Learning to See in the Extremely Dark

Hai Jiang[1,*], Binhao Guan[2,*], Zhen Liu[2], Xiaohong Liu[3], Jian Yu[4], Zheng Liu[4],
Songchen Han[1], Shuaicheng Liu[2,†]

[1]School of Aeronautics and Astronautics, Sichuan University
[2]University of Electronic Science and Technology of China
[3]Shanghai Jiao Tong University [4]National Innovation Center for UHD Video Technology

jianghai@stu.scu.edu.cn, {guanbinhao@std., liushuaicheng@}uestc.edu.cn

Abstract

Learning-based methods have made promising advances in low-light RAW image enhancement, while their capability to extremely dark scenes where the environmental illuminance drops as low as 0.0001 lux remains to be explored due to the lack of corresponding datasets. To this end, we propose a paired-to-paired data synthesis pipeline capable of generating well-calibrated extremely low-light RAW images at three precise illuminance ranges of 0.01-0.1 lux, 0.001-0.01 lux, and 0.0001-0.001 lux, together with high-quality sRGB references to comprise a large-scale paired dataset named See-in-the-Extremely-Dark (SIED) to benchmark low-light RAW image enhancement approaches. Furthermore, we propose a diffusion-based framework that leverages the generative ability and intrinsic denoising property of diffusion models to restore visually pleasing results from extremely low-SNR RAW inputs, in which an Adaptive Illumination Correction Module (AICM) and a color consistency loss are introduced to ensure accurate exposure correction and color restoration. Extensive experiments on the proposed SIED and publicly available benchmarks demonstrate the effectiveness of our method. The code and dataset are available at https://github.com/JianghaiSCU/SIED.

1. Introduction

Restoring high-quality sharp images from low-light observations is a challenging task, as it requires improving global and local contrast, suppressing amplified noise, and preserving details, which is critical for realistic applications such as night-time photography and surveillance. Recent advances in deep learning-based low-light image enhancement (LLIE) have significantly propelled this field forward.

---
*Equal contribution
†Corresponding author

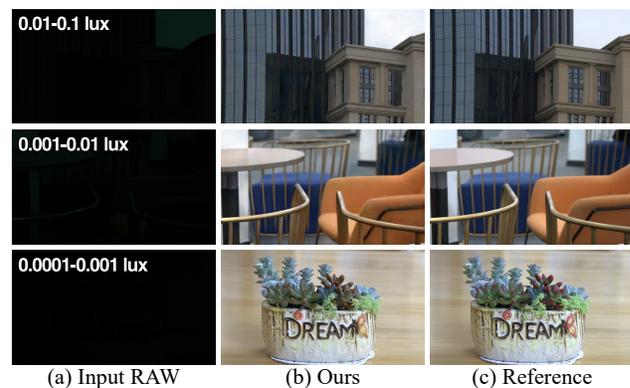(a) Input RAW   (b) Ours   (c) Reference

Figure 1. An illustration of our synthesized low-light RAW images at three extremely low illuminance levels of 0.01-0.1 lux, 0.001-0.01 lux, and 0.0001-0.001 lux are shown in (a), with the restored results from our method in (b) and the reference high-quality sRGB images in (c). The input low-light RAW images are visualized by a simple demosaicing process.

Generally, learning-based methods can be divided into two categories in terms of their input format: sRGB-based methods [1, 8, 12, 14, 20, 30, 33, 37, 38, 40, 42] and RAW-based methods [2, 4, 6, 15, 17, 18, 23, 31, 35]. Compared to the extensively studied sRGB-based LLIE methods, RAW-based approaches have garnered growing attention due to their ability to exploit more informative low-intensity RAW signals and superior noise modeling capabilities.

To this date, numerous efforts have concentrated on transforming low-light noisy RAW images into normal-light clean sRGB images, employing either single-stage or multi-stage training strategies. Single-stage methods [2, 7, 17, 18, 23] attempt to model both the noisy-to-clean and RAW-to-sRGB transformations within a single network. However, the hybrid mappings across these two distinct domains inevitably mislead the enhancement process, causing domain ambiguity issues [15]. Furthermore, the complexity of multiple nonlinear processes in the RAW-to-sRGB trans-

formation exacerbates the problem, often resulting in color distortion and blurry details. To overcome these limitations, multi-stage methods [4, 6, 13, 15, 36, 43] decouple the tasks using intermediate supervision across distinct domains and fully exploit the unique characteristics of RAW and sRGB formats to achieve improved results.

On the other hand, qualified datasets are indispensable for advancing learning-based low-light RAW enhancement methods. Existing datasets [2, 3, 6, 28] typically collect paired images in dark environments via varying camera parameters, where the normal-light reference images are obtained by extending exposure times. However, due to the influence of environmental light sources and brightness fluctuations during data collection, these datasets offer only coarse illuminance levels for the captured low-light RAW images (e.g., 0.2-5.0 lux for indoor scenes and 0.03-0.3 lux for outdoor scenes in the SID [2] dataset). Moreover, in truly extreme dark conditions, such as environments with illuminance as low as 0.0001 lux, obtaining well-exposed reference images through long exposure time is unattainable, which would result in residual noise and motion blur.

To this end, we propose a paired-to-paired data synthesis pipeline to prepare a more challenging dataset specifically designed for extremely low-light RAW image enhancement, named **S**ee-**i**n-the-**E**xtremely-**D**ark (**SIED**), which generates low-light RAW images at three well-calibrated extremely low illuminance levels, along with sharp sRGB references, through three aspects: 1) collecting qualified low-light RAW images with three precise illuminance ranges of 0.01-0.1 lux, 0.001-0.01 lux, and 0.0001-0.001 lux in a professional optical laboratory, 2) capturing paired low-light and normal-light images across diverse real-world scenes by adjusting the camera parameters in line with previous data collection strategies [2, 6] where the normal-light images serve as high-quality references, 3) mapping the illumination of the captured low-light images to the above three distinct illuminance ranges to align with the standard data collected in the laboratory and employing the calibrated sensor noise at each range to the adjusted low-light images to emulate realistic dark environments. With professional laboratory calibration and our paired-to-paired data synthesis strategy, our pipeline ensures realistic low-light RAW inputs for extremely dark scenes and corresponding high-quality sRGB references, as shown in Fig. 1(a) and (c).

In addition to the dataset, we propose a diffusion-based framework trained with a multi-stage strategy, which leverages the generative ability and intrinsic denoising capability of diffusion models for extremely low-light RAW image enhancement. Specifically, we first convert the paired low-light RAW image and normal-light sRGB image into latent space, where the encoded RAW feature is sent to the designed Adaptive Illumination Correction Module (AICM) for exposure correction, which differs from the previous

methods [2, 4, 7, 15, 23] employing the exposure information of reference images to perform pre-amplification, aiming to achieve better restoration results and avoid the exposure bias in the following diffusion processes. Subsequently, the encoded sRGB feature serves as input for the diffusion model for restoration with the guidance of the amplified RAW feature, in which a color consistency loss is further proposed to facilitate the diffusion model to generate reconstructed sRGB features with accurate color mapping. As shown in Fig. 1(b), our method effectively improves global and local contrast, presents vivid color, and avoids noise amplification, resulting in visually satisfactory images in extremely dark environments.

To summarize, our main contributions are as follows:
- We introduce a novel paired-to-paired data synthesis pipeline capable of generating low-light RAW images at three distinct illuminance levels, along with high-quality sRGB images, to form the large-scale SIED dataset for extremely low-light RAW image enhancement.
- We propose a diffusion-based framework that leverages the generative capability and intrinsic denoising property of diffusion models to restore visually pleasing sharp images from extremely low-light RAW inputs.
- Extensive experiments show that our method outperforms existing state-of-the-art single-stage and multi-stage competitors both qualitatively and quantitatively.

2. Related Work

Low-light RAW Image Enhancement methods. RAW images in comparison to sRGB images are more informative and have thus been widely used for image enhancement in low-light conditions. With the development of deep learning, numerous efforts have been made to transform the low-light RAW domain into the sharp sRGB domain with deep neural networks, which can be divided into single-stage and multi-stage methods. Single-stage methods [2, 7, 17, 18, 23] typically promote the network to generate high-quality sRGB images from input RAW images through well-designed optimization objectives directly. However, it is difficult to learn multiple nonlinear transformations with a single network, making them often insufficient for accurate detail reconstruction and color mapping. Multi-stage approaches [4, 6, 13, 15, 36, 43] are designed to overcome the above limitations, achieving improved results by decoupling tasks and thereby effectively reducing ambiguities between different domains.

Low-light RAW Image Datasets. High-quality paired datasets are indispensable for advancing learning-based methods. The SID [2] dataset contains paired low-/normal-light images captured by Sony and Fuji cameras, where the environmental illuminance is 0.2-5.0 lux for indoor scenes and 0.03-0.3 lux for outdoors. The SDSD dataset [28] provides spatially aligned video pairs from dynamic scenes for
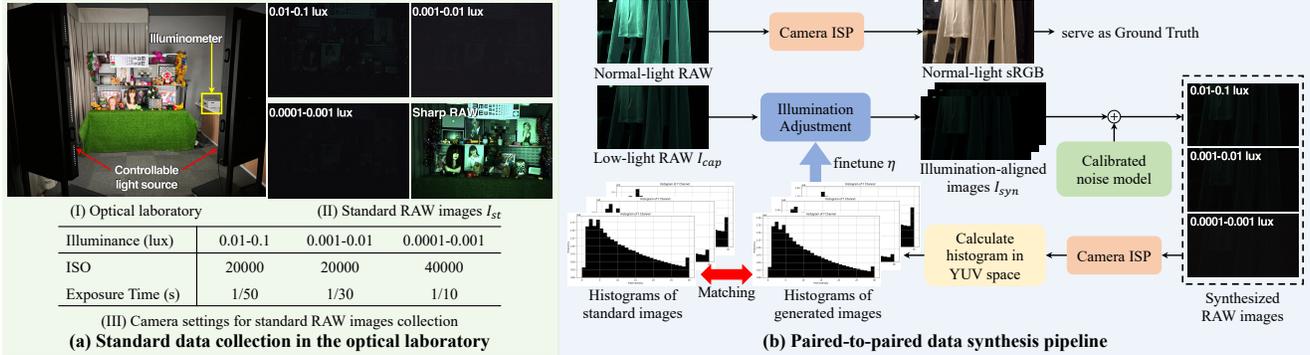
Figure 2. (a) illustrates the optical laboratory, collected standard RAW images, and camera settings adopted for image capturing. (b) presents our proposed paired-to-paired data synthesis strategy. The RAW images are visualized by a simple demosaicing process.

low-light RAW video enhancement, with the illuminance of approximately 0.8–2.0 lux. The SMID dataset [3] includes static videos with ground truth and dynamic videos without ground truth, where most scenes are in the 0.5-5.0 lux range. Due to the environmental complexity and variations in data collection strategies, the above datasets only provide rough illuminance ranges and lack data captured in darker scenes. In this paper, we design a new paired-to-paired data synthesis pipeline to prepare a more challenging dataset that contains low-light RAW images with three well-calibrated illuminance ranges as low as 0.0001 lux.

3. See-in-the-Extremely-Dark Dataset

To generate low-light RAW images with precise illumination ranges, we first use Sony α7RIII and Canon EOS R to capture qualified low-light RAW images with illumination ranges of 0.01-0.1 lux, 0.001-0.01 lux, and 0.0001-0.001 lux in an optical laboratory with controllable light sources and a professional illuminometer (PHOTO-2000μ), as shown in Fig. 2(a). Subsequently, instead of synthesizing low-light images from normal-light images as in the previous methods, we propose a paired-to-paired synthesis strategy to enable the synthesized images to approximate realistic images, as shown in Fig. 2(b), which consists of three steps: 1) collecting paired low-light RAW and sharp sRGB images with the above two cameras in various scenes, 2) manually adjusting the illumination of the low-light images to produce images with specific illuminance levels that align with the standard laboratory data, 3) adding the calibrated noise model to satisfy the realistic dark environments.

Realistic Paired Data Collection. We collect paired low-light RAW images and normal-light sRGB images across various real-world static scenes by varying the camera parameters following [2, 28], where the exposure times for reference images range between 1/10 and 1/200 second, which is 20 to 200 times longer than for low-light images. In each scene, we adapt camera settings such as aperture,

Table 1. The KL divergence ($\downarrow$) of illumination histogram distributions between our synthesized images and laboratory images.

	0.01-0.1 lux	0.001-0.01 lux	0.0001-0.001 lux
Canon	0.017	0.018	0.011
Sony	0.011	0.009	0.059

ISO, and focal length to maximize the quality of the reference image and fix the settings when capturing the low-light image. To ensure content consistency, we adopt tripods to mount the cameras with remote applications to control exposure settings and capture. To improve the scene richness of the dataset, we crop the original full-resolution images, i.e., 7,952×5,304 for Sony and 6,720×4,480 for Canon, to images with 3,840×2,160 resolution, resulting in 1,680 paired images in each subset for subsequent synthesis.

Illumination Alignment. Since the illumination information in RAW images is linearly correlated with the photon intensity, we employ the exposure of standard data I_{st} to adapt the captured low-light images I_{cap}, which is similar to the amplification strategy in previous methods [2, 4, 15], to simulate the realistic illumination degradation as:

$$I_{syn} = I_{cap} * \left(\frac{\text{Expo}(I_{st})}{\text{Expo}(I_{cap})} + \eta \right), \quad (1)$$

where I_{syn} is the synthesized image aligned with the standard data across three illuminance ranges, $\text{Expo}(\cdot)$ calculates the mean value of Bayer channels to characterize illuminance information, and η is the manually defined factor.

Noise Addition. In realistic dark scenes, noise is the inevitable degradation factor. To generate more realistic low-light RAW images, we first estimate the sensor noise model of Canon and Sony cameras in the optical laboratory, fitting the Gaussian and Poisson noise distributions under various ISO values. Moreover, as mentioned in [34, 41], the noise model under extremely low-light conditions should not be considered a pure P+G model. Therefore, we add the dark-
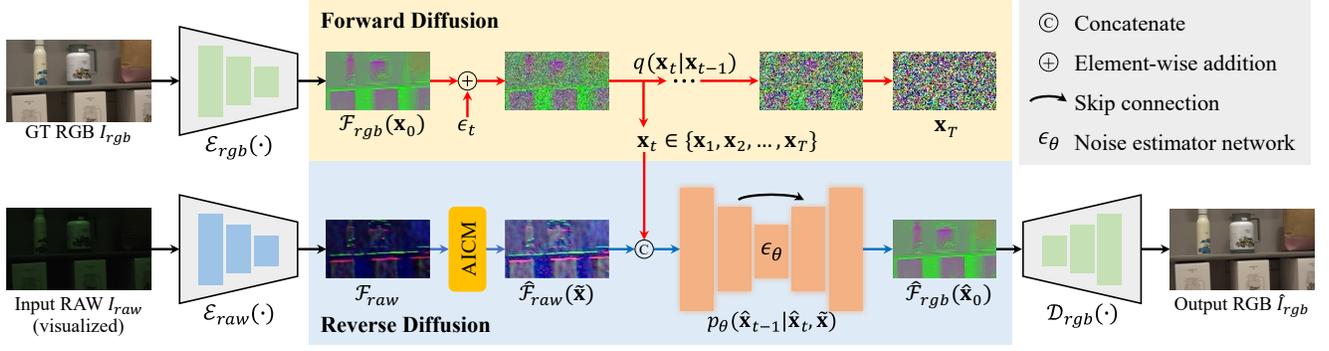
Figure 3. The overall pipeline of our proposed framework. We first employ a RAW encoder $\mathcal{E}_{raw}(\cdot)$ and a sRGB encoder $\mathcal{E}_{rgb}(\cdot)$ to convert the input noisy RAW image I_{raw} and reference GT image I_{rgb} into latent space denoted as $\mathcal{F}_{raw}$ and $\mathcal{F}_{rgb}$. Then, the encoded sRGB feature is taken as the input of the diffusion model to perform the forward diffusion process, while the encoded RAW feature is sent to the proposed Adaptive Illumination Correction Module (AICM) to generate contrast improved feature denoted as $\hat{\mathcal{F}}_{raw}$, aiming to avoid the exposure bias in diffusion processes. In the reverse diffusion process, the refined RAW feature $\hat{\mathcal{F}}_{raw}$ serves as guidance to generate the reconstructed sRGB feature $\hat{\mathcal{F}}_{rgb}$ from the noised tensor $\mathbf{x}_t$, which is replaced by randomly sampled Gaussian noise $\hat{\mathbf{x}}_T$ during inference. Finally, the reconstructed sRGB feature is sent to the sRGB decoder $\mathcal{D}_{rgb}(\cdot)$ to produce the final restored result $\hat{I}_{rgb}$.

Figure 4. Examples from the Canon and Sony subsets of our SIED dataset, with reference sRGB images shown in front and low-light RAW images at three illuminance levels shown in the back.

frame database used for calibrating Gaussian noise to complement the noise types that are hard to model explicitly to better satisfy the characteristics of realistic noise distribution [41], resulting in the calibrated noise model is the combination of the Gaussian, Poisson, and dark-frame distribution. Finally, we adopt the ISO-dependent noise addition strategy [29] to apply the calibrated noise model within an equivalent ISO range between 100 and 20,000 to the illumination-aligned images with the illuminance of 0.01-0.1 lux and 0.001-0.01 lux, while the ISO range is set between 100 and 40,000 for 0.0001-0.001 lux.

To refine the illumination of generated images, we adopt a fixed ISP pipeline to transform the generated RAW image and standard images into YUV space, and manually finetune η to match the illumination histograms of the two images in the Y channel, which represents the illumination information of images [8]. As shown in Table 1, the mean KL divergence of the histogram distributions between our generated images and standard laboratory images is less than 0.06 across three illuminance levels. Overall, the synthesized low-light noisy RAW images with three relatively precise illuminance ranges and the corresponding normal-light sRGB images comprise each sample of our See-in-the-Extremely-Dark (SIED) dataset, containing 1,680 paired images in each illuminance level of Sony and Canon subsets where 1,500 pairs are selected for training and 180 pairs are used for evaluation. We present several samples in Fig. 4 and more details can be found in the supplementary.

4. Methodology

4.1. Overview

Low-light RAW image enhancement presents several critical concerns: contrast enhancement, noise suppression, detail reconstruction, and global color mapping. To overcome these challenges, we propose a diffusion-based framework that leverages the generative ability of diffusion models to achieve visually satisfactory results, as illustrated in Fig. 3. Given a paired low-light RAW image $I_{raw} \in \mathbb{R}^{H \times W \times 1}$ and reference sRGB image $I_{rgb} \in \mathbb{R}^{H \times W \times 3}$, we first employ a RAW encoder $\mathcal{E}_{raw}(\cdot)$ and a sRGB encoder $\mathcal{E}_{rgb}(\cdot)$, to transform the input images into latent space denoted as $\mathcal{F}_{raw} \in \mathbb{R}^{h \times w \times C}$ and $\mathcal{F}_{rgb} \in \mathbb{R}^{h \times w \times C}$. Then, we design an Adaptive Illumination Correction Module (AICM) to improve the contrast of the encoded RAW feature denoted as $\hat{\mathcal{F}}_{raw}$, aiming to avoid the exposure bias in the following diffusion processes and achieve better restoration results. Subsequently, the encoded sRGB feature serves as the input for the diffusion model with the guidance of the amplified RAW feature to generate restored feature $\hat{\mathcal{F}}_{rgb}$. Finally, the restored feature will be sent to a sRGB decoder $\mathcal{D}_{rgb}(\cdot)$ for reconstruction to produce the final restored image $\hat{I}_{rgb}$.
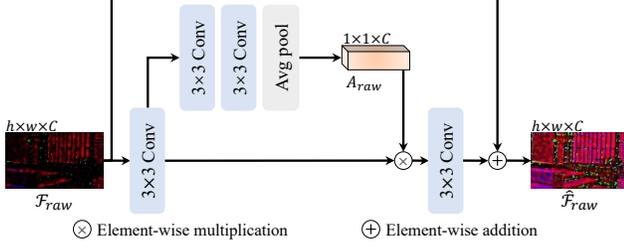
Figure 5. The detailed architecture of our proposed AICM.

4.2. Adaptive Illumination Correction Module

Previous methods [2, 4, 15, 23, 43] have shown that pre-amplifying low-light images can lead to improved restoration results. However, most approaches rely on the exposure values of reference normal-light images as a prior for amplification, which is unlikely to be available in practical applications. Moreover, while diffusion models demonstrate impressive generative capabilities, low-frequency generative biases would be encountered, particularly in exposure [24].

To this end, we propose a simple yet effective Adaptive Illumination Correction Module (AICM), which estimates the amplification factor from the low-light RAW feature to perform illumination correction in the latent space, aiming to avoid exposure bias in diffusion processes and obtain better restoration results. As shown in Fig. 5, we first use a convolutional layer to embed the input feature $\mathcal{F}_{raw}$ followed by cascaded convolutional layers with an adaptive average pooling layer to estimate the per-channel amplification coefficients $A_{raw} \in \mathbb{R}^{1 \times 1 \times C}$, which is used to improve the contrast of the embedded feature. Finally, the refined feature is reconstructed by a convolutional layer with residual content to generate the amplified feature $\hat{\mathcal{F}}_{raw}$.

To enable adaptive exposure correction in AICM and ensure the amplified RAW feature shares content consistency with the input feature, we propose an illumination correction loss $\mathcal{L}_{icl}$ based on the Retinex theory [19] as:

$$\mathcal{L}_{icl} = \|\mathbf{L}_{\hat{\mathcal{F}}_{raw}} - \mathbf{L}_{\tilde{\mathcal{F}}_{raw}}\|_1 + \|\mathbf{R}_{\hat{\mathcal{F}}_{raw}} - \mathbf{R}_{\mathcal{F}_{raw}}\|_1, \quad (2)$$

where the $\tilde{\mathcal{F}}_{raw}$ is the encoded feature of the normal-light RAW image. $\mathbf{L}$ and $\mathbf{R}$ represent the illumination map and the reflectance map, respectively, which are obtained through the decomposition strategy in [9].

4.3. Diffusion-based RAW-to-sRGB Reconstruction

The main concern after obtaining the amplified feature is to convert it into a high-quality sRGB feature. To achieve this, we propose to leverage the generative ability and intrinsic denoising capability of diffusion models to facilitate the reconstruction from noisy RAW data to clean sRGB representation. Our approach follows standard diffusion models [5, 11, 27] that perform forward diffusion and reverse diffusion processes to generate reconstructed results.

Forward Diffusion. The encoded sRGB feature $\mathcal{F}_{rgb}$ is taken as the input of the diffusion model, denoted as $\mathbf{x}_0$, to perform the forward diffusion process, in which a pre-defined variance schedule $\{\beta_1, \beta_2, ..., \beta_T\}$ is employed to progressively convert $\mathbf{x}_0$ into Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ over T steps, which can be formulated as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad (3)$$

where $\mathbf{x}_t$ indicates the corrupted noisy data at time-step $t \in [0, T]$. With parameter renormalization, we can obtain $\mathbf{x}_t$ directly from the input $\mathbf{x}_0$ and thereby simplify Eq.(3) into a closed expression as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}_t$, where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i$, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Reverse Diffusion. The reverse diffusion process learns to gradually denoise a randomly sampled Gaussian noise $\hat{\mathbf{x}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ into a sharp result $\hat{\mathbf{x}}_0$ conforming to the target sRGB distribution. To strengthen the controllability of the generation procedure, we apply the conditional mechanism [5] to improve the fidelity of the reconstructed results conditioned on the amplified RAW feature $\hat{\mathcal{F}}_{raw}$, denoted as $\tilde{\mathbf{x}}$. The reverse diffusion process can be formulated as:

$$p_\theta(\hat{\mathbf{x}}_{t-1}|\hat{\mathbf{x}}_t, \tilde{\mathbf{x}}) = \mathcal{N}(\hat{\mathbf{x}}_{t-1}; \boldsymbol{\mu}_\theta(\hat{\mathbf{x}}_t, \tilde{\mathbf{x}}, t), \sigma_t^2\mathbf{I}), \quad (4)$$

where $\sigma_t^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t$ is the variance and $\boldsymbol{\mu}_\theta(\hat{\mathbf{x}}_t, \tilde{\mathbf{x}}, t) = \frac{1}{\sqrt{\alpha_t}}(\hat{\mathbf{x}}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\hat{\mathbf{x}}_t, \tilde{\mathbf{x}}, t))$ is the mean value.

In the training phase, instead of optimizing the parameters of the network $\boldsymbol{\epsilon}_\theta$ to promote the estimated noise vector close to Gaussian noise, we follow [21, 22, 26] to generate the clean sRGB feature $\hat{\mathbf{x}}_0 = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \tilde{\mathbf{x}})$, i.e., $\hat{\mathcal{F}}_{rgb}$, and employ the content diffusion loss $\mathcal{L}_{cdl}$ for optimization as:

$$\mathcal{L}_{cdl} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0|\tilde{\mathbf{x}}), t \sim [1,T]} [\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_1]. \quad (5)$$

Color Consistency Loss. The color information in RAW images is presented by unique color arrangements within a single channel, which makes the RAW to sRGB conversion potentially vulnerable to unstable color mapping. To this end, we further propose a color consistency loss $\mathcal{L}_{ccl}$ that optimizes the color histogram [10] of the generated sRGB feature denoted as $\mathcal{H}_{\hat{\mathcal{F}}_{rgb}}$ to align with the counterpart of the encoded reference sRGB feature as $\mathcal{H}_{\mathcal{F}_{rgb}}$, aiming to facilitate the diffusion model to generate reconstructed sRGB features with vivid color. We employ the KL divergence to quantify the distribution disparity instead of directly minimizing the difference between the $\mathcal{H}_{\hat{\mathcal{F}}_{rgb}}$ and $\mathcal{H}_{\mathcal{F}_{rgb}}$, since the color histogram primarily captures the proportion of various colors across the entire image without accounting for spatial arrangement. Thus, the $\mathcal{L}_{ccl}$ is formulated as:

$$\mathcal{L}_{ccl} = \sum_{c \in [0,C)} \mathcal{H}_{\hat{\mathcal{F}}_{rgb}^c} \log(\frac{\mathcal{H}_{\hat{\mathcal{F}}_{rgb}^c}}{\mathcal{H}_{\mathcal{F}_{rgb}^c} + \tau}), \quad (6)$$

Table 2. Quantitative comparisons on the Canon and Sony subsets of the proposed SIED dataset, where the low-light RAW images are with three relatively precise illuminance ranges. The best results are highlighted in **bold** and the second-best results are in underlined. We retrain all comparison methods on the training set of each subset using their officially released codes for fair comparison.

| | | | Canon subset | | | | | | | | |
| | | | 0.01 lux-0.1 lux | | | 0.001 lux-0.01 lux | | | 0.0001 lux-0.001 lux | | |
Category	Method	Reference	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
	SID [2]	CVPR' 18	20.69	0.811	0.428	20.34	0.799	0.450	19.28	0.764	<u>0.497</u>
	DID [23]	ICME' 19	20.34	0.806	0.422	20.05	0.798	<u>0.447</u>	18.84	0.760	0.520
Single-stage	SGN [7]	ICCV' 19	21.79	0.813	<u>0.421</u>	21.07	<u>0.800</u>	<u>0.447</u>	19.42	0.762	0.514
	LLPackNet [18]	BMVC' 20	20.64	0.770	0.562	20.16	0.764	0.574	18.91	0.739	0.613
	RRT [17]	CVPR' 21	20.13	0.766	0.506	19.76	0.753	0.539	18.40	0.727	0.591
	LDC [36]	CVPR' 20	20.51	0.785	0.537	19.97	0.741	0.625	18.93	0.720	0.666
	MCR [6]	CVPR' 22	21.76	0.811	0.445	21.60	0.796	0.474	19.87	0.763	0.531
Multi-stage	DNF [15]	CVPR' 23	<u>24.03</u>	<u>0.813</u>	0.456	<u>23.47</u>	0.796	0.486	<u>21.63</u>	<u>0.769</u>	0.522
	RAWMamba [4]	arXiv' 24	22.63	0.791	0.461	21.99	0.782	0.482	21.05	0.757	0.521
	Ours	-	**24.85**	**0.849**	**0.360**	**24.02**	**0.839**	**0.379**	**22.52**	**0.811**	**0.435**

| | | | Sony subset | | | | | | | | |
| | | | 0.01 lux-0.1 lux | | | 0.001 lux-0.01 lux | | | 0.0001 lux-0.001 lux | | |
Category	Method	Reference	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
	SID [2]	CVPR' 18	21.93	0.811	0.471	21.68	0.804	<u>0.483</u>	20.23	0.774	<u>0.514</u>
	DID [23]	ICME' 19	21.64	0.802	0.478	21.24	0.788	0.502	20.46	0.767	0.532
Single-stage	SGN [7]	ICCV' 19	22.41	<u>0.815</u>	<u>0.466</u>	22.13	0.806	<u>0.483</u>	20.77	0.781	0.523
	LLPackNet [18]	BMVC' 20	21.57	0.787	0.554	21.37	0.782	0.568	20.33	0.763	0.596
	RRT [17]	CVPR' 21	21.11	0.773	0.540	20.95	0.777	0.552	20.49	0.756	0.600
	LDC [36]	CVPR' 20	21.74	0.797	0.517	21.20	0.785	0.555	19.59	0.749	0.637
	MCR [6]	CVPR' 22	22.87	0.810	0.492	22.26	0.806	0.502	20.95	0.771	0.542
Multi-stage	DNF [15]	CVPR' 23	<u>24.24</u>	0.814	0.473	<u>23.91</u>	<u>0.807</u>	0.497	<u>22.42</u>	<u>0.785</u>	0.531
	RAWMamba [4]	arXiv' 24	23.94	0.805	0.489	23.64	0.801	0.499	21.88	0.771	0.548
	Ours	-	**24.98**	**0.837**	**0.425**	**24.51**	**0.830**	**0.433**	**23.10**	**0.814**	**0.466**

where τ is a small constant to avoid the zero denominator.

In the inference phase, we derive the restored sRGB feature $\hat{\mathcal{F}}_{rgb}$ from learned distribution through the reverse diffusion process following [27], and then send it to the sRGB decoder $\mathcal{D}_{rgb}(\cdot)$ to produce the final sRGB image $\hat{I}_{rgb}$.

4.4. Network Training

Our approach employs a two-stage training strategy. In the first stage, we use the paired low-light RAW images and reference normal-light RAW/sRGB images to optimize the RAW encoder-decoder ($\mathcal{E}_{raw}$ and $\mathcal{D}_{raw}$), sRGB encoder-decoder ($\mathcal{E}_{rgb}$ and $\mathcal{D}_{rgb}$), and AICM, while freezing the parameters of the diffusion model. The encoders and decoders are optimized with the content loss $\mathcal{L}_{con}$ as:

$$\mathcal{L}_{con} = \sum_{i=\{raw,rgb\}} ||I_i - \mathcal{D}_i(\mathcal{E}_i(I_i))||_1. \quad (7)$$

The optimization objective in the first stage is formulated as $\mathcal{L}_{stage1} = \mathcal{L}_{con} + \mathcal{L}_{icl}$. In the second stage, we optimize the diffusion model through $\mathcal{L}_{stage2} = \mathcal{L}_{cdl} + \lambda\mathcal{L}_{ccl}$, while freezing the parameters of other modules.

5. Experiments

5.1. Experimental Settings

Implementation Details. We implement the proposed method with PyTorch on one NVIDIA A100 GPU, where the batch size and patch size are set to 1 and 512×512. The networks can be converged after training in two stages with 2×10^5 and 4×10^5 iterations, respectively. We employ the Adam optimizer [16] for optimization with the initial learning rate set to 1×10^{-4} in the first stage and decays by a factor of 0.8 while reinitializing it to a fixed value of 8×10^{-5} in the second stage. The hyper-parameter λ is set to 0.1. For the diffusion model, we adopt the U-Net [25] architecture as the noise estimator network with the time step and sampling step set to 1000 and 20, respectively.

Datasets and Metrics. To evaluate the performance of our method, we conducted experiments on the Canon and Sony subsets of our proposed SIED dataset, which contain low-light RAW images under three different illumination conditions, i.e., 0.01-0.1 lux, 0.001-0.01 lux, and 0.0001-0.001 lux. Moreover, we also conduct experiments on the SID [2] dataset which contains low-light RAW images with normal-light reference images captured by Sony and Fuji cameras.
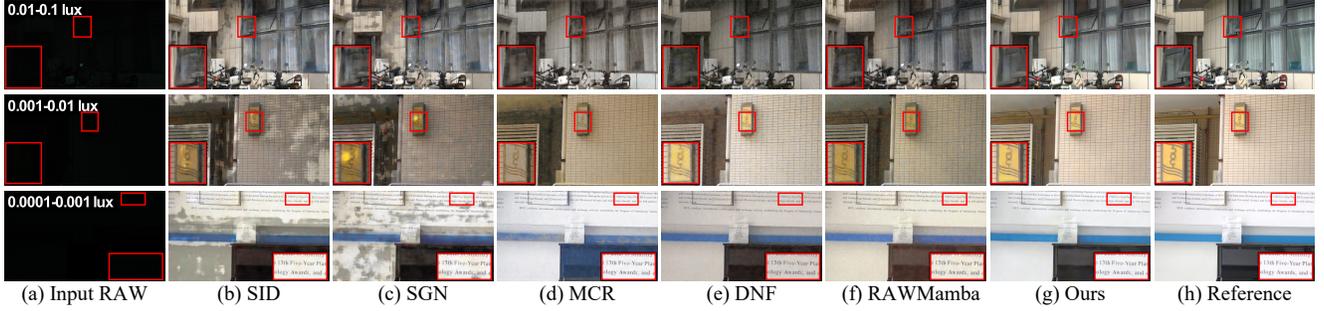
Figure 6. Qualitative comparison of our method and competitive methods [2, 4, 6, 7, 15] on the Canon subset of the proposed SIED dataset. The input RAW images are visualized by a simple demosaicing process. Best viewed by zooming in.
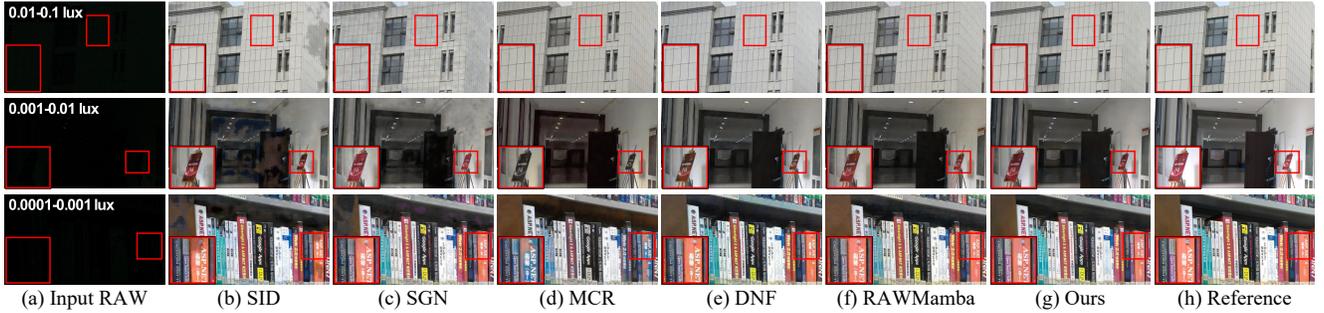

Figure 7. Qualitative comparison of our method and competitive methods [2, 4, 6, 7, 15] on the Sony subset of the proposed SIED dataset. The input RAW images are visualized by a simple demosaicing process. Best viewed by zooming in.

Table 3. Quantitative comparisons on the Sony and Fuji subsets of SID [2] dataset. The best results are highlighted in **bold** and the second-best results are in underlined. "-" indicates the result is not available since the pre-trained model has not been released.

	Method	Sony			Fuji		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Sing-stage	SID [2]	28.96	0.787	0.356	26.66	0.709	0.432
	DID [23]	29.16	0.785	0.368	-	-	-
	SGN [7]	29.28	0.790	0.370	27.41	0.720	0.430
	LLPackNet [18]	27.83	0.755	0.541	-	-	-
	RRT [17]	28.66	0.790	0.397	26.94	0.712	0.446
Mlti-stage	LDC [36]	29.56	<u>0.799</u>	0.359	27.18	0.703	0.446
	MCR [6]	29.65	0.797	0.348	-	-	-
	DNF [15]	<u>30.62</u>	0.797	<u>0.343</u>	<u>28.71</u>	**0.726**	<u>0.391</u>
	RAWMamba [4]	<u>30.62</u>	0.794	0.350	28.65	<u>0.723</u>	0.460
	Ours	**31.20**	**0.801**	**0.339**	**28.90**	**0.726**	**0.390**

Two distortion metrics PSNR and SSIM [32], and a perceptual metric LPIPS [39] are adopted for evaluation.

5.2. Comparison with Existing Methods

Comparison Methods. We compare our method with existing RAW-based LLIE methods including single-stage methods SID [2], DID [23], SGN [23], LLPackNet [18], and RRT [17], as well as multi-stage methods LDC [36], MCR [6], DNF [15], and RAWMamba [4]. For our SIED dataset, we retrain all comparison methods on the training set of each subset for fair comparison. For the SID dataset, we retrain our method on it and adopt the released pre-trained weights of the comparison methods for evaluation.

Quantitative Comparison. We first compare the proposed method with all comparison methods on the Canon and Sony subsets of our proposed SIED dataset. As shown in Table 2, multi-stage methods typically present superior performance than single-stage methods, where our method achieves state-of-the-art performance in terms of both distortion metrics and perceptual metrics across various illumination conditions. To further validate the effectiveness of our method, we also compare our method with comparison methods on the Sony and Fuji subsets of the SID [2] dataset. As shown in Table 3, our method also outperforms previous competitors in all metrics. It is worth noting that, on the SID dataset, previous methods typically use the GT exposure to pre-amplify the low-light RAW image, while our method estimates the amplification coefficients within the proposed AICM, further proving the effectiveness of our method.

Qualitative Comparison. We present visual comparisons of our method and competitive methods on the Canon and Sony subsets of our proposed SIED dataset in Fig. 6 and Fig. 7, where the input RAW images in rows 1-3 are with the illuminance of 0.01-0.1 lux, 0.001-0.01 lux, and 0.0001-0.001 lux, respectively. Single-stage methods suffer from color distortion and unexpected artifacts, while multi-stage

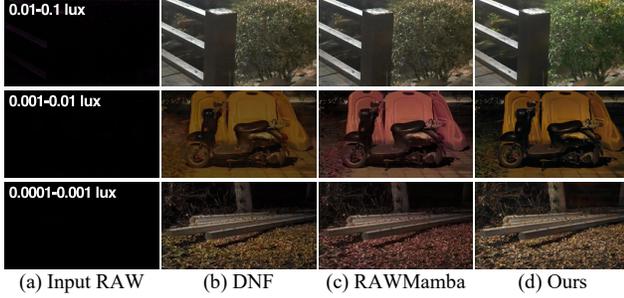|			
(a) Input RAW	(b) DNF	(c) RAWMamba	(d) Ours

Figure 8. Qualitative comparison of our method and competitive methods [4, 15] on real-world extremely dark scenes.

competitors present blurred details, color deviation, or noise amplification. In contrast, our method properly improves contrast, reconstructs sharper details, presents vivid color, and suppresses noise, resulting in visually pleasing results.

Real-world Generalization. To validate the generalization ability of our method, we collect several images in realistic extremely dark scenes using the Sony camera and determine their illuminance level according to the matching of the illumination histograms in the Y-channel to standard laboratory images. As shown in Fig. 8, we use our method and two competitive methods [4, 15] trained on the Sony subset of the proposed SIED dataset to restore the collected realistic images, where all methods can transform extremely low-light images into normal-light counterparts while our method performs better, proving the effectiveness of the proposed method and our synthesized dataset is capable of supporting methods to generalize to real-world scenes.

5.3. Ablation Studies

In this section, we conduct a series of ablation studies to validate the impact of different component choices. The quantitative results for the 0.01-0.1 lux illuminance in the Canon subset of our SIED dataset are illustrated in Table 4.

Training Strategy. To validate the effectiveness of our adopted multi-stage training strategy, we conduct an experiment by optimizing the encoder-decoders, AICM, and diffusion model simultaneously. As shown in row 1 of Table 4, the single-stage training strategy causes overall performance degradation since the encoded features in the early training phase are not favorable for the diffusion model to learn desired target distributions. Nevertheless, our method trained with the single-stage strategy also outperforms previous single-stage competitors in Table 2.

Module Effectiveness. To validate the effectiveness of our proposed AICM, we conduct experiments by removing it from the overall framework as well as replacing it with fixed amplification factors (Amp.) of 100, 200, and 300, respectively. As shown in rows 2-5 of Table 4, with the adaptive exposure improvement realized by our AICM, our method

Table 4. Quantitative results of ablation studies, please refer to the text for more details. 'w/o' denotes without.

	Method	PSNR ↑	SSIM ↑	LPIPS ↓
1)	Single-stage	$22.96_{(-1.89)}$	$0.809_{(-0.040)}$	$0.431_{(+0.071)}$
2)	w/o AICM	$23.23_{(-1.62)}$	$0.839_{(-0.010)}$	$0.378_{(+0.018)}$
3)	Amp. $= 100$	$23.74_{(-1.11)}$	$0.841_{(-0.008)}$	$0.373_{(+0.013)}$
4)	Amp. $= 200$	$23.48_{(-1.37)}$	$0.844_{(-0.005)}$	$0.371_{(+0.011)}$
5)	Amp. $= 300$	$23.18_{(-1.67)}$	$0.838_{(-0.011)}$	$0.382_{(+0.022)}$
6)	w/o $\mathcal{L}_{ccl}$	$24.53_{(-0.32)}$	$0.836_{(-0.013)}$	$0.389_{(+0.029)}$
7)	default	$24.85_{(+0.00)}$	$0.849_{(+0.000)}$	$0.360_{(+0.000)}$

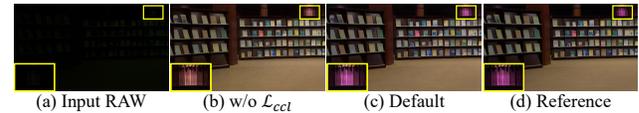|			
(a) Input RAW	(b) w/o $\mathcal{L}_{ccl}$	(c) Default	(d) Reference

Figure 9. Visual results of the ablation study about our proposed color consistency loss $\mathcal{L}_{ccl}$. 'w/o' denotes without.

achieves significant performance gains in terms of PSNR compared to fixed amplification factors since the illumination degradation in realistic scenes is diverse and unknown.

Loss Function. To validate the effectiveness of the proposed color consistency loss $\mathcal{L}_{ccl}$, we conduct an experiment to remove it from the object function utilized to optimize the diffusion model. As reported in rows 6-7 of Table 4, the incorporation of $\mathcal{L}_{ccl}$ results in performance superiority especially in terms of structural similarity and perceptual quality. As shown in Fig. 9, the proposed $\mathcal{L}_{ccl}$ is helpful to achieve accurate color mapping.

6. Conclusion

We have presented a paired-to-paired data synthesis pipeline to prepare a more challenging dataset named SIED for extremely low-light RAW image enhancement, containing low-light RAW images at three well-calibrated illuminance levels as low as 0.0001 lux, along with high-quality reference sRGB images. Moreover, we propose a diffusion-based multi-stage framework that leverages the generative ability and denoising property of diffusion models to restore visually pleasing results from extremely low-light inputs. Technically, we propose an adaptive illumination correction module that performs illumination pre-amplification within the latent space, aiming to obtain better results and avoid exposure bias in diffusion processes. Moreover, we propose a color consistency loss based on the color histogram to promote the diffusion model to generate reconstructed sRGB images with accurate color mapping. Experimental results demonstrate the superiority of our method.

References

[1] 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 *ICCV*, pages 12504–12513, 2023. 1

[2] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, pages 3291–3300, 2018. 1, 2, 3, 5, 6, 7

[3] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *ICCV*, pages 3185–3194, 2019. 2, 3

[4] Xianmin Chen, Peiliang Huang, Xiaoxu Feng, Dingwen Zhang, Longfei Han, and Junwei Han. Retinex-rawmamba: Bridging demosaicing and denoising for low-light raw image enhancement. *arXiv preprint arXiv:2409.07040*, 2024. 1, 2, 3, 5, 6, 7, 8

[5] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *CVPR*, pages 12413–12422, 2022. 5

[6] Xingbo Dong, Wanyan Xu, Zhihui Miao, Lan Ma, Chao Zhang, Jiewen Yang, Zhe Jin, Andrew Beng Jin Teoh, and Jiajun Shen. Abandoning the bayer-filter to see in the dark. In *CVPR*, pages 17431–17440, 2022. 1, 2, 6, 7

[7] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *ICCV*, pages 2511–2520, 2019. 1, 2, 6, 7

[8] Xiaojie Guo and Qiming Hu. Low-light image enhancement via breaking down the darkness. *IJCV*, 131(1):48–66, 2023. 1, 4

[9] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2016. 5

[10] Ju Han and Kai-Kuang Ma. Fuzzy color histogram and its use in color image retrieval. *IEEE TIP*, 11(8):944–952, 2002. 5

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 5

[12] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. In *NeurIPS*, pages 79734–79747, 2023. 1

[13] Haofeng Huang, Wenhan Yang, Yueyu Hu, Jiaying Liu, and Ling-Yu Duan. Towards low light enhancement with raw images. *IEEE TIP*, 31:1391–1405, 2022. 2

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

[15] Xin Jin, Ling-Hao Han, Zhen Li, Chun-Le Guo, Zhi Chai, and Chongyi Li. Dnf: Decouple and feedback network for seeing in the dark. In *CVPR*, pages 18135–18144, 2023. 1, 2, 3, 5, 6, 7, 8

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Mohit Lamba and Kaushik Mitra. Restoring extremely dark images in real time. In *CVPR*, pages 3487–3497, 2021. 1, 2, 6, 7

[18] Mohit Lamba, Atul Balaji, and Kaushik Mitra. Towards fast and light-weight restoration of dark images. *arXiv preprint arXiv:2011.14133*, 2020. 1, 2, 6, 7

[19] Edwin H Land. The retinex theory of color vision. *Scientific American*, 237(6):108–129, 1977. 5

[20] 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. In *ICLR*, 2023. 1

[21] Haipeng Li, Hai Jiang, Ao Luo, Ping Tan, Haoqiang Fan, Bing Zeng, and Shuaicheng Liu. Dmhomo: Learning homography with diffusion models. *ACM TOG*, 43(3):1–16, 2024. 5

[22] Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Flowdiffuser: Advancing optical flow estimation with diffusion models. In *CVPR*, pages 19167–19176, 2024. 5

[23] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via residual learning. In *ICME*, pages 916–921, 2019. 1, 2, 5, 6, 7

[24] Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *ICLR*, 2024. 5

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 6

[26] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2022. 5

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5, 6

[28] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *ICCV*, pages 9700–9709, 2021. 2, 3

[29] Yuzhi Wang, Haibin Huang, Qin Xu, Jiaming Liu, Yiqun Liu, and Jue Wang. Practical deep raw image denoising on mobile devices. In *ECCV*, pages 1–16, 2020. 4

[30] Yinglong Wang, Zhen Liu, Jianzhuang Liu, Songcen Xu, and Shuaicheng Liu. Low-light image enhancement with illumination-aware gamma correction and complete image modelling network. In *ICCV*, pages 13128–13137, 2023. 1

[31] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Exposurediffusion: Learning to expose for low-light image enhancement. In *ICCV*, pages 12438–12448, 2023. 1

[32] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 7

[33] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 1

[34] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE TPAMI*, 44(11):8520–8537, 2021. 3

[35] Qiang Wen, Yazhou Xing, Zhefan Rao, and Qifeng Chen. Ldm-isp: Enhancing neural isp for low light with latent diffusion models. *arXiv preprint arXiv:2312.01027*, 2023. 1

[36] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *CVPR*, pages 2281–2290, 2020. 2, 6, 7

[37] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE TIP*, 30:2072–2086, 2021. 1

[38] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *ICCV*, pages 12302–12311, 2023. 1

[39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 7

[40] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *IJCV*, 129: 1013–1037, 2021. 1

[41] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denoising. In *ICCV*, pages 4593–4601, 2021. 3, 4

[42] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. In *IJCAI*, 2023. 1

[43] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *AAAI*, pages 13106–13113, 2020. 2, 5