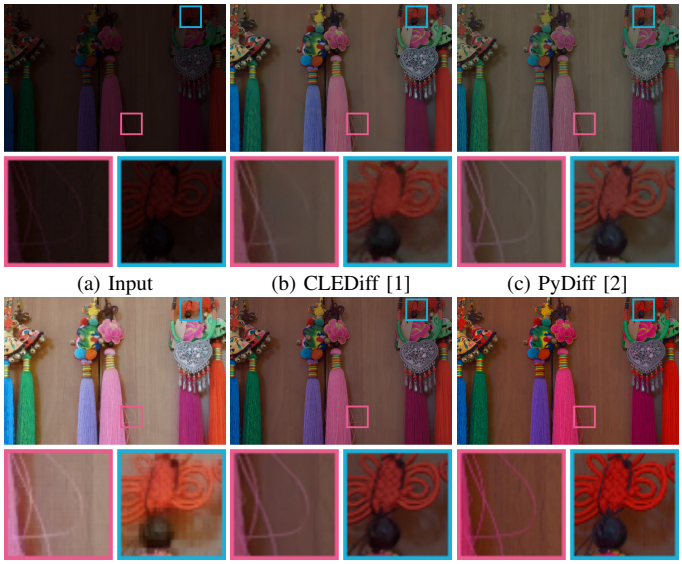
Detail-Preserving Diffusion Models for Low-Light Image Enhancement

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 incrementally remove 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 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 datasets demonstrate that our method achieves state-of-the-art results compared to over twenty existing methods in terms of both perceptual quality (LPIPS) and distortion metrics (PSNR and SSIM). The code is available at: https://github.com/CSYanH/DePDiff.**

***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 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 photographs [4], [5] but also impedes critical vision tasks like detection and tracking [6], [7].

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], [5], [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 [13]–[29]. 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 map 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 [30]–[32]. 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

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

Yan Huang, Xiaoshan Liao, and Yong Xu are with the Guangdong Provincial Key Laboratory of Multimodal Big Data Intelligent Analysis, 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 State Key Laboratory of Multimedia Information Processing and the 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).

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 [33] 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 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 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 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 leverage statistical properties and established image priors, offering computational efficiency. Histogram equalization methods enhance contrast by redistributing pixel intensities across the dynamic range [10]. While recent advances integrate intuitionistic fuzzy set theory [34], these methods often suffer from noise amplification [9], [10]. Retinex-based methods, grounded in color vision theory, focus on illumination enhancement [11], exemplified by adaptive gamma correction in Retinex decomposition [12]. Nonlinear transformation methods frame LLIE as a direct mapping between lighting conditions [9], with gamma correction serving as a classic example [12]. Despite their efficiency, these traditional approaches struggle to handle complex lighting scenarios [5], [35].

B. Regression LLIE methods

Deep learning-based approaches offer an effective alternative to traditional methods, enabling enhanced LLIE performance [4]. These approaches learn direct mappings between low-light and normal-light images [31], [36], leveraging advanced architectures such as transformers [3] and multiscale networks [20], [24] to handle complex degradation patterns. Recent architectural innovations have further advanced the field. HWMNet [13] introduces a half-wavelet attention mechanism that effectively captures multi-scale features, while the lightweight illumination-adaptive Transformer (IAT) [14] demonstrates promising results in downstream tasks such as object detection and semantic segmentation. RUAS [37] combines classical Retinex theory with neural architecture search

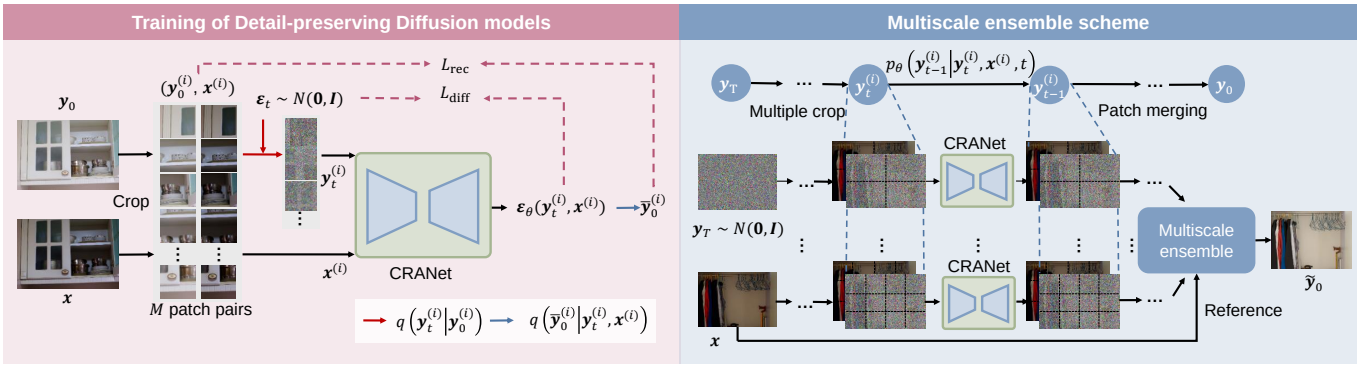


Fig. 2. Overview of DePDiff's training and multiscale ensemble process. Left: Conditioned on $\boldsymbol{x}^{(i)}$ extracted from low-light image $\boldsymbol{x}$, $\boldsymbol{y}_0^{(i)}$ extracted from target image $\boldsymbol{y}_0$ are gradually transitioning into normal-distributed noise during forward diffusion $q(\boldsymbol{y}_t^{(i)}|\boldsymbol{y}_0^{(i)})$ (red arrow). Reconstruction loss $\mathcal{L}_{\text{rec}}$ and denoising loss $\mathcal{L}_{\text{diff}}$ are jointly optimized through reverse diffusion $q(\overline{\boldsymbol{y}}_0^{(i)}|\boldsymbol{y}_t^{(i)}, \boldsymbol{x}^{(i)})$ (blue arrow). Right: For each low-light patches $\boldsymbol{x}^{(i)}$, CRANet estimates noise $p_\theta(\boldsymbol{y}_{t-1}^{(i)}|\boldsymbol{y}_t^{(i)}, \boldsymbol{x}^{(i)}, t)$ to progressively denoise the randomly sampled normal-distributed noise patches $\boldsymbol{y}_T^{(i)}$ into the final output $\boldsymbol{y}_0$. The final enhanced image $\widetilde{\boldsymbol{y}}_0$ is composed by merging these denoised patches processed at different scales.

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

Method	Pixelwise reconstruction	Locality-based brightness adaptability	Scale-adaptive sampling
WeatherDiff [8]	✗	✗	✓
CLEDiff [1]	✗	✓	✗
DiffLL [3]	✗	✗	✗
PyDiff [2]	✗	✗	✗
Ours	✓	✓	✓

to optimize network design, and unsupervised approaches using pseudo-labels [15] have emerged to address the scarcity of paired training data. The techniques developed for LLIE share methodological similarities with related image enhancement tasks, including super-resolution [38] and underwater image enhancement [39]–[41].

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], [32]. For instance, EnlightenGAN integrates attention mechanisms with image-specific regularization within a GAN framework [32]. Cross-image disentanglement for low-light enhancement [19] leverages weak supervision in GANs to achieve enhancement without paired training data. LLFlow [42] demonstrates the potential of normalizing flows. Despite their effectiveness, GANs face challenges like training instability and artifact introduction, while normalizing flows show restricted expressiveness in modeling complex image distributions.

1) Diffusion models: DMs have recently revolutionized image generation. The genetic framework of DMs includes Denoising Diffusion Probabilistic Models (DDPMs) [33], Stochastic Differential Equations (SDE) [43], and Noise Conditional Score Networks (NCSN) [44]. 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 [45]. Existing variants demonstrate the versatility of DMs in addressing various computer vision tasks [46]–[50]. For example, WeatherDiff, a patch-based diffusion model, was developed for image restoration in adverse weather conditions [8], and ShadowDiffusion was proposed for image shadow removal [51]. Luo *et al*. designed a latent diffusion model for low-resolution latent space diffusion [52].

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 [33] use a noise estimation network for supervised reverse process learning but lack spatial adaptation, potentially failing to preserve fine details in complex textures. WeatherDiff [8] 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 [53]. Through integration with the image degradation and priors, a diffusion-based LLIE method (LLDiffusion) was designed [54]. These different variants of DMs show promising prospects. By contrast, our method uniquely integrates a pixel-wise 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 compare 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 learn a parametric approximation 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 [8], which learn a Markov Chain to gradually convert Gaussian noise into the data distribution of target images y, conditioned on input images $\boldsymbol{x}$.

A crucial aspect of effectively leveraging the capabilities of DMs in LLIE is 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 DePDiff, which employs reverse diffusion-based reconstruction loss (left of Fig. 2) and smoothed noise estimation for overlapping patches using a multiscale ensemble scheme (right of Fig. 2) in a patch-wise, scale-agnostic manner. Our network architecture with simplified channel-spatial attention, CRANet, is tailored for content- and region-aware noise estimation (Fig. 3).

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 $(\boldsymbol{x}, \boldsymbol{y})$ where $\boldsymbol{P}^{(i)}$ is a binary mask matrix indicating the i-th patch location, and $\mathrm{Crop}(\cdot)$ extracts the specified patch. DePDiff extends conditional patch-based DDPMs [8], generating a target image patch in T diffusion time steps from 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 $p_\theta(\boldsymbol{y}_{t-1}^{(i)}|\boldsymbol{y}_t^{(i)}, \boldsymbol{x}^{(i)})$. We omit the patch location index i in this subsection.

1) Forward diffusion process: The forward diffusion process incrementally adds Gaussian noise to $\boldsymbol{y}_0$ according to a variance schedule $\beta_1, \cdots, \beta_T$, formulated as $q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1})$. This process, denoted by $q(\boldsymbol{y}_{1:T}|\boldsymbol{y}_0)$, is 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) [8], [55], we adopt a non-Markovian forward diffusion process for deterministic sampling acceleration. The generalized non-Markovian forward process is formulated as:

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

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:

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

the diffusion process formulated by Eq. (5) can not only become 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:

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

A deterministic implicit sampling approach can be achieved by setting $\lambda_t^2 = 0$ [8], [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 $\boldsymbol{y}_0$ through a reverse Markov chain conditioned on $\boldsymbol{x}$. This process involves iteratively reconstructing the signal from noise, to convert the diffusive noise back to $\boldsymbol{y}_0$ using a noise estimator $\boldsymbol{\epsilon}_\theta$. Unlike [33], the reverse process of our model is conditioned on the patch-based low-light image $\boldsymbol{x}$. The conditioning on $\boldsymbol{x}$ leverages the general features of the low-light image to improve image quality, providing a more detailed and fine-grained learning approach for better image enhancement. 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:

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

Algorithm 1 Training of DePDiff

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

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

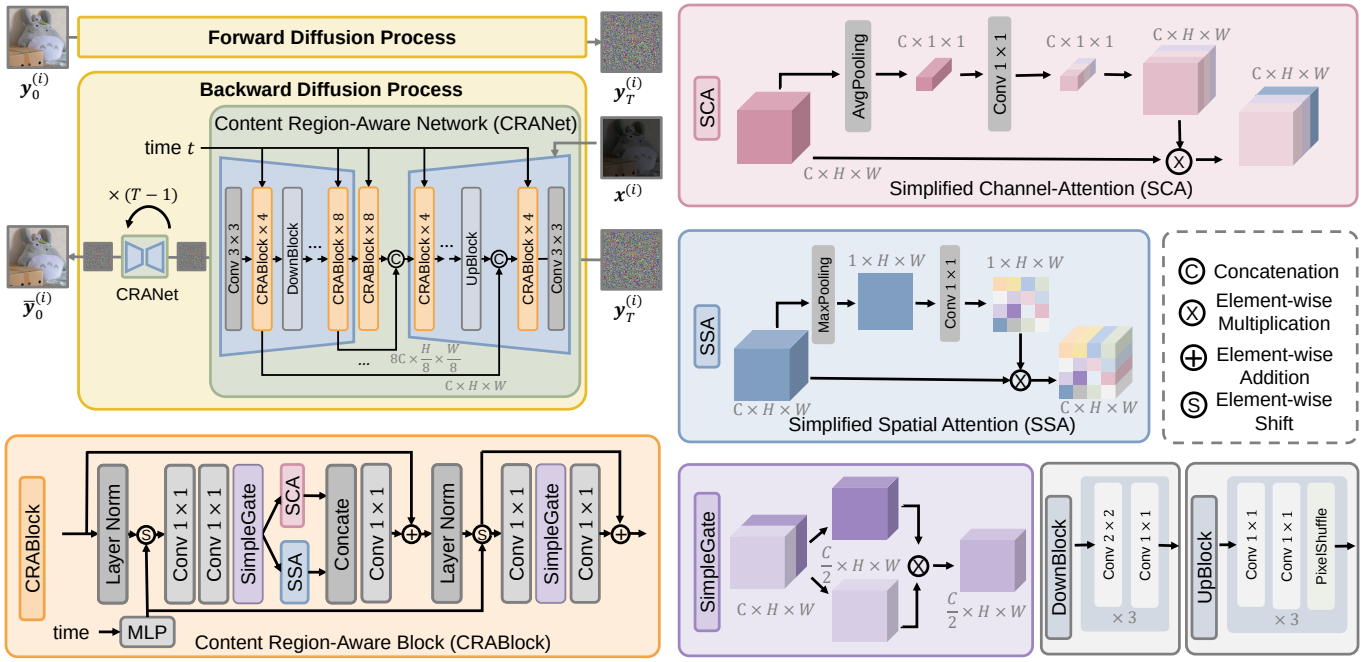


Fig. 3. Architecture of CRANet. Based on U-Net [56] with light configuration from [57], the network removes nonlinear activation functions for computational efficiency. The Simplified Spatial Attention (SSA) module enables spatial adaptability, while the Simplified Channel-Attention (SCA) enhances the noise estimation process through dual-dimension guidance. A multilayer perceptron (MLP) performs time embedding.

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

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 the mean estimated noise for each pixel across overlapping patch regions at denoising time step t. This method ensures enhanced fidelity throughout the reverse sampling process, harmonizing the contributions from all adjacent patches. At each time step t during sampling, we calculate the additive noise for every overlapping patch location i using $\boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t^{(i)},\boldsymbol{x}^{(i)},t)$. 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. With DDIM for accelerated deterministic sampling 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, enabling the training of the noise estimation network $\boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t,\boldsymbol{x},t)$. The denoising loss is realized by 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 $\mathcal{L}_{\text{diff}}$:

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

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 the enhanced image $\overline{\boldsymbol{y}}_0$ and the ground truth $\boldsymbol{y}_0$. $\overline{\boldsymbol{y}}_0$ 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. 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 non-uniform brightness distributions across scenes. It demands adaptive methods that distinguish between noise and genuine details based on local content and regional characteristics. To address this, our network design incorporates channel attention mechanisms to selectively amplify 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 enhance 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 [33] by incorporating an activation-free structure and a streamlined channel-spatial attention mechanism [57]. This results in our 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 maintains the core U-Net [56] while introducing several key modifications. CRANet incorporates the configuration from [57], removing nonlinear activation functions to reduce computational cost. Unlike [57], our architecture uses a multilayer perceptron (MLP) for time embedding and introduces a new simplified spatial attention (SSA) mechanism, combined with the simplified channel-attention (SCA) to guide the noise estimation process.

During inference, the trained CRANet $\boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t^{(i)}, \boldsymbol{x}^{(i)}, t)$ processes individual image patches through the denoising pipeline. 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.

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 $\widetilde{\boldsymbol{y}}_0$

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.

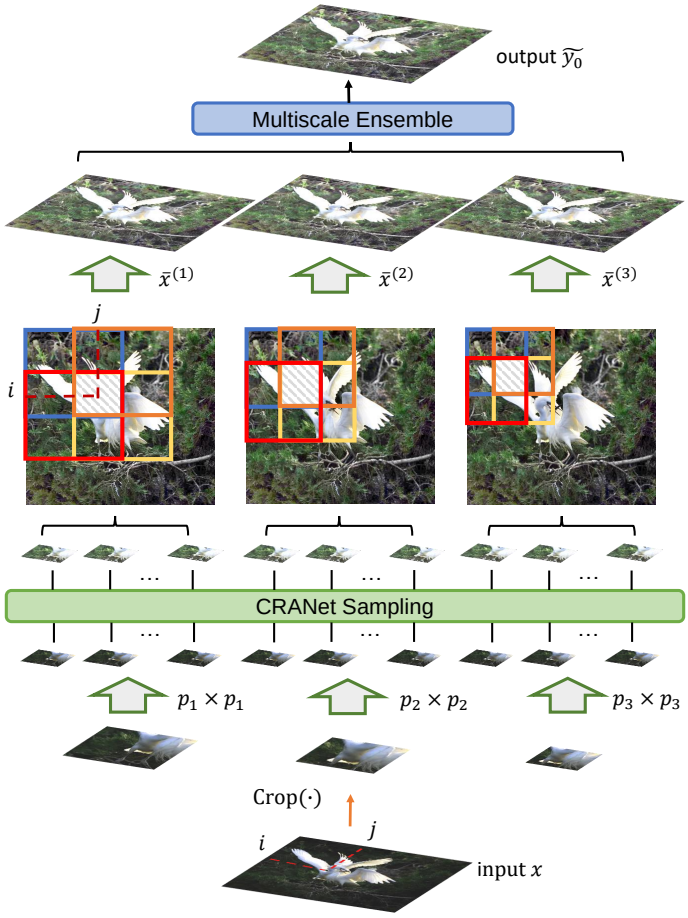


Fig. 4. Pipeline of multiscale patch processing. The input image is divided into patches of varying sizes, processed through CRANet ϵ_θ, and reconstructed via sliding window averaging. Enhanced images from different scales are integrated through the multiscale ensemble scheme.

For the m-th model $(m \in 1, 2, \cdots, M)$, the enhanced image is computed as:

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

The weighted coefficients η are optimized iteratively using the loss function $\mathcal{L}$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. During inference, each base model processes samples of different sizes to produce N weighted images. The M models generate corresponding sets of combination weights, yielding M fused images, as shown in Fig. 4. Taking advantage of patch size-agnostic image enhancement, images of different patch sizes can be generated inexpensively. 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 achieving 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 analyzing histogram differences between the enhanced images and the original low-light image $\boldsymbol{x}$. Image histograms, representing pixel intensity distributions, serve as an efficient and invariant measure for assessing light intensity variations. The image among $\widetilde{\boldsymbol{x}}^{(1)}, \widetilde{\boldsymbol{x}}^{(2)}, ..., \widetilde{\boldsymbol{x}}^{(M)}$ with the maximum histogram difference from $\boldsymbol{x}$ is selected as the final enhanced image $\widetilde{\boldsymbol{y}}_0$:

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

where

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

with $\text{Hist}(\widetilde{\boldsymbol{x}}^{(m)})$ 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 experiments are divided into three parts: within-dataset experiments comparing our method against state-of-the-art approaches; cross-dataset validation to assess generalization capability; and ablation studies examining the contribution of each proposed component.

A. Experimental settings

1) Datasets: A total of eight datasets are utilized in experiments: LOL, LOL-v1, LOL-v2 real, NPE, DICM, MEF, LIME, and VV datasets [35], [58], [59]. Among these, only LOL, LOL-v1, and LOL v2-real datasets [35], [58] provide paired normal-light reference images. For within-dataset evaluation on LOL and LOL-v1, we follow the standard splits for training and testing. Cross-dataset evaluation on the real-world datasets (DICM, MEF, NPE, LIME, VV, and LOL-v2 real) is conducted using the model trained on LOL. For the ensemble scheme, the models are trained on the LOL dataset. All ablation studies are conducted on the LOL dataset to maintain consistency in analysis.

2) Evaluation metrics: For paired datasets (LOL, LOL-v1, and LOL-v2 real), we employ standard metrics: peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and learned perceptual image patch similarity (LPIPS). For unpaired datasets (DICM, MEF, NPE, LIME, and VV), we use a naturalness image quality evaluator (NIQE). Higher PSNR and SSIM values indicate better results, while lower LPIPS and NIQE values indicate better quality.

3) Implementation details: In all experiments, the training patch size is set to 64×64 and the batch size is set to 4. The LOL dataset training requires 800,000 iterations. The LOL-v1 dataset requires 600,000 iterations. We use Adam [60] the optimizer with a fixed learning rate of 0.00003. The diffusion time step T is set to 1000 for training, and the implicit sampling step S is set to 20 for inference. Our multiscale ensemble patch sizes 64×64, 96×96, 128×128, 160×160, 192×192, 225×225 and 256×256. We implement our method using PyTorch and run experiments on a single NVIDIA GTX 3090 Ti GPU.

(a) Input (b) LLFlow [42] (c) WeatherDiff [8] (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. ↑(↓) MEANS HIGHER (LOWER) IS BETTER.

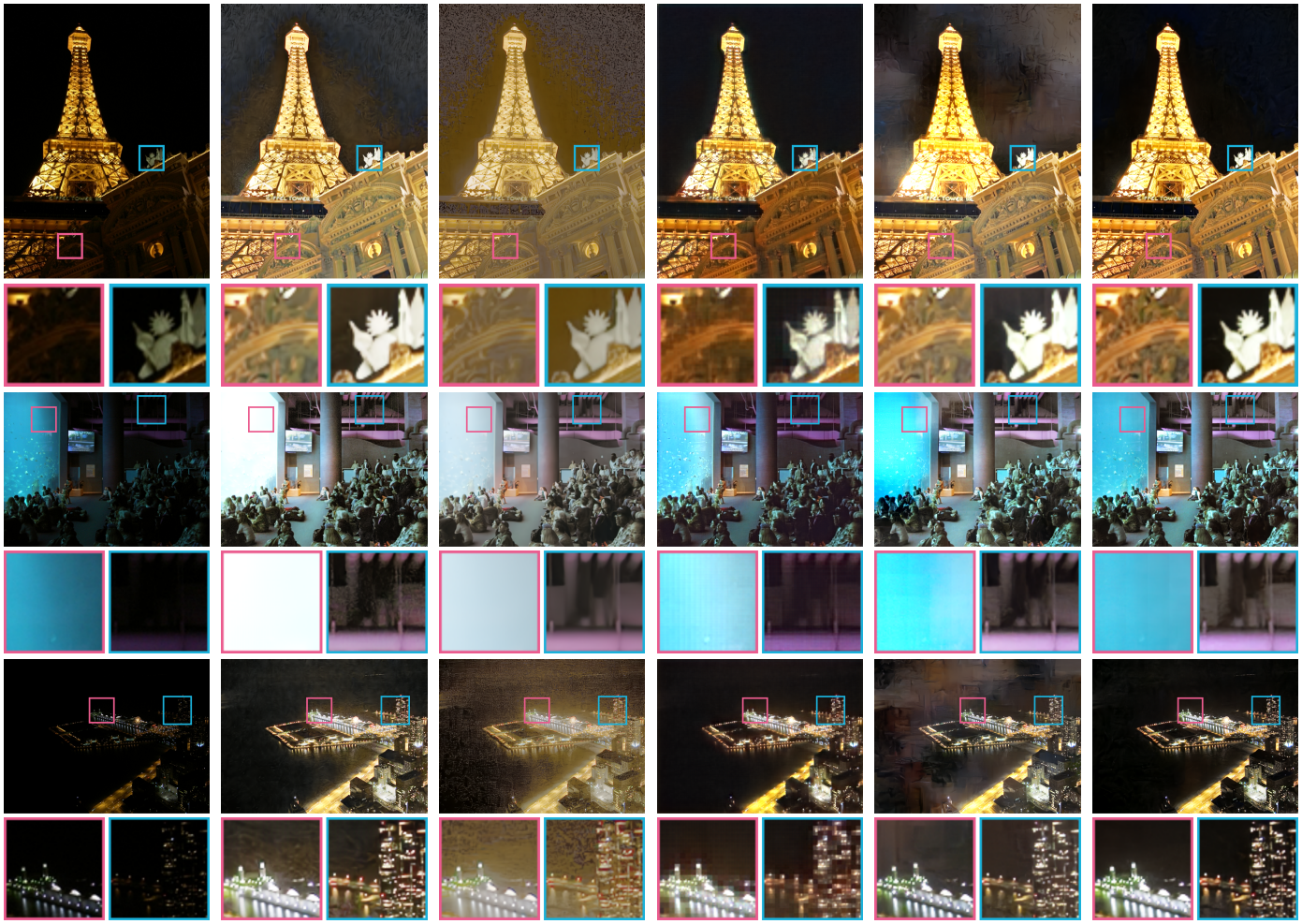
Method	Source	PSNR↑	SSIM↑	LPIPS↓
KinD [61]	MM'19	19.19	0.815	0.170
Zero-DCE [36]	CVPR'20	14.73	0.509	0.401
EnlightenGAN [32]	TIP'21	17.48	0.650	0.320
KinD++ [62]	IJCV'21	20.90	0.823	0.164
Bread [63]	IJCV'22	22.96	0.838	0.160
Uformer [64]	CVPR'22	18.96	0.778	0.505
Restormer [65]	CVPR'22	22.17	0.819	0.149
IAT [14]	BMVC'22	23.38	0.861	0.216
HWMNet [13]	ICIP'22	24.24	0.922	0.113
LLFlow [42]	AAAI'22	24.99	0.923	0.116
SMG-LLIE [66]	CVPR'23	23.85	0.893	0.131
PairLIE [67]	CVPR'23	19.51	0.736	0.248
NeRCo [68]	ICCV'23	19.84	0.771	0.315
RetinexFormer [69]	ICCV'23	25.16	0.845	0.129
RQ-LLIE [70]	ICCV'23	25.24	0.855	0.250
STGNet [59]	TCSVT'23	22.03	0.838	<u>0.101</u>
WeatherDiff [8]	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
DePDiff (Ours)	-	**27.44**	**0.939**	**0.085**

4) Comparing methods: We select various state-of-the-art learning-based methods from the past five 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 [36], DRBN [31], RUAS [37], KinD [61], KinD++ [62], STGNet [59], Bread [63], SNR-Net [71], Zero-DCE++ [72], Restormer [65], Uformer [64], RetinexFormer [69], SMG-LLIE [66] and PairLIE [67]. 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 [8], EnlightenGAN [32], PyDiff [2], CLEDiff [1], DiffLL [3], LLFlow [42], NeRCo [68] and RQ-LLIE [70]. All these methods would be used to conduct the first two parts of experiments to verify the effectiveness of the proposed method within and across datasets. For convincing 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 quantitative results of the within-dataset evaluation are summarized in Tables II and III. In Tab. II, we retrained DiffLL on the LOL dataset for a fair comparison, while STGNet results are cited from their original publications. Our proposed method outperforms all comparison methods in both PSNR and LPIPS metrics. While PSNR evaluates pixel-wise accuracy and LPIPS measures perceptual similarity, achieving superior performance in both metrics demonstrates our method's effectiveness in both objective and perceptual quality enhancement. Regarding SSIM, our method achieves the best performance on the LOL dataset but ranks second on the LOL-v1 dataset with a marginal difference.

This performance variation can be attributed to two key factors: First, the LOL dataset comprises exclusively real-world low-light images, whereas LOL-v1 contains both synthetic and real-world images. This mixed composition creates distribution inconsistencies that particularly challenge diffusion-based methods, which rely on learning mappings between



(a) Input (b) LLFlow [42] (c) PyDiff [2] (d) DiffLL [3] (e) WeatherDiff [8] (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.

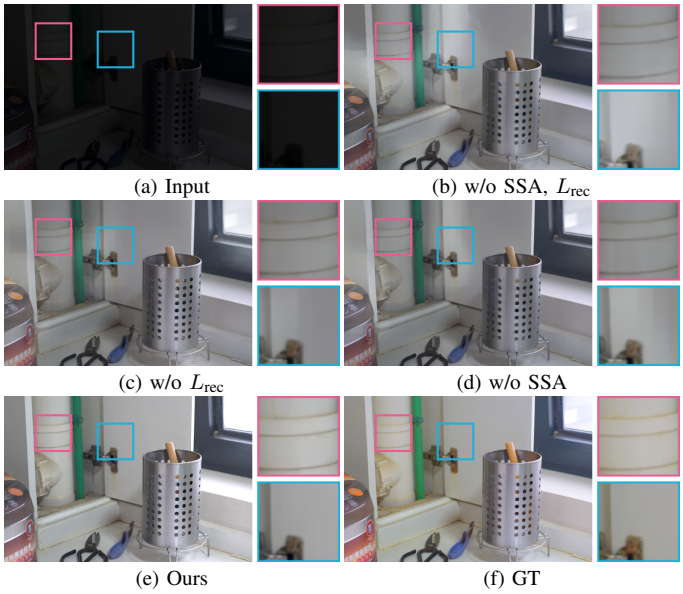
Method	Source	PSNR↑	SSIM↑	LPIPS↓
KinD [61]	MM'19	22.15	0.853	0.257
Zero-DCE [36]	CVPR'20	20.54	0.778	0.331
EnlightenGAN [32]	TIP'21	17.60	0.653	0.372
Uformer [64]	CVPR'22	19.00	0.741	0.354
Restormer [65]	CVPR'22	20.61	0.797	0.288
RUAS [37]	CVPR'22	16.40	0.503	0.364
SNRNet [71]	CVPR'22	24.61	0.842	0.233
IAT [14]	BMVC'22	21.25	0.844	0.255
HWMNet [13]	ICIP'22	19.62	0.862	0.271
LLFlow [42]	AAAI'22	26.02	**0.926**	<u>0.100</u>
SMG-LLIE [66]	CVPR'23	24.03	0.878	0.144
PairLIE [67]	CVPR'23	24.02	0.803	0.118
NeRCo [68]	ICCV'23	25.17	0.833	0.160
RetinexFormer [69]	ICCV'23	**27.69**	0.856	0.166
RQ-LLIE [70]	ICCV'23	22.37	0.854	0.228
WeatherDiff [8]	TPAMI'23	17.91	0.811	0.272
DiffLL [3]	TOG'23	26.33	0.845	0.217
DePDiff (Ours)	-	<u>26.52</u>	<u>0.922</u>	**0.098**

noise and image distributions. Second, our method's patch-based approach, while effective for local detail enhancement, may influence the learning of global image structure. Since SSIM emphasizes structural information and is less sensitive to minor perceptual distortions, this patch-based strategy could impact SSIM performance. This explains why both our method and DiffLL achieve lower SSIM scores compared to LLFlow, a non-diffusion-based approach. Nevertheless, the performance across different metrics demonstrates that our method achieves superior overall performance in preserving perceptual quality, pixel-level accuracy, and structural details.

2) Qualitative results: The visual results are compared in Figs. 1, 5 and 6. Compared to the quantitative results, these qualitative visual results can intuitively demonstrate the effectiveness and practicality of the proposed method. Real-world low-light conditions present various challenges including overexposure and underexposure, or saturated pixel areas caused by nighttime light sources. Fig. 1 demonstrates an input low-light image containing both overexposed and underexposed regions. CNN-based HWMNet, Transformer-based IAT, normalizing flow-based LLFlow, and diffusion-based PyDiff struggle with these extreme regions. However, the proposed method successfully enhances both regions simultaneously. Fig. 5 and Fig. 6 demonstrate our method's effectiveness across various image conditions, regardless of whether there are pixel-saturated areas in the image or whether it has different resolution sizes.

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 [61]	MM'19	24.05	0.917	0.1140
Zero-DCE [36]	CVPR'20	18.05	0.580	0.352
EnlightenGAN [32]	TIP'21	18.67	0.678	0.364
KinD++ [62]	IJCV'21	22.21	0.885	0.174
Bread [63]	IJCV'22	23.69	0.912	0.155
Uformer [64]	CVPR'22	18.44	0.759	0.347
Restormer [65]	CVPR'22	24.91	0.851	0.264
RUAS [37]	CVPR'22	15.35	0.495	0.395
SNRNet [71]	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 [42]	AAAI'22	28.35	0.945	0.076
SMG-LLIE [66]	CVPR'23	25.62	0.905	0.131
PairLIE [67]	CVPR'23	19.88	0.841	0.234
NeRCo [68]	ICCV'23	15.67	0.684	0.409
RetinexFormer [69]	ICCV'23	28.99	0.939	0.106
RQ-LLIE [70]	ICCV'23	25.94	0.941	0.219
WeatherDiff [8]	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**
DePDiff (Ours)	-	**33.87**	<u>0.947</u>	<u>0.067</u>



(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.

C. Cross-dataset performance comparisons

The quantitative results of the cross-dataset evaluation are summarized in Tables IV and V 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 images across datasets and has good generalization performance on cross-domain datasets. Table V demonstrates 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, CLEDiff, SNRNet, and SMG-LLIE were pre-trained and tested exclusively on paired training sets with fixed-size input 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 study is conducted by comparing the results with and without the reverse diffusion-based reconstruction loss using both ℓ_1-norm and ℓ_2-norm expression. Table VI reveals two key findings: First, the ℓ_1-norm-based reverse diffusion-based reconstruction loss $\mathcal{L}_{\text{rec}}$ underperforms compared to the ℓ_2-norm variant. Second, the ℓ_2-norm achieves optimal results when combined with our CRANet backbone. These demonstrate the effectiveness and compatibility of our reconstruction loss with the designed architecture.

2) The effectiveness of CRANet architecture: From the perspective of model architecture, our method's main contribution lies in the design of the noise estimation network. We conduct ablation studies by replacing our backbone network with basic (vanilla) U-Net, NAFNet, and CRANet without SSA module or $\mathcal{L}_{\text{rec}}$. We maintain consistent parameter counts across all variants and exclude post-processing. Table VII demonstrates that both the SSA module and $\mathcal{L}_{\text{rec}}$ improve the model performance.

To further verify the effectiveness of the proposed model architecture, visual comparisons are shown in Fig. 7 to demonstrate the impact of removing SSA or $\mathcal{L}$rec. Without $\mathcal{L}_{\text{rec}}$, patch-based models would produce obvious artifacts or inconsistencies in the image. The proposed reverse diffusion-based reconstruction loss effectively compensates for patch-based learning limitations. However, without SSA, the reverse diffusion-based reconstruction loss alone cannot achieve optimal results. The SSA module enables fine-grained enhancement at different spatial scales, combining direct pixel-level supervision with channel-spatial attention-based learning for smooth and consistent image effects.

Moreover, compared to U-Net in DDPMs (419MB parameters), our CRANet-based model achieves superior results with 14% fewer parameters. Table IX demonstrates reduced computational cost (FLOPs) compared to diffusion-based LLIE methods WeatherDiff [8] and PyDiff [2] while maintaining best perceptual performance (LPIPS). Our patch-based sampling scheme offers significant memory efficiency, making it accessible for users with limited computational resources. Processing time averages 3.8s for a 600×400 resolution image (averaged over 15 runs on a single RTX 3090 Ti GPU).

3) The effectiveness of multiscale ensemble scheme: As shown in Fig. 8, inappropriate image patch size disrupts the continuity of the image structure, which may lead to inconsistent brightness or blurring (as shown in the first row of Fig. 9). This observation motivates our multiscale ensemble approach, which adaptively integrates information from multiple scales to ensure robust performance across diverse image content and lighting conditions. Table VIII shows the multiscale ensemble scheme can significantly improve the enhancement effect of low-light images, regardless of which of the three backbone

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 [61]	MM'19	5.28	5.61	5.06	6.14	4.25	5.26
Zero-DCE [36]	CVPR'20	4.58	4.93	4.57	5.82	4.81	4.94
EnlightenGAN [32]	TIP'21	4.82	5.01	5.26	5.11	3.85	4.81
KinD++ [62]	IJCV'21	5.29	6.23	4.56	7.20	4.87	5.63
Bread [63]	IJCV'22	4.78	4.93	4.91	5.07	3.86	4.71
Uformer [64]	CVPR'22	11.29	35.56	37.68	14.73	11.79	22.21
Restormer [65]	CVPR'22	12.12	13.22	11.93	14.01	10.29	12.31
RUAS [37]	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 [42]	AAAI'22	**4.46**	4.80	4.78	5.83	3.60	4.69
STGNet [59]	TCSVT'23	9.95	10.11	11.80	10.01	8.00	9.97
PairLIE [67]	CVPR'23	5.15	5.03	5.47	4.98	4.30	4.98
RetinexFormer [69]	ICCV'23	4.19	4.12	4.20	4.88	3.66	4.21
WeatherDiff [8]	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
DePDiff (Ours)	-	4.47	**4.20**	4.51	4.41	**3.17**	**4.15**

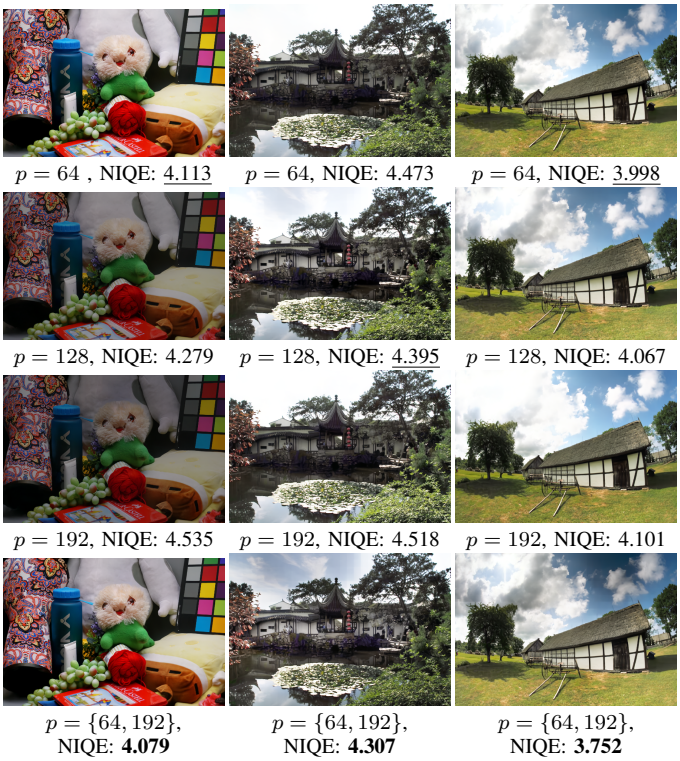


Fig. 8. Effect of patch size (p) on enhancement quality. Images sampled with different patch sizes demonstrate varying optimal sizes across scenes. Best NIQE scores (lower is better) are shown in bold, second-best underlined.

networks is used. Fig. 9 compares the visual results with the fixed patch-sized and multiscale ensemble. All these results demonstrate the wide applicability and superiority of the proposed multiscale ensemble scheme as a post-processing scheme, which effectively compensates for the limitations by learning to fuse images generated using different patch sizes.

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 [56]	$\mathcal{L}_{\text{diff}}$	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 (Ours)	$\mathcal{L}_{\text{diff}}$	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 [56]	-	19.74	0.908	0.113
NAFNet [57]	-	23.76	0.926	0.121
CRANet (Ours)	w/o SSA, $\mathcal{L}_{\text{rec}}$	24.75	0.931	0.101
	w/o $\mathcal{L}_{\text{rec}}$	24.69	0.930	0.101
	w/o SSA	24.81	0.931	0.100
	w/ SSA, w/ $\mathcal{L}_{\text{rec}}$	**26.33**	**0.936**	**0.089**

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 DePDiff 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, a 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 highlight the superiority of our proposed diffusion-based LLIE method in achieving both realism and detail preservation in image enhancement.

While our approach demonstrates significant improvements, several limitations remain. The multiscale ensemble scheme can introduce additional computational complexity, making the method less efficient for real-time applications. Moreover, the patch-based learning strategy may limit the ability to capture global image structures, which could affect the performance of structural similarity. Future work will explore more efficient implementations of the multiscale ensemble scheme and enhance 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.

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 [56]	w/o	22.04	0.923	0.105
	w/	24.73	0.933	0.088
NAFNet [57]	w/o	19.60	0.897	0.170
	w/	24.41	0.931	0.092
CRANet (Ours)	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 [61]	0.170	34.99	8.02	1.50
Zero-DCE [36]	0.401	15.59	0.08	0.01
EnlightenGAN [32]	0.320	114.35	67.80	0.34
KinD++ [62]	0.164	40.93	21.11	4.50
Bread [63]	0.160	106.96	2.15	0.10
Uformer [64]	0.505	12.00	5.29	0.50
Restormer [65]	0.149	144.25	26.13	0.12
SNRNet [71]	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 [42]	0.116	358.40	17.42	0.40
SMG-LLIE [66]	0.131	92.66	19.35	0.10
PairLIE [67]	0.248	20.81	0.35	0.15
NeRCo [68]	0.315	130.70	25.80	0.34
RetinexFormer [69]	0.129	15.85	1.53	0.21
WeatherDiff [8]	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
DePDiff (Ours)	0.085	640.40	94.01	3.80

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 ACM MM*, 2023, pp. 8145–8156.

[2] D. Zhou, Z. Yang, and Y. Yang, "Pyramid diffusion models for low-light image enhancement," in *Proceedings of the 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. 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.

[6] 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.

[7] 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.

[8] 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.

[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 *Proceedings of the IEEE International Conference on Image Processing*. 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 *Proceedings of the 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] C. Zhou, M. Teng, J. Han, J. Liang, C. Xu, G. Cao, and B. Shi, "Deblurring Low-Light Images with Events," *International Journal of Computer Vision*, vol. 131, no. 5, pp. 1284–1298, May 2023.

[18] J. Liang, Y. Yang, B. Li, P. Duan, Y. Xu, and B. Shi, "Coherent Event Guided Low-Light Video Enhancement," in *Proceedings of the*

IEEE/CVF Conference on International Conference on Computer Vision, 2023, pp. 10 615–10 625.

[19] 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.

[20] 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.

[21] 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 ACM MM*, 2022, pp. 1484–1492.

[22] 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.

[23] 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 *Proceedings of the IEEE International Conference on Multimedia and Expo*. IEEE, 2021, pp. 1–6.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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 Conference on International Conference on Computer Vision*, 2023, pp. 12 438–12 448.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] 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.

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

[56] 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.

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

[58] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex Decomposition for Low-Light Enhancement," in *Proceedings of the British Machine Vision Conference*, Aug. 2018.

[59] 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.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *2014, arXiv:1412.6980*, pp. 1–11, 2014.
[61] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of ACM MM*, 2019, pp. 1632–1640.
[62] 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.
[63] 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.
[64] 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.
[65] 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.
[66] 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.
[67] 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.
[68] 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 Conference on International Conference on Computer Vision*, 2023, pp. 12 918–12 927.
[69] 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 Conference on International Conference on Computer Vision*, 2023, pp. 12 504–12 513.
[70] 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 Conference on International Conference on Computer Vision*, 2023, pp. 12 140–12 149.
[71] 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.
[72] 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.

Yan Huang received the B.Sc. degree in Intelligence Science and Technology from Hunan University in 2013, and the Ph.D. degree in Computer Science and Technology from South China University of Technology (SCUT) in 2018. She worked as a postdoctoral research fellow at SCUT from 2018 to 2020. She is currently an associate professor at the School of Computer Science and Engineering in SCUT. Her research interests include computer vision and deep learning.

Xiaoshan Liao received the B.E. degree from Nanjing University of Science and Technology, Nanjing, China, in 2021. She is currently pursuing an M.E. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Her research interests include image processing, image restoration, and computer vision.

Jinxiu Liang (Member, IEEE) received the B.E. and Ph.D. degrees from the South China University of Technology, Guangzhou, China, in 2016 and 2021, respectively. She is currently a Postdoctoral Research Fellow in Computer Science at Peking University, Beijing, China. Her main research interests include computational photography and computer vision. Her paper was awarded as Best Paper, Runners-Up at CVPR 2024. She serves as a reviewer for several computer vision (*e.g.*, CVPR, ICCV, ECCV, IJCV) and machine learning (*e.g.*, NeurIPS, ICLR) venues, and was recognized as an Outstanding Reviewer for IJCV 2023 and a Top Reviewer at NeurIPS 2024. She is a member of IEEE.

Boxin Shi (Senior Member, IEEE) received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Associate Professor (with tenure) and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper, Runners-Up at CVPR 2024, ICCP 2015, and selected as Best Paper candidate at ICCV 2015. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV/ECCV. He is a senior member of IEEE.

Yong Xu (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in mathematics from Nanjing University, Nanjing, China, in 1993, 1996, and 1999, respectively. He was a Postdoctoral Research Fellow of computer science with the South China University of Technology, Guangzhou, China, from 1999 to 2001, where he became a Faculty Member and is currently a professor at the School of Computer Science and Engineering. He is the Dean of the Guangdong Big Data Analysis and Processing Engineering & Technology Research Center. His current research interests include computer vision, pattern recognition, image processing, and big data. He is a senior member of the IEEE Computer Society and the ACM. He has received the New Century Excellent Talent Program of MOE Award.

Patrick Le Callet (Fellow, IEEE) is Full Professor at Ecole polytechnique de l'Université de Nantes (Engineering School) in the Electrical Engineering departement, and senior member of Institut Universitaire de France (IUF). He co-steering the CNRS LS2N lab (500+ researchers), as a representative of Polytech Nantes. He was the scientific director of the cluster "Ouest Industries Créatives", a cluster gathering more than 10 institutions (including 3 universities). "Ouest Industries Créatives" aims to strengthen the Research, Education & Innovation of the Region Pays de Loire in the field of Creative Industries. He is mostly engaged in research dealing with the application of human vision modeling in image and video processing. His current centers of interest are Quality of Experience assessment, Visual Attention modeling and applications, Perceptual Video Coding, and Immersive Media Processing. He is a co-author of more than 400 publications and communications and co-inventor of 16 international patents on these topics. He serves or has been served as associate editor or guest editor for several Journals such as IEEE Signal Processing Magazine, IEEE TIP, IEEE STSP, IEEE TCSVT, SPRINGER EURASIP Journal on Image and Video Processing, and SPIE JEI. He is serving in IEEE IVMSP-TC (2015- to present) and IEEE MMSP-TC (2015-to present) and is one the founding members of EURASIP SAT (Special Areas Team) on Image and Video Processing. He is a co-recipient of an Emmy Award in 2020 for his work on the development of Perceptual metrics for video encoding optimization.