DiffLight: Integrating Content and Detail for Low-light Image Enhancement

Yixu Feng[1] Shuo Hou[1] Haotian Lin[1] Yu Zhu[1] Peng Wu[1] Wei Dong[2]
Jinqiu Sun[1] Qingsen Yan[1†] Yanning Zhang[1]
[1]Northwestern Polytechnical University
[2]Xi'an University of Architecture and Technology

Abstract

The Low Light Image Enhancement (LLIE) task has been a hotspot in low-level computer vision research. The camera sensor can only capture a small amount of ambient light signal in low-light condition, resulting in significant noise black pseudo artifacts in images, which not only degrade visual quality but also affect the performance of downstream visual tasks. However, current methods often produce overly smoothed and distorted results, or introduce strong noise artifacts. Moreover, for recent UHD high-definition low-light images, due to GPU memory limitations, LLIE must be conducted in patches, leading to block artifacts. Faced with these challenges, we propose a dual-branch pipeline called DiffLight. Specifically, it consists of the Denoising Enhancement (DE) branch and the Detail Preservation (DP) branch. The DE-branch adopts a combination of DiffIR and LEDNet to reduce noise and enhance brightness, while the DP-branch utilizes a novel Light Full-Former (LFF) method, which comprises 20 Full-Attention (LFA) modules to preserve full-scale image details. To tackle block artifacts, we further introduce Progressive Patch Fusion (PPF) for patch fusion. Experimental results demonstrate that our approach is high-ranked in the CVPR2024 NTIRE Low Light Enhancement challenge and produced state-of-the (SOTA) results on other datasets.

1. Introduction

In low-light conditions, sensors typically capture only a tiny amount of light signal, leading to a significant amount of noise in the resulting images, consequently impairing the performance of downstream visual tasks such as image classification [25] and automatic driving [22]. The Low-light Image Enhancement (LLIE) task is to brighten images, enhance contrast, mitigate image degradation, and alleviate severe image noise caused by low-light conditions, aiming to restore the original image information and improve perceptual quality [50]. In addition, high-resolution imaging has experienced significant advancements due to the emergence of sophisticated imaging sensors and displays in recent years. The Ultra High Definition (UHD) LLIE task becoming a research focus, gradually.

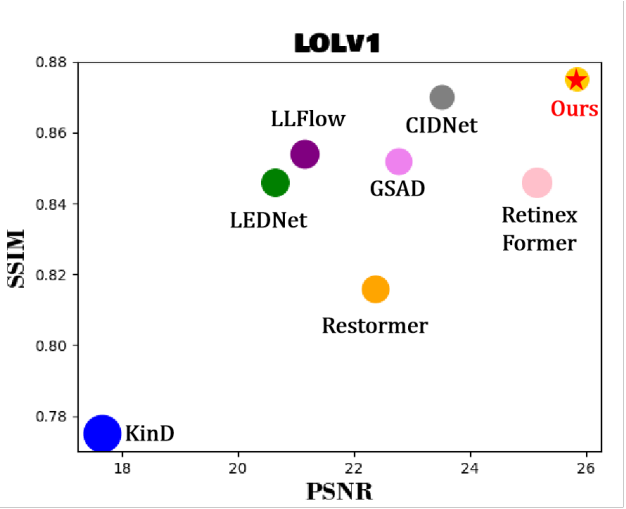

Figure 1. Comparison with recent SOTA methods in LOLv1 [38] dataset. We choose PSNR/SSIM↑ and LPIPS↓ for the measure metrics. The circle size represent LPIPS of each method. The result shows DiffLight have the best performance between SOTAs.

Traditional LLIE methods, like histogram equalization [1, 3, 4] and gamma correction [14, 30, 37], while capable of enhancing image brightness and restoring some details, often lead to significant color artifacts and are unable to recover a pure black area filled with noise or underexposure.

Currently, with the strong generalization capability of deep learning, a plethora of solutions based on deep convolutional neural networks (CNNs) have been proposed to address some low-light degradation issues. Typically, these

†Qingsen Yan (qingsenyan@nwpu.edu.cn) is the corresponding author. This work is supported by NSFC of China 62301432, 62306240, Natural Science Basic Research Program of Shaanxi No. 2023-JC-QN-0685, QCYRCXM-2023-057, the Fundamental Research Funds for the Central Universities No. D5000220444.

methods directly learn the mapping function between low-light degraded images and high-quality normal illumination images. These approaches can be roughly categorized into two types: end-to-end networks [38, 43, 48] and image generative methods based on the Diffusion Model [16, 44].

The latest end-to-end single stage networks are mostly based on the multi-head attention mechanism of the Vision Transformer model [8] to capture pixel-level long-range dependencies and non-local self-similarity in images [7, 43]. These methods have indeed addressed the issue of low-light artifacts, but they still fall short in restoring authentic details in extremely low-light and highly noisy regions.

To address the end-to-end single stage problems, the diffusion model comes and has consistently demonstrated outstanding performance in image denoising, repainting, and other fields [23]. Recent efforts have adapted the Diffusion model to the domain of LLIE [13, 16, 40, 51], successfully addressing the restoration of extremely low-light regions. However, it leads to two new issues: (a) Images restored by the large models (*e.g*. diffusion model) often exhibit excessive smoothing and loss of details. (b) The computational cost of the diffusion model is typically high, especially for UHD images, potentially causing GPU out-of-memory errors. These challenges bring us to develop a method that aims to produce low noise without sacrificing details.

At present, high-resolution images are mostly divided into small patches, which are input into the network separately, and the brightened images are finally output and stitched together [21, 34, 39]. This process leads to the generation of block artifacts within the image, significantly impacting the visual quality of the image.

To sum up, we propose a novel pipeline, named DiffLight, which consists of two branches. The Denoising Enhancement (DE) branch employs the diffusion method from DiffIR [41] to remove noise from low-light images while mapping them from low-light to normal-light, thereby preliminarily improving image brightness. Subsequently, the images are input into LEDNet [52] for refinement, enhancing image details and correcting any existing color deviations. Another branch, the Detail Preservation (DP) branch, utilizes our designed Light Full-Former (LFF) network. This network comprises 20 Light Full-Attention (LFA) modules embedded in a UNet structure. We input a low-light image into both branches and generate two brightened images from each branch, which are then weighted fused to produce the final output. Finally, for high-resolution images, we employ the Progressive Patch Fusion (PPF) method in testing, applying progressive weight handling at the edges to effectively address block artifacts and significantly enhance visual quality.

The proposed DiffLight method achived **rank #4** in the CVPR2024 NTIRE Workshop Low Light Enhancement Challenge [42]. Experimental results conducted in various challenging scenarios demonstrate that DiffLight approach has achieved state-of-the-art (SOTA) results in both quantitative and visual evaluations for the LLIE task. The main contributions of this paper are threefold:

- We propose a dual-branch pipeline called DiffLight. One branch employs the diffusion model, focusing on image noise removal, while the other branch utilizes an end-to-end UNet Transformer for low-light detail restoration.
- We propose a new transformer method Light Full-Former, which use the novel Light Full-Attention (LFA) module.
- We proposed the Progressive Patch Fusion (PPF) method that addresses block artifacts and yields favorable visual perceptual results.

2. Relative Work

2.1. Low-light Image Enhancement

Traditional Methods. Early work primarily relied on various heuristic algorithms to improve image quality. For instance, histogram equalization [15], which effectively redistributes the brightness of an image to enhance its global contrast. Methods based on Retinex theory [11, 29] enhance low-light images by decomposing the image into reflectance and illumination components. LIME [11] estimates the illumination intensity of each pixel and refines the initial illumination map through structural priors to improve image quality. Although traditional methods typically do not require training data, they still have limitations in detail preservation and noise control.

Deep Learning Methods. Recently, deep learning-based methods have garnered attention for their accuracy, robustness, and speed, achieving state-of-the-art performance in a wide range of image enhancement tasks. Restormer [46] model achieves high-resolution image restoration through an effective Transformer architecture, while Zhou *et al.* [52] focuses on joint low-light enhancement and deblurring, introducing a large-scale dataset LOL-Blur and demonstrating effectiveness on both synthetic and real-world datasets. Additionally, methods based on Retinex theory and deep learning [2, 24, 38, 49] enhance images by optimizing the reflectance and illumination components through the network. In recent times, multi-stage networks [32, 45] have effectively addressed the issues of single-stage models by processing and refining images in stages, achieving impressive results.

2.2. Diffusion Model

With the advancement of Denoising Diffusion Probabilistic Models (DDPMs) [12], diffusion-based generative models have achieved remarkable results in the domain of image generation. Wang *et al.* [23] have summarized recent diffusion-based image restoration techniques, which are primarily categorized into two types: **(1)** Training dif-

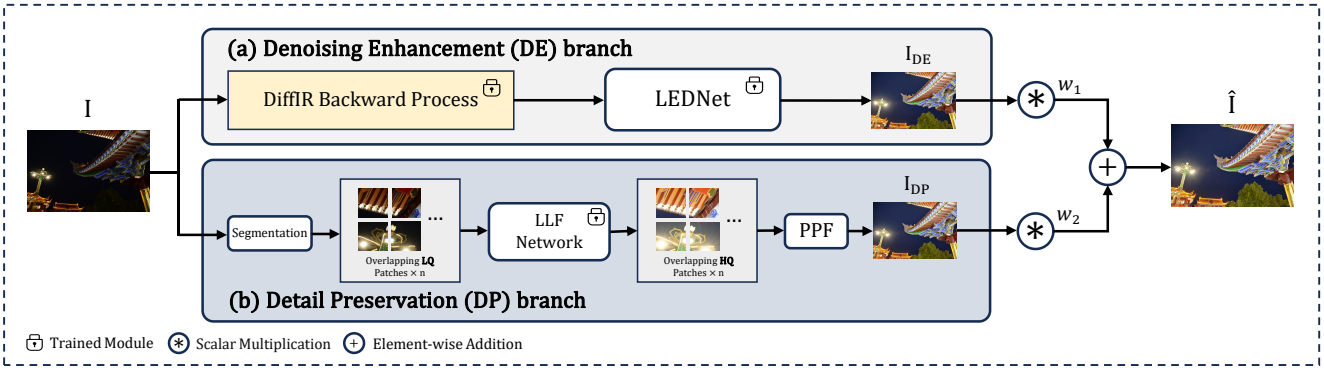

Figure 2. DiffLight pipeline. DiffLight is composed of two branches: (a) Denoising Enhancement (DE) branch and (b) Detail Preservation (DP) branch. For **higher PSNR** (only for NTIRE Competition), we opted for **non-**patch inference in the DE-branch.

fusion models from scratch through supervised learning to adapt to various image restoration tasks [16, 18, 23, 28, 40, 51]. DiffPIR [53] innovatively integrates traditional plug-and-play image restoration methods with diffusion model, achieved the state-of-the-art reconstruction fidelity and perceptual quality, while maintaining a low count of neural function evaluations. **(2)** Zero-shot [5, 6, 27], where pre-trained generative models are considered repositories of structures and textures constructed from extensive real datasets, thus leveraging pre-trained diffusion models to acquire structural and textural priors for image restoration.

3. Method

For the Challenge UHD dataset, we combine diffusion model and transformer-based model that could yield better results, which is introduced in Sec. 3.1. As shown in Fig. 2, DiffLight is separated by Denoising Enhancement (DE) branch and Detail Preservation (DP) branch. Additionally, to alleviate the block artifacts for UHD image cut-patch problem, we furhter proposed the progressive patch fusion (PPF) method in Sec. 3.4 .

3.1. Overview Pipeline

Given a low-light image $\mathbf{I}$ as input, it will be fed into the two branches separately. In the DE branch (Fig. 2 (a)), the low-quality image is processed sequentially by DiffIR [41] and LEDNet [52] to get the enhanced image $\mathbf{I}_{DE}$ as

$$\mathbf{I}_{DE} = \text{LEDNet}(\text{Diff}_B(\mathbf{I})), \tag{1}$$

where $\text{Diff}_B(\cdot)$ represents the the Backward Process of DiffIR in the inference stage (see Sec. 3.2). The enhanced image, $\mathbf{I}_{DE}$, has fairly low noise, minimal color deviation, as well as highly-increased brightness, but there is an excessive loss of details.

In the DP-branch (Fig. 2 (b)), the Light Full-Former (LFF) network and the method for high-resolution image restoration, PPF, are proposed. To better recover the details from high-resolution images, the image is divided into n small overlapping patches in the segmentation process, the patch size is adapted to the input size for training. Each patch is individually processed through LLF. The image enhanced by DP-branch $\mathbf{I}_{DP}$ as

$$\mathbf{I}_{DP} = \text{PPF}(\text{LLF}(\text{Seg}(\mathbf{I}))), \tag{2}$$

where Seg represent the segmentation process. The block artifacts produced by the traditional fusion method is removed by PPF, providing good visual quality and rich details. Finally, $\mathbf{I}_{DE}$ and $\mathbf{I}_{DP}$ are multiplied by customized weights w_1 and w_2 respectively as

$$\hat{\mathbf{I}} = w_1\mathbf{I}_{DE} + w_2\mathbf{I}_{DP}, \tag{3}$$

where $\hat{\mathbf{I}}$ denotes the final enhanced image, w_2 is typically set to $1 - w_1$.

Modules in two branches are trained during their respective training stages. Details of branches are described in the followed two sections.

3.2. Denoising Enhancement Branch

Denoising Stage. The primary task is to train DiffIR[1] and input the training set into the trained DiffIR for inference, resulting in an enhanced training set. The training for DiffIR consits of two phases. As show in Fig. 3 1) a), The purpose of the first phase (i.e., Training DiffIR_{S1}) is to obtain the pre-trained IPR Generator, which guides the training of the second phase (i.e., Training DiffIR_{S2}). IPR Generator is composed of CPEN_{S1} module, which takes Ground Truth (GT) $\mathbf{I}_{GT}$ and Low Quality (LQ) image $\mathbf{I}_{LQ}$ as input to generate IPR $\mathbf{Z}$:

$$\mathbf{Z} = \text{CPEN}_{S1}(\text{Unshuffle}(\text{Concat}(\mathbf{I}_{GT}, \mathbf{I}_{LQ}))), \tag{4}$$

[1]For more details about the modules and terms in DiffIR such as IPR and CPEN, please refer to [41].

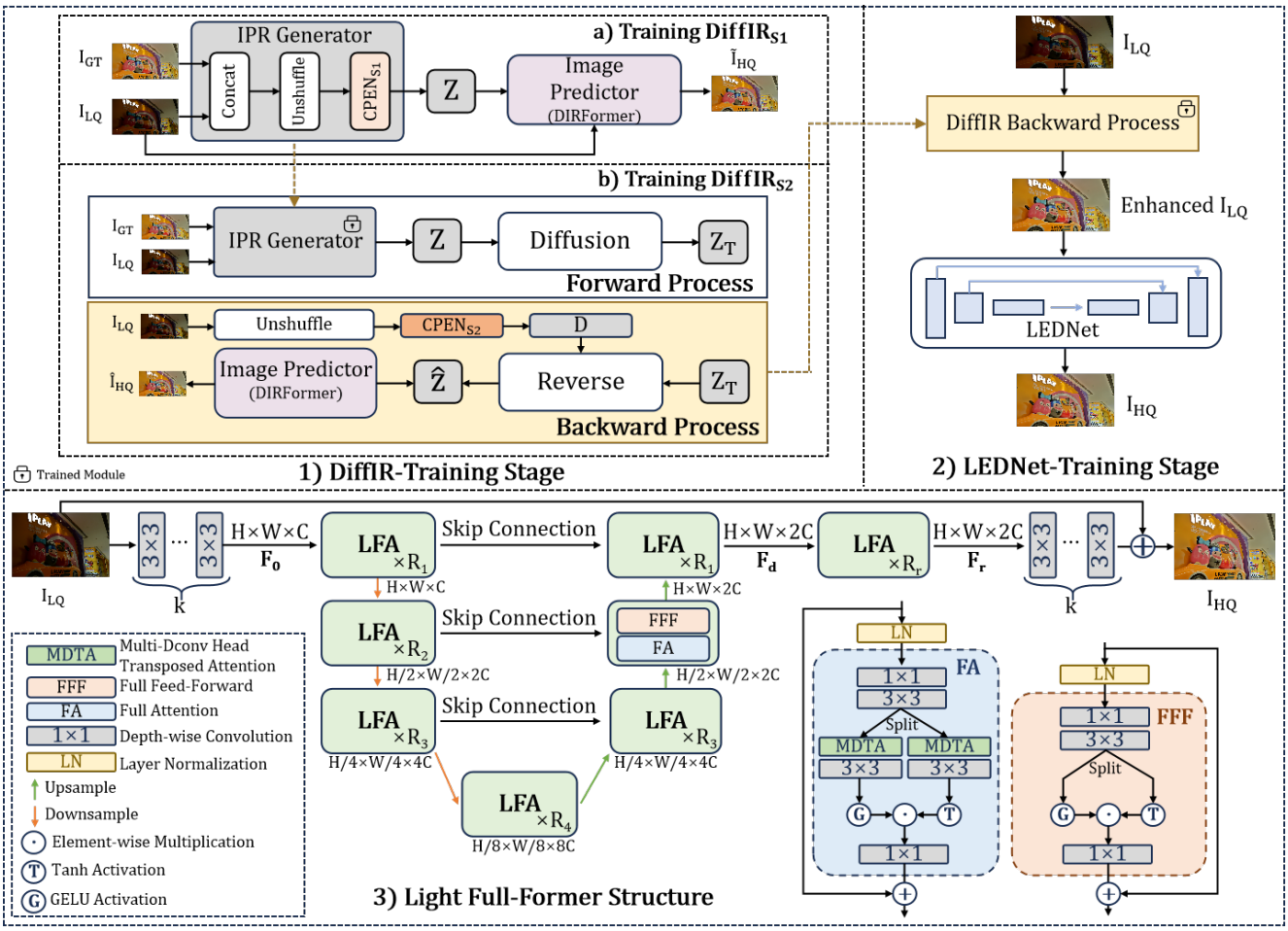

Figure 3. **1)** DiffIR-Training Stage [41]. **2)** LEDNet-Training Stage [52]. **3)** Our proposed Light Full-Former (LFF), which contains 20 Light Full-Attention (LFA) blocks, where $R_{1,r} = 2$, $R_{2,3} = 4$, and $R_4 = 8$. **(Zoom in for the best view.)**

where Unshuffle$(\cdot)$ denotes the pixel unshuffle operation, and Concat$(\cdot)$ denotes the concatenation operation. This generated IPR is then fed into Image Predictor (i.e., DIRFormer), to produce High Quality (HQ) image $\mathbf{I}_{HQ}$:

$$\tilde{\mathbf{I}}_{HQ} = \text{DIRFormer}(\mathbf{Z}, \mathbf{I}_{LQ}). \tag{5}$$

These two modules are jointly trained to fully leverage CEPN$_{S1}$ in generating high-quality IPR $\mathbf{Z}$.

As depicted in Fig. 3 1) b), the second phase consists of the Forward and Backward Process. The Forward Process guides Backward Process, which is trained for final inference. The Forward process employs the IPR Generator trained in the first phase, and the generated IPR $\mathbf{Z}$ in Eq.4 is fed into the diffusion module to obtain the noised vector $\mathbf{Z_T}$:

$$\mathbf{Z}_T = \text{Diffusion}(\mathbf{Z}). \tag{6}$$

In the Backward Process, $\mathbf{Z_T}$ is fed into Reverse module along with the conditional vector $\mathbf{D}$ produced by the CEPN$_{S2}$ module (structurally identical to CEPN$_{S1}$), which takes $\mathbf{I}_{LQ}$ as input, to produce the estimated IPR $\hat{\mathbf{Z}}$. Subsequently, the Image Predictor takes $\hat{\mathbf{Z}}$ and $\mathbf{I}_{LQ}$ as input, yielding the estimated high-quality image $\hat{\mathbf{I}}_{HQ}$. This phase can be described as follows:

$$\begin{aligned} \mathbf{D} &= \text{CPEN}_{S2}(\text{Unshuffle}(\mathbf{I}_{LQ})), \\ \hat{\mathbf{Z}} &= \text{Reverse}(\mathbf{Z}_T, \mathbf{D}), \\ \hat{\mathbf{I}}_{HQ} &= \text{DIRFormer}(\hat{\mathbf{Z}}, \mathbf{I}_{LQ}). \end{aligned} \tag{7}$$

During this training phase, CEPN$_{S2}$, the Reverse module, and DIRFormer are jointly optimized.

Enhancement Stage. In order to implement LEDNet's enhancement of DIffIR output images, the enhanced training set must first be obtained through DIffIR and used as input to LEDNet. During the inference stage in DiffIR, only the Backward Process is used (the bottom part of Fig. 3 1) b)). As shown in Fig. 3 2), given a low-quality image $\mathbf{I}_{LQ}$ and a randomly sampled Gaussian noise $\mathbf{Z}_T$, the En-

hanced $\mathbf{I}_{LQ}$ can be derived using Eq. 7. Subsequently,The Enhanced $\mathbf{I}_{LQ}$ is fed into LEDNet to obtain $\mathbf{I}_{HQ}$, which can be succinctly formulated as

$$\mathbf{I}_{HQ} = \text{LEDNet}(\text{Diff}_B(\mathbf{I}_{LQ})), \tag{8}$$

where $\text{Diff}_B(\cdot)$ corresponds to its representation in Eq. 1, and it can be regarded as the formulation of Eq. 7 with the input $\mathbf{Z}_T$ omitted, which is implicitly implied in the inference stage.

3.3. Detail Preservation branch

Inspired by Restormer [46], we have designed a Light Full-Former (LLF) Transformer architecture UNet on the DP-branch. As illustrated in Fig. 3 3), this architecture comprises 20 Light Full-Attention (LFA) modules.

Starting with a low-light image $\mathbf{I}_{LQ} \in \mathbb{R}^{H\times W\times 3}$, the image undergoes feature extraction using k times 3×3 depth-wise convolution layers to generate the $\mathbf{F}_0 \in \mathbb{R}^{H\times W\times C}$ feature map. The rationale for reusing multiple layers of small-kernel depth convolutional layers is to reduce parameters and computations while enhancing receptive fields and expanding non-linear fitting capabilities. The $\mathbf{F}_0$ is then fed into a UNet network with full LFA, resulting in a new brightened feature $\mathbf{F}_d \in \mathbb{R}^{H\times W\times 2C}$. Subsequently, $\mathbf{F}_d$ is passed through R_r times LFAs for refinement, producing a new feature $\mathbf{F}_r \in \mathbb{R}^{H\times W\times 2C}$. Finally, after compressing the $\mathbf{F}_r$ feature map with k times 3×3 depth-wise convolution layers, a residual image $\Delta\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ of low-light and highlights is generated, which is added back to the original low-light image to obtain the ultimate enhanced image as

$$\mathbf{I}_{HQ} = \mathbf{I}_{LQ} + \Delta\mathbf{I}, \tag{9}$$

where $\mathbf{I}_{HQ}$ is the well-enhanced image in DP-branch.

Full Attention. One LFA contains a Full Attention (FA) and a Full Feed-Forward (FFF). Given a embedded feature tensor $\mathbf{F}_k \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$ for input, FA first generated the depth feature $\mathbf{F'}_k \in \mathbb{R}^{\hat{H}\times\hat{W}\times 3\hat{C}}$ by the 1×1 and 3×3 depth-wise convolution layers. The $\mathbf{F'}_k$ feature is divided into two equal parts based on the number of channels and then input into two separate MDTA modules that are derived from the modules in Restormer [46]. These two sub-branch output features further embeds by a 3×3 depth-wise convolution layer. Sub-branches pass through GELU activation and Tanh activation as outputting feature $\mathbf{F}_G$ and $\mathbf{F}_T$, respectively. The final output of FA can be measured as

$$\hat{\mathbf{F}} = W_d(\mathbf{F}_G \odot \mathbf{F}_T), \tag{10}$$

where $\odot$ denotes the element-wise multiplication, W_d represents the 1×1 depth-wise convolution layer, and $\hat{\mathbf{F}}$ represent the final output feature of FA module.

Full Feed-Forward. Similar to the FA module, the input features are compressed by two deep-wise convolution layers and split into two groups of features. However, FFF module removed MDTA and a 3×3 convolution layer, which directly embeds the FFF output by Eq. 10.

Activations. In the LFA module, we apply the features through both GELU and Tanh activation functions simultaneously, followed by element-wise multiplication. This approach is chosen to maintain the intrinsic integrity of features (**fully** enhancing the original details of image features), as well as to benefit from the gating property provided by the GELU function (constraining noise in the original image). Additionally, the Tanh activation function helps restrict feature values to the range of -1 to 1, leading to faster convergence speed of the model.

Algorithm 1 Progressive Patch Fusion

Input: Image $\mathbf{I}$, Patch size p, Stride s, Model $model$
Output: restored image $\mathbf{I}'$
$\mathbf{I}' \leftarrow$ empty tensor
$patch_{row} \leftarrow$ empty tensor
for $i = 0$ to $overlap - 1$ **do**
  $w_{factor} \leftarrow \frac{i}{overlap}$
  for $j = 0$ to $p - 1$ **do**
    $weight_1[i, j] \leftarrow 1 - w_{factor}$
    $weight_2[i, j] \leftarrow w_{factor}$
    $weight_3[j, i] \leftarrow 1 - w_{factor}$
    $weight_4[j, i] \leftarrow w_{factor}$
  end for
end for
for each position (h_i, w_i) in $\mathbf{I}$ with step s **do**
  $patch \leftarrow \mathbf{I}[h_i : h_i + p, w_i : w_i + p]$
  $patch' \leftarrow model(patch)$
  if $w_i = 0$ **then**
    $patch_{row} \leftarrow patch'$
  else
    $patch_{row} \leftarrow (patch_{row} \cdot weight_1 + patch' \cdot weight_2)$
  end if
  if $h_i \neq 0$ **then**
    $\mathbf{I}'[h_i : h_i + p, :] \leftarrow (\mathbf{I}'[h_i : h_i + p, :] \cdot weight_3 + patch_{row} \cdot weight_4)$
  else
    $\mathbf{I}'[h_i : h_i + p, :] \leftarrow patch_{row}$
  end if
end for
return $\mathbf{I}'$

3.4. Progressive Patch Fusion

For high-resolution images, we utilize the Progressive Patch Fusion (PPF) method during testing. This approach incorporates progressive weight management at the Over-

lapping areas to effectively mitigate edge and substantially improve visual fidelity. As represented in Alg. 1, PPF is performed by these steps as follows:

1. The input image $\mathbf{I}$ is segmented into several patches with patch size p and stride s.
2. Each patch, after model inference, is added to the restored image $\mathbf{I}'$. The overlapping parts between patches are processed with four weight matrices $weight_{1,2,3,4}$ linearly varying from 0 to 1.
3. Specifically, when patches overlap horizontally, $weight_{1,2}$ are used to blend the overlapping edges, ensuring a smooth and seamless fusion. Similarly, when patches overlap vertically, $weight_{3,4}$ are utilized for blending.
4. By applying these weights row by row and column by column, a seamless large image $\mathbf{I}'$ is restored.

The PPF method allows for a smooth transition in the overlapping areas of the image patches, avoiding hard edges and resulting in a more natural-looking image.

3.5. Training Loss Functions

DE-branch. We follow the loss functions in the original paper of DiffIR [41] and LEDNet [52].

DP-branch. We trained LFF with the combination losses $\mathcal{L}(\hat{x}, x)$. Given a Ground Truth x and a restored image $\hat{x}$, we employ L1 loss $\mathcal{L}_1$, edge loss $\mathcal{L}_e$ [31] and perceptual loss $\mathcal{L}_p$ [17] at sRGB space in DP-branch as

$$\mathcal{L}(\hat{x}, x) = \mathcal{L}_1(\hat{x}, x) + \lambda_e \cdot \mathcal{L}_e + \lambda_p \cdot \mathcal{L}_p(\hat{x}, x), \quad (11)$$

where λ_e, λ_p are 50 and 0.01, respectively.

4. Experiment

4.1. Experimental Settings

Dataset. We use the NTIRE2024 Low Light Enhancement Challenge dastest [42] to train and test the proposed method. It is a UHD dataset containing images with resolutions up to 4K and beyond, which comprises 230 training scenes, along with 35 validation and 35 testing ones. Moreover, in order to verify the effectiveness of our method on well-known publicly available datasets, we also conducted experiments on the LOLv1 dataset [38]. Besides, to illustrate the strength of PPM, we randomly sample 50 paired pictures from UHD-LOL4K dataset [33].

Evaluation Metrics. For all datasets, we adopt the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [36] as the distortion metrics. To evaluate the perceptual quality of restored images, we report Learned Perceptual Image Patch Similarity (LPIPS) [47] by using AlexNet [20] for references as a perceptual metric.

4.2. Implementation details

Denoising Enhancement. For the DE-branch, we train DiffIR [41] and LEDNet [52] separately, both are trained on two V100 GPUs.

DiffIR [41] adopts a four-level encoder-decoder structure. From level-1 to level-4, the attention heads in DMTA are 1, 2, 4, 8, the number of channels is setting to $48, 96, 192, 384$, and the number of dynamic transformer blocks to 3, 5, 6, 6.

In training DiffIR [41], total timesteps T are set to 4, and β_t linearly increase from $\beta_1 = 0.9$ to $\beta_T = 0.99$. DiffIR$_{S1}$ are trained for $300K$ iterations with the initial learning rate 2×10^{-4} gradually reduced with the cosine annealing. And For DiffIR$_{S2}$, we train $300K$ iterations with initial learning rate 2×10^{-4} and gamma 0.5 with the MultiStepLR scheduler. For both training stage, we progressively increase patch size and decrease batch size. Specifically, during iterative training, the patch size and batch size pair are set to respectively train for $(92K)$, $(80K)$, $(38K)$, $(90K)$ iterations under the configurations of $(192, 8)$, $(256, 4)$, $(320, 2)$, $(400, 1)$.

LEDNet [52] model is trained on the inference results on Train set produced by DiffIR, and the Ground Truth remains unchanged. We train LEDNet using Adam [19] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$ for a total of 300k iterations. The initial learning rate is set to 1×10^{-4} and updated with cosine annealing strategy [26]. For NTIRE2024 dataset, we still adopt a progressive training approach, the patch size and batch size pair are set to train for $(90K)$, $(70K)$, $(70K)$, $(70K)$ iterations respectively under the configurations of $(256, 8)$, $(512, 4)$, $(1024, 1)$, $(1320, 1)$. As for LOLv1 dataset, LEDNet are trained with 400 patch size and 10 batch size for $100K$ iterations.

Detail Preservation. We implement our the DP-branch by PyTorch. The model is trained with the Adam [19] optimizer (β_1 = 0.9 and β_2 = 0.999) for at least 300 epochs by using a single NIVIDA 3090 GPU. The learning rate is initially set to 1×10^{-4} and then steadily decreased to 1×10^{-7} by the cosine annealing scheme [26] during the training process. We randomly crop the image to 256 $\times$ 256 on NTIRE2024 dataset and 80 $\times$ 80 on LOLv1 dataset [38] for patch size and set batch size to 8. When testing the high-resolution images, we use our proposed PPF that set p as 256 and s as 128.

4.3. Qualitative Evaluations

As illustrated in Fig. 4, we compared our method with five other state-of-the-art (SOTA) methods on the LOLv1 dataset. Visually, our method exhibits smaller color deviations, significantly reduces noise while preserving detail accuracy, maintains normal contours, and greatly reduces artifacts, approaching closer to the Ground Truth. Moreover, we conducted PSNR and SSIM comparisons between

(a) Input PSNR↑/SSIM↑ | (b) RetinexNet [38] 15.59/0.801 | (c) RUAS [24] 11.75/0.733 | (d) LLFlow [35] 11.86/0.779 | (e) KinD [48] 16.96/0.832

(f) SNR-Aware [43] 15.89/0.865 | (g) CIDNet [9] 20.10/0.930 | (h) GSAD [13] 22.33/0.938 | (i) Ours 27.21/0.947 | (j) Ground Truth ∞/1.000

Figure 4. The visual quality comparison results on LOLv1 dataset with various SOTA methods. The quantitative comparisons PSNR/SSIM↑ is also followed.

Method	PSNR↑	SSIM↑	LPIPS↓	FLOPs (G)	Type
RetinexNet [38]	16.77	0.419	0.474	584.5	CNN
KinD [48]	17.65	0.775	0.207	35.0	CNN
ZeroDCE [10]	14.86	0.559	0.335	4.8	Zero-shot
RUAS [24]	16.41	0.500	-	0.8	Unsupervised
LLFlow [35]	21.15	0.854	0.119	358.4	Flow
Restormer [46]	22.37	0.816	0.108	144.3	Transformer
LEDNet [52]	20.63	0.823	0.118	35.9	CNN
Retinexformer [2]	25.15	0.846	0.131	15.85	Transformer
GSAD [13]	22.77	0.852	0.102	-	Diffusion
Diff-LLE [44]	22.24	0.792	-	56.86	Diffusion
HVI-CIDNet [9]	23.50	0.870	0.086	7.57	Transformer
Ours	25.85	0.876	0.082	168.3	Mixed

Table 1. Quantitative comparisons PSNR/SSIM↑ and LPIPS↓ on LOLv1 dataset. The highest result is in red color while the second highest result is in cyan color.

the selected image and its Ground Truth, showing that our DiffLight method outperforms in both metrics.

4.4. Quantitative Evaluations

To further validate the effectiveness of our method, we conducted additional performance testing on the standard-resolution dataset LOLv1 outside the NTIRE competition. We selected 11 state-of-the-art (SOTA) models, with the "type" column indicating the network architecture type of each model. We assessed the performance using three metrics: PSNR, SSIM, and LPIPS.

As shown in Tab. 1, our method excels among recent state-of-the-art (SOTA) methods, achieving the best values in all three metrics. Specifically, our method outperforms Retinexformer by **0.70 dB** in PSNR, surpasses the latest CIDNet by **0.06** in SSIM, and surpasses CIDNet by **0.04** in LPIPS. These results further demonstrate that our DiffLight method not only accurately restores the brightness information of images in low-light conditions but also significantly enhances the subjective visual perception of the images. Despite our relatively high FLOPs, the other three metrics are all optimal, and achieve the best visual results.

4.5. Ablation Study

PPF Method. To better demonstrate the generalization ability of the PPF, we employ our PPF with LLFormer on UHD-LOL4K dataset [33] using pre-trained weights[2]. We randomly sample 50 paired pictures with random-cropped size [1500, 2000] to evaluate the performance. The visual result (Fig. 5) and quantitative evaluations (Tab. 2) under the following three conditions are presented. It is noted that different patch size and stride for overlapping patches are evaluated for (b) and (c).

(a) Without patch. It denotes inference on the entire image. Due to GPU memory size limitations, we typically inference full UHD image on only CPU, which takes much longer time and limit result compared with GPU inference.

[2]The pre-trained weights (with best PSNR) is obtained from https://github.com/TaoWangzj/LLFormer

(a) w/o Patch (b) Patches w/o PPF (c) Patches w PPF

Figure 5. The visual comparisons on UHD-LOL4K dataset. (a) Full image inferenced directly by CPU. (b) and (c) are both segmented into overlapping patches with patch size 512 and stride 256, but (c) is employed with PPF method. It can be seen that the block artifacts are eliminated after using the PPF method, with better visual effects obtained. **(Zoom in for the best view.)**

Patch size / Stride (px)	PPF	PSNR↑	SSIM↑	CPU/GPU	Time (s)↓
Without Patch	/	36.78	0.9886	CPU	467.38
256 / 128	✗	36.98	0.9903	GPU	23.39
	✓	36.90	**0.9907**	GPU	24.58
384 / 192	✗	37.52	0.9902	GPU	12.46
	✓	37.53	0.9905	GPU	12.48
512 / 256	✗	37.69	0.9901	GPU	**10.89**
	✓	**37.73**	0.9903	GPU	**10.89**
640 / 320	✗	37.60	0.9899	GPU	12.96
	✓	37.67	0.9901	GPU	12.93

Table 2. The quantitative comparisons of PSNR↑ and SSIM↑ on w or w/o PPF and different patch size/stride. Using CPU/GPU and inference time↓ per image for reference. The best result is **bolded**.

(b) Patches without PPF. Overlapping enhanced patches are reconstructed through only averaging pixel values within overlapping regions. It improves performance in terms of PSNR and SSIM on UHD image restoration, but produce obvious block artifacts, thereby impacting the visual perception.

(c) Patches with PPF. Overlapping enhanced patches fused through PPF, which retain higher numerical values while achieving superior visual effects by moving the block artifacts in (b).

DiffLight Pipeline. To validate the effectiveness of the weighted combination of the two branches, we conducted with different values of w_1 and w_2 on LOLv1 dataset. The experimental results are shown in Fig. 6. In the experiments, w_1 is varied from 0 to 1 with a step of 0.1, and the corresponding w_2 is set to $1 - w_1$. The results indicate that the weighted averaging of the two branches improves PSNR and SSIM, and for the LOLv1 dataset, the best results are obtained with $w_1 = 0.4$ and $w_2 = 0.6$.

We further conducted ablation experiments on different modules in DiffLight on LOLv1 dataset. As shown in Tab. 3, the results indicate that: if only the DE-branch is used for output, losing the influence of DiffIR is more significant than losing LEDNet because using only CNN methods for end-to-end training may result in inaccurate noise and color artifacts. If output is solely from the DP-branch, removing LFA module leads a decrease of 3.48 dB in PSNR and 0.058 in SSIM. The performance metrics of separately outputting the two branches are lower than the complete DiffLight, further demonstrating the superiority of our pipeline.

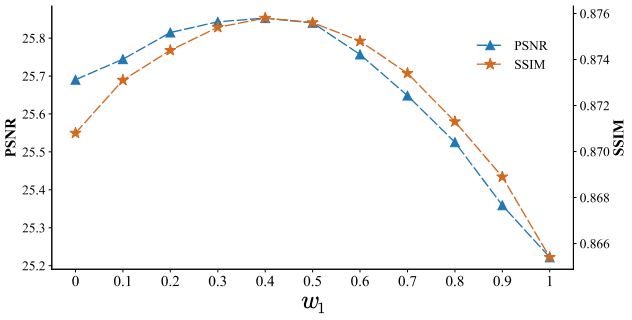

Figure 6. The dotted line illustrates the impact of w_1 vs. metrics on the quantitative results of LOLv1 dataset.

Method	PSNR↑	SSIM↑	Params (M)	FLOPs (G)
Only DiffIR	23.72	0.851	26.0	51.2
Only LEDNet	20.56	0.825	7.1	35.9
Full DE-branch	25.22	0.865	33.1	87.1
w/o LFA	22.21	0.813	3.79	21.6
Full DP-branch	25.69	0.871	18.6	81.2
Full Pipeline	**25.85**	**0.876**	51.7	168.3

Table 3. The quantitative comparative analysis of each branch within DiffLight on LOLv1 dataset. The FLOPs is tested on a 256×256 image. The best result is **bolded**.

5. Conclusion

In this paper, we propose a dual-branch pipeline DiffLight for Low Light Image Enhancement task. The proposed method performs better in NTIRE, LOLv1, and UHD-LOL dataset. Specifically, it consists of the Denoising Enhancement (DE) branch for removing noise and color bias, and the Detail Preservation (DP) branch to full recover the normal-light details. Moreover, we design a Light Full-Former (LFF) that comprises 20 Full-Attention (LFA) modules in DP-branch to preserve full-scale image details. Finally, we introduce the Progressive Patch Fusion (PPF) for better Ultra High Definitio (UHD) image patches fusion. We compared our method to several state-of-the-art (SOTA) approaches obtaining a better performance.

References

[1] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan, and Oksam Chae. A dynamic histogram equalization for image contrast enhancement. *IEEE transactions on consumer electronics*, 53(2):593–600, 2007. 1

[2] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12504–12513, 2023. 2, 7

[3] Turgay Celik and Tardi Tjahjadi. Contextual and variational contrast enhancement. *IEEE Transactions on Image Processing*, 20(12):3431–3441, 2011. 1

[4] Heng-Da Cheng and XJ Shi. A simple and effective histogram equalization approach to image enhancement. *Digital signal processing*, 14(2):158–170, 2004. 1

[5] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14347–14356, 2021. 3

[6] Hyungjin Chung, Byeongsu Sim, and Jong-Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12403–12412, 2021. 3

[7] Jiachen Dang, Zehao Li, Yong Zhong, and Lishun Wang. Wavenet: Wave-aware image enhancement. In *Proc. Pacific Conf. Comput. Graph. Appl*, pages 21–29, 2023. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Yixu Feng, Cheng Zhang, Pei Wang, Peng Wu, Qingsen Yan, and Yanning Zhang. You only need one color space: An efficient network for low-light image enhancement. *arXiv preprint arXiv:2402.05809*, 2024. 7

[10] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020. 7

[11] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017. 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[13] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 7

[14] Shih-Chia Huang, Fan-Chieh Cheng, and Yi-Sheng Chiu. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE transactions on image processing*, 22(3):1032–1041, 2012. 1

[15] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007. 2

[16] Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023. 2, 3

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 6

[18] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022. 3

[19] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[20] Alex Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25(2), 2012. 6

[21] Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. *arXiv preprint arXiv:2302.11831*, 2023. 2

[22] Guofa Li, Yifan Yang, Xingda Qu, Dongpu Cao, and Keqiang Li. A deep learning based image enhancement approach for autonomous driving at night. *Knowledge-Based Systems*, 213:106617, 2021. 1

[23] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023. 2, 3

[24] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan 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*, pages 10561–10570, 2021. 2, 7

[25] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. 1

[26] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6

[27] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11461, 2022. 3

[28] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. pages 23045–23066, 2023. 3

[29] Michael K. Ng and Wei Wang. A total variation model for retinex. *SIAM Journal on Imaging Sciences*, 4(1):345–365, 2011. 2

[30] Shanto Rahman, Md Mostafijur Rahman, Mohammad Abdullah-Al-Wadud, Golam Dastegir Al-Quaderi, and Mohammad Shoyaib. An adaptive gamma correction for image enhancement. *EURASIP Journal on Image and Video Processing*, 2016:1–13, 2016. 1

[31] George Seif and Dimitrios Androutsos. Edge-based loss function for single image super-resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1468–1472, 2018. 6

[32] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5759–5770, 2022. 2

[33] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2654–2662, 2023. 6, 7

[34] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2654–2662, 2023. 2

[35] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2604–2612, 2022. 7

[36] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

[37] Zhi-Guo Wang, Zhi-Hu Liang, and Chun-Liang Liu. A real-time image processor with combining dynamic contrast ratio enhancement and inverse gamma correction for pdp. *Displays*, 30(3):133–139, 2009. 1

[38] C Wei, W Wang, W Yang, and J Liu. Deep retinex decomposition for low-light enhancement. arxiv 2018. *arXiv preprint arXiv:1808.04560*, 1808. 1, 2, 6, 7

[39] Chen Wu, Zhuoran Zheng, Xiuyi Jia, and Wenqi Ren. Mixnet: Towards effective and efficient uhd low-light image enhancement. *arXiv preprint arXiv:2401.10666*, 2024. 2

[40] Yuhui Wu, Guoqing Wang, Zhiwen Wang, Yang Yang, Tianyu Li, Peng Wang, Chongyi Li, and Heng Tao Shen. Reco-diff: Explore retinex-based condition strategy in diffusion model for low-light image enhancement. *arXiv preprint arXiv:2312.12826*, 2023. 2, 3

[41] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13095–13105, 2023. 2, 3, 4, 6

[42] Ao Li Florin-Alexandru Vasluianu Yulun Zhang Shuhang Gu Le Zhang Ce Zhu Radu Timofte Xiaoning Liu, Zongwei Wu. NTIRE 2024 challenge on low light enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 2, 6

[43] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17714–17724, 2022. 2, 7

[44] Shuzhou Yang, Xuanyu Zhang, Yinhuai Wang, Jiwen Yu, Yuhan Wang, and Jian Zhang. Diffflle: Diffusion-guided domain calibration for unsupervised low-light image enhancement. *arXiv preprint arXiv:2308.09279*, 2023. 2, 7

[45] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14816–14826, 2021. 2

[46] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 2, 5, 7

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[48] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1632–1640, 2019. 2, 7

[49] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, and Jiawan Zhang. Beyond brightening low-light images. *International Journal of Computer Vision*, 129:1013–1037, 2021. 2

[50] Shen Zheng, Yiling Ma, Jinqian Pan, Changjie Lu, and Gaurav Gupta. Low-light image and video enhancement: A comprehensive survey and beyond. *arXiv preprint arXiv:2212.10772*, 2022. 1

[51] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. *arXiv preprint arXiv:2305.10028*, 2023. 2, 3

[52] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *European conference on computer vision*, pages 573–589. Springer, 2022. 2, 3, 4, 6, 7

[53] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1219–1229, 2023. 3