L^{2}DM: A Diffusion Model for Low-Light Image Enhancement

Xingguo Lv[1], Xingbo Dong[2(✉)], Zhe Jin[2], Hui Zhang[1], Siyi Song[1], and Xuejun Li[1]

[1] Anhui Provincial International Joint Research Center for Advanced Technology in Medical Imaging, School of Computer Science and Technology, Anhui University, Hefei 230093, China

[2] Anhui Provincial Key Laboratory of Secure Artificial Intelligence, School of Artificial Intelligence, Anhui University, Hefei 230093, China

xingbo.dong@ahu.edu.cn

Abstract. Low-light image enhancement is a challenging yet beneficial task in computer vision that aims to improve the quality of images captured under poor illumination conditions. It involves addressing difficulties such as color distortions and noise, which often degrade the visual fidelity of low-light images. Although tremendous CNN-based and ViT-based approaches have been proposed, the potential of diffusion models in this domain remains unexplored. This paper presents L^{2}DM, a novel framework for low-light image enhancement using diffusion models. Since L^{2}DM falls into the category of latent diffusion models, it can reduce computational requirements through denoising and the diffusion process in latent space. Conditioning inputs are essential for guiding the enhancement process, therefore, a new ViT-based network called ViTCondNet is introduced to efficiently incorporate conditioning low-light inputs into the image generation pipeline. Extensive experiments on benchmark LOL datasets demonstrate L^{2}DM's state-of-the-art performance compared to diffusion-based counterparts. The L^{2}DM source code is available on GitHub for reproducibility and further research.

Keywords: Computational photography · Low-light image enhancement · Diffusion models · Latent diffusion models

1 Introduction

Computational photography is a field that merges digital image processing and computer vision methods to enhance and manipulate photographs. One crucial aspect of computational photography is low-light image enhancement. When capturing digital images in conditions of poor illumination, such as indoors, at night, or with improper camera exposure settings, issues like color distortions and noise can arise, diminishing the overall quality of the image. To enhance low-light images, tremendous Low-light Image Enhancement (LLIE) solutions

Q. Liu et al. (Eds.): PRCV 2023, LNCS 14435, pp. 130–145, 2024.
https://doi.org/10.1007/978-981-99-8552-4_11

have been proposed, which can be generally classified into traditional approaches, convolutional neural network (CNN) approaches and vision transformer (ViT) approaches.

As a traditional approach, histogram equalization (HE) techniques [1,23,49] aim to rearrange pixel values in order to achieve a uniform distribution. Methods based on dehaze models [33] adjust pixel values to conform to a natural distribution. Retinex-theory-based approaches [10,43,47,65] utilize models that assume an image can be decomposed into illumination and reflectance components. These methods are generally regarded as traditional approaches.

In the last decade, Convolutional Neural Networks (CNNs) have achieved impressive progress in low-level image processing applications [11,20,50,68]. For instance, Chen et al. [3] employed a U-Net [35] to process RAW sensor data and generate high-perceptual-quality RGB images. Xu et al. [50] proposed a pipeline for low-light image enhancement that incorporates a frequency-based decomposition and enhancement model. Similarly, Dong et al. [7] introduced a CNN-based network based on the U-Net architecture, aiming to virtually eliminate the color filter and achieve improved image processing performance.

In recent years, Transformers have demonstrated significant advantages compared to CNN-based approaches in the field of low-level vision tasks. This is primarily due to their ability to leverage spatial and channel-wise attention mechanisms, as originally introduced in [41]. One example is the framework proposed by Xu et al. in [53], which exploits the Signal-to-Noise Ratio (SNR) as prior information to guide the process of feature fusion. By incorporating a novel self-attention model, this SNR-aware transformer effectively avoids incorporating tokens from image regions with significant noise or very low SNR values. Consequently, it dynamically enhances pixels using spatial-varying operations, resulting in state-of-the-art performance in low-level vision tasks.

Recently, diffusion models [37] have gained significant attention as a type of probabilistic generative model capable of generating high-resolution images with a wide range of diversity [6]. These models have shown great promise in image reconstruction and offer easier training and improved sample quality compared to traditional generative models and variational autoencoders (VAEs). However, due to the nature of probabilistic generative models, which aim to learn a representative distribution rather than a deterministic solution like the combination of U-Net and L2 loss, they often exhibit inferior performance on traditional metrics such as PSNR/SSIM for image enhancement tasks. Consequently, the use of diffusion models for low-light image enhancement remains relatively unexplored.

Motivated by the preceding discussion, we present a diffusion model for low-light image enhancement, namely L^{2}DM (Low-Light Diffusion Model). L^{2}DM draws inspiration from latent diffusion models (LDMs) [34], which leverage denoising and diffusion models in the latent space of autoencoders to enhance image synthesis and substantially decrease computational demands. To address the task of low-light image enhancement and mitigate potential degradation in peak signal-to-noise ratio (PSNR) performance, we further introduce a novel

ViT-based network, namely ViTCondNet, to handle low-light images conditioning inputs. The contributions of our work can be summarized as follows:

1. We proposed a Diffusion Model enabled framework (i.e. L^2DM) for low-light image enhancement. The proposed framework attempts latent diffusion in low-light image enhancement task, which results in the reduction of the computational overhead, particularly beneficial to the higher dimensional data. Simultaneously offering more faithful and detailed reconstructions.
2. We introduced a ViT-based network, namely ViTCondNet to effectively handles conditioning inputs. ViTCondNet manages to achieve a decent low-light image enhancement performance and improves poor PSNR performance while allowing for the synthesis of normal light images.
3. We conducted the extensive experiments on publicly available benchmark datasets and the results demonstrate that L^2DM exhibits state-of-the-art performance compared to diffusion-based counterparts. The PSNR on the LOL dataset reaches 24.54, while on the LOLV2-real and synthetic subsets, it achieves 24.80 and 23.96, respectively. The source code of our work is available at github.com/Yore0/L2DM.

2 Related Work

In this section, we primarily concentrate on data-driven methods that employ convolutional neural networks (CNN) and vision transformers (ViT) to enhance low-light images.

With the advancement of deep learning technology, CNN-based approaches have been at the forefront, where researchers have leveraged the hierarchical structure of convolutional layers to effectively extract pertinent features from low-light images. In the Zero-DCE method [11], a lightweight CNN network is proposed to estimate pixel-wise and high-order curves for image enhancement. Zero-DCE achieves efficient and effective enhancement across various lighting conditions without the need for reference images. It outperforms state-of-the-art methods and shows potential applications in dark face detection. However, compared to supervised approaches, Zero-DCE exhibits a performance gap. In [30], DeepLPF is introduced as a deep neural network approach for automatic image enhancement. It uses spatially local filters (Elliptical, Graduated, Polynomial) and achieves superior results on benchmark datasets with fewer parameters. In [55], DRBN is proposed for low-light image enhancement, using paired low/normal-light images and adversarial learning. DSLR [25] utilizes Laplacian pyramid for global and local enhancement. EnlighenGAN [17] enhances low-light images using unsupervised GANs without paired training data.

In a recent work by Xu et al. [54], a novel framework is introduced for enhancing low-light images by simultaneously addressing their appearance and structure. This approach incorporates edge detection to effectively model the image structure and employs a structure-guided enhancement module to enhance the overall image quality. In another study by Wang et al. [45], a normalizing flow

model called LLFlow is proposed to tackle the challenging task of enhancing low-light images. By modeling the one-to-many relationship between low-light and normally exposed images, LLFlow achieves notable improvements in terms of brightness enhancement, noise reduction, artifact removal, and color enhancement.

The CNN-based models generally produce visually satisfactory enhancement results in most scenarios. However, when it comes to global modeling, the ViT architecture outperforms CNNs by focusing on the entire image rather than just local regions. In a study by Xu et al. [53], the authors exploit the Signal-to-Noise Ratio (SNR) as prior information to guide the feature fusion process. As a result, it dynamically enhances pixels using spatial-varying operations, leading to state-of-the-art performance in low-level vision tasks. In another work by Yuan et al. [59], an end-to-end low-light image enhancement network combining transformers and CNNs is proposed. This network accurately captures both global and local features to restore normal light images, resulting in improved brightness, reduced noise, and preserved texture and color information. Similarly, in a study by Jiang et al. [16], a Stage-Transformer-Guided Network (STGNet) is introduced to effectively enhance low-light images by addressing region-specific distortions. It employs a multi-stage approach with efficient transformers that capture degradation distributions at different scales and orientations. Learnable degradation queries are used for adaptive feature selection. Additionally, a histogram loss and other loss functions are utilized to exploit global contrast and local details, further enhancing the quality of the enhanced images.

Recently, denoising diffusion probabilistic models (DDPM) [37] have garnered considerable attention as a type of probabilistic generative model capable of generating high-resolution images. These models have found applications in various domains, including image generation [34], inpainting [36], colorization [36], and image segmentation [34]. Reader is suggested to refer to [4] for a comprehensive survey.

While DDPMs have achieved impressive results, they typically require a high number of iterations, leading to slow performance. In contrast, Latent diffusion models (LDMs) [34] address this issue by training the diffusion process in the latent space of pre-trained autoencoders. Furthermore, LDMs incorporate cross-attention layers to achieve near-optimal complexity reduction and preserve fine details. This enhancement in visual fidelity not only improves the performance but also mitigates the slow performance associated with traditional DDPMs.

In a nutshell, DDPMs have shown promise in image reconstruction, offering easier training and improved sample quality compared to traditional generative models and VAEs. However, as probabilistic generative models learn representative distributions rather than deterministic solutions like U-Net with L2 loss, they often underperform on traditional metrics like PSNR/SSIM for image enhancement. Consequently, the exploration of diffusion models for low-light image enhancement remains limited. Given their impressive capabilities in image generation, in this work we investigate their potential in enhancing low-light images.

3 Proposed Method

3.1 Preliminaries

Diffusion models draw inspiration from non-equilibrium thermodynamics principles. They utilize a Markov chain of diffusion processes to introduce random noise to data and subsequently learn a reverse process that reconstructs the desired data samples from the noise. Numerous generative models based on diffusion have been proposed, such as diffusion probabilistic models [37], noise-conditioned score network [39], and denoising diffusion probabilistic models (DDPM) [14].

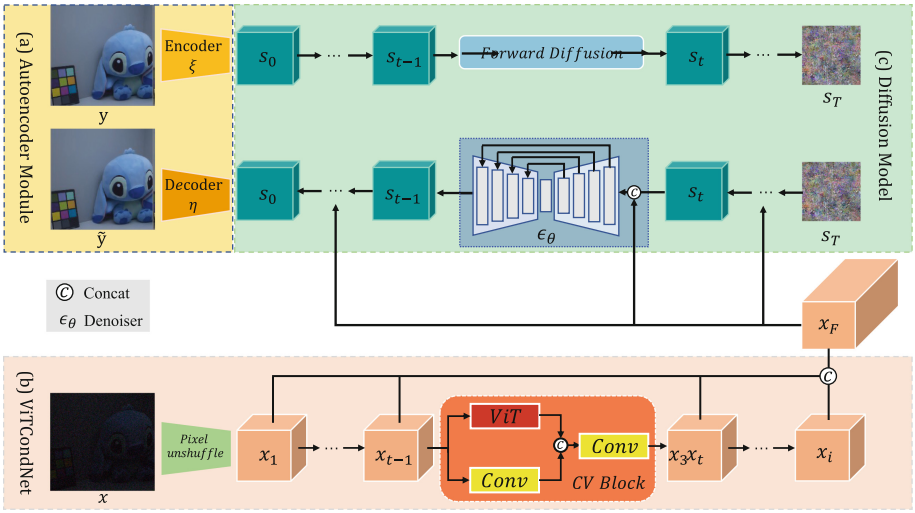

Fig. 1. Overview of our proposed model, L^{2}DM. (a) Autoencoder module is a trainable component used to transform images from the pixel space to the latent space, significantly reducing computational requirements during model training. (b) Condition module is a newly proposed trainable framework that achieves excellent feature extraction capabilities while significantly reducing the number of parameters. (c) Extracted condition feature map is concatenated with the s_t and fed into the time-conditional U-Net.

In the forward diffusion process, an image data sampled from a real data distribution undergoes degradation over T timesteps. At each timestep, Gaussian noise is added incrementally, gradually reducing the quality of the image. This process can be formulated as:

$$q(y_t|y_{t-1}) = \mathcal{N}(y_t; \sqrt{\beta}y_{t-1}, (1-\beta_t)I), \tag{1}$$

$$q(y_t|y_0) = \mathcal{N}(y_t; \sqrt{\bar{\alpha}}y_0, (1-\bar{\alpha}_t)I), \tag{2}$$

where $\{\beta_1, \ldots, \beta_T\}$ is a prescribed variance schedule defined over timestep T, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$. By iteratively introducing noise steps with small

variances, the image data progressively converges to a distribution that closely resembles a standard Gaussian in form of:

$$q(y_T|y_0) = \mathcal{N}(y_t; 0, I). \tag{3}$$

In the reverse diffusion process, we start with a completely random standard normal distribution, denoted as y_t, and gradually remove the Gaussian noise to recover y_0 using t intermediate steps. Each reverse step utilizes a distribution p modeled by a neural network with parameters θ, which can be expressed as follows:

$$p_\theta(y_{t-1}|y_t) = \mathcal{N}(y_{t-1}; \mu_\theta(y_t, t), \widetilde{\beta}_t I), \tag{4}$$

where $\widetilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{\bar{\alpha}_t}\beta_t$. The mean μ_θ can be written as:

$$\mu_\theta(y_t, t) = \frac{1}{\sqrt{\alpha_t}}(y_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(y_t, t)), \tag{5}$$

where ϵ_θ represents the estimated residual noise. To determine the parameters of the distribution $p_\theta(\cdot)$, we optimize the variational lower bound of the log-likelihood of $p_\theta(y_0)$. This can be expressed as follows:

$$L_{DM} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),t\sim[1,T]}[\|\epsilon - \epsilon_\theta(y_t, t)\|_2^2], \tag{6}$$

The network p_θ is trained to model the Gaussian distribution $\mathcal{N}(0, I)$ at each timestep t.

3.2 Autoencoder Module

Training a full-resolution diffusion model is impractical due to its extremely high computational demands, requiring days of training on multiple GPUs. Taking inspiration from [34], instead of processing the original image input in the pixel space, which is only suitable for small-resolution image processing, we employ an autoencoder that compresses the pixel-space features into a latent feature space. The autoencoder is trained using a combination of perceptual loss [63], patch-based [15] adversarial objectives [8,58].

More precisely, given an input image $y \in \mathbb{R}^{H\times W\times 3}$ in RGB space, the encoder ξ encodes x into a latent representation $s = \xi(y)$, where $s \in \mathbb{R}^{h\times w\times 3}$. We set the encoder downsampling factors $f = H/h = W/w = 4$. The decoder η reconstructs the representation from the latent $\widetilde{y} = \eta(s)$, and note that $\widetilde{y}$ has the same shape as y. Within the decoder, vector quantization (VQ) regularization, which uses a vector quantization layer [40] is imposed in order to avoid arbitrarily high-variance latent spaces. The autoencoder model is trained in an adversarial manner following [8]. A patch-based discriminator D_ω is optimized to differentiate original images y from reconstructions $\widetilde{y} = \eta(\xi(y))$. To avoid arbitrarily scaled latent spaces, we regularize the latent s to be zero centered and obtain small variance by introducing an regularizing loss term L_{reg}. We regularize the

latent space with a vector quantization layer by learning a high codebook dimensionality $|\mathscr{I}|$ [40]. The full objective L_{ae} to train the autoencoder model can be written as:

$$L_{ae} = \min_{\eta,\xi} \max_{\omega} \big(L_{rec}(y, \eta(\xi(y))) - L_{adv}(\eta(\xi(y))) + \log D_{\omega}(y) + L_{reg}(y; \eta, \xi)\big), \quad (7)$$

where $L_{rec} = ||y - \widetilde{y}||_2^2$ is a reconstruction loss.

3.3 ViTCondNet

Similar to other types of generative models [29,38], diffusion models have the inherent capability to model conditional distributions in the form of $p(y|x)$. This can be accomplished by introducing a conditional variable x to control the synthesis process using condition inputs, such as text [32], semantic maps [15,31], or images [15]. For the specific task of low-light image enhancement, we utilize the captured low-light image as the condition input.

In this work, a new conditional feature extraction framework is designed to extract the representations of conditional images and fuse them with Diffusion Model. As shown in Fig. 1(b), layered structure of our condition module is based on Convolution-ViT (CV) block [52] and a concatenation that concatenates all the hierarchical features processed by CV blocks. Given an input of condition image $x \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the spatial height and width of the condition image. A Pixel-unshuffle Layer unshuffles x into 48 channels by setting the downscale factor to 4, to construct $x_1 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 48}$. Then the same operation is performed on x_1 with CV block to obtain x_2, and so on to obtain x_i. Finally, $x_1, x_2, \cdots, x_i$ are concatenated in the channel dimension by concat operation, and a feature map $x_F \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times (48 * i)}$ is obtained. Through a series of ablation experiments, we have demonstrated that this feature extraction framework can achieve excellent feature extraction results even with minimal parameters, making it highly effective and easy to train.

3.4 Main Architecture

The backbone of $\epsilon_\theta(x, y_t, t)$ in Eq. 6 is implemented by a time-conditional U-Net [35]. With a well-trained autoencoder model composed of ξ and η, we have gained access to an efficient and low-dimensional latent space, in which high-frequency and imperceptible details are abstracted. As shown in Fig. 1(a)(c), the ground truth image is compressed into s_0 in the latent space by the encoder module ξ. During the forward diffusion process, Gaussian noise is gradually added to s_0 for T iterations, resulting in a final image, s_T, which is completely Gaussian-distributed. s_T is concatenated with the features x_F, extracted by our ViTCondNet in the channel dimension. The concatenated data is then fed into the U-Net, and the desired data distribution is gradually restored through loss calculation with respect to the standard normal distribution. The loss function of our network is:

$$L = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I), t \sim [1,T]}[||\epsilon - \epsilon_\theta(x, y_t, t)||_1], \quad (8)$$

where x represents the conditional features obtained after passing through the ViTCondNet. We abandon the widely used L2 loss in diffusion models and instead employ the L1 loss. Considering the image enhancement task and the significant amount of noise pixels introduced during the forward diffusion process, we anticipate the presence of more outlier pixels. L1 loss, being sensitive to outliers, is capable of penalizing these exceptional pixels effectively. Therefore, by employing L1 loss, we can precisely control the extent of penalization for outlier pixels, leading to improved enhancement results. This facilitates better modeling and handling of outlier pixels, ultimately contributing to superior image enhancement outcomes. The experimental results also validate the effectiveness of our approach.

4 Experiments

4.1 Setup

Schedules. First, we set timestep $T = 1000$ and the base learning rate of the model to 1×10^{-5}. We follow a linear increase for the variance schedule β_t, ranging from 0.0015 to 0.0155. To assess the system's performance, we evaluate it based on two metrics: peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [46]. For the number of concatenation features x_i in ViTCondNet, we empirically set i to 4.

Datasets. We conduct experiments on LOL [47] and LOLV2 [57] datasets. The LOL dataset contains 485 training and 15 testing paired images, with each pair comprising a low-light and a normal-light image. The LOLV2 dataset contains two parts, Real-captured and Synthetic. Real part has noise distributions not present in the LOL dataset, and Synthetic part has illumination distributions not present in the LOL dataset.

Training. During the training phase, we randomly cropped pairs of low-light and normal-light images to a size of 256×256. For LOL and LOLV2-real datasets, we set the batch size to 16. For the LOLV2-sync dataset, we use a larger batch size of 32. All training processes are completed on NVIDIA GeForce RTX 3090.

Evaluation. Combined with DPM-solver [27], sampling process can be completed within 20 steps. We must note that due to the diffusion model denoising from a completely random Gaussian distribution, each generated sample will not be exactly the same. This leads to fluctuations in the assessment of the results within a range. During the validation process, we define the batch size as $b = 10$, which represents the number of repetitions of the same image. We select one high quality sample from each batch and calculate the average PSNR and SSIM over 5 batches. The images in the LOL and LOL-real datasets have a resolution of 400×600, which doesn't match the size required by our network. Therefore, during the validation phase, we uniformly resize the images in the test set to a size of 384×576 while maintaining the aspect ratio.

Other Details. We pre-trained the complete architecture of L^{2}DM on the COCO dataset. We excluded images from the COCO training set that had inappropriate sizes, resulting in a set of 117,188 images, which we used as the ground truth. We then applied darkening transformations to these images and added random-sized noise to each pixel. These processed images were used as the conditional inputs. After training the model weights on the COCO dataset, we only need to fine-tune the model on the target dataset to achieve excellent results.

Table 1. Quantitative results on the **LOL dataset** in terms of PSNR and SSIM. ↑ denotes that larger values lead to better quality. Models marked with an asterisk ∗ indicate that the model is implemented based on the diffusion model. The best results are shown in **bold**, and the second-best results are underlined.

Method	Zero-DCE [11]	LIME [13]	RetinexNet [47]	EG [17]	RUAS [26]	DRBN [55]
PSNR↑	14.86	16.76	16.77	17.48	18.23	20.13
SSIM↑	0.54	0.56	0.56	0.65	0.72	0.83
Method	KinD [65]	KinD++ [64]	NE [18]	Bread [12]	RCTNet [21]	IAT [5]
PSNR↑	20.87	21.30	21.52	22.96	22.67	23.38
SSIM↑	0.80	0.82	0.76	0.82	0.79	0.81
Method	HWMNet [9]	LLFlow [45]	StarDiffusion* [60]	DDRM* [19]	LDM* [34]	L^{2}DM*(Ours)
PSNR↑	24.24	**24.99**	20.77	16.41	21.41	24.54
SSIM↑	0.83	**0.92**	0.80	0.65	0.75	0.83

Table 2. Quantitative comparison on the **LOLV2-real dataset**. ↑ denotes that larger values lead to better quality. Models marked with an asterisk ∗ indicate that the model is implemented based on the diffusion model. The best results are shown in **bold**, and the second-best results are underlined.

Methods	DeepUPE [43]	RF [22]	DeepLPF [30]	KIND [65]	FIDE [50]	LPNet [24]
PSNR↑	13.27	14.05	14.10	14.74	16.85	17.80
SSIM↑	0.452	0.458	0.480	0.641	0.678	0.792
Methods	3DLUT [62]	MIR-Net [61]	UNIE [18]	LCDR [42]	LLFlow [45]	A3DLUT [44]
PSNR↑	17.59	20.02	20.85	18.57	19.36	18.19
SSIM↑	0.721	0.820	0.724	0.641	0.705	0.745
Methods	Band [56]	EG [17]	Retinex [26]	Sparse [57]	DSN [66]	RCTNet [21]
PSNR↑	20.29	18.23	18.37	20.06	19.23	20.51
SSIM↑	0.831	0.617	0.723	0.815	0.736	0.831
Methods	UTVNet [67]	SCI [28]	Uretinex [48]	SNR [53]	SMG [54]	L^{2}DM*(Ours)
PSNR↑	20.37	20.28	21.16	21.48	24.62	**24.80**
SSIM↑	0.834	0.752	0.840	0.849	**0.867**	0.817

Table 3. Quantitative comparison on the **LOLV2-synthetic dataset**. ↑ denotes that larger values lead to better quality. Models marked with an asterisk ∗ indicate that the model is implemented based on the diffusion model. The best results are shown in **bold**, and the second-best results are underlined.

Methods	KIND [65]	SID [3]	DeepUPE [43]	FIDE [51]	RF [22]	DeepLPF [30]
PSNR↑	13.29	15.04	15.08	15.20	15.97	16.02
SSIM↑	0.578	0.610	0.623	0.612	0.632	0.587
Methods	Retinex [26]	3DLUT [62]	HWMNet [9]	LCDR [42]	A3DLUT [44]	Bread [12]
PSNR↑	16.55	18.04	18.79	18.91	18.92	19.28
SSIM↑	0.652	0.800	0.817	0.825	0.838	0.831
Methods	LPNet [24]	LLFlow [45]	DSN [66]	UTVNet [67]	MIR-Net [61]	UNIE [18]
PSNR↑	19.51	19.69	21.22	21.62	21.94	21.84
SSIM↑	0.846	0.871	0.827	0.904	0.876	0.884
Methods	Sparse [57]	SCI [28]	RCTNet [21]	Uretinex [48]	Band [56]	L^{2}DM*(Ours)
PSNR↑	22.05	22.20	22.44	22.89	23.22	**23.96**
SSIM↑	0.905	0.887	0.891	0.895	**0.927**	0.786

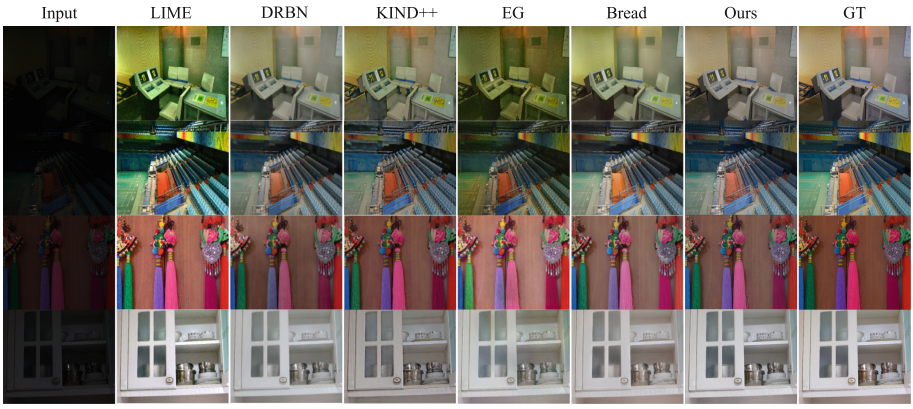

Fig. 2. Qualitative comparison with state-of-the-art methods on the **LOL dataset**.

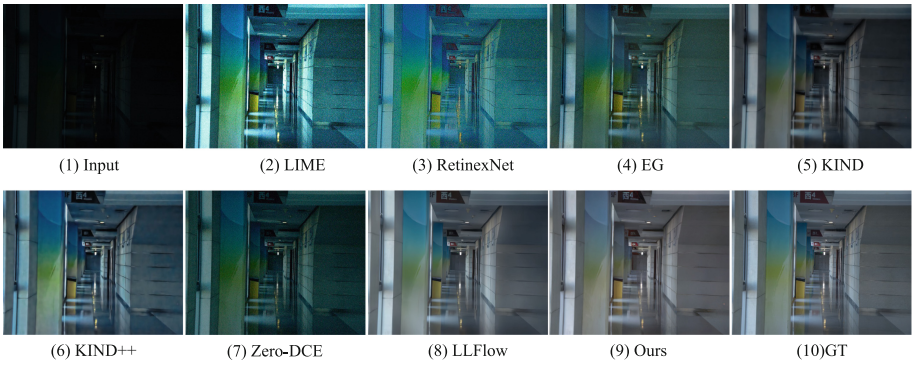

Fig. 3. Qualitative comparison with state-of-the-art methods on the **LOLV2-real dataset**.

4.2 Comparsion with SOTA Methods

LOL Dataset. We first compare L^{2}DM with SOTA methods on the LOL dataset. As mentioned in the previous subsection, the generated samples exhibit uncertainty. Therefore, our validation results are averaged over 5 tests. As shown in Table 1, we differentiate between methods based on conventional U-Net architecture and those based on the diffusion model. Our proposed model achieves state-of-the-art performance among all diffusion-based methods, surpassing the second-best method by a significant margin of 3.13 in terms of PSNR. Figure 2 shows qualitative comparisons with other methods, the differences between our generated images and the ground truth images are barely discernible.

LOLV2 Dataset. For the LOLV2 dataset, since there are limited results available for diffusion-based methods, we compare our method with state-of-the-art approaches based on CNN and Transformer models. The quantitative results of LOLV2 real part and synthetic are shown in Table 2 and Table 3, respectively. Our proposed model achieves the top position in terms of PSNR on the LOLV2 dataset. However, there is still a gap compared to mainstream methods in terms of SSIM. This is an area that we aim to improve in our future work. Figure 3 presents the qualitative comparison results on the LOLV2-real dataset, where our method showcases excellent enhancement performance.

4.3 Ablation Studies

In this subsection, we conduct ablation studies on the main components of L^{2}DM to better demonstrate the effectiveness of each module of our system. The experiments were conducted on the LOL dataset.

Table 4. Ablation study on the L^{2}DM using LOL dataset. *Param* stands for the parameter size of the condition module, while "–" means the hyperparameter variation does not affect the number of parameters in the conditional module.

	Metrics		
	PSNR	SSIM	*Param*(M)
Baseline	24.54	0.83	0.07
KL-reg	19.83	0.76	–
SpatialRescaler	17.51	0.65	None
ResNet50	23.21	0.80	23.51
U-Net-encoder	22.02	0.77	20.64
L2	23.89	0.81	–
Charbonnier	24.10	0.81	–

Regularization. In the autoencoder module, we replaced VQ regularization with Kullback-Leibler (KL) regularization. KL regularization applies a slight

KL penalty to the learned latent towards a standard normal distribution. From Table 4, we can observe that using VQ regularization achieves better generation results.

Condition Module. We compared our proposed condition module ViTCondNet ($i = 4$) with some conditional modules, such as SpatialRescaler in LDM [34], ResNet and U-Net. We utilize the encoder component of a U-Net architecture that incorporates an Attention mechanism to extract features. The comparison results are shown in Table 4. Our condition module outperforms others in terms of both model parameter size and image generation quality.

Loss Function. We replaced the L1 loss with the L2 loss and the Charbonnier loss [2]. However, our experiments revealed that the L1 loss outperforms both the L2 and Charbonnier losses.

5 Conclusion

In this paper, we present L^{2}DM, an end-to-end trainable low-light image enhancement model based on the diffusion model, showcasing exceptional image generation performance. L^{2}DM incorporates an autoencoder module that significantly reduces the computational demands of the diffusion model, along with an innovative lightweight conditional module known as ViTCondNet. This combination empowers the model to process low-light images as input and generate high-quality enhanced images. Furthermore, the underwhelming performance of diffusion models in terms of PSNR on low-light enhancement datasets has resulted in their limited adoption in the field. Our approach addresses this issue by demonstrating that diffusion-based methods can achieve remarkable outcomes across diverse datasets. Through extensive experiments and comparisons on publicly available datasets, we provide further validation of our approach's performance. It is our intention for L^{2}DM to serve as a baseline for low-light image enhancement and to encourage the application of diffusion models in low-level visual tasks.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62376003, 62306003, 62372004, 62302005).

References

1. Arici, T., Dikbas, S., Altunbasak, Y.: A histogram modification framework and its application for image contrast enhancement. IEEE Trans. Image Process. **18**(9), 1921–1935 (2009)
2. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of IEEE International Conference on Image Processing, vol. 2, pp. 168–172. IEEE (1994)

3. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3291–3300 (2018)
4. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: a survey. IEEE Trans. Pattern Anal. Mach. Intelli. (2023)
5. Cui, Z., et al.: Illumination adaptive transformer. arXiv preprint arXiv:2205.14871 (2022)
6. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Adv. Neural. Inf. Process. Syst. **34**, 8780–8794 (2021)
7. Dong, X., et al.: Abandoning the Bayer-filter to see in the dark. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17431–17440 (2022)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12873–12883 (2021)
9. Fan, C.M., Liu, T.J., Liu, K.H.: Half wavelet attention on M-Net+ for low-light image enhancement. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 3878–3882. IEEE (2022)
10. Fan, M., Wang, W., Yang, W., Liu, J.: Integrating semantic segmentation and retinex model for low-light image enhancement. In: Proceedings of the 28th ACM International Conference on Multimedia (ACMMM). pp. 2317–2325 (2020)
11. Guo, C., et al.: Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1780–1789 (2020)
12. Guo, X., Hu, Q.: Low-light image enhancement via breaking down the darkness. Int. J. Comput. Vision **131**(1), 48–66 (2023)
13. Guo, X., Li, Y., Ling, H.: LIME: low-light image enhancement via illumination map estimation. IEEE Trans. Image Process. **26**(2), 982–993 (2016)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
16. Jiang, N., Lin, J., Zhang, T., Zheng, H., Zhao, T.: Low-light image enhancement via stage-transformer-guided network. IEEE Trans. Circuits Syst. Video Technol. (2023)
17. Jiang, Y., et al.: EnlightenGAN: deep light enhancement without paired supervision. IEEE Trans. Image Process. **30**, 2340–2349 (2021)
18. Jin, Y., Yang, W., Tan, R.T.: Unsupervised night image enhancement: when layer decomposition meets light-effects suppression. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022, Part XXXVII. LNCS, vol. 13697, pp. 404–421. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19836-6_23
19. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. arXiv preprint arXiv:2201.11793 (2022)
20. Kim, G., Kwon, D., Kwon, J.: Low-lightgan: low-light enhancement via advanced generative adversarial network with task-driven training. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 2811–2815. IEEE (2019)
21. Kim, H., Choi, S.M., Kim, C.S., Koh, Y.J.: Representative color transform for image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4459–4468 (2021)

22. Kosugi, S., Yamasaki, T.: Unpaired image enhancement featuring reinforcement-learning-controlled image editing software. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11296–11303 (2020)
23. Lee, C., Lee, C., Kim, C.S.: Contrast enhancement based on layered difference representation of 2D histograms. IEEE Trans. Image Process. **22**(12), 5372–5384 (2013)
24. Li, J., Li, J., Fang, F., Li, F., Zhang, G.: Luminance-aware pyramid network for low-light image enhancement. IEEE Trans. Multimedia **23**, 3153–3165 (2020)
25. Lim, S., Kim, W.: DSLR: deep stacked laplacian restorer for low-light image enhancement. IEEE Trans. Multimedia **23**, 4272–4284 (2020)
26. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10561–10570 (2021)
27. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-solver: a fast ode solver for diffusion probabilistic model sampling in around 10 steps. arXiv preprint arXiv:2206.00927 (2022)
28. Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z.: Toward fast, flexible, and robust low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5637–5646 (2022)
29. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
30. Moran, S., Marza, P., McDonagh, S., Parisot, S., Slabaugh, G.: DeepLPF: deep local parametric filters for image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12826–12835 (2020)
31. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2337–2346 (2019)
32. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning, pp. 1060–1069. PMLR (2016)
33. Ren, W., et al.: Deep video dehazing with semantic segmentation. IEEE Trans. Image Process. **28**(4), 1895–1908 (2018)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
35. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
36. Saharia, C., et al.: Palette: image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–10 (2022)
37. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR (2015)
38. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
39. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

40. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
41. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
42. Wang, H., Xu, K., Lau, R.W.: Local color distributions prior for image enhancement. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022, Part XVIII. LNCS, vol. 13678, pp. 343–359. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19797-0_20
43. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6849–6857 (2019)
44. Wang, T., Li, Y., Peng, J., Ma, Y., Wang, X., Song, F., Yan, Y.: Real-time image enhancer via learnable spatial-aware 3D lookup tables. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2471–2480 (2021)
45. Wang, Y., Wan, R., Yang, W., Li, H., Chau, L.P., Kot, A.: Low-light image enhancement with normalizing flow. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2604–2612 (2022)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
47. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018)
48. Wu, W., Weng, J., Zhang, P., Wang, X., Yang, W., Jiang, J.: URetinex-Net: Retinex-based deep unfolding network for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5901–5910 (2022)
49. Wu, X., Liu, X., Hiramatsu, K., Kashino, K.: Contrast-accumulated histogram equalization for image enhancement. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3190–3194. IEEE (2017)
50. Xu, K., Yang, X., Yin, B., Lau, R.W.: Learning to restore low-light images via decomposition-and-enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2281–2290 (2020)
51. Xu, K., Yang, X., Yin, B., Lau, R.W.: Learning to restore low-light images via decomposition-and-enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2281–2290 (2020)
52. Xu, W., Dong, X., Ma, L., Teoh, A.B.J., Lin, Z.: Rawformer: an efficient vision transformer for low-light raw image enhancement. IEEE Signal Process. Lett. **29**, 2677–2681 (2022)
53. Xu, X., Wang, R., Fu, C.W., Jia, J.: SNR-aware low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17714–17724 (2022)
54. Xu, X., Wang, R., Lu, J.: Low-light image enhancement via structure modeling and guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9893–9903 (2023)
55. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: From fidelity to perceptual quality: a semi-supervised approach for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3063–3072 (2020)

56. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: Band representation-based semi-supervised low-light image enhancement: bridging the gap between signal fidelity and perceptual quality. IEEE Trans. Image Process. **30**, 3461–3473 (2021)
57. Yang, W., Wang, W., Huang, H., Wang, S., Liu, J.: Sparse gradient regularized deep retinex network for robust low-light image enhancement. IEEE Trans. Image Process. **30**, 2072–2086 (2021)
58. Yu, J., et al.: Vector-quantized image modeling with improved VQGAN. arXiv preprint arXiv:2110.04627 (2021)
59. Yuan, N., et al.: Low-light image enhancement by combining transformer and convolutional neural network. Mathematics **11**(7), 1657 (2023)
60. Yuan, Y., et al.: Learning to kindle the starlight. arXiv preprint arXiv:2211.09206 (2022)
61. Zamir, S.W., et al.: Learning enriched features for real image restoration and enhancement. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12370, pp. 492–511. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58595-2_30
62. Zeng, H., Cai, J., Li, L., Cao, Z., Zhang, L.: Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time. IEEE Trans. Pattern Anal. Mach. Intell. **44**(4), 2058–2073 (2020)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
64. Zhang, Y., Guo, X., Ma, J., Liu, W., Zhang, J.: Beyond brightening low-light images. Int. J. Comput. Vision **129**, 1013–1037 (2021)
65. Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: a practical low-light image enhancer. In: Proceedings of the 27th ACM International Conference on Multimedia (ACMMM), pp. 1632–1640 (2019)
66. Zhao, L., Lu, S.P., Chen, T., Yang, Z., Shamir, A.: Deep symmetric network for underexposed image enhancement with recurrent attentional learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12075–12084 (2021)
67. Zheng, C., Shi, D., Shi, W.: Adaptive unfolding total variation network for low-light image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4439–4448 (2021)
68. Zhu, M., Pan, P., Chen, W., Yang, Y.: EEMEFN: low-light image enhancement via edge-enhanced multi-exposure fusion network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13106–13113 (2020)