Back Projection Generative Strategy for Low and Normal Light Image Pairs With Enhanced Statistical Fidelity and Diversity

Cheuk-Yiu Chan, *Student Member, IEEE*, Wan-Chi Siu, *Life Fellow, IEEE*, Yuk-Hee Chan, *Member, IEEE*, and H. Anthony Chan, *Life Fellow, IEEE*

***Abstract*—Low light image enhancement (LLIE) using supervised deep learning is limited by the scarcity of matched low/normal light image pairs. We propose Back Projection Normal-to-Low Diffusion Model (N2LDiff-BP), a novel diffusion-based generative model that realistically transforms normal-light images into diverse low-light counterparts. By injecting noise perturbations over multiple timesteps, our model synthesizes low-light images with authentic noise, blur, and color distortions. We introduce innovative architectural components - Back Projection Attention, BP^2 Feedforward, and BP Transformer Blocks - that integrate back projection to model the narrow dynamic range and nuanced noise of real low-light images. Experiment and results show N2LDiff-BP significantly outperforms prior augmentation techniques, enabling effective data augmentation for robust LLIE. We also introduce LOL-Diff, a large-scale synthetic low-light dataset. Our novel framework, architectural innovations, and dataset advance deep learning for low-light vision tasks by addressing data scarcity. N2LDiff-BP establishes a new state-of-the-art in realistic low-light image synthesis for LLIE.**

***Index Terms*—Low light image enhancement, image synthesis, generative model, diffusion, data augmentation.**

I. Introduction

THE RAPID progress in camera technology has greatly enhanced the ability of individuals to document important events and experiences in their daily lives. Nevertheless, a significant obstacle encountered by photographers of all skill levels is the deterioration of image quality in low illumination settings. Photographs captured under these circumstances frequently suffer from poor brightness and excessive noise, obscuring details and diminishing the aesthetic value of the image.

Low light image enhancement (LLIE) is an active area of research that has garnered significant attention from the computer vision community. With the rise of deep learning, researchers have increasingly sought neural network-based solutions for enhancing images captured in low light conditions. However, a key challenge is the limited availability of matched low/normal light image pairs required to train deep networks. Some researchers have attempted to synthesize more normal/low training images by applying simple linear and gamma transformations to normal-exposed images [1], such as VOC2007 [2] and ILSVRC2012 [3]. Others have used photo editing software like Adobe Lightroom to manipulate image brightness [4]. Some publicly available dataset like LOL-v2 [5] and VE-LOL [6] also employed low light image synthesis to enlarge the dataset volume. While these techniques do alter illumination and darkness, the resulting synthetic images may not match the true pixel distribution and image characteristics of real low light photos. More advanced rendering and augmentation procedures are needed to accurately model noise, blur, and color shifts induced by low light capture, remaining an open research direction on generating realistic training data that facilitate development of learning-based low light enhancement algorithms.

Diffusion probabilistic models [7], [8], [9], [10] have recently gained prominence as a highly effective generative modeling technique for various image processing applications, including image generation and super resolution. These models operate by progressively adding noise to an image over a series of timesteps, enabling the synthesis of visually compelling images that effectively capture intricate image statistics. The foundational concept behind this paper was initially presented in a 3-page conference paper at ICCE2024 [11]. However, the current version has undergone a comprehensive rewrite and incorporates substantial enhancements, leveraging the power of diffusion models to make the following key contributions (Code is available at https://github.com/allanchan339/N2LDiff-BP):

- We develop a novel Back Projection (BP) **N**ormal-to-**L**ow **Diff**usion Model (N2LDiff-BP) model from [11] that leverages diffusion processes to generate high fidelity low light images with realistic color distortions, noise, and blurring. Our model can generate multiple

Received 3 May 2024; revised 30 November 2024; accepted 30 November 2024. Date of publication 11 December 2024; date of current version 14 August 2025. This work was supported in part by the Saint Francis University (1SG200206), UGC under Grant UGC/FDS11/E05/22, and in part by The Hong Kong Polytechnic University, Hong Kong, SAR. *(Corresponding author: Wan-Chi Siu.)*

Cheuk-Yiu Chan and Wan-Chi Siu are with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China, and also with the School of Computing and Information Sciences, Saint Francis University, Hong Kong, China (e-mail: cy-allan.chan@connect.polyu.hk; enwcsiu@polyu.edu.hk).

Yuk-Hee Chan is with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: enyhchan@polyu.edu.hk).

H. Anthony Chan is with the School of Computing and Information Sciences, Saint Francis University, Hong Kong, China (e-mail: hhchan@sfu.edu.hk).

Digital Object Identifier 10.1109/TCE.2024.3516366

varied low light versions from an identical normal light image, enabling effective data augmentation for low light enhancement tasks.

- We introduce three novel components - Back Projection Attention Block (BPAttn), BP^2 Feedforward Block (BP^2F), and BP Transformer Block (BPT) - that integrate back projection techniques into attention, feedforward, and transformer architectures. Our specialized blocks leverage back projection principles to navigate the narrow dynamic range and complex noise characteristics inherent to low light images, generating more realistic underexposed representations.
- We construct a large-scale generative dataset, LOL-Diff, derived from applying our N2LDiff-BP model on VOC2007 [2] and UHD4K [12] datasets. This expands the limited availability of real-world low light data that constrains supervised learning approaches. Our generated images exhibit realistic noise, blurring and color distortions induced by low lighting. This high quality low light dataset will promote the development and benchmarking of deep learning models across tasks like image enhancement, classification, detection and segmentation under low illumination.

II. Preliminaries

A. Low Light Image Enhancement and Synthesis

Recent progress in low-light image enhancement (LLIE) [13], [14] has been driven by deep learning. Early methods applied retinex theory for illumination correction and artifact suppression [15], [16], but had limited generalization. More recent convolutional neural network (CNN) approaches have demonstrated superior perceptual quality, using Autoencoders [17], multi-scale fusion [18], [19], [20], [21], retinex [22] and CNNs parameterized by Gaussian kernels [22], [23], or two-stream architectures [15], [16]. Generative Adversarial Networks have also been explored for LLIE [24].

While deep learning methods have shown great promise, they require vast amounts of supervised training data. However, acquiring real-world low-light images is challenging due to practical constraints. To address this, prior work resorted to synthetically distorting normal exposures through linear adjustment and gamma correction [17]. Others leveraged commercial photo editors for manual brightness manipulation [4]. In the work of Lv et al. [1], low light images were artificially generated by applying linear and gamma adjustments to normally exposed images. This process can be mathematically expressed as follows:

$$I_{out}^{(i)} = \beta \times \left(\alpha \times I_{in}^{(i)}\right)^{\gamma}, i \in \{R, G, B\} \quad (1)$$

where $I_{in}^{(i)}$ and $I_{out}^{(i)}$ represent the input and output pixel values for each color channel i {Red, Green, and Blue}, respectively. The linear factors α and β, along with the gamma correction parameter γ, are employed to modify the pixel intensities. These parameters are randomly selected from uniform distributions, with α drawn from $\mathcal{U}(0.9, 1)$, β from $\mathcal{U}(0.5, 1)$, and γ from $\mathcal{U}(1.5, 5)$.

Existing deep learning-based low-light enhancement methods like AGLLNet [1] and DLN [25] typically rely on artificially darkening normal-exposed images to synthetically create more low-light training data. However, such a simplistic simulation approach may fail to faithfully replicate the nuanced statistics of real low-illumination photographs. Advancing physically-based rendering techniques could potentially generate more realistic training data and further boost the performance of learning-based enhancement methods, by taking into account the complex light transport characteristics under low-light conditions.

B. Evolution and Challenges in Low-Light Image Enhancement

Low-Light Image Enhancement (LLIE) has evolved significantly over the past decades. Traditional approaches, focusing on global adjustments and statistical methods [26], [27], often struggled with noise amplification and detail loss in extreme low-light conditions.

The emergence of deep learning techniques revolutionized Low-Light Image Enhancement, enabling more sophisticated and context-aware enhancements [16], [17]. However, LLIE still faces critical challenges, primarily due to the scarcity of large-scale, diverse pair datasets representing real-world low-light scenarios with normal-light counterparts. While datasets like ImageNet 2017 for object detection contain 14M images, the number of low light image datasets are extremely limited. The widely used LOL Dataset (proposed in 2018) [16] contains only 500 image pairs, while LOLv2 (proposed in 2021) expanded this to 789 pairs (exclude synthesis subset). This data scarcity significantly impedes the development of models capable of generalizing across varied real-world scenarios.

Existing synthetic data generation approaches, while valuable, often fail to accurately model the intricate characteristics of real low-light images, such as spatially-varying noise, complex color distortions, and subtle degradations inherent in authentic low-light photography. This disparity between synthetic and real data limits the effectiveness of data augmentation strategies and hinders LLIE model performance in practical applications. Addressing these challenges requires innovative approaches to generate realistic low-light images, potentially leveraging advanced techniques like generative models, e.g., N2LDiff-BP.

C. Diffusion Model

Modern generative techniques, especially diffusion models, have become a groundbreaking approach in artificial intelligence, significantly advancing the field of generative modeling. Drawing inspiration from physical processes like the diffusion of particles, these models work by gradually adding random noise to data, much like how particles disperse in a fluid over time. Introduced by Sohl-Dickstein et al. [28], diffusion models simulate this noisy transformation in a forward process, effectively breaking down complex data into simpler, noisier versions step by step. The innovative aspect lies in learning how to reverse this process: the model is trained to carefully remove the added noise in stages, reconstructing

the original data from its increasingly blurred states. This method transforms the daunting task of generating high-quality data into a series of manageable refinements, akin to gradually clarifying a fogged-up image.

Notably, diffusion models such as those in [7], [8], [29], [30], [31], [32], [33] have demonstrated the ability to generate high-quality, diverse samples and have achieved state-of-the-art results in many benchmarks. Their success stems from stable training processes that avoid common pitfalls like producing repetitive results, their capacity to understand and replicate complex patterns within the data, and their versatility in incorporating additional information, such as labels or descriptions, to guide the generation process. Furthermore, recent advancements have introduced techniques to accelerate the sampling process, addressing earlier challenges related to the iterative nature of these models. In essence, diffusion models bridge the gap between chaotic noise and coherent data through a series of thoughtful, incremental steps, embodying a natural and intuitive transformation that mirrors physical diffusion, and establishing themselves as a leading method in the ever-evolving landscape of artificial intelligence.

More specifically, diffusion models adopt a Markov chain approach, beginning from an input data point and gradually introducing stochastic disturbances over a series of timesteps. This "forward diffusion" process enables the training of a noise prediction network to learn the disturbances' probabilistic distribution. Concretely, it incrementally injects noise into the target domain as:

$$q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \tag{2}$$

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t \tag{3}$$

where $\boldsymbol{x}_t$ signifies the data observed at temporal index t. Parameters α_t and β_t are components of the noise schedule, adhering to the condition $\alpha_t + \beta_t = 1$. T is the maximum timestep in noise schedule. The variable $\boldsymbol{\epsilon}_t$ represents the noise added to the image at each timestep t, which is randomly sampled from a standard normal distribution $\mathcal{N}(0, \boldsymbol{I})$. As demonstrated by Ho et al. [8], the forward diffusion process at any given timestep t can be concisely expressed in the following manner:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{4}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$. The learning paradigm is thus cast as a task of predicting noise. In this framework, a noise predictor network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ is utilized to approximate the conditional probability distribution $p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. This probability is used for the reverse diffusion process that aims to restore the original data $\boldsymbol{x}_0$ from the noisy data $\boldsymbol{x}_T$. The objective is to refine the noise prediction to facilitate this reconstruction, as follows:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|_2^2, \text{ where } t \sim \mathcal{U}(1, T) \tag{5}$$

The noise predictor network, $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$, makes use of current noisy data $\boldsymbol{x}_t$ along with the time step t to estimate the noise perturbation $\boldsymbol{\epsilon}$, which has been incorporated into $\boldsymbol{x}_t$ through the forward process. In order to reverse this noise addition and reconstruct the original image — a process, termed as the reverse process, has been suggested making use of the following reverse equation:

$$\boldsymbol{x}_{t-1} = \mathcal{N}\left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{(1-\bar{\alpha}_t)}\boldsymbol{x}_0, \tilde{\beta}_t\boldsymbol{I}\right) \tag{6}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{7}$$

By exploiting the forward diffusion equation presented in Eq. (4), the predicted mean $\bar{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ is designed to accurately approximate the original data $\boldsymbol{x}_0$. This is achieved through the following formulation:

$$\bar{\boldsymbol{\mu}}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right) \tag{8}$$

By substituting Eq. (8) into Eq. (6) such that $\boldsymbol{x}_0 \coloneqq \bar{\boldsymbol{\mu}}_\theta(\boldsymbol{x}_t, t)$, the final reverse diffusion equation can be derived as follows:

$$\boldsymbol{x}_{t-1} = \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{(1-\bar{\alpha}_t)}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right), \tilde{\beta}_t\boldsymbol{I}\right) \tag{9}$$

Through the execution of the reverse process, the diffusion model is capable of retrieving the uncontaminated original data, $\boldsymbol{x}_0$, starting from the entirely noisy data $\boldsymbol{x}_T$ which is assumed to follow a normal distribution $N(0, \boldsymbol{I})$. The entire sequence, encompassing both the forward and reverse transitions, can be refined in an end-to-end manner using neural networks that encapsulate the parameters governing these processes.

D. Back Projection (BP)

Back-projection (BP) is a technique that has been widely employed in various image processing tasks, such as Super-Resolution (SR), Remote Sensory, and Low-Light Image Enhancement (LLIE) [25], [34], [35], [36], [37]. This section explores the application of BP in these domains and highlights the potential for extending the BP concept to novel frameworks that integrate with diffusion models for generating realistic representations of underexposed images.

In the context of SR, the BP technique traditionally utilizes multiple low-resolution (LR) images to reconstruct a single high-resolution (HR) image. The Deep Back-Projection Network (DBPN) [34] enhances the SR image quality by iteratively applying BP blocks, which aim to minimize the discrepancy between the LR images and their down-sampled SR counterparts. The BP block firstly down-samples the intermediate SR image $\hat{Y}_t$ and computes the residual with respect to the LR image X. The residual is then up-sampled to estimate the difference between the intermediate SR image and the true HR image. The SR image is refined by adding the scaled up-sampled residual to produce the updated SR image $\hat{Y}_{t+1}$:

$$\hat{Y}_{t+1} = \hat{Y}_t + \lambda U\Big(X - D(\hat{Y}_t)\Big) \tag{10}$$

where $D(\cdot)$ and $U(\cdot)$ represent the down-sampling and up-sampling operations, respectively, and $\lambda \in \mathbb{R}$ is a balance coefficient.

(a) Seed 42 (b) Seed 0

Fig. 1. Qualitative comparison on different seeds, which shows the stochastic nature of our generative model for data augmentation, with changes in output images corresponding to variations in input noise.

In the domain of LLIE, Wang et al. [25] introduced the Lightening Back-Projection (LBP) block, which adapts the BP framework for enhancing images captured in low-light conditions. The LBP block aims to estimate the difference between the input low-light image and its enhanced version by applying a series of lightening and darkening operations. The enhanced image $\hat{Y}$ is obtained by adding the scaled residual to the lightened input image:

$$\hat{Y} = L_2(X - D(L_1(X))) + L_1(X) \tag{11}$$

where $L_1(\cdot)$ and $L_2(\cdot)$ are designed as lightening operators, and $D(\cdot)$ represents the darkening operation. The lightening operators $L_1(\cdot)$ and $L_2(\cdot)$ are responsible for enhancing the low-light image, while the darkening operator $D(\cdot)$ is used to estimate the residual between the lightened image and the original input. By iteratively applying the LBP block, the network can effectively learn to enhance low-light images, preserving the overall structure and details while improving the visibility of the scene.

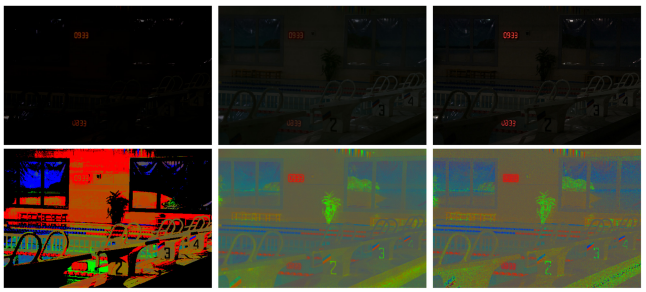
The success of BP in SR and LLIE tasks demonstrates its potential for application in other image processing domains. By incorporating the BP technique into novel frameworks, such as those integrating diffusion models, researchers can explore new avenues for generating high-quality, realistic representations of images captured in challenging conditions, such as low-light image synthesis.

III. Methodology

A. Back Projection Normal-to-Low Diffusion Model (N2LDiff-BP)

Our method employs a generative approach to synthesize diverse and realistic under-exposed images from normal-exposed images, offering two significant advantages over linear/gamma transformations or deterministic models. First, it eliminates the need for manual parameter adjustment to achieve optimal results, such as Eq. (1). Second, the stochastic nature of diffusion models enables the generation of varied low light augmentations from each input, as illustrated in Fig. 1. This data augmentation is crucial for training robust networks. Consequently, N2LDiff-BP augments the training data, increasing both volume and diversity.

Our approach effectively models the complexities of low-light environments through diffusion, capturing intricate noise patterns and color distortions via Eq. (18). As illustrated in Fig. 2, our model demonstrates its capability to produce



(a) Lv *et al.* [1] (b) N2LDiff-BP (c) Ground Truth

Fig. 2. Sample comparison of Low-Light Image Simulation Methods. Top row: simulated low-light images. Bottom row: visualization of noise and color distortions map.

realistic low-light images, with the top row displaying simulated low-light images and the bottom row visualizing the corresponding noise and color distortions. Specifically, column (a) presents the results from Lv at al.'s method [1], column (b) showcases our proposed N2LDiff-BP results, and column (c) provides the ground truth. Lv at al.'s method [1] (a) does not accurately reproduce the true noise distribution and perform incorrect color shifts seen in low-light conditions, as evidenced by the dominant orange-red color shift in the wall. In contrast, our N2LDiff-BP results (b) more closely match the ground truth (c) by aligning more closely with the ground truth in terms of darkening level, maintaining a similar saturation level that accurately represents the low-light scene effectively. Notably, our approach successfully reproduces intrinsic noise patterns, as indicated by the green tints near the swimming pool starting blocks, replicates subtle color distortions and achieves realistic detail degradation, for which method (a) fails to capture. These shortcomings in simulating realistic low-light conditions lead to suboptimal training of enhancement algorithms. Specifically, inaccurate noise patterns, color distortions, and detail preservation in the simulated images can cause enhancement methods to incorrectly handle noise, misadjust colors, or fail to reconstruct details that are indiscernible in true low-light conditions.

The proposed model consists of two sub-networks: an encoder ϕ_e as illustrated in Fig. 3 and a diffusion noise predictor, as illustrated in Fig. 4. The encoder ϕ_e takes the normal-exposed image x_H as input and outputs a latent feature map ϕ. The diffusion noise predictor then takes both ϕ and the noisy input x_t, which is derived from the normal-exposed image x_H according to Eq. (12), to predict a noise perturbation $\boldsymbol{\epsilon}_\theta$. To generate the low-light image output x_L, the reconstruction equation Eq. (13) is applied repeatedly, utilizing the predicted noise map $\boldsymbol{\epsilon}_\theta$. At each step, we apply the equation to predict and remove the noise from the input, gradually transforming it into a structured output. This reconstruction equation is carefully designed to employ the latent feature map ϕ as an anchor during the image reconstruction process. By incorporating ϕ in this manner, the intermediate output $\boldsymbol{x}_{t-1}$ is better anchored, leading to improved image quality in the final low-light image output.

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_H + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t^\star; \quad \boldsymbol{\epsilon}_t^\star \sim \mathcal{N}\left(\frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\phi, \boldsymbol{I}\right) \tag{12}$$

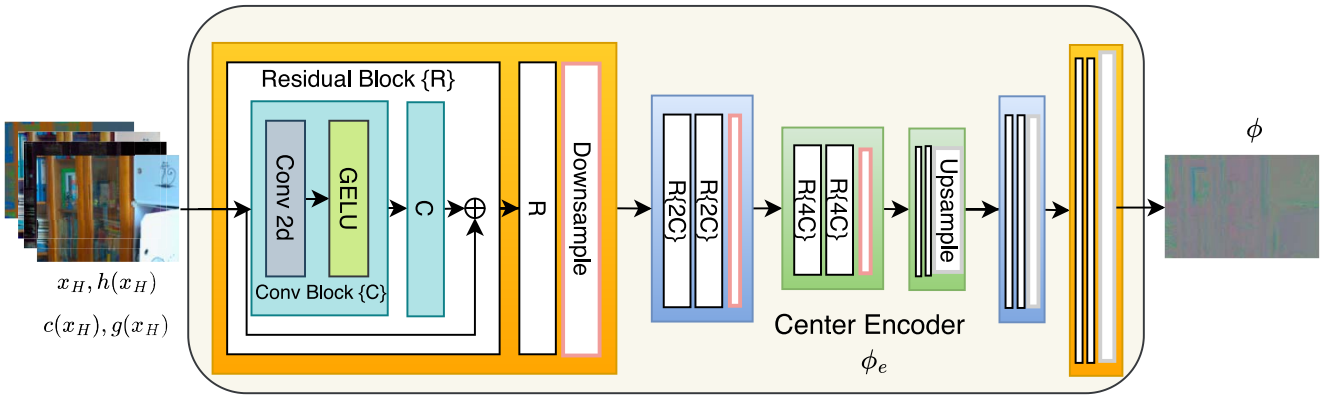


Fig. 3. Architecture of N2LDiff-BP center encoder ϕ_e, where the center encoder shares a UNet structure. The notation c, 2c and 4c after the block name means the channel size of each block w.r.t c. "Conv Block", "Res Block", "Downsample" and "Upsample" denote 2D-Convolution block, residual block, downsampling layer and upsampling layer respectively.

$$\boldsymbol{x}_{t-1} = \mathcal{N}\Bigg(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{x}_H)\right) + \frac{1-\bar{\alpha}_t + \sqrt{\bar{\alpha}_{t-1}}(\alpha_t - 1) + \sqrt{\alpha_t}(\bar{\alpha}_{t-1} - 1)}{1-\bar{\alpha}_t}\phi, \boldsymbol{I}\Bigg) \tag{13}$$

While the encoder in Fig. 3 and noise predictor sub-networks in Fig. 4 share some similarities in their overall design, there is a notable difference in latent level. The noise predictor incorporates a Back Projection Vision Transformer (BPViT) which sets it apart from the encoder. BPViT is specifically designed to better capture and address the unique noise characteristics inherent in low-light images. The BPViT within the noise predictor consists of three novel components: the Back Projection Attention Block, the BP^2 Feedforward block, and the BP Transformer Block. These components are discussed in detail in the following sections.

B. Back Projection Attention Block (BPAttn)

Forward propagation is initially performed to transfer from the input domain (e.g., normal light image x_H) to the output domain (e.g., low light image x_L) via the forward mapping function $x_L = f(x_H)$. However, this approach omits the fact that the input domain likely contains useful information for enhancing the output. Thus, we consider the back-projection theory, which applies reverse mapping, $x_H = r(f(x_H))$, to transfer the output back to the input domain, as illustrated in depicted blue block in Fig. 5. The goal is to minimize the discrepancy between the original input and the input's reconstruction. In an ideal scenario, the residual error $x_H - r(f(x_H)) = 0$. However, in practice, a non-zero residual indicates lost detail information during forward mapping to be recovered. Therefore, we can introduce a residual mapping function $g(r(f(x_H)))$ that transforms residual information back to the output domain. The improved forward mapping becomes:

$$x_L = f(x_H) + \lambda * g(r(f(x_H))) \tag{14}$$

where λ controls the degree of residual incorporation. This allows us to utilize the residual domain to enhance the forward mapping results in a principled manner.

The Back Projection Attention Block (BPAttn), as illustrated in white block of Fig. 4, is a specialized component designed to enhance the performance of low-light image synthesis tasks. Unlike traditional attention blocks that only contains forward mapping, BPAttn accounts for the nuanced dynamics of low-light imaging with enhanced reverse mapping mechanism by back projection theory to narrow the dynamic range of normal light image, which is adeptly employed the flash attention [38] within neural networks.

More specifically, unlike traditional attention blocks, BPAttn incorporates an additional step that aims to capture and correct the discrepancies between the input and their predicted representations. This is akin to a form of error correction, where the attention mechanism is not only focusing on relevant features but also on rectifying the information that may have been lost or misrepresented in earlier layers of the network.

The BPAttn approach employs a reverse mapping function, which can be conceptually represented as $x'_H = r(f(x_H))$, where x_H symbolizes the normal light image and f denotes the forward mapping function that typically transforms the image into a low-illumination counterpart. The reverse mapping r attempts to project the counterpart back into the normal light domain. Ideally, if the darkening (D) and lightening operations (L) perfectly preserved all relevant information, the identity $x_H = r(f(x_H))$ holds true. In practice, however, there is often a residual difference, denoted as $\gamma = x_H - r(f(x_H))$, indicating residual information loss. BPAttn targets this residual by introducing a residual mapping function $g(\gamma)$, which is designed to re-incorporate lost details back into the latent space. The enhanced latent features are then represented as:

$$x'_L = f(x_H) + \lambda * g(\gamma) \tag{15}$$

In this equation, λ is a tunable parameter that controls the degree of residual information reintegrated into the feature space. This innovative mechanism is particularly beneficial in low-light conditions, where it is crucial to capture as much detail as possible to produce a high-quality enhanced image.

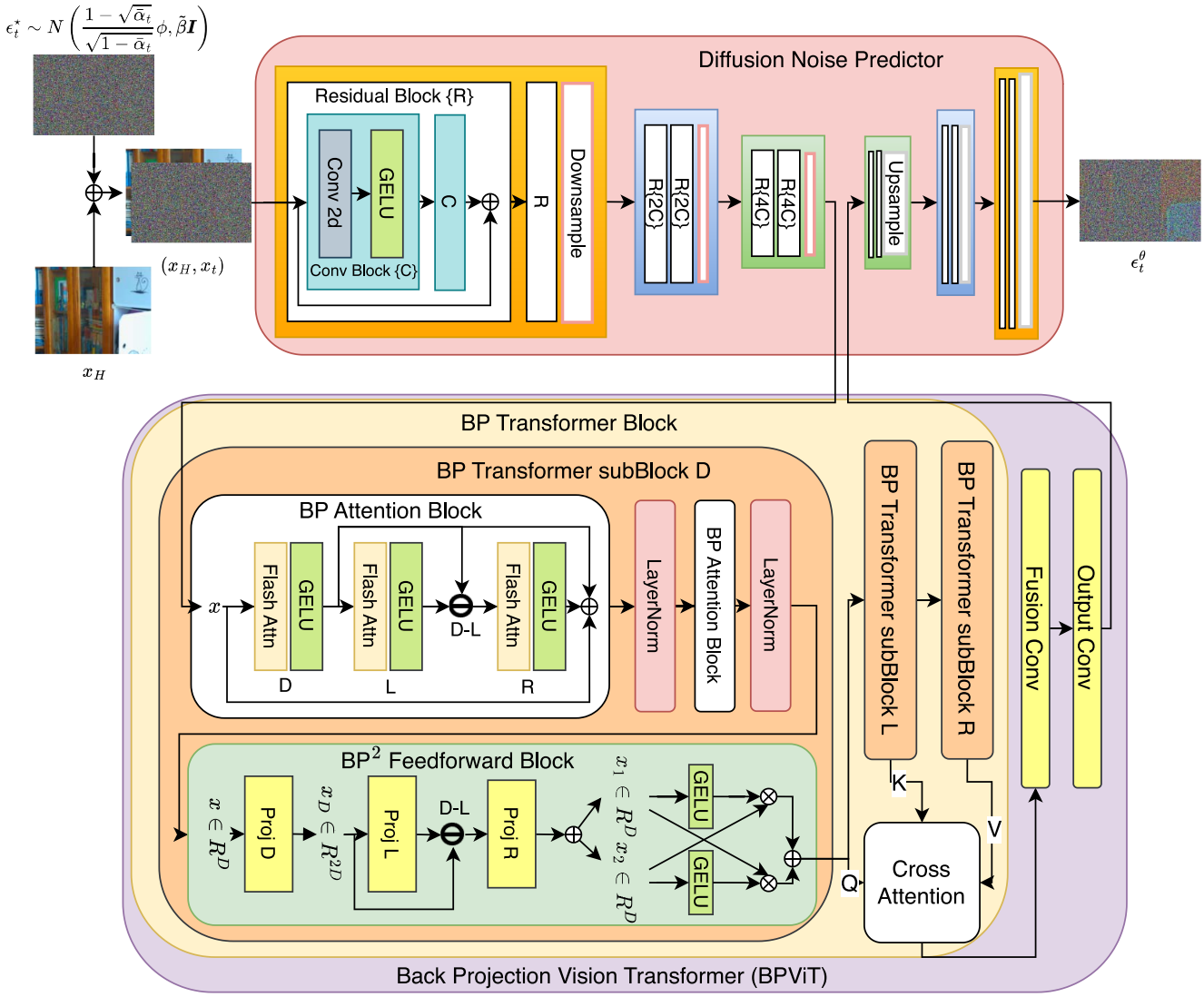


Fig. 4. Architecture of N2LDiff-BP's Diffusion Noise Predictor, which is a UViT structure. The BPViT employs an embedding layer as the latent input, followed by 3 BP Transformer Blocks. Each BP Transformer Block utilizes a BPAttn Block and a BP^2 Feedforward Block to capture more realistic representations of underexposed images. This is achieved by leveraging the BP technique to successfully navigate the narrow dynamic range inherent to low-light imaging scenarios. Lastly, output of 3 BP transformer will be merged and processed by a Cross Attention Block, a Fusion Block and an output Conv Block to generate the predicted noise in latent space.

BPAttn's capability to back-project and refine the attention process ensures that even subtle features are not overlooked, thus enabling the generation of more detailed and realistic low-light images.

C. BP^2 Feedforward (BP^{2}F) Block

We introduce a novel BP^2 Feedforward Block that enhances standard Feed-Forward Networks (FFNs) [39], [40] by incorporating back projection bidirectionally to address their limitations in capturing local context. Unlike traditional FFNs that process information sequentially in a unidirectional flow, our BP^2 Feedforward Block employs a bidirectional approach that performs reverse mapping within the same block.

Specifically, the BP^2 Feedforward Block applies a dual back projection mechanism alongside GELU activation, where GELU is the Gaussian Error Linear Unit activation function [41]. The choice of GELU over ReLU is motivated by its ability to provide smoother gradients and mitigate the "dying ReLU" problem, where neurons can become inactive and stop learning [42]. Also, GELU's non-linear transformation retains more information during forward passes, crucial for capturing subtle details in low-light images. This choice enhances the network's ability to learn and preserve fine-grained features between transitions from normal light to low light images. BP^{2}F fuses useful information from parallel processing paths, as shown in green block of Fig. 4, to reintegrate lost details during the forward pass while retaining the original input representation. This dual integrated approach enables bidirectional refinement of hierarchical features via back projection. The complete mathematical formulation leverages back projection to refine feature transformations, improving context modeling over standard FFNs. Our BP^2 Feedforward Block innovates by enabling joint forward and

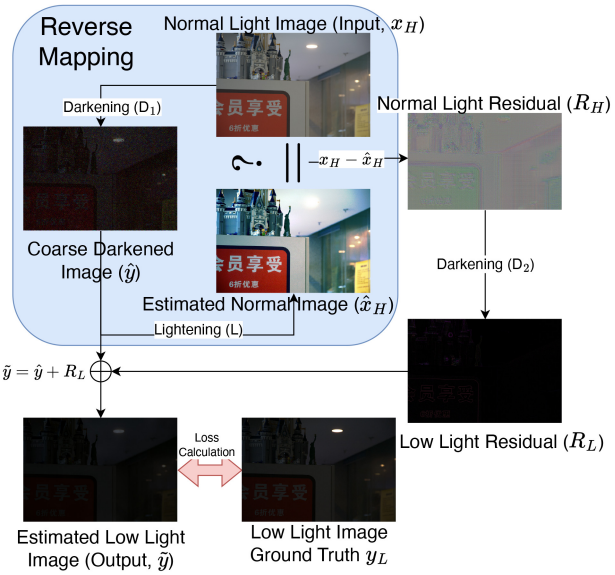


Fig. 5. Illustration of Back Projection and Reverse Mapping on N2LDiff-BP: the back projection capture low lighting nuances by modeling the discrepancy between features and ideal representations, then incorporating this residual information to rectify errors and reconstruct lost details.

reverse processing to bidirectionally enhance representations through an integrated back projection approach. The complete BP^2 Feedforward Block is formulated as:

$$\begin{aligned} BPF &= \big(x_D \odot Proj_R(x_D - x_L)\big) \\ [x_1, x_2] &= \text{split}(BPF(x)) \\ BP^2F &= GELU(x_1) \odot x_2 + GELU(x_2) \odot x_1 \end{aligned} \tag{16}$$

where BP^2F signifies the dual back-projection mechanism, $\odot$ represents the element-wise multiplication, split$(\cdot)$ represents the function that splits the input tensor into two equal parts along the channel dimension, resulting in x_1 and x_2.

D. Back Projection Transformer (BPT) Block

We propose a novel Back Projection Transformer Block that integrates hierarchical features using back projection principles, as shown in orange blocks of Fig. 4. The BPT Block comprises three sub-transformer blocks, each embedded with our proposed BPAttn and BP^2 Feedforward Blocks. This architecture culminates in a cross-attention block for adaptively fusing features across layers.

A key novelty is the application of back projection theory to enable reverse mapping in the transformer architecture. The cross-attention layer incorporates reverse mapping principles, allowing the model to reconstruct lost details and rectify representation errors. It achieves this by modeling the discrepancy between current features and their ideal representations. The cross-attention then uses this residual information to refine the integrated hierarchical features. This application of back projection provides the transformer access to reverse mapping, improving its ability to capture nuances and refine representations.

Another innovation is employing cross-attention in back projection for efficiency, which is critical for complex back projection under constraints. Unlike standard back projection

(a) x_H (b) $h(x_H)$ (c) $c(x_H)$ (d) $g(x_H)$

Fig. 6. Illustration of pre-processing steps for input image x_H for encoder ϕ_e. (a) Original image x_H. (b) Contrast-enhanced image $h(x_H)$. (c) Channel-normalized image $c(x_H)$. (d) High-frequency image $g(x_H)$.

that requires an extra module to transform residual representations, as shown in Fig. 5, the BPT Block implicitly handles reverse mapping and residual integration in its cross-attention. This concurrently fuses features and refines them with residual information, balancing complexity and resource usage. The cross-attention's integrated design leads to savings in memory and computations compared to extra transformation modules in typical back projection. The operation can be mathematically represented as:

$$\begin{aligned} &Y_D, Y_L, Y_R = \text{sub-BPT}(x) \quad \text{for each layer;} \\ &Y_{\text{BPT}} = \text{CrossAttn}(Y_D, Y_L, Y_R) = \text{softmax}\left(\frac{Y_D Y_L^T}{\sqrt{d_k}}\right) Y_R; \end{aligned} \tag{17}$$

where x is the input, Y_i is the feature map from the i-th layer, Y_{BPT} is the output from the cross-attention fusion that processes hierarchical feature with back projection.

IV. Experiments

A. Dataset

We optimized and evaluated our model using the publicly available LOL [16], LOL-v2 [5], and VE-LOL [6] datasets, which comprise real low light images captured in diverse scenarios. Only the real-captured low light images from each dataset were utilized for training and testing based on default splits. Amalgamating multiple real image datasets could expose the model to diverse low light conditions during training.

B. Preprocessing and Implementation Details

We utilized an encoder network ϕ_e to obtain the non-zero mean perturbation vector ϕ, as illustrated in Fig. 3. In this work, we augmented the normal-light input image $\boldsymbol{x}_H$ with multiple pre-processed representations to provide illumination-invariant information to the perturbation encoder ϕ_e. Specifically, we integrated a contrast-enhanced version $h(\boldsymbol{x}_H)$, a channel-normalized version $c(\boldsymbol{x}_H)$, and a high-frequency version $g(\boldsymbol{x}_H)$, as illustrated in Fig. 6. The contrast-enhanced image equalizes the histogram to improve visibility in dark regions. The channel-normalized version re-weighted color channels based on overall pixel intensity. Finally, the high-frequency version computes gradient magnitudes to capture edge details. Fusing these diverse image views provides ϕ_e with richer representations of $\boldsymbol{x}_H$ for determining the perturbation vector, where $c(\boldsymbol{x}_H)$ is defined as:

$$c(x_{i,j}) = \frac{x_{i,j}}{(R_{i,j} + G_{i,j} + B_{i,j})/3} \tag{18}$$

(a) Input (b) α=0.9,β=0.5,γ=1.5 (c) α=0.9,β=0.5,γ=5

(d) α=1,β=1,γ=1.5 (e) Ours (f) Ground Truth

Fig. 7. A qualitative comparison of simulated and generated low light Image from simulated-based approach and N2LDiff.

and variables $R_{i,j}$, $G_{i,j}$, and $B_{i,j}$ represent the red, green, and blue channel values, respectively, for the pixel at row i and column j in the image. Similarly, $g(\boldsymbol{x}_H)$ is defined as:

$$g(x_{i,j}) = \max\{|\nabla_x c(x_{i,j})|, |\nabla_y c(x_{i,j})|\} \tag{19}$$

where ∇_x, and ∇_y are the image gradients in horizontal and vertical direction.

The perturbation vector ϕ was computed by a trainable encoder network ϕ_e that takes the normal-light image $\boldsymbol{x}_H$ and its pre-processed versions as input:

$$\phi = \phi_e(\boldsymbol{x}_H, h(\boldsymbol{x}_H), c(\boldsymbol{x}_H), g(\boldsymbol{x}_H)) \tag{20}$$

where the concatenated input passes through ϕ_e for further encoding. The loss function combines LPIPS [43] and the diffusion process loss, which is defined as:

$$\mathcal{L}(\boldsymbol{x}_0, \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t^\theta) = \mathcal{L}_{\text{LPIPS}}\Big(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t^\theta\Big) \tag{21}$$

where $\boldsymbol{x}_0$ is the original noiseless input, α_t is the noise schedule, $\boldsymbol{\epsilon}_t^\star$ is the sampled noise perturbation in Eq. (12) and $\boldsymbol{\epsilon}_t^\theta$ is the predicted noise perturbation from model in Fig. 4.

C. Training

The training procedure optimized the parameters of the encoder ϕ_e and the diffusion model's noise predictor $\boldsymbol{\epsilon}_\theta$ jointly. Cosine noise schedule was employed [30]. The training process is outlined in Algorithm 1. The maximum number of timesteps was set to $T = 100$. Stochastic gradient descent was utilized with a batch size of 16 examples. Lion optimizer [44] was employed to update weights from each batch. The BPAttn set as $\lambda = 1$. The learning rate was initialized at 0.0004. Two RTX4090 GPUs were leveraged to accelerate the process.

Algorithm 1: Training Procedures of N2LDiff-BP

Input: Low Light (LL) image and its corresponding Normal Light (NL) image pairs $\boldsymbol{P} = \{\boldsymbol{x}_L^i, \boldsymbol{x}_N^i\}_{i=1}^I$, total diffusion step T

Initialize: noise predictor $\boldsymbol{\epsilon}_\theta$ with center encoder ϕ_e randomly;

repeat

Sample $(\boldsymbol{x}_L, \boldsymbol{x}_N) \sim \boldsymbol{P}$; $x_0 = x_N$

Sample $\boldsymbol{\epsilon}_t^\star \sim \mathcal{N}(\frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\phi_e(\boldsymbol{x}_H, \dots), \tilde{\beta}_t\boldsymbol{I})$, where $t \sim \mathcal{U}(1, T)$

Compute $\boldsymbol{\epsilon}_t^\theta = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c} := \boldsymbol{x}_H)$, where $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t^\star$

Take gradient step on $\mathcal{L}(\boldsymbol{x}_0, \boldsymbol{\epsilon}_t^\star, \boldsymbol{\epsilon}_t^\theta)$ with respect to $\boldsymbol{\epsilon}_\theta$ and ϕ_e

until converged;

After training for sufficient epochs, the total time taken to converge was approximately 46 hours.

D. Results

In the context of the uniform distribution $\mathcal{U}(a, b)$ in Eq. (1) [1], which is continuous, we applied interpolation on parameters a and b for a more nuanced comparison. Our experimental evaluation illustrates the effectiveness of our proposed generative strategy for enhancing low-light images. The quantitative results are summarized in Table I, where we compare our method against established simulation-based techniques across a variety of metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [45], and Learned Perceptual Image Patch Similarity

TABLE I
QUANTITATIVE COMPARISON ON LOL [16], LOL-V2 [5] AND VELOL [6] DATASETS IN TERMS OF PSNR, SSIM AND LPIPS. ↑ (↓) DENOTES THAT, LARGER (SMALLER) VALUES LEAD TO BETTER QUALITY. (**BOLD** REPRESENTS THE BEST)

Method	α	β	γ	LOL [16] PSNR ↑	LOL [16] SSIM ↑	LOL [16] LPIPS ↓	LOL-v2 [5] / VELOL [6] PSNR ↑	LOL-v2 [5] / VELOL [6] SSIM ↑	LOL-v2 [5] / VELOL [6] LPIPS ↓
Lv *et al.* [1]	0.9	0.5	1.5	21.418	0.609	0.251	26.578	0.667	0.169
	0.9	0.5	3.25	29.082	0.706	0.159	28.992	0.61	0.193
	0.9	0.5	5	27.096	0.492	0.236	26.435	0.331	0.358
	0.9	0.75	1.5	15.784	0.442	0.36	19.676	0.506	0.272
	0.9	0.75	3.25	25.483	0.702	0.199	28.258	0.672	0.171
	0.9	0.75	5	26.928	0.537	0.225	26.524	0.384	0.319
	0.9	1	1.5	12.333	0.334	0.44	15.689	0.385	0.357
	0.9	1	3.25	22.049	0.666	0.242	26.39	0.692	0.178
	0.9	1	5	25.523	0.557	0.233	26.209	0.422	0.296
	1	0.5	1.5	19.113	0.544	0.292	23.75	0.609	0.206
	1	0.5	3.25	26.195	0.706	0.191	28.523	0.664	0.172
	1	0.5	5	26.461	0.548	0.226	26.452	0.401	0.308
	1	0.75	1.5	13.842	0.381	0.404	17.391	0.439	0.318
	1	0.75	3.25	21.373	0.657	0.252	25.951	0.693	0.181
	1	0.75	5	23.896	0.563	0.248	25.565	0.453	0.283
	1	1	1.5	10.576	0.283	0.483	13.764	0.323	0.405
	1	1	3.25	17.869	0.599	0.307	23.419	0.687	0.209
	1	1	5	21.506	0.556	0.276	24.174	0.484	0.276
Ours				**29.614**	**0.738**	**0.134**	**29.39**	**0.728**	**0.148**

(LPIPS) [43]. These metrics collectively assess the quality, perceptual quality, and accuracy of the output.

The PSNR and SSIM are both essential for evaluating the restoration quality, where higher values of PSNR and SSIM indicate better image reconstruction. Conversely, a lower LPIPS value is preferred as it implies higher perceptual similarity to the ground truth. In our results, the superiority of our approach is evident, with our method outperforming the competing simulation-based techniques across all these metrics, thereby establishing a new state-of-the-art performance in low-light image synthesis.

Visual comparisons are provided in Fig. 7, where the qualitative improvements rendered by our generative model are clearly observable. Our method not only accurately reproduces the low-light pixel distributions but also maintains the integrity of the image content, producing visually pleasing and realistic enhancements. These visual results corroborate the numerical metrics presented in Table I, further validating our approach as a potent alternative to traditional simulation-based augmentation methods.

E. Ablation Study

In this section, we present an ablation study designed to evaluate the contribution of each component within our proposed model. We leverage the Low-Light (LOL) dataset introduced by Wei et al. [16], which can be considered as a benchmark for low-light image enhancement tasks. Our ablation study methodically deconstructs the model by removing individual blocks, thereby isolating the impact of each on the overall performance.

The primary objective of this study is to understand the significance of each component and how they synergize to improve the model's efficacy. We focus on three key metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). These metrics provide a comprehensive assessment of our model's ability to recover high-quality images from low-light conditions.

TABLE II
ABLATION STUDY OF BPATTN BLOCK

Block λ	Metric PSNR ↑	SSIM ↑	LPIPS ↓
$\lambda = 0$	27.751	0.723	0.154
$\lambda = 0.5$	28.465	0.704	0.193
$\lambda = 1$	**29.614**	**0.738**	**0.134**
$\lambda = 1.5$	28.574	0.718	0.172

1) Back Projection Attention Block (BPAttn): Our first ablation focuses on the BPAttn Block, which leverages attention mechanisms and back projection theory to better capture contextual information in low-light images. Table II presents the results of the model with different λ values, including $\lambda = 0$, which effectively disables the BPAttn Block. As demonstrated in the table, the presence and proper tuning of the BPAttn Block lead to substantial improvements in all metrics, underscoring its critical role in enhancing image quality.

Parameter λ in Eq. (15) controls the degree of residual information reintegrated into the feature space. When $\lambda = 0$, the model relies solely on the forward mapping function, while increasing λ incorporates more residual information. The optimal value of $\lambda = 1$ which yields the best results across all metrics, indicating an ideal balance between forward mapping and residual information.

Interestingly, when $\lambda = 1.5$, it shows a slight decrease in performance compared to $\lambda = 1$, suggesting that over-emphasizing residual information may introduce noise. Conversely, $\lambda = 0.5$, it improves upon $\lambda = 0$ but falls short of $\lambda = 1$, indicating that some, but not all, residual information is beneficial for image enhancement.

2) BP^2 Feedforward (BP^2F) Block: Next, let us examine the BP^2 Feedforward Block, which employs a dual back projection feedforward block for efficient feature activation. The ablation results are shown in Table III. The inclusion of

TABLE III
ABLATION STUDY OF BP^2 FEEDFORWARD BLOCK

Block	Metric		
	PSNR ↑	SSIM ↑	LPIPS ↓
w/o BP^2 Feedforward Block	28.026	0.696	0.145
w/ BP^2 Feedforward Block	**29.614**	**0.738**	**0.134**

TABLE IV
ABLATION STUDY OF BP TRANSFORMER FEEDFORWARD BLOCK

Block	Metric		
	PSNR ↑	SSIM ↑	LPIPS ↓
w/o BP Transformer Block	29.056	0.725	0.164
w/ BP Transformer Block	**29.614**	**0.738**	**0.134**

the BP^2 Feedforward Block yields significant improvements in PSNR, SSIM and LPIPS, indicating its effectiveness in the model's architecture.

3) BP Transformer (BPT) Block: Lastly, we consider the BP Transformer Block, which integrates a transformer-based architecture to model long-range dependencies within the images. The results of this ablation are summarized in Table IV. The inclusion of the BP Transformer Block yields significant improvements in all these metrics, indicating of its effectiveness in the transformer design.

V. PROPOSED DATASET

We present LOL-Diff, a large-scale dataset of synthetically generated low light images derived from the VOC2007 [2] and UHD4K [12] datasets. LOL-Diff comprises two subsets: a low resolution subset based on VOC2007 containing a total of 9,965 images, with 5,011 allocated for training/validation and 4,952 for testing; and a high resolution (4K) subset based on UHD4K consisting of 5,999 training images and 2,100 testing images. The generated low light images simulate a wide range of real-world low illumination conditions, as illustrated in Fig. 8.

The LOL-Diff dataset enables the training and benchmarking of deep learning models for tasks such as low light image enhancement, object detection, and semantic segmentation. These synthetically generated images serve to supplement the limited availability of real low light image corpora, providing a scalable data source for model development and evaluation.

To validate the effectiveness of LOL-Diff, we conducted a small comparative experiment using our N2LDiff-BP model with inverted input and output on the LOL dataset, with and without the inclusion of 500 image pairs subset from LOL-Diff. The results demonstrate the significant impact of incorporating LOL-Diff in the training process. When trained with LOL-Diff, the model achieved superior performance metrics: a PSNR of 16.836 dB, SSIM of 0.706, and LPIPS of 0.301. In contrast, without LOL-Diff, the model's performance was lower, with a PSNR of 16.52 dB, SSIM of 0.722, and LPIPS of 0.344. (Note that lower scores of both the SSIM and LPIPS mean the better results.) The improved PSNR and LPIPS indicate that the model trained with LOL-Diff produces enhanced image quality with better perceptual similarity to the ground truth. The extended training data volume demonstrates that LOL-Diff provides richer and more diverse learning examples, allowing for better convergence. These results underscore the value of LOL-Diff in enhancing both the training process and the contribution of our proposed approach towards low-light image enhancement tasks.

Fig. 8. Examples of LOL-Diff dataset with generated low light images (left) and ground truth (right).

The LOL-Diff dataset aims to facilitate further advancements in the application of deep learning techniques to low light computer vision tasks. By providing a diverse and large-scale dataset specifically tailored for low light conditions, LOL-Diff contributes a valuable resource to drive progress in this challenging domain.

VI. CONCLUSION

In this paper, we present N2LDiff-BP, a novel diffusion-based generative model for synthesizing realistic low light images, which surpasses current augmentation techniques in generating diverse, high-quality results. Key innovations include specialized attention and transformer blocks that capture low light nuances by back projection techniques. Additionally, our work introduces the LOL-Diff dataset, expanding training resources for low light imaging tasks. This confirms the potential of diffusion models in realistic image synthesis and sets the stage for future exploration in low light video and other vision applications.

REFERENCES

[1] F. Lv, Y. Li, and F. Lu, "Attention guided low-light image enhancement with a large scale low-light simulation dataset," *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2175–2193, 2021.
[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. "The PASCAL visual object classes challenge 2007 (VOC2007) results." Accessed: Nov. 30, 2024.[Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
[3] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.* vol. 115, no. 3, pp. 211–252, 2015.
[4] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, and T. Lu, "Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method," 2022, *arXiv:2212.11548*.
[5] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep Retinex network for robust low-light image enhancement," *IEEE Trans. Image Process.*, vol. 30, pp. 2072–2086, 2021.
[6] J. Liu, X. Dejia, W. Yang, M. Fan, and H. Huang, "Benchmarking low-light image enhancement and beyond," *Int. J. Comput. Vis.*, vol. 129, pp. 1153–1184, Jan. 2021.
[7] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. 35th Adv. Neural Inf. Process. Syst.*, 2021, pp. 8780–8794.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. 34th Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.
[9] C.-C. Hui, W.-C. Siu, N.-F. Law, and H. A. Chan, "Intelligent painter: New masking strategy and self- referencing with resampling," in *Proc. 24th Int. Conf. Digit. Signal Process. (DSP)*, 2023, pp. 1–5.
[10] C.-Y. Chan, W.-C. Siu, Y.-H. Chan, and H. A. Chan, "AnlightenDiff: Anchoring diffusion probabilistic model on low light image enhancement," *IEEE Trans. Image Process.*, vol. 33, pp. 6324–6339, 2024.
[11] C.-Y. Chan, W.-C. Siu, Y.-H. Chan, and H. A. Chan, "Generative strategy for low and normal light image pairs with enhanced statistical fidelity," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, 2024, pp. 1–3.
[12] K. Zhang et al., "Benchmarking ultra-high-definition image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14749–14758.
[13] Z. Zhang, H. Zheng, R. Hong, J. Fan, Y. Yang, and S. Yan, "FRC-Net: A simple yet effective architecture for low-light image enhancement," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3332–3340, Feb. 2024.
[14] N. Singh and A. K. Bhandari, "Noise aware L2-LP decomposition-based enhancement in extremely low light conditions with Web application," *IEEE Trans. Consum. Electron.*, vol. 68, no. 2, pp. 161–169, May 2022.
[15] Y. Wang et al., "Progressive Retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement," in *Proc. 27th ACM Int. Conf. Multimedia*, 2023, pp. 2015–2023.
[16] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.
[17] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep Autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.
[18] F. Lv, F. Lu, J. Wu, and C. Lim, "MBLLEN: Low-light image/video enhancement using CNNs," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–13.
[19] W. Ren et al., "Low-light image enhancement via a deep hybrid network," *IEEE Trans. Image Process.*, vol. 28, pp. 4364–4375, 2019.
[20] L. Tao, C. Zhu, G. Xiang, Y. Li, H. Jia, and X. Xie, "LLCNN: A convolutional neural network for low-light image enhancement," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, 2017, pp. 1–4.
[21] Z. Ji, H. Zheng, Z. Zhang, Q. Ye, Y. Zhao, and M. Xu, "Multi-scale interaction network for low-light stereo image enhancement," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3626–3634, Feb. 2024.
[22] S. Park, S. Yu, B. Moon, S. Ko, and J. Paik, "Low-light image enhancement using variational optimization-based Retinex model," *IEEE Trans. Consum. Electron.*, vol. 63, no. 2, pp. 178–184, May 2017.
[23] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "MSR-net:Low-light image enhancement using deep convolutional network," 2017, *arXiv:1711.02488*.
[24] Y. Jiang et al., "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.
[25] L.-W. Wang, Z.-S. Liu, W.-C. Siu, and D. P. K. Lun, "Lightening network for low-light image enhancement," *IEEE Trans. Image Process.*, vol. 29, pp. 7984–7996, 2020.
[26] E. H. Land and J. J. McCann, "Lightness and retinex theory," *J. Opt. Soc. America*, vol. 61, no. 1, pp. 1–11, 1971.
[27] R. Dale-Jones and T. Tjahjadi, "A study and modification of the local histogram equalization algorithm," *Pattern Recognit.*, vol. 26, no. 9, pp. 1373–1381, 1993.
[28] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.
[29] H. Li et al., "SRDiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, Mar. 2022.
[30] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.
[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.
[32] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2022, *arXiv:2010.02502*.
[33] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Proc. NeurIPS*, 2021, pp. 1–14. [Online]. Available: https://openreview.net/forum?id=qw8AKxfYbI
[34] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1664–1673.
[35] M. Kim, H. Lim, S. Yu, and J. Paik, "Pan-sharpening of multispectral remote sensing imagery using deep back-projection network," in *Proc. IEEE Int. Conf. Consum. Electron. Asia (ICCE-Asia)*, 2020, pp. 1–2.
[36] Z.-S. Liu, L.-W. Wang, C.-T. Li, and W.-C. Siu, "Hierarchical back projection network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 2041–2050.
[37] Z.-S. Liu, L.-W. Wang, C.-T. Li, W.-C. Siu, and Y.-L. Chan, "Image super-resolution via attention based back projection networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, 2019, pp. 3517–3525.
[38] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–14.
[39] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[40] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image 450 recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 3–7.
[41] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2023, *arXiv:1606.08415*.
[42] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, "Dying ReLU and initialization: Theory and numerical examples," *Commun. Comput. Phys.*, vol. 28, no. 5, pp. 1671–1706, 2020.
[43] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[44] X. Chen et al., "Symbolic discovery of optimization algorithms," 2023, *arXiv:2302.06675*
[45] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.

Cheuk-Yiu Chan (Student Member, IEEE) received the B.Eng. (First-Class Hons.) degree in electronic and information engineering from The Hong Kong Polytechnic University in 2021, where he is currently pursuing the M.Phil. degree in electrical and electronic engineering. Concurrently, he is a Research Assistant with the School of Computing and Information Sciences, Saint Francis University, Hong Kong. His research interests include computer vision, deep learning, and image/video enhancement.

Wan-Chi Siu (Life Fellow, IEEE) received the M.Phil. degree from The Chinese University of Hong Kong in 1977, and the Ph.D. degree from Imperial College London in 1984. He is currently an Emeritus Professor (formerly the Chair Professor, the HoD(EIE) and the Dean of Engineering Faculty) with The Hong Kong Polytechnic University and the Research Professor with St. Francis University, Hong Kong. He was a Vice President, the Chair of Conference Board and core member of Board of Governors of the IEEE SP Society from 2012 to 2014, and the President of APSIPA from 2017 to 2018. He is an Outstanding Scholar with many awards, including the Distinguished Presenter Award, the Best Teacher Award, the Best Faculty Researcher Award (twice) and the IEEE Third Millennium Medal in 2000. He was an APSIPA Distinguished Lecturer from 2021 to 2022, and an Advisor & Distinguished Scientist of the European research project SmartEN (offered by European Commissions). He has been a Keynote Speaker and an Invited Speaker of many conferences, published over 500 research papers (200 appeared in international journals such as IEEE TRANSACTIONS ON IMAGE PROCESSING) in DSP, transforms, fast algorithms, machine learning, deep learning, super-resolution imaging, 2D/3D video coding, object recognition and tracking, and organized IEEE society-sponsored flagship conferences as a TPC Chair of ISCAS1997 and the General Chair of ICASSP2003 and ICIP2010. He was an Independent Non-Executive Director from 2000 to 2015 of a publicly-listed video surveillance company and chaired the First Engineering/IT Panel of the RAE, Hong Kong, from 1992 to 1993. He has been a Guest Editor/Subject Editor/Associate Editor for IEEE Transactions on CAS, IP & CSVT, and Electronics Letters. Recently, he has been a member of the IEEE Educational Activities Board, the IEEE Fourier Award for Signal Processing Committee from 2017 to 2020, the Hong Kong RGC Engineering-JRS Panel from 2020 to 2026, Hong Kong ASTRI Tech Review Panel from 2006 to 2024, and some other IEEE technical committees.

Yuk-Hee Chan (Member, IEEE) received the B.Sc. (Hons.) degree in electronics from the Chinese University of Hong Kong in 1987, and the Ph.D. degree in signal processing from The Hong Kong Polytechnic University in 1992. From 1987 to 1989, he worked as a R&D Engineer with Elec & Eltek Group, Hong Kong. He joined this University in 1992, He is currently an Associate Professor with the Department of Electrical and Electronic Engineering. He has published over 165 research papers in various international journals and conferences. His research interests include image processing and deep learning. He was the Chair of IEEE Hong Kong Section in 2015. He is the Treasurer of Asia-Pacific Signal and Information Processing Association Headquarters.

H. Anthony Chan (Life Fellow, IEEE) received the B.Sc. degree from the University of Hong Kong, the M.Phil. degree from the Chinese University of Hong Kong, and the Ph.D. degree in physics from the University of Maryland. He is currently the Dean of Yam Pak Charitable Foundation, School of Computing and Information, Saint Francis University. He conducted industry research at the former AT&T Bell Labs where he had served as the lead AT&T delegate at 3GPP network standards. He was a Professor with University of Cape Town and then joined Huawei Technologies, USA, to conduct standards and research in 5G Wireless and IETF standards. He has authored/co-authored 30 USA and international patents, over 260 journal/conference papers, a book and five book chapters; edited/authored/contributed to four network standards documents at IEEE and IETF. He has presented over 20 keynotes/invited talks and 40 conference tutorials. He had been a Distinguished Speaker of IEEE ComSoc, IEEE CMPT Society, and IEEE Reliability Society.