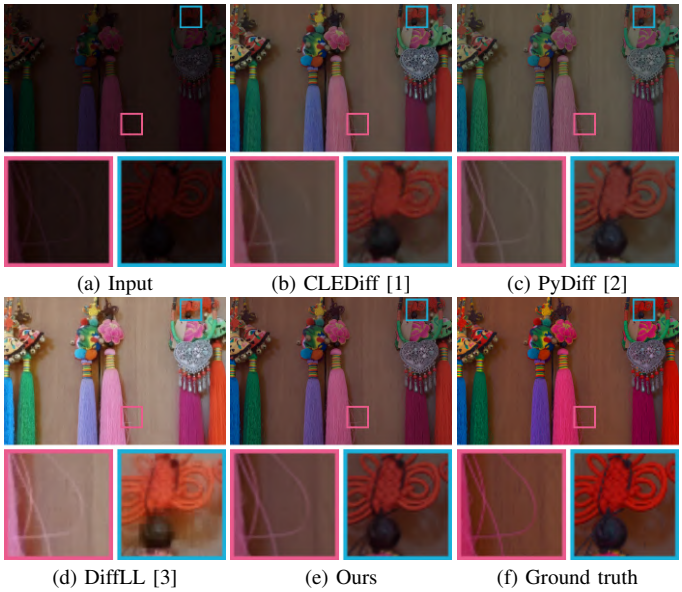
Detail-Preserving Diffusion Models for Low-Light Image Enhancement

Yan Huang, Xiaoshan Liao, Jinxiu Liang, *Member, IEEE*, Boxin Shi, *Senior Member, IEEE*, Yong Xu, *Senior Member, IEEE*, and Patrick Le Callet, *Fellow, IEEE*

***Abstract*—Existing diffusion models for low-light image enhancement typically focus on incrementally removing noise introduced during the forward diffusion process using a denoising loss, with the process being conditioned on input low-light images. While these models demonstrate remarkable abilities in generating realistic high-frequency details, they often struggle to accurately restore fine details that are faithful to the input. To address this, we present a novel detail-preserving diffusion model for realistic and faithful low-light image enhancement. Our approach integrates a size-agnostic diffusion process with a reverse process reconstruction loss, significantly enhancing the fidelity of enhanced images to their low-light counterparts and enabling more accurate recovery of fine details. To ensure the preservation of region- and content-aware details, we employ an efficient noise estimation network with a simplified channel-spatial attention mechanism. Additionally, we propose a multiscale ensemble scheme to maintain detail fidelity across diverse illumination regions. Comprehensive experiments on eight benchmark low-light image enhancement datasets demonstrate that our method achieves state-of-the-art results compared to 20 existing methods in terms of both perceptual quality (LPIPS) and distortion metrics (PSNR and SSIM).**

***Index Terms*—Low-light image enhancement, conditional patch-based diffusion models, detail-preserving, reverse diffusion-based reconstruction, multiscale ensemble scheme.**



(a) Input (b) CLEDiff [1] (c) PyDiff [2]

(d) DiffLL [3] (e) Ours (f) Ground truth

Fig. 1. Visual comparison results of existing diffusion-based LLIE methods and the proposed one. It clearly demonstrates the superior performance of our method in preserving details and handling variations in brightness and noise.

I. INTRODUCTION

Achieving high-quality photography in real-world scenarios frequently confronts the significant challenge of inadequate lighting, particularly in indoor or nighttime settings where illumination is often insufficient. Conventional solutions, such as applying analog or digital gain, tend to amplify noise, while extending exposure time can result in motion blur due to camera shake or subject movement. This issue not only affects the perceptual quality of photograph [4] but also impedes critical vision tasks like detection and tracking [5], [6].

This work was supported in part by grants from the National Natural Science Foundation of China (Nos. 62072188 and 62302019), the 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 China Postdoctoral Science Foundation (No. 2022M720236). (Corresponding author: Jinxiu Liang)

Yan Huang, Xiaoshan Liao, and Yong Xu are with the Guangdong Key Lab of Communication and Computer Network, School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: aihuangy@scut.edu.cn; csxsliao@mail.scut.edu.cn; yxu@scut.edu.cn). Yong Xu is also with the Pazhou Lab, Guangzhou 510005, China.

Jinxiu Liang and Boxin Shi are with the National Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China (e-mail: cssherryliang@pku.edu.cn; shiboxin@pku.edu.cn).

Patrick Le Callet is with the Polytech Nantes, Université de Nantes, Nantes 44306, France (e-mail: patrick.lecallet@univ-nantes.fr).

Low-Light Image Enhancement (LLIE) is dedicated to improving the quality of photographs captured under low-light conditions, characterized by low signal-to-noise ratio (SNR) and poor contrast [4], [7], [8]. The past few decades have seen the emergence of LLIE methods [4], [9], evolving from traditional techniques [10]–[12] to deep learning-based approaches [2], [13]–[27]. Despite the advancements in LLIE methods, significant challenges remain in achieving high-quality image enhancement under low-light conditions. Existing deep learning-based regression methods focus on mapping low-light images to normal-light images using metrics like mean squared error. Although these methods optimize distortion metrics such as PSNR, they tend to produce overly smoothed predictions that lack high-frequency details essential for perceptual realism [28]–[30]. These methods often struggle to maintain the delicate balance between noise reduction and detail preservation, leading to results that may appear visually unrealistic or lacking in fine detail.

Diffusion models (DMs) have recently demonstrated considerable potential in producing perceptually realistic high-frequency details for LLIE [1]–[3]. These methods operate by gradually transforming an image into a normal distribution by adding noise during the forward diffusion process, followed by a reverse denoising step in which a neural network with the

low-light image as a condition is guided by a denoising loss. However, as shown in Fig. 1, existing diffusion-based LLIE methods face several notable challenges in *detail preservation*: *Firstly*, current diffusion-based LLIE methods prioritize optimizing denoising loss rather than predicting normal-light images against ground truth pixel-wisely. Although they excel in data distribution fitting and realistic enhancements, their ability to faithfully recover fine details may be limited. This limitation is critical in applications requiring high-fidelity detail preservation, such as medical imaging and security surveillance. *Secondly*, the brightness distribution in low-light images is highly variable. Existing methods often apply a uniform enhancement approach, which may not adapt well to the varying illumination conditions across different regions of an image. An approach with locality-based brightness adaptability is crucial for accurately identifying and differentiating between noise and fine details in different regions of the input low-light images. *Last but not least*, low-light images exhibit non-uniform noise properties that vary across different scales and regions. Current training and inference schemes in DMs often target whole input images, limiting scale-agnostic detail recovery and failing to address the diversity in real-world textures and patterns. This issue becomes more pronounced in real-world scenarios, where lighting conditions and noise characteristics can vary significantly.

To mitigate these challenges, we introduce Detail-Preserving Diffusion Models (DePDiff) for realistic and faithful low-light image enhancement, which utilize the following strategies for better detail preservation: *i)* **Reverse diffusion-based reconstruction loss**: In the DDIM case, latent noises converted from the input low-light images through a forward diffusion can be nearly perfectly inverted to target normal-light images using a reverse diffusion if the score function for the reverse diffusion is retained the same as that of the forward diffusion. To ensure detail preservation of the predicted normal-light images to the targets, we constrain the faithfulness by a reconstruction loss between them during training, akin to GANs. This loss function helps maintain high-frequency details while reducing noise. *ii)* **Content and region-aware architecture**: To enhance spatial adaptability for distinguishing between relevant image content and noise in challenging low-light conditions, we equip the commonly-used U-Net in DMs [31] with an activation-free architecture and simplified channel-spatial attention, dubbed Content and Region-Aware Network (CRANet). This architecture can adaptively focus on relevant features across both channels and spatial dimensions, selectively enhancing important features while suppressing less relevant information. The integration of channel and spatial attention mechanisms allows the network to better handle varying brightness and noise characteristics across different locations. *iii)* **Multiscale ensemble scheme**: For scale-adaptivity, we adopt a patch-based training approach to guide the denoising process in DMs with adaptive noise estimates for overlapping patches and a multiscale ensemble scheme to aggregate details from various scales. This scheme allows our model to effectively capture and preserve details at multiple scales, addressing non-uniform noise properties and enhancing overall image quality.

By addressing these challenges, our proposed DePDiff method offers a detail-preserving diffusion-based method in the field of low-light image enhancement. Extensive experiments demonstrate that the proposed method effectively balances between noise reduction and detail preservation in low-light images, achieving state-of-the-art performance on various benchmarks. The contributions of our work include:

- Equipping conditional patch-based DMs with multiscale ensemble scheme for scale-adaptive enhancement and detail aggregation in low-light images;
- Proposing a reverse diffusion-based reconstruction loss for more faithful enhancement for low-light image, akin to GAN-based training schemes; and
- Introducing an efficient architecture with channel-spatial attention for precise, localized enhancement from inputs with non-uniform degradation levels.

II. RELATED WORK

The field of low-light image enhancement (LLIE) has seen considerable advancement in the last few decades, evolving from traditional methods to sophisticated deep learning techniques [4], [9], [10]. This section outlines the development in LLIE, focusing initially on non-learning and deep regression methods before delving into the generative approaches, especially the diffusion-based ones.

A. Non-learning LLIE methods

Traditional non-learning LLIE methods, known for their computational efficiency, generally rely on statistical properties or established image priors. They can be broadly categorized into histogram equalization-based [10], [32], Retinex-based, and nonlinear transformation-based methods [9], [12]. LLIE methods based on histogram equalization focus on redistributing pixel intensities across the full dynamic range to improve contrast [10]. An example includes integrating intuitionistic fuzzy set theory into histogram equalization for LLIE [32]. While simple and effective in contrast enhancement, these methods may inadvertently amplify noise [9], [10]. Retinex-based methods, rooted in color vision theory, aim to enhance the illumination component of low-light images [11], such as the use of adaptive gamma correction in Retinex decomposition [12]. Nonlinear transformation-based methods treat LLIE as a mapping from low to normal light conditions [9], with gamma correction as a classic example [12]. Although efficient, these traditional methods often fall short in complex lighting conditions [8], [33].

B. Regression LLIE methods

Deep learning-based approaches offer an effective alternative to traditional methods, enabling enhanced LLIE performance [4]. The learning-based regression methods aim to learning a mapping between low-light images and their corresponding normal-light ones by using existing deep learning methods [29], [34]. In [3], a transformer-based architecture is incorporated to capture different degradation distributions in low-light images. Multiscale network architectures are used

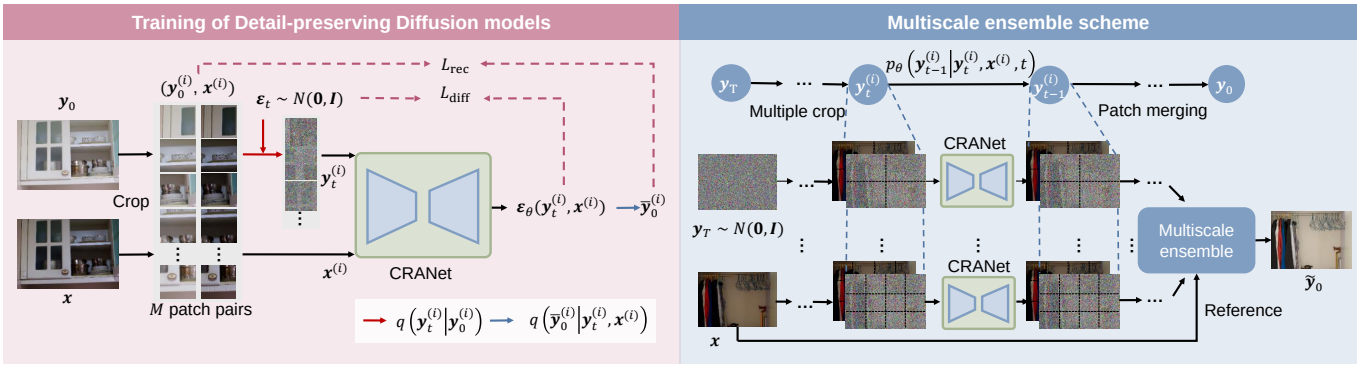


Fig. 2. Illustration of the training and multiscale ensemble process of the proposed DePDiff. Left: Conditioned on $\boldsymbol{x}^{(i)}$ extracted from low-light image $\boldsymbol{x}$, $\boldsymbol{y}_0^{(i)}$ extracted from target image $\boldsymbol{y}_0$ are gradually transitioning into noise drawn from normal distribution during forward diffusion $q(\boldsymbol{y}_t^{(i)}|\boldsymbol{y}_0^{(i)})$ (red arrow). Rreconstruction loss $\mathcal{L}_{\text{rec}}$ is applied in conjunction with the standard denoising loss $\mathcal{L}_{\text{diff}}$ by using the reverse diffusion $q(\overline{\boldsymbol{y}}_0^{(i)}|\boldsymbol{y}_t^{(i)}, \boldsymbol{x}^{(i)})$ (blue arrow). Right: For each low-light patches $\boldsymbol{x}^{(i)}$, the noise estimator CRANet estimates $p_\theta(\boldsymbol{y}_{t-1}^{(i)}|\boldsymbol{y}_t^{(i)}, \boldsymbol{x}^{(i)}, t)$ to progressively denoise random noise patches $\boldsymbol{y}_T^{(i)}$—initially sampled from a normal distribution—into the final output $\boldsymbol{y}_0$, which adheres to the data distribution. The ultimate enhanced image, $\widetilde{\boldsymbol{y}}_0$, is composed by merging these denoised patches, each processed at different scales.

combined with techniques like frequency domain optimization and illumination constraints to enhance low-light images [18], [22]. Fan *et al*. [13] integrated the half-wavelet attention mechanism with CNN to design a hierarchical model (dobbed HWMNet) for LLIE tasks. Based on the famous Transformer, a lightweight illumination-adaptive Transformer (IAT) was proposed and applied to object detection and semantic segmentation under different light conditions [14]. A neural-architecture-search-based LLIE method (RUAS) was designed based on Retinex theory [35]. An unsupervised LLIE method is developed in [15] by proposing a pseudo-supervised training strategy, which relies on pseudo-labels for training. In addition, there are some image enhancement methods, such as super-resolution [36], underwater image enhancement [37]–[39], etc., which can also be referred to for low-light image enhancement tasks.

C. Generative LLIE methods

Generative methods, known for their exceptional perceptual quality, are adept at producing high-frequency details reminiscent of the input low-light images [2], [30]. These methods differ mainly in their underlying generative models and learning principles. For instance, EnlightenGAN integrates attention mechanisms with image-related regularization in a GAN-based approach for effective LLIE [30]. A GAN-based approach under weak supervision is employed for cross-image disentanglement for low-light enhancement [17]. Yin *et al*. proposed CLEDiff, a controllable light enhancement diffusion model, offering both enhancement and region-specific controllability [1]. Normalizing flow models have also been utilized, as demonstrated by LLFlow [40]. Despite their effectiveness, GANs face challenges like training instability and artifact introduction, while normalizing flows have limitations in their expressive capacity due to restrictive invertible building blocks.

1) Diffusion models: DMs have recently revolutionized image generation. The genetic framework of DMs includes three alternative formulations: Denoising Diffusion Probabilistic Models (DDPMs) [31], Stochastic Differential Equations (SDE) [41], and Noise Conditional Score Networks (NCSN) [42]. DDPMs are inspired by non-equilibrium thermodynamics, consisting of a noise-added diffusion process and a noise-removal-based reverse process. NCSN models focus on score-based generative modeling for denoising and image enhancement. SDE-based models generalize these concepts through forward and reverse stochastic differential equations [43]. Existing variants demonstrate the versatility of DMs in addressing various computer vision tasks [44]–[48]. For example, WeatherDiff, a patch-based diffusion model, was developed for image restoration in adverse weather conditions [7], and ShadowDiffusion was proposed for image shadow removal [49]. Luo *et al*. designed a latent diffusion model for low-resolution latent space diffusion [50].

2) Diffusion-based LLIE: Focusing on diffusion-based LLIE, various approaches have been developed [1]–[3]. These models enhance images by employing a noise estimation network for the reverse process. The basic DDPMs [31] focus on using a noise estimation network for supervised reverse process learning, but lacks spatial adaptation, potentially failing to preserve fine details in complex textures. WeatherDiff [7] is designed for image restoration in adverse weather conditions, which can also applied to LLIE. CLEDiff [1] introduces a controllable light enhancement diffusion model that offers region-specific controllability. DiffLL [3] employs a wavelet-based conditional diffusion model to enhance low-light images using wavelet transformation; however, it does not address non-uniform noise properties effectively due to its global approach to noise estimation. PyDiff [2] enhances low-light images by progressively increasing resolution and globally correcting degradation. By using existing LLIE methods, a diffusion-based post-processing framework was proposed [51]. Through integration with the image degradation and priors, a diffusion-based LLIE method (LLDiffusion) was designed [52]. These different variants of DMs show the promising prospects in LLIE applications. In contrast, our proposed diffusion-based

TABLE I
COMPARISON OF PREVIOUS DIFFUSION-BASED LLIE METHODS AND THE PROPOSED ONE.

Method	Pixelwise reconstruction	Locality-based brightness adaptability	Scale-adaptive sampling
WeatherDiff [7]	✗	✗	✓
CLEDiff [1]	✗	✓	✗
DiffLL [3]	✗	✗	✗
PyDiff [2]	✗	✗	✗
Ours	✓	✓	✓

LLIE method uniquely integrates a pixelwise reconstruction loss to ensure detailed preservation, employs content and region-aware attention mechanisms to improve locality-based brightness adaptability, and uses a scale-adaptive sampling scheme to enhance robustness to noise. Table I compares our proposed methods with previous works.

III. METHOD

To model the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ for image enhancement tasks involving a one-to-many mapping inherently, we focus on learning a parametric approximation of $p(\boldsymbol{y}|\boldsymbol{x})$ through a stochastic iterative process that maps a low-light image $\boldsymbol{x}$ to a normal-light image $\boldsymbol{y}$. Our approach employs conditional patch-based DDPMs [7], which learn a Markov Chain to gradually convert Gaussian noise into the data distribution of target images y, conditioned on input images $\boldsymbol{x}$. This process operates locally on patches from input images, and it merges the results optimally.

A crucial aspect of designing conditional diffusion methods for LLIE is to harness the ability of DMs to generate perceptually realistic high-frequency details while also faithfully recovering fine details inherent in the input low-light images. To this end, we introduce Detail-Preserving Diffusion Models (DePDiff) for realistic and faithful low-light image enhancement. DePDiff is trained in a patch-wise, scale-agnostic manner with reverse diffusion-based reconstruction loss (left side of Fig.2), and employs smoothed noise estimation for overlapping patches using a multiscale ensemble scheme (right side of Fig.2). We also propose a network architecture with simplified channel-spatial attention, CRANet, tailored for content- and region-aware noise estimation (Fig. 3). The following sections detail the technical aspects of our method.

A. Detail-preserving diffusion models

Considering an arbitrary-sized ground truth normal-light image $\boldsymbol{y}$ and its corresponding low-light image $\boldsymbol{x}$, we define $\boldsymbol{y}^{(i)} = \mathrm{Crop}(\boldsymbol{P}^{(i)} \circ \boldsymbol{y})$ and $\boldsymbol{x}^{(i)} = \mathrm{Crop}(\boldsymbol{P}^{(i)} \circ \boldsymbol{x})$ as $p \times p$ patches from the training image pair $(\boldsymbol{x}, \boldsymbol{y})$. Here, $\boldsymbol{P}^{(i)}$ is a binary mask matrix indicating the i-th patch location, and $\mathrm{Crop}(\cdot)$ extracts the specified patch. DePDiff is built on conditional patch-based DDPMs [7], generating a target image patch in T diffusion time steps from a starting point of pure noise $\boldsymbol{y}_T^{(i)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The model iteratively refines the output image to eventually achieve $\boldsymbol{y}_0 \sim p(\boldsymbol{y}|\boldsymbol{x})$ through learned conditional distributions $p_\theta(\boldsymbol{y}_{t-1}^{(i)}|\boldsymbol{y}_t^{(i)}, \boldsymbol{x}^{(i)})$. For the sake of clarity, the patch location index i is omitted in this subsection.

1) Forward diffusion process: The forward diffusion process incrementally adds Gaussian noise to the output $\boldsymbol{y}_0$ according to a variance schedule $\beta_1, \cdots, \beta_T$, formulated as $q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1})$. This process, denoted by $q(\boldsymbol{y}_{1:T}|\boldsymbol{y}_0)$, can be represented as a Markov chain:

$$q(\boldsymbol{y}_{1:T}|\boldsymbol{y}_0) = \prod_{t=1}^{T} q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}), \tag{1}$$

$$q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}) = \mathcal{N}(\boldsymbol{y}_t; \sqrt{1-\beta_t}\boldsymbol{y}_{t-1}, \beta_t \boldsymbol{I}). \tag{2}$$

With $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{j=0}^{t} \alpha_j$, the state $\boldsymbol{y}_t$ at time step t is given by:

$$q(\boldsymbol{y}_t|\boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{y}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{3}$$

which also can be expressed in closed form:

$$\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, \tag{4}$$

with $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ as noise from a normal distribution.

Utilizing the denoising diffusion implicit model (DDIM) [7], [55], we adopt a non-Markovian forward diffusion process for deterministic sampling acceleration. Implicit sampling exploits a generalized non-Markovian forward process formulation:

$$\begin{aligned} q(\boldsymbol{y}_{1:T}|\boldsymbol{y}_0) &= q(\boldsymbol{y}_t|\boldsymbol{y}_0) \prod_{t=2}^{T} q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{y}_0), \\ q_\lambda(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{y}_0) &= \mathcal{N}(\boldsymbol{y}_{t-1}; \widetilde{\boldsymbol{\mu}}_t(\boldsymbol{y}_t, \boldsymbol{y}_0, t), \lambda_t^2 \boldsymbol{I}), \end{aligned} \tag{5}$$

with the mean value $\widetilde{\boldsymbol{\mu}}_t(\boldsymbol{y}_t, \boldsymbol{y}_0, t)$ derived as:

$$\widetilde{\boldsymbol{\mu}}_t(\boldsymbol{y}_t, \boldsymbol{y}_0, t) = \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_{t-1}-\lambda_t^2}\boldsymbol{\epsilon}_t. \tag{6}$$

When λ_t^2 is expressed by:

$$\lambda_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \tag{7}$$

the diffusion process formulated by Eq. (5) can not only becomes Markov, but also maintain the same training objective as the diffusion process formulated by Eq. (2). According to Eq. (4) and Eq. (6), the mean value $\widetilde{\boldsymbol{\mu}}_t(\boldsymbol{y}_t, \boldsymbol{y}_0, t)$ can be finally derived as:

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}_t(\boldsymbol{y}_t, \boldsymbol{y}_0, t) =& \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\boldsymbol{y}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t}{\sqrt{\bar{\alpha}_t}}\right) \\ &+ \sqrt{1-\bar{\alpha}_{t-1}-\lambda_t^2}\boldsymbol{\epsilon}_t. \end{aligned} \tag{8}$$

A deterministic implicit sampling approach can be achieved by setting $\lambda_t^2 = 0$ [7], [55], which, following the generation of an initial $\boldsymbol{y}_T$ from normal distribution, renders subsequent sampling deterministic.

2) Reverse diffusion process: Our model reverses the Gaussian diffusion process to regenerate the target image $\boldsymbol{y}_0$ through a reverse Markov chain conditioned on $\boldsymbol{x}$. This process involves iteratively reconstructing the signal from noise, with the aim of converting the diffusive noise back to $\boldsymbol{y}_0$ using a noise estimator $\boldsymbol{\epsilon}_\theta$. The reverse process of our model is conditioned on the patch-based low-light image $\boldsymbol{x}$, in contrast to the reverse process that uses unconditioned noise alone [31]. The conditioning on $\boldsymbol{x}$ leverages the general features of the

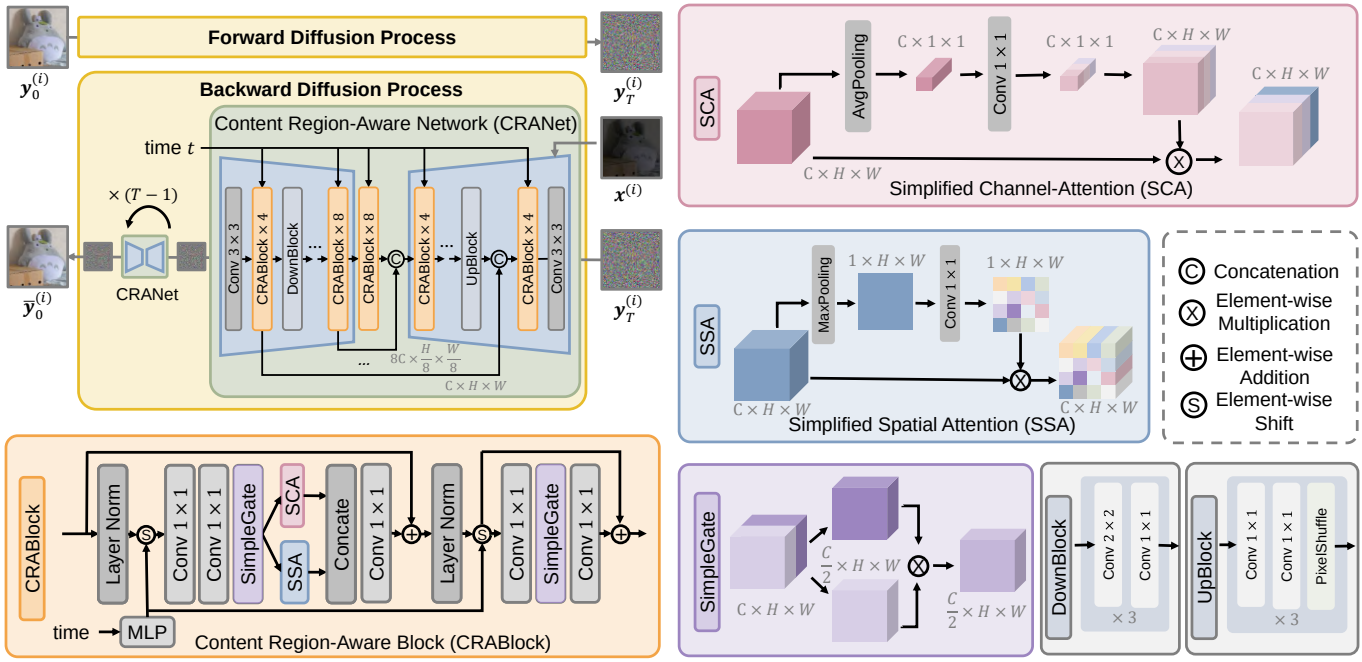


Fig. 3. Illustration of the proposed CRANet. Built upon U-Net [53], it integrates the light configuration from [54] by omitting nonlinear activation functions to reduce computational overhead. The Simplified Spatial Attention (SSA) module is designed for spatial adaptability. This is complemented by Simplified Channel-Attention (SCA), together enhancing the noise estimation process by providing focused guidance from both spatial and channel dimensions. Furthermore, it incorporates a multilayer perceptron (MLP) for efficient time embedding.

low-light image to improve image quality, providing a more detailed and fine-grained learning approach for better image enhancement. Specifically, the reverse diffusion process begins with an initial value $p(\boldsymbol{y}_T) = \mathcal{N}(\boldsymbol{y}_T; \boldsymbol{0}, \boldsymbol{I})$. We define the conditional reverse process $p_\theta(\boldsymbol{y}_{0:T}|\boldsymbol{x})$ as a Markov chain with learned Gaussian transitions:

$$p_\theta(\boldsymbol{y}_{0:T}|\boldsymbol{x}) = p(\boldsymbol{y}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}, t), \tag{9}$$

$$\begin{aligned} &p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}, t) = \\ &\mathcal{N}(\boldsymbol{y}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t)), \end{aligned} \tag{10}$$

For simplicity, $\boldsymbol{\Sigma}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t) = \sigma_t^2 \boldsymbol{I}$, and $\boldsymbol{\mu}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t)$ are parameterized by a neural network with parameters θ.

To prevent the generation of differing normal-light image patches for overlapping grid cells during conditional reverse sampling from neighboring overlapping low-light image patches, we adopt a strategy that utilizes the mean estimated noise for each pixel across overlapping patch regions at any given denoising time step t. This method ensures enhanced fidelity throughout the reverse sampling process, harmonizing the contributions from all adjacent patches. Specifically, at each time step t during sampling, we calculate the additive noise for every overlapping patch location i using $\boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t^{(i)}, \boldsymbol{x}^{(i)}, t)$. Subsequently, these overlapping noise estimates at their corresponding patch locations are aggregated into a matrix $\widetilde{\boldsymbol{\epsilon}}_t$ of the same size as the entire low-light image $\boldsymbol{x}$, which is then normalized based on the count of estimates received per pixel. Using DDIM for accelerated deterministic sampling in this reverse process by setting $\lambda_t^2 = 0$ in Eq. (8), sample $\boldsymbol{y}_{t-1} \sim p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}, t)$ is formulated as follows:

$$\boldsymbol{y}_{t-1} = \sqrt{\overline{\alpha}_{t-1}}\left(\frac{\boldsymbol{y}_t - \sqrt{1-\overline{\alpha}_t}\widetilde{\boldsymbol{\epsilon}}_t}{\sqrt{\overline{\alpha}_t}}\right) + \sqrt{1-\overline{\alpha}_{t-1}}\widetilde{\boldsymbol{\epsilon}}_t, \tag{11}$$

which starts from $\boldsymbol{y}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and is updated using the smoothed whole-image noise estimate $\widetilde{\boldsymbol{\epsilon}}_t$. To expedite the sampling process, we use a sub-sequence with equal intervals from the overall sequence $t_1, t_2, ..., t_S \subseteq 1, 2, ..., T$:

$$t_j = (j-1)\cdot T/S + 1, \ \ j = 1, ..., S, \tag{12}$$

where t_1 denotes the final step of reverse sampling.

3) Optimizing with reverse diffusion-based reconstruction: Unlike other generative models such as GANs, DMs prioritize optimizing denoising loss rather than predicting normal-light images against ground truth. For a given diffusion process under implicit deterministic sampling, the noise $\boldsymbol{\epsilon}_t$ added at each diffusion step t is deterministic, which can then be used to train the noise estimation network $\boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t)$. The denoising loss can then be realized from optimizing the variational bound on negative data log likelihood $\mathbb{E}_{q(\boldsymbol{y}_0)}[-\log p_\theta(\boldsymbol{y}_0|\boldsymbol{y}_t, \boldsymbol{x})]$, which is equivalent to optimizing the objective $\mathcal{L}_{\text{diff}}$:

$$\mathcal{L}_{\text{diff}} = \|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t)\|_2^2, \tag{13}$$

where

$$\boldsymbol{y}_t = \sqrt{\overline{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}_t. \tag{14}$$

DMs trained with such denoising loss excel in data distribution fitting and realistic enhancements; however, their capability in faithfully recovering fine details may be limited. To address this, we introduce a reconstruction loss $\mathcal{L}_{\text{rec}}$ between

Algorithm 1 Training of DePDiff

Input: Dataset containing low-light images $\boldsymbol{x}$ and normal-light images $\boldsymbol{y}_0$.

repeat
  Randomly sample a binary patch mask $\boldsymbol{P}^{(i)}$
  $\boldsymbol{y}_0^{(i)} = \mathrm{Crop}(\boldsymbol{P}^{(i)} \circ \boldsymbol{y}_0), \boldsymbol{x}^{(i)} = \mathrm{Crop}(\boldsymbol{P}^{(i)} \circ \boldsymbol{x})$
  $t \in \mathrm{Uniform}\{1, \cdots, T\}$
  $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  Take gradient descent step on $\nabla_\theta \mathcal{L}_{\text{train}}$ using Eq. (17)
until converged
return θ

the enhanced image $\overline{\boldsymbol{y}}_0$ and the ground truth $\boldsymbol{y}_0$. The enhanced image $\overline{\boldsymbol{y}}_0$ is derived directly from $\boldsymbol{y}_t$ and the learned noise estimator $\boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t)$ in the reverse diffusion process:

$$\overline{\boldsymbol{y}}_0 = \frac{\boldsymbol{y}_t - \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t)}{\sqrt{\overline{\alpha}_t}}. \tag{15}$$

This formulation allows direct evaluation of the difference between enhanced images and original normal-light ones:

$$\mathcal{L}_{\text{rec}} = \|\overline{\boldsymbol{y}}_0 - \boldsymbol{y}_0\|_2^2, \tag{16}$$

optimizing the noise estimator in an image enhancement-oriented supervised manner.

Algorithm 1 outlines the training procedure of DePDiff with reverse diffusion-based reconstruction. The DePDiff optimizes both the denoising loss and the reverse diffusion-based reconstruction, making it more effective for LLIE. The overall training loss is a weighted sum of $\mathcal{L}_{\text{diff}}$ and $\mathcal{L}_{\text{rec}}$:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{diff}} + \gamma \mathcal{L}_{\text{rec}}. \tag{17}$$

where γ is a weighted coefficient.

B. Content and region-aware network for noise estimation

The inherent challenge in enhancing low-light images stems from their highly variable brightness distribution, which demands a method attuned to content- and region-specific characteristics for accurate differentiation between noise and fine details. To address this, our network design incorporates channel attention mechanisms. These mechanisms are adept at selectively amplifying task-relevant features, particularly useful in enhancing underexposed areas and recovering details obscured in shadows. By weighting the channels according to their importance, the network can focus more on features that contribute to enhancing under-exposed areas or details lost in shadows. It helps in understanding the global context of the image, which is crucial for content-aware processing, ensuring that the enhancements are uniform and coherent across the entire image. Furthermore, spatial attention mechanisms are integrated to enable the network to focus on specific image regions that require enhanced processing. In low-light conditions, this translates to the network dedicating more resources to darker or shadowed areas that need brightness adjustments, while conservatively handling well-lit sections. This approach is particularly beneficial for identifying noisy regions and applying targeted noise reduction, thereby preserving the integrity of smoother areas in the image. Motivated by these considerations, we have tailored the U-Net architecture within the DDPMs framework [31] to include an activation-free structure and a streamlined channel-spatial attention mechanism [54]. This adaptation results in our novel content and region-aware network, CRANet, for noise estimation in DMs, specifically designed to tackle the unique challenges posed by low-light image enhancement.

As depicted in Fig. 3, CRANet follows the main architecture of U-Net [53] to fuse multiscale feature information for noise estimation. Unlike basic U-Net, CRANet incorporates the light configuration from [54], removing nonlinear activation functions throughout the network to reduce computational cost. Unlike [54], it uses a multilayer perceptron (MLP) for time embedding and introduces a new simplified spatial attention (SSA) mechanism for spatial adaption, combined with the simplified channel-attention (SCA) to guide the noise estimation process from both spatial and channel perspectives.

After training CRANet as the noise estimation network $\boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t^{(i)}, \boldsymbol{x}^{(i)}, t)$, it is used to denoize and enhance each image patch. The denoised patches are then combined to construct the whole image, using the mean estimated noise for pixels within overlapping patches to perform reverse sampling for the entire image enhancement.

C. Multiscale ensemble scheme

The varying noise properties and natural image patch scales in low-light images necessitate adaptive receptive fields. Traditional training and inference schemes in DMs, which often focus on entire images, are limited in their ability to recover details across different scales and fail to capture the diversity in real-world textures and patterns. To address this, we utilize a multiscale ensemble scheme, allowing for the effective aggregation of details from various scales and enhancing the overall image quality, as shown in Fig. 4.

1) Multiscale ensemble-based image fusion: The core of multiscale image fusion involves performing a weighted sum on images generated at different patch sizes, thereby achieving image enhancement. We employ a bagging-based ensemble scheme for this purpose. Suppose that there are N image patches of different sizes $(p_1, p_2, ..., p_N)$ extracted from a low-light image $\boldsymbol{x}$. Pre-trained diffusion models are used on the training set to generate N types of enhanced collections. For the low-light image $\boldsymbol{x}$, the corresponding enhanced images using N different patch sizes are denoted as $\overline{\boldsymbol{x}}^{(1)}, \overline{\boldsymbol{x}}^{(2)}, \cdots, \overline{\boldsymbol{x}}^{(N)}$. The enhanced images in the training set are randomly selected with a certain probability σ to form a bootstrap sample that consists of images using N different patch sizes $(p_1, p_2, ..., p_N)$.

After obtaining M bootstrap samples, they are used to independently train M base models, simplified by using weighted sum operation. Each bootstrap sample is used to train a base model, essentially learning optimal weighting coefficients η. Without loss of generality, the m-th $(m \in 1, 2, \cdots, M)$

Algorithm 2 Multiscale Ensemble

Input: Dataset $\mathbb{D}$ containing low-light image $\boldsymbol{x}$, pretrained $\boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t)$, number of implicit sampling steps S, N sampling scales, number of bootstrap sample M, dictionary of D overlapping patch locations.

for $\boldsymbol{x} \in \mathbb{D}$ **do**
  for $n = 1 ..., N$ **do**
    $\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
    for $j = S ..., 1$ **do**
      $t = (j-1) \cdot T/S + 1$
      $t_{\text{next}} = (j-2) \cdot T/S + 1$ if $j > 1$ else 0
      $\boldsymbol{M} = \mathbf{0}$, $\widetilde{\boldsymbol{\epsilon}}_t = \mathbf{0}$
      for $i = 1 ..., D$ **do**
        $\boldsymbol{x}^{(i)} = \text{Crop}(\boldsymbol{P}^{(i)} \circ \boldsymbol{x})$
        $\boldsymbol{y}_t^{(i)} = \text{Crop}(\boldsymbol{P}^{(i)} \circ \boldsymbol{y}_t)$
        $\widetilde{\boldsymbol{\epsilon}}_t = \widetilde{\boldsymbol{\epsilon}} + \boldsymbol{P}^{(i)} \circ \boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t^{(i)}, \boldsymbol{x}^{(i)}, t)$
        $\boldsymbol{M} = \boldsymbol{M} + \boldsymbol{P}^{(i)}$
      end for
      $\widetilde{\boldsymbol{\epsilon}}_t = \widetilde{\boldsymbol{\epsilon}}_t \oslash \boldsymbol{M}$
      compute $\boldsymbol{y}_{t_{\text{next}}}$ using Eq. (11)
    end for
    $\overline{\boldsymbol{x}}^{(n)} = \widetilde{\boldsymbol{y}}_0$
  end for
end for
for $m = 1 ..., M$ **do**
  create bootstrap sample $\mathbb{D}_m$ containing $\{\overline{\boldsymbol{x}}^{(n)}\}_{n=1}^N$
  optimize $\{\eta_{m,n}\}_{n=1}^N$ using Eq. (18) and Eq. (19)
  compute $\widetilde{\boldsymbol{x}}^{(m)}$ using Eq. (18)
end for
compute $\widetilde{\boldsymbol{y}}_0$ from $\widetilde{\boldsymbol{x}}^{(m)}$ using Eq. (20) and Eq. (21)
return $\widetilde{\boldsymbol{y}}_0$

base model is used for illustration. The normal-light images obtained by the m-th base model are expressed as follows:

$$\widetilde{\boldsymbol{x}}^{(m)} = \eta_{m,1}\overline{\boldsymbol{x}}^{(1)} + \eta_{m,2}\overline{\boldsymbol{x}}^{(2)} + ... + \eta_{m,N}\overline{\boldsymbol{x}}^{(N)}. \tag{18}$$

The weighted coefficients η are iteratively computed by optimizing the loss function $\mathcal{L}_{\text{mul}}$:

$$\mathcal{L}_{\text{mul}} = \|\boldsymbol{y}_0 - \widetilde{\boldsymbol{x}}^{(m)}\|_2^2. \tag{19}$$

Compared to randomly generated parameters, the parameters predicted by our lightweight network avoid significant random errors. Therefore, during testing, samples of different sizes are input into the weak learners, and each set of combination weights weights the N images to produce a fused image. The M weight generators output M sets of combination weights, resulting in M fused images. The process is illustrated in Fig. 4. Taking advantage of patch size-agnostic image enhancement, images of different patch sizes can be generated in an inexpensive manner. The designed learning strategy can then effectively utilize the information from the different patch size-based generated images to learn and fuse details from different scales, ultimately achieve faithful enhancement.

2) Image histogram difference-based aggregation prediction: Selecting the final enhanced image from the M outputs of the base models is achieved by comparing the histogram

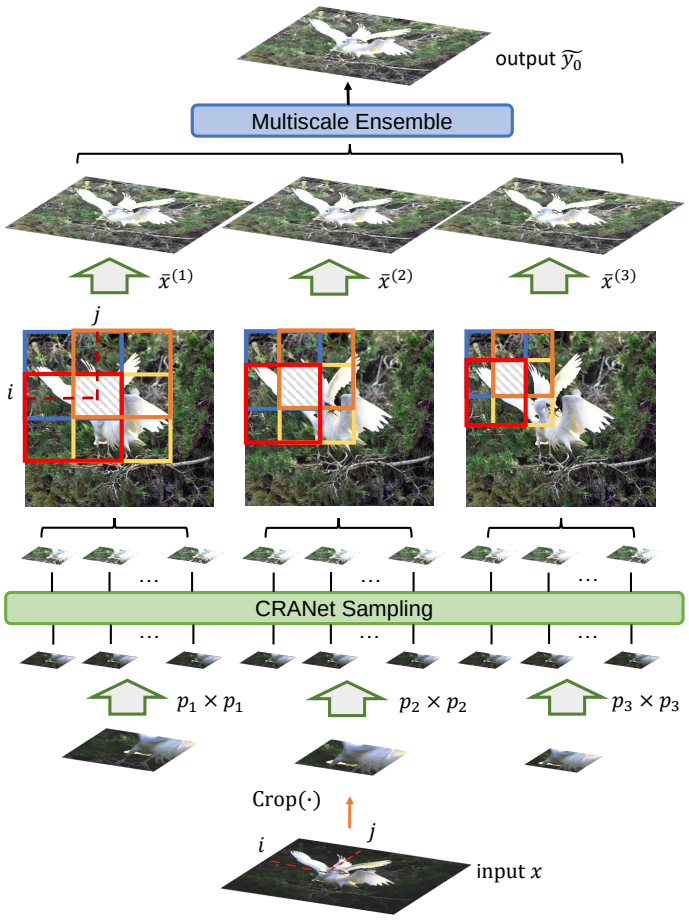


Fig. 4. Given a low-light image, the image is cropped using multiple patch sizes to obtain several sets of patches. These patches are fed into CRANet $\boldsymbol{\epsilon}_\theta$ for sampling the corresponding enhanced patches. Then each set of patches is combined in a sliding window way with averaging the overlapping areas, to output the same size as the input. Finally, these enhanced images are integrated by the multiscale ensemble scheme to produce the optimal final output.

differences between these images and the original low-light image $\boldsymbol{x}$. The histogram represents the distribution of pixel intensities in an image and serves as an efficient and invariant measure for assessing the light intensity of different images. The image among $\widetilde{\boldsymbol{x}}^{(1)}, \widetilde{\boldsymbol{x}}^{(2)}, ..., \widetilde{\boldsymbol{x}}^{(M)}$ with the largest histogram difference from the low-light $\boldsymbol{x}$ is selected as the final enhanced image $\widetilde{\boldsymbol{y}}_0$:

$$\widetilde{\boldsymbol{y}}_0 = \widetilde{\boldsymbol{x}}^{(m^*)}, \tag{20}$$

with

$$m^* = \underset{m \in [1,M]}{\text{argmax}}\, \Delta(\text{Hist}(\widetilde{\boldsymbol{x}}^{(m)}), \text{Hist}(\boldsymbol{x})), \tag{21}$$

where $\text{Hist}(\widetilde{\boldsymbol{x}}^{(m)})$ and $\text{Hist}(\boldsymbol{x}))$ denote the histograms of $\widetilde{\boldsymbol{x}}^{(m)}$ and $\boldsymbol{x}$, respectively, and $\Delta(.)$ calculates the histogram difference. Algorithm 2 outlines the multiscale ensemble scheme.

IV. Experiments

Our experimental setup is divided into three parts. Firstly, we evaluate the efficacy of the proposed method by training and testing it on the same dataset and comparing it with several state-of-the-art methods. The second part involves cross-dataset evaluations, where the proposed method is trained on

(a) Input (b) LLFlow [40] (c) WeatherDiff [7] (d) CLEDiff [1]

(e) PyDiff [2] (f) DiffLL [3] (g) Ours (h) Ground truth

Fig. 5. Qualitative comparison results of several state-of-the-art LLIE methods and the proposed one.

TABLE II
QUANTITATIVE RESULTS ON THE LOL DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD AND THE SECOND ONES ARE UNDERLINED. * REFERS TO THE METHOD THAT IS RETRAINED ON GIVEN DATASET. ↑(↓) MEANS HIGHER (LOWER) IS BETTER.

Method	Source	PSNR↑	SSIM↑	LPIPS↓
KinD [56]	MM'19	19.19	0.815	0.170
Zero-DCE [34]	CVPR'20	14.73	0.509	0.401
EnlightenGAN [30]	TIP'21	17.48	0.650	0.320
KinD++ [57]	IJCV'21	20.90	0.823	0.164
Bread [58]	IJCV'22	22.96	0.838	0.160
Uformer [59]	CVPR'22	18.96	0.778	0.505
Restormer [60]	CVPR'22	22.17	0.819	0.149
IAT [14]	BMVC'22	23.38	0.861	0.216
HWMNet [13]	ICIP'22	24.24	0.922	0.113
LLFlow [40]	AAAI'22	24.99	0.923	0.116
SMG-LLIE [61]	CVPR'23	23.85	0.893	0.131
PairLIE [62]	CVPR'23	19.51	0.736	0.248
NeRCo [63]	ICCV'23	19.84	0.771	0.315
RetinexFormer [64]	ICCV'23	25.16	0.845	0.129
RQ-LLIE [65]	ICCV'23	25.24	0.855	0.250
STGNet [66]	TCSVT'23	22.03	0.838	<u>0.101</u>
WeatherDiff [7]	TPAMI'23	19.73	0.908	0.112
CLEDiff [1]	MM'23	25.50	0.907	0.163
DiffLL [3]	TOG'23	21.84	0.871	0.201
PyDiff [2]	IJCAI'23	<u>27.09</u>	<u>0.930</u>	0.109
Ours	-	**27.44**	**0.939**	**0.085**

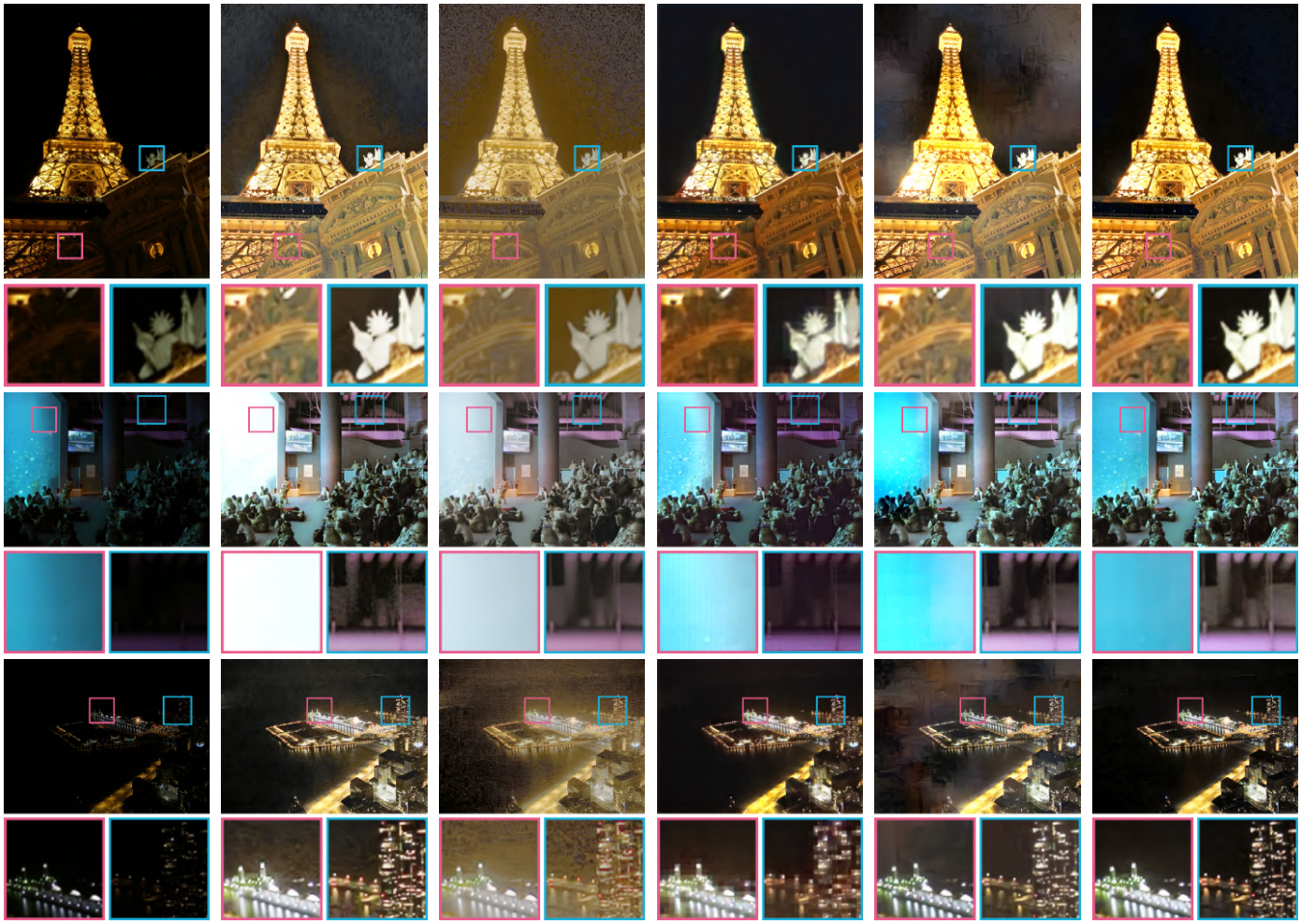
one dataset and tested on others. Lastly, ablation studies are conducted to underscore the significance of each component in the proposed method.

A. Experimental settings

1) Datasets: A total of eight datasets are used in experiments, including LOL, LOL-v1, LOL-v2 real, NPE, DICM, MEF, LIME, and VV datasets [33], [66], [67]. Except for LOL, LOL-v1 and LOL v2-real datasets [33], [67], other datasets have no normal-light images corresponding to the low-light ones. The training and testing sets of LOL and LOL-v1 are split in proportion to 485:15 and 689:100. The DePDiff model is trained on the LOL and LOL-v1 datasets. For the ensemble scheme experiment, the models are trained only on the LOL dataset but tested on six real-world datasets: DICM, MEF, NPE, LIME, VV, and LOL-v2 real. The ablation experiment is conducted on the LOL dataset.

2) Evaluation metrics: For the paired datasets, namely LOL, LOL-v1 and LOL-v2 real, we use the commonly used metrics, including peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and learned perceptual image patch similarity (LPIPS). For the DICM, MEF, NPE, LIME and VV datasets without ground truth, we select the commonly used naturalness image quality evaluator (NIQE). Theoretically, the larger the values of PSNR and SSIM, the better the corresponding results. Conversely, the smaller the values of LPIPS and NIQE, the better the corresponding result.

3) Implementation details: In all experiments, the training patch size is set to 64×64 and the batch size is set to 4. For LOL dataset, the number of iterations is set to 800,000 for modeling training. For LOL-v1 dataset, the corresponding number of iteration is 600,000. The Adam [68] is used as the optimizer with a fixed learning rate of 0.00003. The time step T of diffusion is set to 1000 for the training stage, and the implicit sampling step S is set to 20 for the inference stage. The multiscale ensemble patch sizes include 64×64, 96×96, 128×128, 160×160, 192×192, 225×225 and 256×256. All datasets are implemented using PyTorch and run on a single



(a) Input (b) LLFlow [40] (c) PyDiff [2] (d) DiffLL [3] (e) WeatherDiff [7] (f) Ours

Fig. 6. Qualitative comparison results of several state-of-the-art LLIE methods and the proposed one on real-world datasets without ground truth.

TABLE III
QUANTITATIVE RESULTS ON THE LOL-V1 DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD AND THE SECOND ONES ARE UNDERLINED. ↑(↓) MEANS HIGHER (LOWER) IS BETTER.

Method	Source	PSNR↑	SSIM↑	LPIPS↓
KinD [56]	MM'19	22.15	0.853	0.257
Zero-DCE [34]	CVPR'20	20.54	0.778	0.331
EnlightenGAN [30]	TIP'21	17.60	0.653	0.372
Uformer [59]	CVPR'22	19.00	0.741	0.354
Restormer [60]	CVPR'22	20.61	0.797	0.288
RUAS [35]	CVPR'22	16.40	0.503	0.364
SNRNet [69]	CVPR'22	24.61	0.842	0.233
IAT [14]	BMVC'22	21.25	0.844	0.255
HWMNet [13]	ICIP'22	19.62	0.862	0.271
LLFlow [40]	AAAI'22	26.02	**0.926**	<u>0.100</u>
SMG-LLIE [61]	CVPR'23	24.03	0.878	0.144
PairLIE [62]	CVPR'23	24.02	0.803	0.118
NeRCo [63]	ICCV'23	25.17	0.833	0.160
RetinexFormer [64]	ICCV'23	**27.69**	0.856	0.166
RQ-LLIE [65]	ICCV'23	22.37	0.854	0.228
WeatherDiff [7]	TPAMI'23	17.91	0.811	0.272
DiffLL [3]	TOG'23	26.33	0.845	0.217
Ours	-	<u>26.52</u>	<u>0.922</u>	**0.098**

NVIDIA GTX 3090 Ti GPU. The source code would be available from https://github.com/CSYanH/DePDiff.

4) Comparing methods: We select various state-of-the-art learning-based methods from the past 3-5 years for comparison, divided into regression LLIE methods and generative LLIE methods. Regression LLIE methods consists of CNN-based and Transformer-based methods, including IAT [14], HWMNet [13], Zero-DCE [34], DRBN [29], RUAS [35], KinD [56], KinD++ [57], STGNet [66], Bread [58], SNR-Net [69], Zero-DCE++ [70], Restormer [60], Uformer [59], RetinexFormer [64], SMG-LLIE [61] and PairLIE [62]. As for generative LLIE methods, resently proposed GAN-based, normalizing flow-based, VAE-based methods and diffusion-based methods are used for comparison, including WeatherDiff [7], EnlightenGAN [30], PyDiff [2], CLEDiff [1], DiffLL [3], LLFlow [40], NeRCo [63] and RQ-LLIE [65]. All these methods would be used to conduct the first two parts of experiments for verifying the effectiveness of the proposed method within and across datasets. For convinced comparison, all results are directly from published works or tested based on the source codes of published works.

B. Performance comparisons

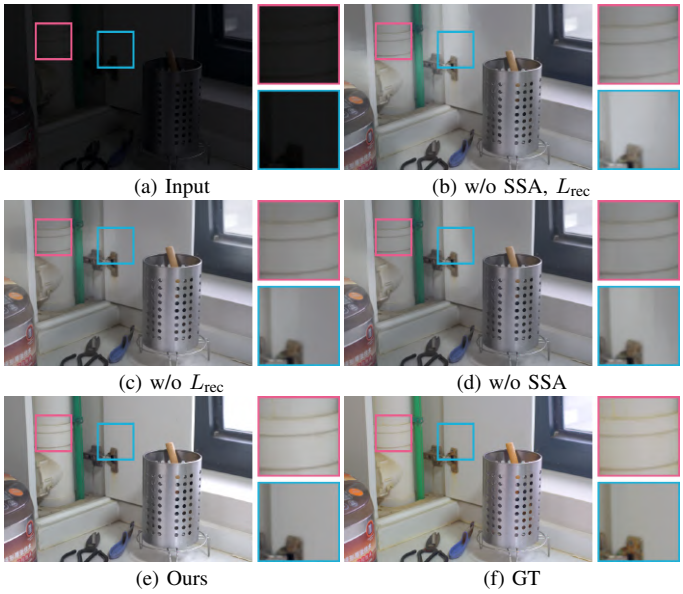
1) Quantitative results: The experimental results of the first part are compared in Table II and III. From the point of view

TABLE IV
QUANTITATIVE RESULTS ON THE LOL-V2 REAL DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD AND THE SECOND ONES ARE UNDERLINED. ↑(↓) MEANS HIGHER (LOWER) IS BETTER.

Method	Source	PSNR↑	SSIM↑	LPIPS↓
KinD [56]	MM'19	24.05	0.917	0.1140
Zero-DCE [34]	CVPR'20	18.05	0.580	0.352
EnlightenGAN [30]	TIP'21	18.67	0.678	0.364
KinD++ [57]	IJCV'21	22.21	0.885	0.174
Bread [58]	IJCV'22	23.69	0.912	0.155
Uformer [59]	CVPR'22	18.44	0.759	0.347
Restormer [60]	CVPR'22	24.91	0.851	0.264
RUAS [35]	CVPR'22	15.35	0.495	0.395
SNRNet [69]	CVPR'22	21.48	0.849	0.237
IAT [14]	BMVC'22	26.45	0.895	0.170
HWMNet [13]	ICIP'22	30.29	0.937	0.080
LLFlow [40]	AAAI'22	28.35	0.945	0.076
SMG-LLIE [61]	CVPR'23	25.62	0.905	0.131
PairLIE [62]	CVPR'23	19.88	0.841	0.234
NeRCo [63]	ICCV'23	15.67	0.684	0.409
RetinexFormer [64]	ICCV'23	28.99	0.939	0.106
RQ-LLIE [65]	ICCV'23	25.94	0.941	0.219
WeatherDiff [7]	TPAMI'23	15.86	0.801	0.272
DiffLL [3]	TOG'23	28.85	0.876	0.207
PyDiff [2]	IJCAI'23	<u>33.40</u>	**0.949**	**0.065**
Ours	-	**33.87**	<u>0.947</u>	<u>0.067</u>

of PSNR and LPIPS, the proposed method can outperform all comparison methods. The PSNR focuses primarily on pixel-wise differences without considering perceptual factors, whereas LPIPS emphasizes perceptual similarity. Hence, the proposed method can achieve the best results in both perceptual performance and pixel-level image enhancement effects. From the SSIM point of view, the proposed method can achieve the best results in the LOL dataset, but achieve the second-best result on LOL-v1 dataset with a slight disadvantage. The possible reasons are illustrated as the following two aspects.

From the perspective of datasets and comparison methods themselves, the LOL dataset consists of real-world low-light images, whereas the LOL-v1 dataset contains both synthetic and real-world images. There may be some differences between synthetic and real images, which may lead to inconsistent image distribution. Considering that diffusion-based methods achieve image restoration by learning mapping relationship between the noise and image distributions, the inconsistency of image distributions in LOL-v1 dataset may pose additional challenges to diffusion-based methods, compared to other types of methods. From the perspective of SSIM of the proposed method, SSIM considers structural information and is less sensitive to small, perceptually insignificant distortions. However, the proposed method achieves image enhancement by dividing the whole image into smaller patches. This patch-based learning strategy can potentially affect the learning of the complete image structure, which could subsequently impact its performance on the SSIM metric. As a result, the SSIM of both the proposed method and DIffLL is not as good as that of LLFlow, which is not a diffusion-based solution. These comparative results of the different indicators show that the proposed method has the best overall performance in terms of perceptual, pixel-level, and structural details.



(a) Input (b) w/o SSA, L_{rec} (c) w/o L_{rec} (d) w/o SSA (e) Ours (f) GT

Fig. 7. Visual comparison results of the proposed method with or without the SSA module and the proposed reverse diffusion-based reconstruction loss.

2) Qualitative results: The visual results are compared in Fig. 1, Fig. 5 and Fig. 6. Compared to the quantitative results, these qualitative visual results can intuitively demonstrate the effectiveness and practicality of the proposed method. In practical applications, different low-light conditions and environments can increase the difficulty of LLIE, such as overexposure and underexposure in low-light images, or saturated pixel areas caused by nighttime light sources. As shown in Fig. 1, the input low-light image contains both overexposed and underexposed regions. All the CNN-based HWMNet, Transformer-based IAT, normalizing flow-based LLFlow, and diffusion-based PyDiff cannot effectively handle either the overexposed or underexposed regions. However, the proposed method can simultaneously enhance the low-light image in these regions and achieve the best visual effect. As shown in Fig. 5 and Fig. 6, the proposed method can effectively enhance low-light images regardless of whether there are pixel-saturated areas in the image or whether it has different resolution sizes.

C. Cross-dataset performance comparisons

The experimental results of the second part are summarized in Table IV and Table V. The comparative results in this part are basically comparable to those in the first part and can further demonstrate the effectiveness of the proposed method. As shown in Table IV, the proposed method can achieve the best overall performance than other comparison methods. It clearly shows that the proposed method can effectively enhance low-light image across datasets and has good generalization performance on cross-domain datasets. The experimental results in Table V demonstrate the practicality of our method in effectively handling LLIE problems in real-world scenarios without the guidance of normal-light images. It noted that NeRCo, RQ-LLIE, and SMG-LLIE were pre-trained and tested exclusively on paired training sets with fixed-size input

TABLE V
QUANTITATIVE RESULTS OF NIQE ACROSS FIVE REAL-WORLD DATASETS WITHOUT GROUND TRUTH. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD AND THE SECOND ONES ARE UNDERLINED. LOWER IS BETTER.

Method	Source	Datasets					AVG
		DICM	MEF	NPE	LIME	VV	
KinD [56]	MM'19	5.28	5.61	5.06	6.14	4.25	5.26
Zero-DCE [34]	CVPR'20	4.58	4.93	4.57	5.82	4.81	4.94
EnlightenGAN [30]	TIP'21	4.82	5.01	5.26	5.11	3.85	4.81
KinD++ [57]	IJCV'21	5.29	6.23	4.56	7.20	4.87	5.63
Bread [58]	IJCV'22	4.78	4.93	4.91	5.07	3.86	4.71
Uformer [59]	CVPR'22	11.29	35.56	37.68	14.73	11.79	22.21
Restormer [60]	CVPR'22	12.12	13.22	11.93	14.01	10.29	12.31
RUAS [35]	CVPR'22	7.31	5.44	7.20	5.32	4.99	6.05
IAT [14]	BMVC'22	7.92	4.65	4.65	4.76	3.25	5.04
HWMNet [13]	ICIP'22	5.48	4.98	**4.48**	OOM	OOM	4.98
LLFlow* [40]	AAAI'22	**4.46**	4.80	4.78	5.83	3.60	4.69
STGNet [66]	TCSVT'23	9.95	10.11	11.80	10.01	8.00	9.97
PairLIE [62]	CVPR'23	5.15	5.03	5.47	4.98	4.30	4.98
RetinexFormer [64]	ICCV'23	4.19	4.12	4.20	4.88	3.66	4.21
WeatherDiff [7]	TPAMI'23	4.75	4.57	4.68	4.62	3.38	4.40
DiffLL [3]	TOG'23	4.56	4.54	4.54	**4.34**	3.67	4.33
PyDiff [2]	IJCAI'23	5.00	4.87	5.01	OOM	OOM	4.96
Ours	-	4.47	**4.20**	4.51	4.41	**3.17**	**4.15**

TABLE VI
ABLATION STUDIES ON THE REVERSE DIFFUSION-BASED RECONSTRUCTION LOSS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. ↑(↓) MEANS HIGHER (LOWER) IS BETTER.

Backbone	Loss	PSNR↑	SSIM↑	LPIPS↓
U-Net [53]	$\mathcal{L}_{\text{diff}}$	19.74	0.908	0.113
	$\mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{rec}}(\ell_1)$	21.02	0.916	0.136
	$\mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{rec}}(\ell_2)$	21.63	0.918	0.129
CRANet	$\mathcal{L}_{\text{diff}}$	24.69	0.930	0.101
	$\mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{rec}}(\ell_1)$	24.04	0.930	0.107
	$\mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{rec}}(\ell_2)$	**26.33**	**0.936**	**0.089**

TABLE VII
ABLATION STUDIES ON THE MODEL ARCHITECTURE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. ↑(↓) MEANS HIGHER (LOWER) IS BETTER.

Backbone	Setting	PSNR↑	SSIM↑	LPIPS↓
U-Net [53]	-	19.74	0.908	0.113
NAFNet [54]	-	23.76	0.926	0.121
CRANet	w/o SSA, L_{rec}	24.75	0.931	0.101
	w/o L_{rec}	24.69	0.930	0.101
	w/o SSA	24.81	0.931	0.100
	w/ SSA, w/ L_{rec}	**26.33**	**0.936**	**0.089**

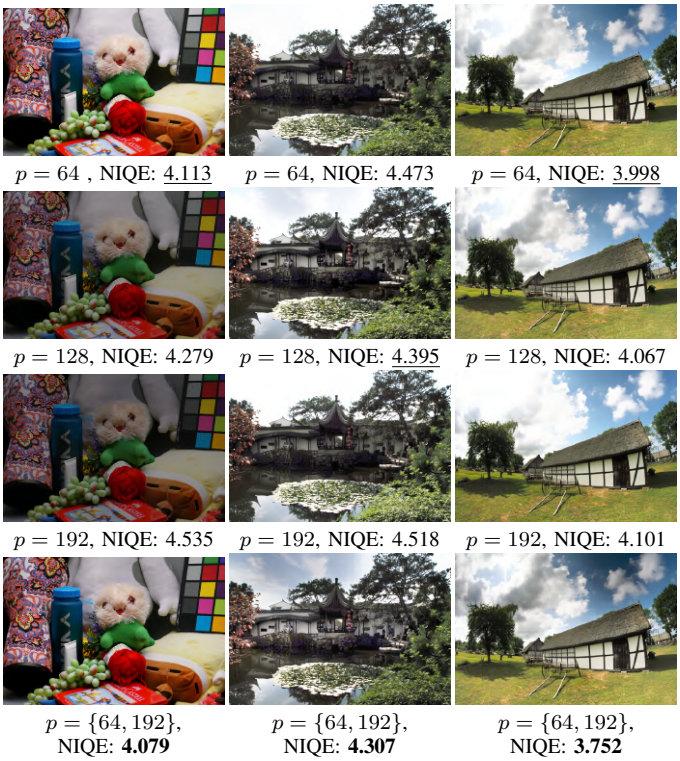


Fig. 8. Visual comparison results of sampling using different patch sizes (p). Each column represents a distinct image sampled using different patch sizes, highlighting how the optimal patch size varies for each image. Results with the best NIQE (lower is better) are highlighted in bold and the second ones are underlined. This variability underscores the necessity of employing a multiscale ensemble scheme for optimal low-light image enhancement across diverse image scenarios.

images; therefore, their performance on unpaired test sets with varying image resolutions is not included in this comparison.

D. Ablation studies

1) The effectiveness of reverse diffusion-based reconstruction loss: This ablation experiment is conducted by comparing the results of the presence and absence of the additional reverse diffusion-based reconstruction loss based on the ℓ_1-norm and ℓ_2-norm expression, respectively. According to Table VI, at least two conclusions can be drawn. First, the ℓ_1-norm-based reverse diffusion-based reconstruction loss $\mathcal{L}_{\text{rec}}$ is less effective than ℓ_2-norm-based one. Second, the ℓ_2-norm-based reverse diffusion-based reconstruction loss can achieve the best results by using the designed CRANet as the backbone of the proposed method. In summary, the reverse diffusion-based reconstruction loss is effective and highly compatible with the designed CRANet.

2) The effectiveness of the CRANet network architecture: From the perspective of model architecture, the main contributions of the proposed method lie in the design of its noise estimation network and reverse diffusion-based reconstruction loss. To verify the effectiveness of the designed architecture, ablation experiments are conducted to replace the backbone network (noise estimation network) of the proposed method with the basic (vanilla) U-Net, NAFNet, and CRANet without the designed SSA module or $\mathcal{L}_{\text{rec}}$. For fair comparison, the corresponding parameters of all methods are the same, and none of them includes the proposed post-processing module. As shown in Table VII, both the designed SSA module and $\mathcal{L}_{\text{rec}}$ can improve the model performance, while removing them can lead to a decrease.

To further verify the effectiveness of the proposed model architecture, visual comparison experiments are conducted under the condition that there is no SSA or $\mathcal{L}_{\text{rec}}$, or both. As shown in Fig. 7, when there is no $\mathcal{L}_{\text{rec}}$, conditional patch-based models would produce obvious artifacts or inconsistencies of the image. It clearly shows that the proposed reverse diffusion-

Fig. 9. Visual comparison results of the proposed method with or without multiscale ensemble. Top: The results of the proposed method using a fixed patch size. Middle: Results of the proposed method using multiscale ensemble. Bottom: Ground truth.

TABLE VIII
ABLATION STUDY ON THE MULTISCALE ENSEMBLE SCHEME. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. ↑(↓) MEANS HIGHER (LOWER) IS BETTER.

Backbone	Multiscale	PSNR↑	SSIM↑	LPIPS↓
U-Net [53]	w/o	22.04	0.923	0.105
	w/	24.73	0.933	0.088
NAFNet [54]	w/o	19.60	0.897	0.170
	w/	24.41	0.931	0.092
CRANet	w/o	26.33	0.936	0.089
	w/	**27.44**	**0.937**	**0.083**

TABLE IX
COMPUTATIONAL COMPLEXITY OF DIFFERENT METHODS.

Method	LPIPS	FLOPs (G)	Parms. (M)	Runtime (s)
KinD [56]	0.170	34.99	8.02	1.50
Zero-DCE [34]	0.401	15.59	0.08	0.01
EnlightenGAN [30]	0.320	114.35	67.80	0.34
KinD++ [57]	0.164	40.93	21.11	4.50
Bread [58]	0.160	106.96	2.15	0.10
Uformer [59]	0.505	12.00	5.29	0.50
Restormer [60]	0.149	144.25	26.13	0.12
SNRNet [69]	0.237	26.35	4.01	0.31
IAT [14]	0.216	87.21	0.09	2.50
HWMNet [13]	0.113	943.39	66.56	0.30
LLFlow [40]	0.116	358.40	17.42	0.40
SMG-LLIE [61]	0.131	92.66	19.35	0.10
PairLIE [62]	0.248	20.81	0.35	0.15
NeRCo [63]	0.315	130.70	25.80	0.34
RetinexFormer [64]	0.129	15.85	1.53	0.21
WeatherDiff [7]	0.112	726.20	109.68	15.00
DiffLL [3]	0.201	702.60	22.15	0.19
PyDiff [2]	0.109	708.68	97.19	0.23
Ours	0.085	640.40	94.01	3.80

based reconstruction loss can effectively compensate for the shortcomings of patch-based learning. However, without the assistance of SSA, the reverse diffusion-based reconstruction loss alone is not sufficient. The SSA module can help achieve fine-grained low-light enhancement at a spatial scale. With the help of direct and global pixel-level image supervision and channel-spatial attention-based learning, the proposed method can achieve smooth and consistent image effects.

Moreover, compared to the U-Net in DDPMs with 419MB parameters, the proposed CRANet-based model can achieve better results with a 14% reduction in parameters. As shown in Table IX, the proposed method has a reduced computational cost in terms of FLOPs compared to previous diffusion-based LLIE methods WeatherDiff [7] and PyDiff [2] while maintaining the best perceptual performance in terms of LPIPS. Thanks to the patch-based sampling scheme, our method has significant advantages in terms of memory efficiency, which is more accessible for users with limited computational resources. The proposed method takes about 3.8s to process an input image of 600×400 resolution. The test time is obtained by averaging from 15 runs on a single RTX 3090 Ti GPU.

3) The effectiveness of multiscale ensemble scheme: As shown in Fig. 8, an inappropriate image patch size may disrupt the continuity of the image structure, which may lead to inconsistent brightness or blurring of the image as shown in the first row of Fig. 9. As shown in Table VIII, the multiscale ensemble scheme can significantly improve the enhancement effect of low-light images, regardless of which of the three backbone networks is used. Fig. 9 compares the visual results of DePDiff with fixed patch-sized and multiscale ensemble. All these results clearly demonstrates the wide applicability and superiority of the proposed multiscale ensemble scheme as post-processing scheme. These results can be illustrated as follows. With the help of multiscale ensemble scheme, however, proposed method can effectively compensate for these shortcomings and obtain better LLIE results by learning to fuse images generated using different patch sizes.

V. CONCLUSION

This paper addresses the challenges in diffusion-based low-light image enhancement methods, which struggle with preserving fine details due to their denoising-centric training schemes and the varying brightness and noise characteristics of low-light images. We propose detail-preserving diffusion models specifically tailored for realistic and faithful enhancement of low-light images. Our method capitalizes on a patch-based

denoising process, integrated with a reverse process reconstruction loss that enhances fidelity to the original low-light images, facilitating more precise detail recovery. The development of an efficient noise estimation network, equipped with a content and region-aware attention mechanism, contributes significantly to retaining crucial details in the enhanced images. Furthermore, multiscale ensemble scheme helps ensure the preservation of detail fidelity in both well-lit and shadowed areas. The efficacy of our approach is demonstrated through extensive experiments, which clearly highlight the superiority of our proposed diffusion-based LLIE method in achieving both realism and detail preservation in image enhancement.

While our proposed detail-preserving diffusion models for low-light image enhancement demonstrate significant improvements over existing methods, there are some limitations. The multiscale ensemble scheme, though effective in preserving details across varying illumination regions, can introduce additional computational complexity, making the method less efficient for real-time applications. Moreover, the patch-based learning strategy, while beneficial for detail preservation, may impact the ability to fully capture global image structures, which could affect the performance of structural similarity. Future work will focus on addressing these limitations by exploring more efficient implementations of the multiscale ensemble scheme, and enhancing the model's capability to capture global image structures without compromising detail preservation. These improvements aim to make the method more practical for a wider range of applications.

REFERENCES

[1] Y. Yin, D. Xu, C. Tan, P. Liu, Y. Zhao, and Y. Wei, "Cle diffusion: Controllable light enhancement diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8145–8156.

[2] D. Zhou, Z. Yang, and Y. Yang, "Pyramid diffusion models for low-light image enhancement," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 1795–1803.

[3] H. Jiang, A. Luo, S. Han, H. Fan, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Transactions on Graphics*, vol. 42, no. 6, pp. 1–15, 2023.

[4] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9396–9416, 2022.

[5] J. Ye, C. Fu, Z. Cao, S. An, G. Zheng, and B. Li, "Tracker Meets Night: A Transformer Enhancer for UAV Tracking," *IEEE Robotics and Automation Letters*, 2022.

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

[7] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2023.

[8] J. Liang, Y. Xu, Y. Quan, B. Shi, and H. Ji, "Self-supervised low-light image enhancement using discrepant untrained network priors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7332–7345, 2022.

[9] M. T. Rasheed, D. Shi, and H. Khan, "A comprehensive experiment-based review of low-light image enhancement methods and benchmarking low-light image quality assessment," *Signal Processing*, vol. 204, p. 108821, 2023.

[10] J. Liu, D. Xu, W. Yang, M. Fan, and H. Huang, "Benchmarking low-light image enhancement and beyond," *International Journal of Computer Vision*, vol. 129, pp. 1153–1184, 2021.

[11] F. Jia, H. S. Wong, T. Wang, and T. Zeng, "A reflectance re-weighted retinex model for non-uniform and low-light image enhancement," *Pattern Recognition*, vol. 144, pp. 109 823–109 837, 2023.

[12] J. Yang, Y. Xu, H. Yue, Z. Jiang, and K. Li, "Low-light image enhancement based on retinex decomposition and adaptive gamma correction," *IET image processing*, vol. 15, no. 5, pp. 1189–1202, 2021.

[13] C.-M. Fan, T.-J. Liu, and K.-H. Liu, "Half wavelet attention on m-net+ for low-light image enhancement," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 3878–3882.

[14] Z. Cui, K. Li, L. Gu, S. Su, P. Gao, Z. Jiang, Y. Qiao, and T. Harada, "You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction." in *The 33rd British Machine Vision Conference*, 2022, pp. 238–255.

[15] Y. Luo, B. You, G. Yue, and J. Ling, "Pseudo-supervised Low-light Image Enhancement with Mutual Learning," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[16] Z. Zhao, B. Xiong, L. Wang, Q. Ou, L. Yu, and F. Kuang, "Retinexdip: A unified deep framework for low-light image enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1076–1088, 2021.

[17] L. Guo, R. Wan, W. Yang, A. Kot, and B. Wen, "Cross-Image Disentanglement for Low-Light Enhancement in Real World," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[18] Z. He, W. Ran, S. Liu, K. Li, J. Lu, C. Xie, Y. Liu, and H. Lu, "Low-Light Image Enhancement with Multi-Scale Attention and Frequency-Domain Optimization," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[19] Z. Ni, W. Yang, H. Wang, S. Wang, L. Ma, and S. Kwong, "Cycle-interactive generative adversarial network for robust unsupervised low-light enhancement," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1484–1492.

[20] Q. Jiang, Y. Mao, R. Cong, W. Ren, C. Huang, and F. Shao, "Unsupervised decomposition and correction network for low-light image enhancement," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19 440–19 455, 2022.

[21] Z. Zhang, W. Sun, X. Min, W. Zhu, T. Wang, W. Lu, and G. Zhai, "A no-reference evaluation metric for low-light image enhancement," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[22] G.-D. Fan, B. Fan, M. Gan, G.-Y. Chen, and C. L. P. Chen, "Multiscale low-light image enhancement network with illumination constraint," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7403–7417, 2022.

[23] Z. Rahman, P. Yi-Fei, M. Aamir, S. Wali, and Y. Guan, "Efficient image enhancement model for correcting uneven illumination images," *IEEE Access*, vol. 8, pp. 109 038–109 053, 2020.

[24] Z. Rahman, Y.-F. Pu, M. Aamir, and S. Wali, "Structure revealing of low-light images using wavelet transform based on fractional-order denoising and multiscale decomposition," *The Visual Computer*, vol. 37, no. 5, pp. 865–880, 2021.

[25] Z. Rahman, Z. Ali, I. Khan, M. I. Uddin, Y. Guan, and Z. Hu, "Diverse image enhancer for complex underexposed image," *Journal of Electronic Imaging*, vol. 31, no. 4, pp. 041 213–041 213, 2022.

[26] Z. Rahman, M. Aamir, Z. Ali, A. K. J. Saudagar, A. AlTameem, and K. Muhammad, "Efficient contrast adjustment and fusion method for underexposed images in industrial cyber-physical systems," *IEEE Systems Journal*, 2023.

[27] Z. Rahman, J. A. Bhutto, M. Aamir, Z. A. Dayo, and Y. Guan, "Exploring a radically new exponential retinex model for multi-task environments," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 7, p. 101635, 2023.

[28] C. Liu, F. Wu, and X. Wang, "Efinet: Restoration for low-light images via enhancement-fusion iterative network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8486–8499, 2022.

[29] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3063–3072.

[30] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Transactions on Image Processing*, vol. 30, pp. 2340–2349, 2021.

[31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[32] J. R. Jebadass and P. Balasubramaniam, "Low light enhancement algorithm for color images using intuitionistic fuzzy sets with histogram equalization," *Multimedia Tools and Applications*, vol. 81, no. 6, pp. 8093–8106, 2022.

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

[34] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1780–1789.

[35] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 561–10 570.

[36] Q. Jiang, Z. Liu, K. Gu, F. Shao, X. Zhang, H. Liu, and W. Lin, "Single image super-resolution quality assessment: a real-world dataset, subjective studies, and an objective metric," *IEEE Transactions on Image Processing*, vol. 31, pp. 2279–2294, 2022.

[37] Q. Jiang, Y. Gu, C. Li, R. Cong, and F. Shao, "Underwater image enhancement quality evaluation: Benchmark dataset and objective metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5959–5974, 2022.

[38] Y. Kang, Q. Jiang, C. Li, W. Ren, H. Liu, and P. Wang, "A perception-aware decomposition and fusion framework for underwater image enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 988–1002, 2022.

[39] Q. Jiang, Y. Kang, Z. Wang, W. Ren, and C. Li, "Perception-driven deep underwater image enhancement without paired supervision," *IEEE Transactions on Multimedia*, 2023.

[40] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. Kot, "Low-light image enhancement with normalizing flow," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2604–2612.

[41] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proceedings of the International Conference on Learning Representations*, 2021, pp. 1–12.

[42] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1–13, 2019.

[43] C.-W. Huang, J. H. Lim, and A. C. Courville, "A variational perspective on diffusion-based generative models and score matching," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 863–22 876, 2021.

[44] Y. Wang, Y. Yu, W. Yang, L. Guo, L.-P. Chau, A. C. Kot, and B. Wen, "Exposurediffusion: Learning to expose for low-light image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 438–12 448.

[45] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving diffusion models for inverse problems using manifold constraints," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 683–25 696, 2022.

[46] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 593–23 606, 2022.

[47] Y. Zhu, K. Zhang, J. Liang, J. Cao, B. Wen, R. Timofte, and L. Van Gool, "Denoising diffusion models for plug-and-play image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1219–1229.

[48] H. Chung, J. Kim, S. Kim, and J. C. Ye, "Parallel diffusion models of operator and image for blind inverse problems," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6059–6069.

[49] L. Guo, C. Wang, W. Yang, S. Huang, Y. Wang, H. Pfister, and B. Wen, "Shadowdiffusion: When degradation prior meets diffusion model for shadow removal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 049–14 058.

[50] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1680–1691.

[51] S. Panagiotou and A. S. Bosman, "Denoising diffusion post-processing for low-light image enhancement," *2023, arXiv:2303.09627*, pp. 1–11, 2023.

[52] T. Wang, K. Zhang, Z. Shao, W. Luo, B. Stenger, T.-K. Kim, W. Liu, and H. Li, "Lldiffusion: Learning degradation representations in diffusion models for low-light image enhancement," *2023, arXiv:2307.14659*, pp. 1–16, 2023.

[53] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[54] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European Conference on Computer Vision*. Springer, 2022, pp. 17–33.

[55] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021, pp. 1–20.

[56] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1632–1640.

[57] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *International Journal of Computer Vision*, vol. 129, pp. 1013–1037, 2021.

[58] X. Guo and Q. Hu, "Low-light image enhancement via breaking down the darkness," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 48–66, 2023.

[59] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 683–17 693.

[60] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.

[61] X. Xu, R. Wang, and J. Lu, "Low-light image enhancement via structure modeling and guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9893–9903.

[62] Z. Fu, Y. Yang, X. Tu, Y. Huang, X. Ding, and K.-K. Ma, "Learning a simple low-light image enhancer from paired low-light instances," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 252–22 261.

[63] S. Yang, M. Ding, Y. Wu, Z. Li, and J. Zhang, "Implicit neural representation for cooperative low-light image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 918–12 927.

[64] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 504–12 513.

[65] Y. Liu, T. Huang, W. Dong, F. Wu, X. Li, and G. Shi, "Low-light image enhancement with multi-stage residue quantization and brightness-aware attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 140–12 149.

[66] N. Jiang, J. Lin, T. Zhang, H. Zheng, and T. Zhao, "Low-light image enhancement via stage-transformer-guided network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3701–3712, 2023.

[67] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex Decomposition for Low-Light Enhancement," in *British Machine Vision Conference (BMVC)*, Aug. 2018.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *2014, arXiv:1412.6980*, pp. 1–11, 2014.

[69] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 714–17 724.

[70] C. Li, C. Guo, and C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4225–4238, 2022.