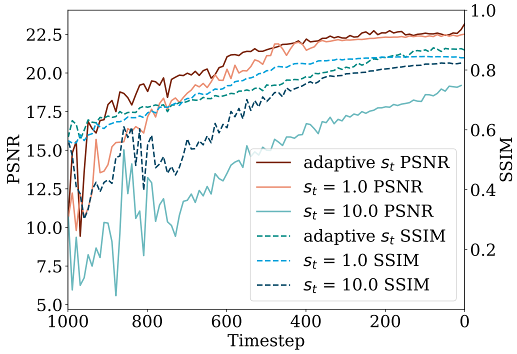


Figure 9: PSNR and SSIM values for intermediate normal-light image estimations $\mathcal{D}(Z_{0|t})$ at different timesteps (from $t = 1000$ to $t = 0$) using various strategies of guidance scale s_t during iterative denoising.

Effect of channel consistency loss $\mathcal{L}_{\text{col}}$ To address potential color shifts, we introduce a channel consistency loss $\mathcal{L}_{\text{col}}$. While $\mathcal{L}_{\text{col}}$ shows minimal impact on conventional quantitative metrics (PSNR, SSIM), its effects are particularly notable in preserving white balance and preventing color over-saturation. We conduct ablation studies to analyze its effectiveness across various lighting conditions and scene types. As shown in Figure 10, the channel consistency loss significantly reduces color distortions, especially in challenging scenarios with extreme illumination variations.

Effect of the brightness channel L We evaluate the effectiveness of our bright channel-based degradation model in Figure 11. The heatmap visualizations reveal that our method achieves significantly closer alignment with ground truth illumination patterns, particularly in preserving local contrast variations and avoiding over-saturation artifacts com-

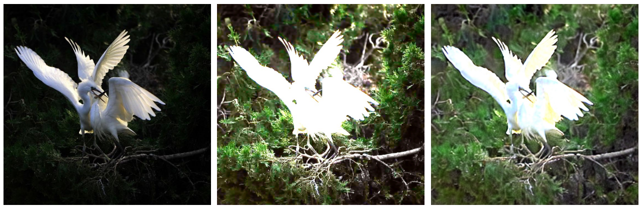


Figure 10: Analysis of channel consistency loss: (a) Input, (b) w/o $\mathcal{L}_{\text{col}}$, (c) w/ $\mathcal{L}_{\text{col}}$. The improved color rendition demonstrates its effectiveness.

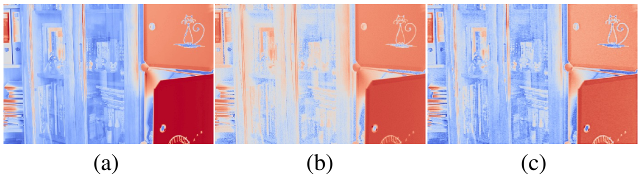


Figure 11: Analysis of bright channel-based degradation model, which shows heatmap visualizations of estimated illumination maps using (a) ground truth from paired data, (b) gamma correction ($\gamma = 0.7$), and (c) our proposed bright channel-based model.

mon in gamma correction.

Effect of adaptive exposure β We conduct an ablation study to investigate the effectiveness of the adaptive exposure β in modeling the degradation process. Different constant β values (0.4, 0.5, 0.6, 0.7, 0.8, and 0.9) are compared against the proposed adaptive approach. Figure 12 presents the qualitative results. When β is less than 0.8, image quality improves as β increases. However, when β exceeds 0.8, the image quality declines, and overexposure becomes apparent. The adaptive β consistently produces superior results, preserving the local structure, contrast, and overall visual quality. Quantitative results are provided in Table 6. The adaptive β significantly outperforms the constant β settings across all metrics. This highlights the efficacy of the adaptive approach in leveraging local information from the low-light input images to preserve structure and contrast.

Effect of low-frequency extraction in $\mathcal{L}_{\text{latent}}$ We examine the effect of the low-frequency extraction operation $\mathcal{G}$ within the $\mathcal{L}_{\text{latent}}$ regularization term in Eq. (20). $\mathcal{G}$ aids in removing noise while preserving texture in the latent space. We experiment with different kernel sizes for $\mathcal{G}$: 3×3, 5×5, 7×7, and 9×9. Figure 13 demonstrates that the 5×5 kernel size we used achieves the best balance between noise suppression and texture preservation across various input image sizes.

Analysis to input spatial resolution The impact of input image resolution on the quality of the enhanced outputs is analyzed with and without low-frequency extraction. Figure 14 shows that resizing the input image to 256×256 results in reduced texture clarity and noise removal compared to processing at the original 512×512 resolution. This highlights the importance of processing images at appropriate resolutions to fully leverage the pre-trained model's capabilities. The results with or without the low-frequency extraction for different input sizes are also shown, which demonstrates that it is more effective when the input size is larger.

Table 6: Quantitative comparison of the proposed adaptive exposure β against constant exposure mask values.

Metric	PSNR↑	SSIM↑	LPIPS↓
$\beta = 0.4$	10.32	0.556	0.585
$\beta = 0.5$	11.76	0.624	0.547
$\beta = 0.6$	13.68	0.671	0.535
$\beta = 0.7$	16.17	0.725	0.515
$\beta = 0.8$	17.57	0.750	0.496
$\beta = 0.9$	16.88	0.734	0.515
Adaptive β	19.15	0.767	0.452

More qualitative comparison Figure 15 show another visual comparison in LOL-v2 dataset (Yang et al. 2020). We provide visual comparisons on datasets DICM (Lee, Lee, and Kim 2013), MEF (Ma, Zeng, and Wang 2015), NPE (Wang et al. 2013), and LIME (Guo, Li, and Ling 2017) that are without ground truth normal-light images: Figure 16 for DICM, Figs. Figures 17 and 18 for MEF, Figure 19 for NPE, Figs. Figures 20 and 21 for LIME. These datasets encompass a wide spectrum of challenging scenarios, including varying illumination conditions, scene compositions, and object complexities. These results demonstrate the robustness and generalizability of our proposed LLIE method across diverse low-light conditions and image characteristics. To further validate real-world applicability, we perform comparative analysis on challenging scenes from the MEF (Ma, Zeng, and Wang 2015) and NPE (Lee, Lee, and Kim 2013), as shown in Figure 22. While existing approaches often introduce artifacts or inappropriate exposure adjustments in homogeneous regions when attempting to enhance structural details, our method maintains a natural appearance while preserving both global illumination consistency and local detail fidelity.

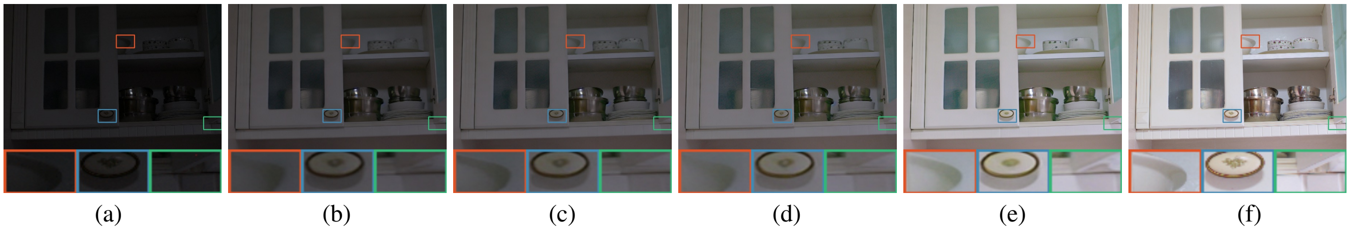


Figure 12: Qualitative comparison of adaptive exposure (β) against constant exposure masks on the LOL-v1 dataset [wei2018deep]. (a) Input, (b) $\beta = 0.4$, (c) $\beta = 0.5$, (d) $\beta = 0.6$, (e) Adaptive β, (f) Ground truth. Increasing the fixed β value primarily improves brightness without significantly enhancing contrast. Adaptive β shows significant improvements in both brightness and contrast, evident in the clearer bowl edges (red box) and the enhanced light-dark contrast at the cabinet corner (green box).

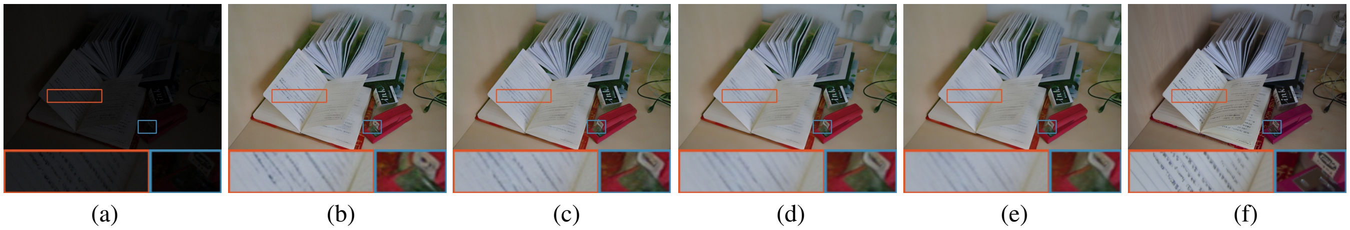


Figure 13: Ablation study on the impact of different kernel sizes for the low-frequency extraction operation ($\mathcal{G}$) within the regularization term $\mathcal{L}_{\text{latent}}$, evaluated on the LOL-v1 dataset (Wei et al. 2018). (a) Input, (b) $\mathcal{G}$ size = 3×3, (c) $\mathcal{G}$ size = 5×5, (d) $\mathcal{G}$ size = 7×7, (e) $\mathcal{G}$ size = 9×9, (f) Ground truth. Larger $\mathcal{G}$ sizes effectively remove noise but introduce unwanted blurring of details. As our method utilizes a pseudo-normalized light image with rich texture details as guidance, the Gaussian blur applied in the latent space aims to primarily remove noise while preserving essential low-frequency information. Kernel sizes of $\mathcal{G} = 5 \times 5$ or $\mathcal{G} = 7 \times 7$ achieve the best balance between noise removal and detail preservation.

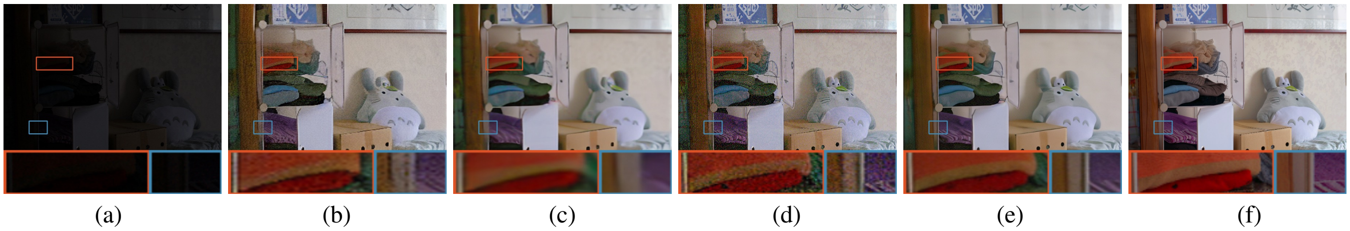
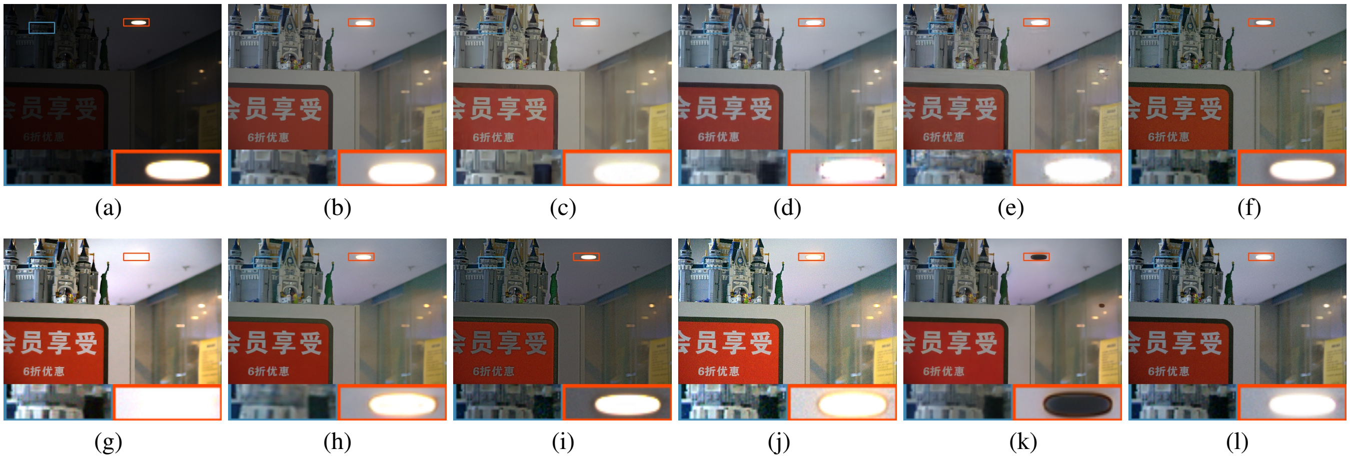


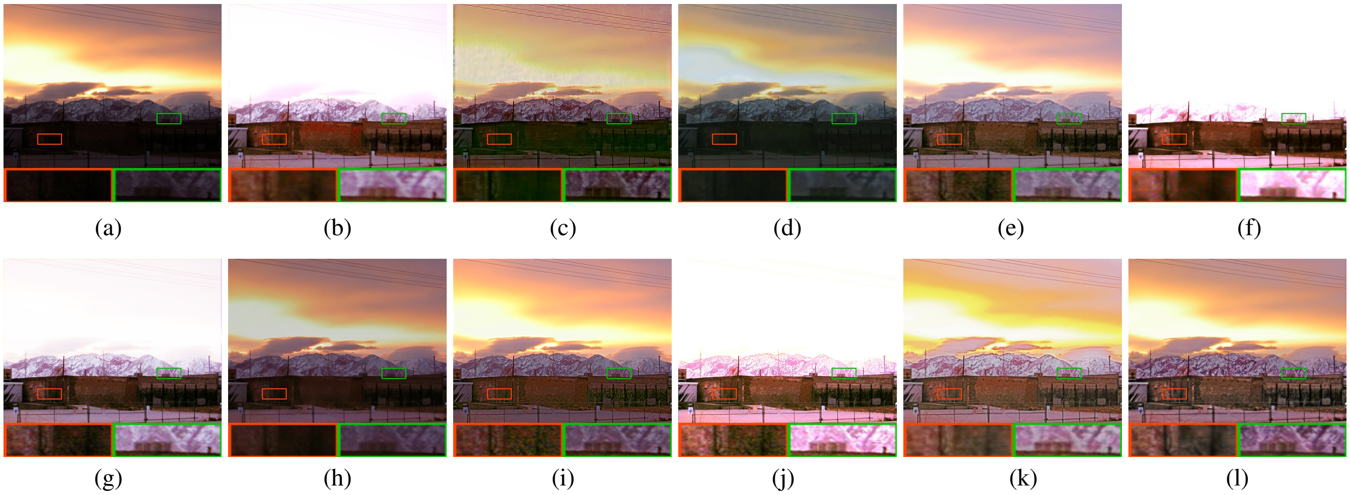
Figure 14: Visual comparison of different input image sizes and the effect of using $\mathcal{G}$ within the regularization term $\mathcal{L}_{\text{latent}}$ on enhancement quality on the LOL-v1 datasett (Wei et al. 2018). (a) Input, (b) w/o $\mathcal{G}$ on 256×256, (c) w/ $\mathcal{G}$ on 256×256, (d) w/o $\mathcal{G}$ on 512×512, (e) w/ $\mathcal{G}$ on 512×512, (f) Ground truth. (256×256 and 512×512 represent the pixel processing sizes of pre-trained models). The comparison between (b) and (c), and between (d) and (e) demonstrates the denoising benefit of incorporating the guidance term in the latent space. Furthermore, comparing (c) and (d) highlights the positive correlation between input image size and the sampling quality of Stable Diffusion; larger input images lead to higher sampling quality.



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

Figure 15: Comparison of enhancement results on the LOL-v2 dataset (Yang et al. 2020). (a) Input, (b) Ground truth, (c) GSAD (Hou et al. 2024), (d) DiffLL (Jiang et al. 2023), (e) NeRCo (Yang et al. 2023), (f) Zero-DCE (Guo et al. 2020), (g) RUAS (Liu et al. 2021b), (h) SCI (Ma et al. 2022), (i) GDP (Fei et al. 2023), (j) ZeroIG (Shi et al. 2024), (k) QuadPrior (Wang et al. 2024), (l) Ours. Our approach excels in reconstructing fine textures (e.g., the castle toy in the blue box) and mitigating noise in challenging areas (red box). Specifically, our method outperforms both supervised and unsupervised techniques in preserving intricate line details and effectively removing noise while avoiding overexposure and artifacts often present in alternative approaches.



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

Figure 16: Comparison results on real data in the DICM dataset (Lee, Lee, and Kim 2013). (a) Input, (b) KinDPlus (Zhang et al. 2021), (c) DiffLL (Jiang et al. 2023), (d) NeRCo (Yang et al. 2023), (e) Zero-DCE (Guo et al. 2020), (f) RUAS (Liu et al. 2021b), (g) SCI (Ma et al. 2022), (h) QuadPrior (Wang et al. 2024), (i) CLIP-LIT (Liang et al. 2023b), (j) ZeroIG (Shi et al. 2024), (k) PairLIE (Fu et al. 2023), (l) Ours. Our proposed method effectively addresses the overexposure and color shift issues commonly observed in some supervised and zero-shot approaches, resulting in enhanced visual quality.

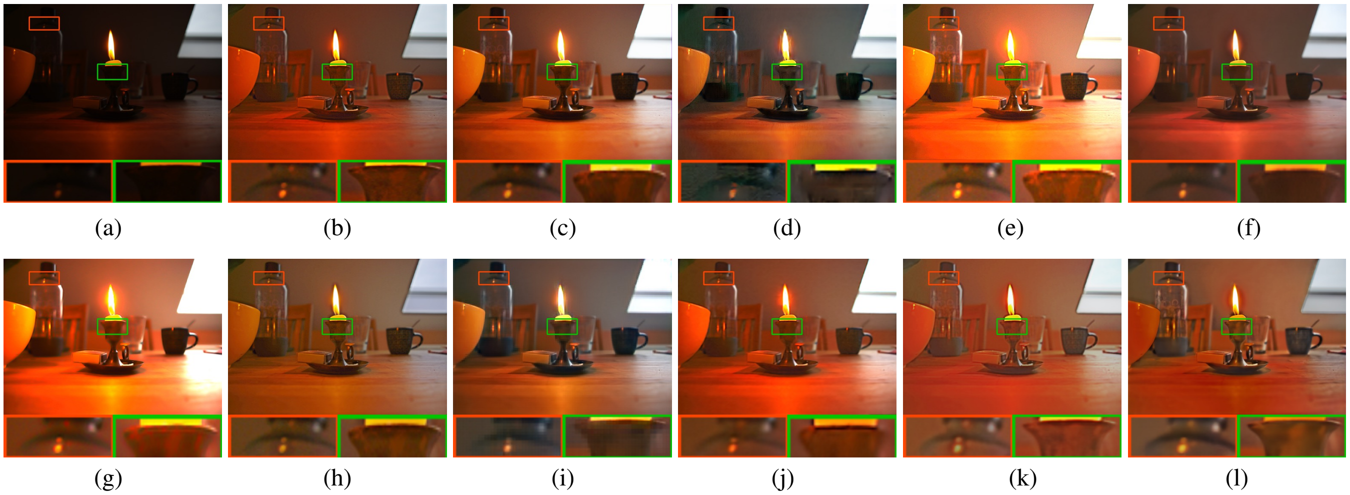


Figure 17: Comparison results on real data in the MEF dataset (Ma, Zeng, and Wang 2015). (a) Input, (b) KinDPlus (Zhang et al. 2021), (c) DiffLL (Jiang et al. 2023), (d) NeRCo (Yang et al. 2023), (e) Zero-DCE (Guo et al. 2020), (f) RUAS (Liu et al. 2021b), (g) SCI (Ma et al. 2022), (h) QuadPrior (Wang et al. 2024), (i) CLIP-LIT (Liang et al. 2023b), (j) ZeroIG (Shi et al. 2024), (k) PairLIE (Fu et al. 2023), (l) Ours.

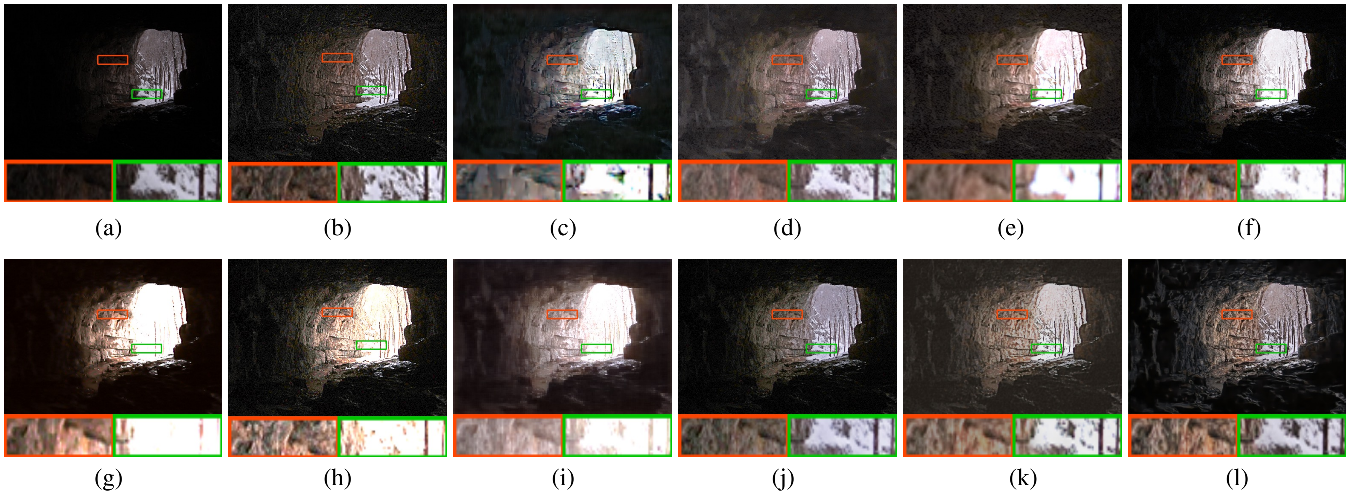


Figure 18: Comparison results on real data in the MEF dataset (Ma, Zeng, and Wang 2015). (a) Input, (b) KinDPlus (Zhang et al. 2021), (c) DiffLL (Jiang et al. 2023), (d) NeRCo (Yang et al. 2023), (e) Zero-DCE (Guo et al. 2020), (f) RUAS (Liu et al. 2021b), (g) SCI (Ma et al. 2022), (h) QuadPrior (Wang et al. 2024), (i) CLIP-LIT (Liang et al. 2023b), (j) ZeroIG (Shi et al. 2024), (k) PairLIE (Fu et al. 2023), (l) Ours.

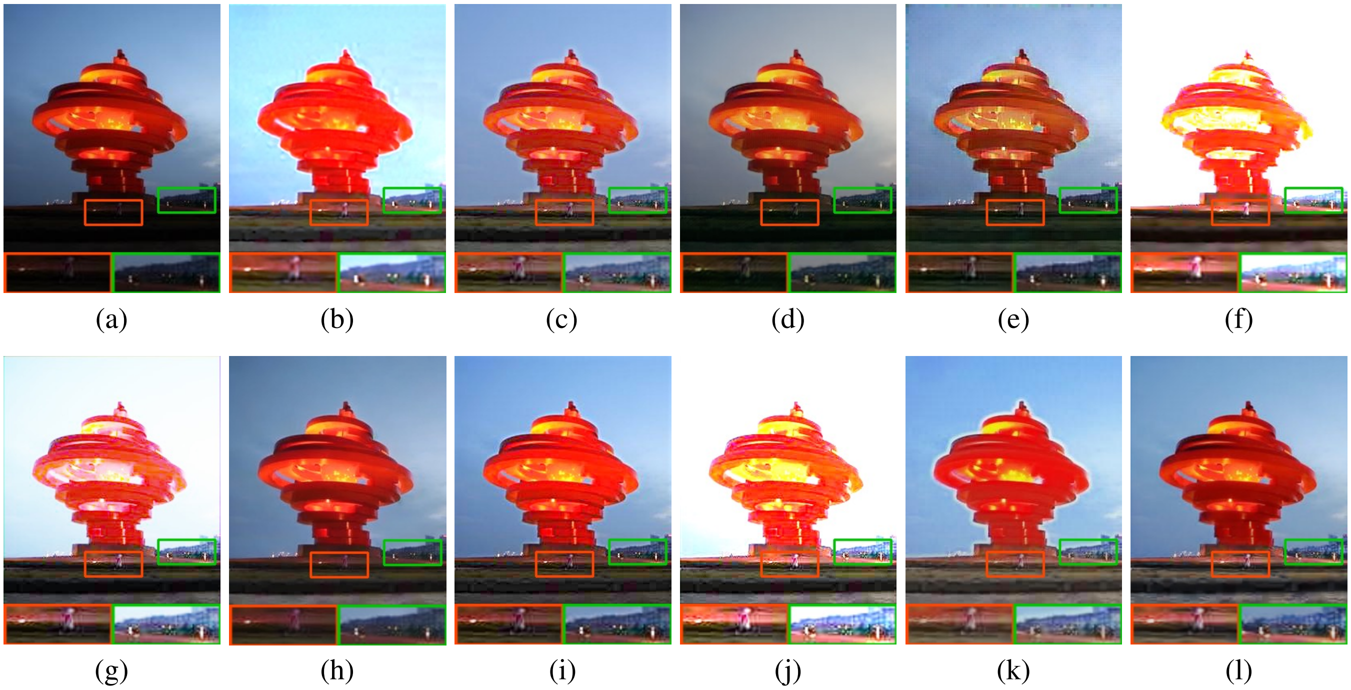


Figure 19: Comparison results on real data in the NPE dataset (Wang et al. 2013). (a) Input, (b) KinDPlus (Zhang et al. 2021), (c) DiffLL (Jiang et al. 2023), (d) NeRCo (Yang et al. 2023), (e) Zero-DCE (Guo et al. 2020), (f) RUAS (Liu et al. 2021b), (g) SCI (Ma et al. 2022), (h) QuadPrior (Wang et al. 2024), (i) CLIP-LIT (Liang et al. 2023b), (j) ZeroIG (Shi et al. 2024), (k) PairLIE (Fu et al. 2023), (l) Ours. Our method effectively enhances image quality while avoiding overexposure and color shifts, demonstrating superior noise removal capabilities compared to other techniques.

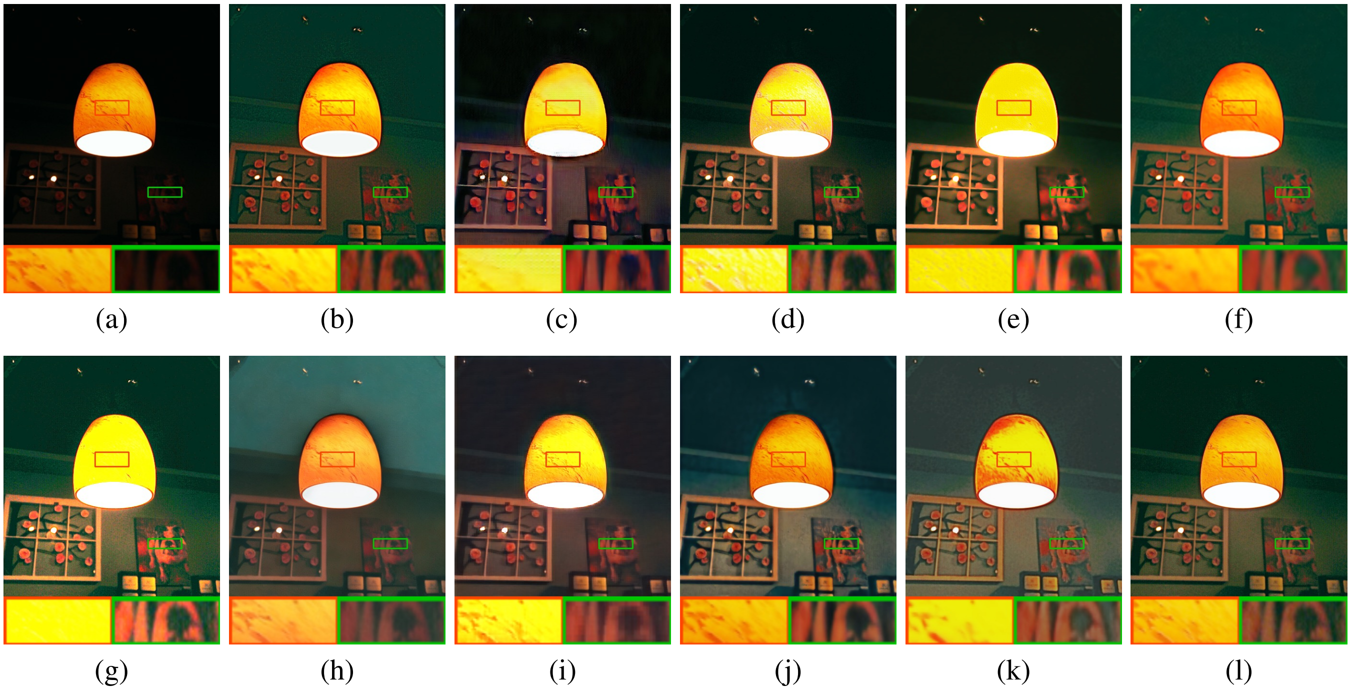


Figure 20: Comparison results on real data in the LIME dataset (Guo, Li, and Ling 2017). (a) Input, (b) KinDPlus (Zhang et al. 2021), (c) DiffLL (Jiang et al. 2023), (d) NeRCo (Yang et al. 2023), (e) Zero-DCE (Guo et al. 2020), (f) RUAS (Liu et al. 2021b), (g) SCI (Ma et al. 2022), (h) QuadPrior (Wang et al. 2024), (i) CLIP-LIT (Liang et al. 2023b), (j) ZeroIG (Shi et al. 2024), (k) PairLIE (Fu et al. 2023), (l) Ours.

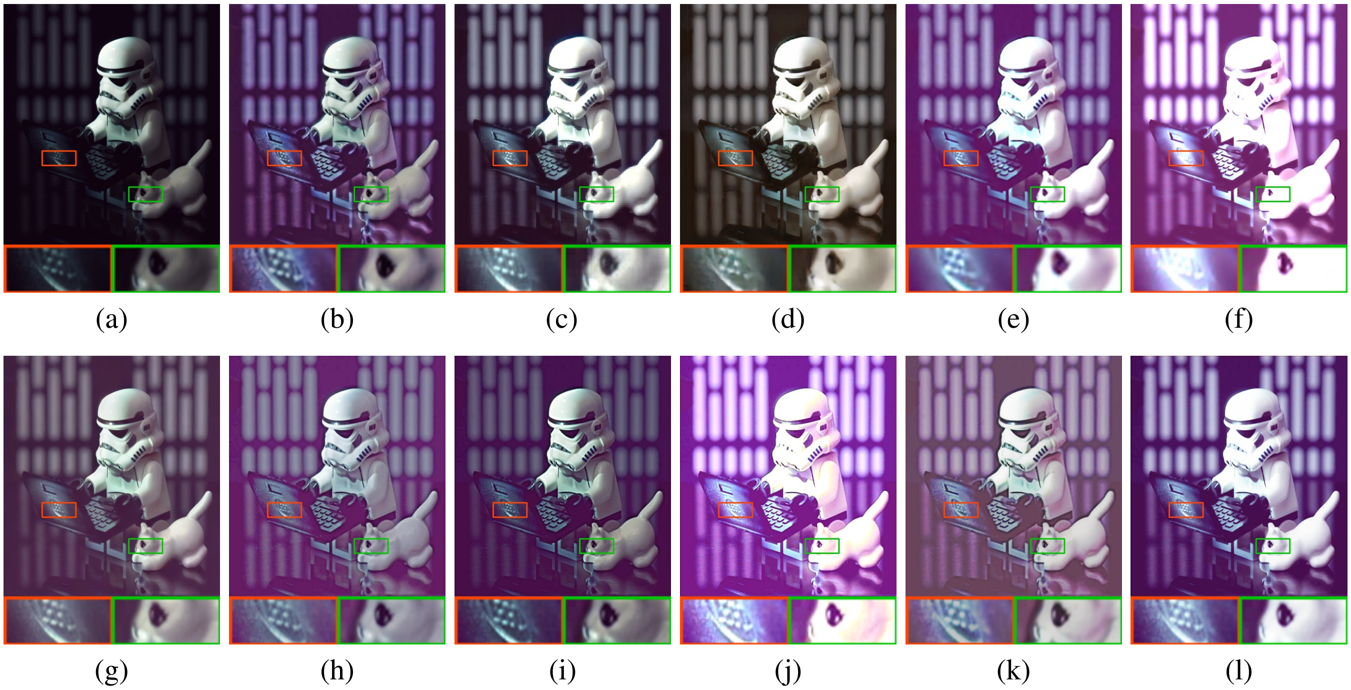


Figure 21: Comparison results on real data in the LIME dataset (Guo, Li, and Ling 2017). (a) Input, (b) KinDPlus (Zhang et al. 2021), (c) DiffLL (Jiang et al. 2023), (d) NeRCo (Yang et al. 2023), (e) Zero-DCE (Guo et al. 2020), (f) RUAS (Liu et al. 2021b), (g) SCI (Ma et al. 2022), (h) QuadPrior (Wang et al. 2024), (i) CLIP-LIT (Liang et al. 2023b), (j) ZeroIG (Shi et al. 2024), (k) PairLIE (Fu et al. 2023), (l) Ours.

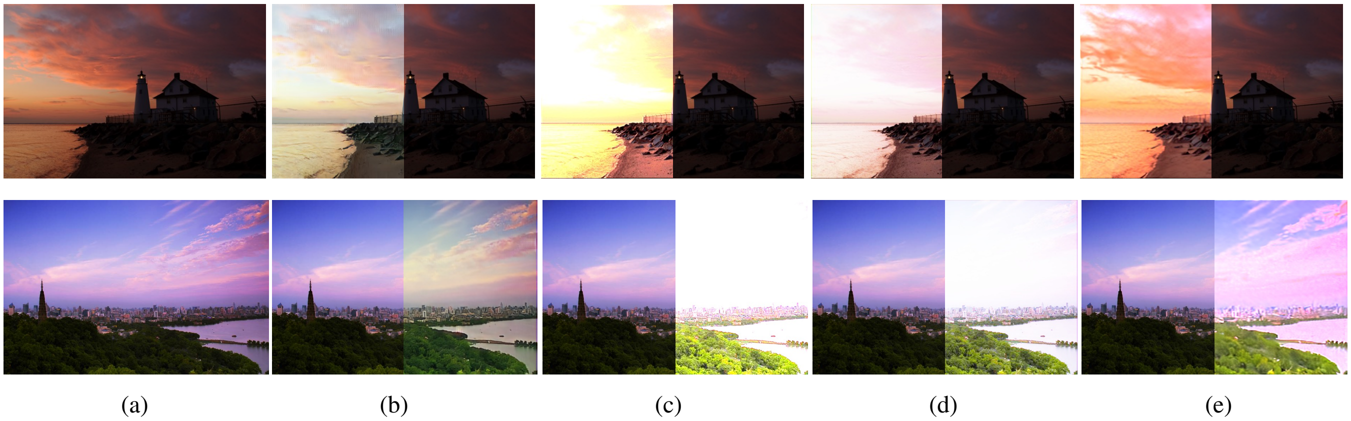


Figure 22: Comparison results between different unsupervised LLIE methods on challenging real-world scenes from MEF (Ma, Zeng, and Wang 2015) (top row) and NPE (Lee, Lee, and Kim 2013) (bottom row) datasets. (a) Input, (b) NeRCo (Yang et al. 2023), (c) SCI (Ma et al. 2022), (d) ZeroIG (Shi et al. 2024), (e) Ours. For each scene, we present split-view comparisons where one half shows the original input and the other half shows the enhanced result.