Learning Novel View Synthesis from Heterogeneous Low-light Captures

Quan Zheng[1], Hao Sun[2], Huiyao Xu[3], Fanjiang Xu[1]

[1] Institute of Software, Chinese Academy of Sciences, [2] UCAS, [3] Zhejiang University

Abstract

Neural radiance field has achieved fundamental success in novel view synthesis from input views with the same brightness level captured under fixed normal lighting. Unfortunately, synthesizing novel views remains to be a challenge for input views with heterogeneous brightness level captured under low-light condition. The condition is pretty common in the real world. It causes low-contrast images where details are concealed in the darkness and camera sensor noise significantly degrades the image quality. To tackle this problem, we propose to learn to decompose illumination, reflectance, and noise from input views according to that reflectance remains invariant across heterogeneous views. To cope with heterogeneous brightness and noise levels across multi-views, we learn an illumination embedding and optimize a noise map individually for each view. To allow intuitive editing of the illumination, we design an illumination adjustment module to enable either brightening or darkening of the illumination component. Comprehensive experiments demonstrate that this approach enables effective intrinsic decomposition for low-light multi-view noisy images and achieves superior visual quality and numerical performance for synthesizing novel views compared to state-of-the-art methods.

1. Introduction

Neural radiance field (NeRF) [30] has recently become a new paradigm for novel view synthesis from a corpus of input images of a scene. NeRF assumes that all input images are captured under sufficient lighting and each image has the same brightness level. Thus, it has difficulty in learning from input images captured under a low-light condition. The difficult case, however, is pretty common in the real world. Imagine the scenario to capture a low-light scene from multiple views with a smartphone.

In the above low-light scenario, multi-view images are expected to be captured in a relatively short exposure time to simplify the operation and reduce the total time cost of the capturing. These images pose two challenges for NeRF training. First, the images are featured with vary-

Figure 1. Images rendered from NeRF [30] trained on the input views with heterogeneous brightness (*left*) present unevenly illuminated artifacts (*middle* and *right*).

ing low brightness which violates the assumption of the vanilla NeRF. Second, the images are plagued with significant noise which causes distraction for NeRF training (Fig. 1).

Aleth-NeRF [7] proposes to learn an albedo and a concealing field for a low-light image, but this method requires that all input images have the same brightness levels. NeR-Factor [50] factorizes a scene into lighting, normals, albedo, and materials and assumes that multi-view images share the same brightness. For images with varying brightness, NeRF-W [29] proposes to encode the varying image appearance with view-wise appearance embedding. ExtremeN-eRF [20] proposes to decompose a normal-light image into albedo and shading. All of them do not consider the noise issue which is non-negligible for real-world low-light images.

Inspired by the property that the intrinsic reflectance of a scene remains illumination-invariant across multiple views, we propose to decompose the input views into reflectance, illumination, and noise in a self-supervised manner, according to the generalized Retinex theory [19]. The decomposition allows to edit the illumination component and eliminate the impact of noise. Yet, the decomposition is an ill-posed problem due to the ambiguity to explain the image with the three decomposed components. For instance, a dark pixel may be caused by low reflectance, low illumination, or even a noise value. To mitigate the ambiguity and form a plausible decomposition, we incorporate into the decomposition several priors that the reflectance is multi-view consistent, the reflectance value ranges from 0 to 1, and the illumination is locally smooth. Specifically, we design con-

straints on the reflectance, illumination, and noise components based on the priors.

In addition, we introduce an illumination embedding to encode the view-wise illumination. Considering the possible dynamic range of the illumination, we first learn high dynamic range (HDR) illumination, followed by a learnable tone mapping module to simulate the effect of converting HDR illumination to LDR values. Meanwhile, to cope with cross-view varying noise levels, we propose to learn an individual noise map for each view.

To allow an intuitive enhancement operation for synthesizing a novel view, we design an illumination adjustment module to learn the illumination adjustment operation. The module takes as input the decomposed illumination and an adjustment ratio, which serves as an intuitive interface to edit the illumination, without altering the intrinsic reflectance component.

To summarize, this paper has the following contributions:

- We propose an unsupervised scheme to decompose real-world low-light captures into reflectance, illumination, and noise, and the decomposition enables the novel view synthesis from low-light noisy images with varying brightness.
- We propose to learn illumination embedding and individual noise map to cope with view-wise heterogeneous brightness and noise levels.
- We further design an illumination adjustment module to allow intuitive editing of the illumination of novel views.

2. Related work

Our work explores novel view synthesis and brightness enhancement from low-light inputs. In the following, we focus on the majority of research works that are close to our work, and refer readers to [39] for a broad survey on neural rendering and [22] on low-light image enhancement.

Novel View Synthesis. Neural radiance field (NeRF) [1, 8, 30, 32] leverages coordinate-based neural networks to learn a continuous representation and enables to synthesize novel views. However, NeRF assumes input views with homogeneous brightness and bakes the illumination of a scene into its radiance field, thus it does not allow editing the illumination. To provide the editability, a bunch of works have expanded NeRF to handle varying illumination and transient occluders [6, 29, 38]. In terms of varying exposures, Huang et al. [12] and [17] learned a high dynamic range radiance field and a tone mapper for the conversion of dynamic ranges, but they are not designed for low-light captures with the sensor noise. To deal with the noise issue, Pearl et al. [34] propose NeRF-based burst denoising to obtain clean photographs. RawNeRF [31] proposed to learn a denoised high dynamic range radiance field from noisy RAW images. However, the multi-view RAW for-

mat images incur relatively high storage cost on mobile devices, thus reducing the practicality. In contrast, we propose to build a noise-aware neural radiance field from the commonly used sRGB mutli-view low-light images with varying brightness.

Intrinsic Decomposition and Factorization. Instead of learning the radiance, recent works [4, 25, 40, 47] proposed to learn decomposed factors like albedo, roughness, and normals from multi-view images based on the assumed rendering models. Similarly, other works [3, 37, 49, 50] leverage the inverse rendering framework to recover material and geometric properties [14, 33]. However, these methods generally assume the scene is captured under sufficient lighting, different from the tough low-light condition like ours. For low-light captures, camera sensor noise is non-negligible, but these methods do not have mechanisms to process such noise. A recent work [41] learns neural representation from low-light images via intrinsic decomposition, but its decomposition model includes only the reflectance and illumination. Also, it is not designed for input images with varying illumination. In contrast, our approach learns to decompose the image into the reflectance, illumination and noise, which enables denoising and provides an intuitive control for the illumination.

2D Low-light Enhancement. Low-light Image Enhancement (LLIE) strives to improve the perception of images captured in poorly or unevenly lit environments. Early work are based on histogram equalization [35] and the Retinex model-based iterative optimization [15, 16]. Guo et al. [11] introduced structural prior along with the Retinex decomposition to estimate the illumination. Lore et al. [26] started a new genre of LLIE methods with the autoencoder architecture. Afterwards, recent years have witnessed the rapid development of deep learning-based LLIE methods based on diverse learning schemes, including supervised learning [5, 27, 42, 51, 53], semi-supervised learning [45], unsupervised learning [13], and reinforcement learning [48]. Zero-DCE [10] and its follow up [21] proposed to learn pixel-wise curve to adjust the dynamic ranges. Also, several works [36, 43, 46, 52] incorporate the Retinex decomposition into the design of neural network architectures. Recently, unrolling optimization with neural networks and weight sharing [24, 28, 44] are also intensively explored to accelerate the speed of LLIE. Fu et al. [9] decompose the Retinex components from a pair of monocular low-light images, but the noise is not considered. Different from the above 2D LLIE methods, we focus on learning 3D disentangled neural representations from multi-view low-light images to allow synthesis and enhancement of novel views.
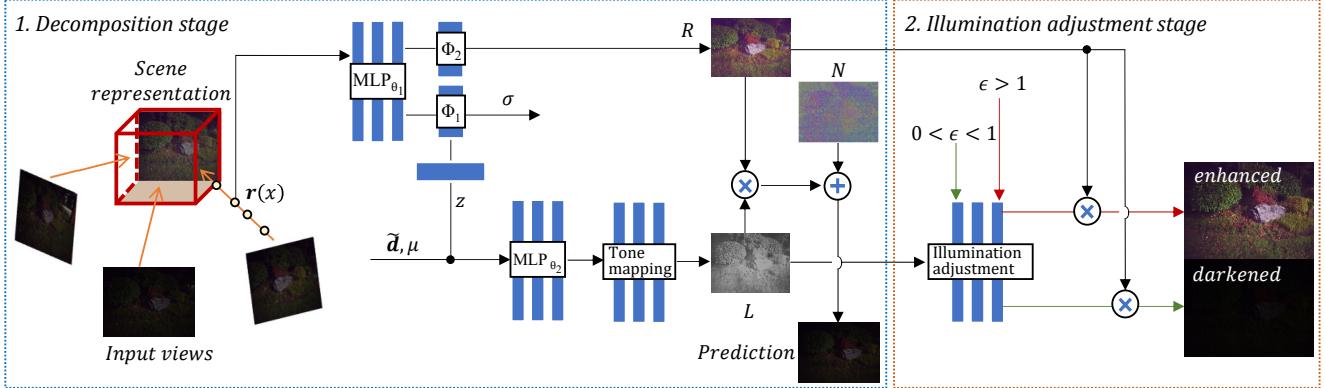
Figure 2. An overview of the proposed framework. In the first stage, our method learns disentangled reflectance, illumination, and noise components from low-light input images with varying brightness based on the Retinex theory. In the second stage, our method can robustly enhance or darken the illumination component. The adjusted image is synthesized by the product of the decomposed reflectance and the adjusted illumination.

3. Preliminaries

Neural Radiance Fields. NeRF learns a continuous volumetric representation of a scene through a series of posed images. Its 5D inputs, consisting of a spatial location and a viewing direction, are mapped to volume density and color through a multilayer perceptron (MLP), and the image is calculated by volume rendering. NeRF emits a ray $r(t) = o + t \cdot d$ from the camera projection center o with the distance t along the direction d. Sampling on the ray is performed between the near plane t_n and the far plane t_f. Feeding 5D inputs directly to NeRF causes low quality at representing high-frequency content, so *positional encoding* on the location and direction coordinates is employed.

The incident radiance of ray r can be computed with

$$\hat{C}(r) = \int_{t_n}^{t_f} T(r(t)) \sigma(r(t)) c(r(t), -d) \, dt, \quad (1)$$

where T is the transmittance, σ denotes the volume density, and c is the direction-dependent color. The integral can be evaluated with numerical quadrature. It firstly casts rays and draws point samples along each ray to get density and color. The approximate radiance of each ray then can be computed by $\hat{C}(r) = \sum_{k=1}^{K} T(t_k) \alpha(\sigma(t_k)\delta_k)c(t_k)$, with transmittance $T(t_k) = \exp\left(-\sum_{j=1}^{k-1} \sigma(t_j)\delta_j\right)$. $\hat{C}(r)$ is the final predicted color of the pixel, where $\alpha(x) = 1 - \exp(-x)$, and $\delta_k = t_{k+1} - t_k$ is the distance between adjacent point samples.

Robust Retinex decomposition. Robust Retinex model [23] assumes that an image can be decomposed into reflectance, illumination, and noise components. Let S denote the source image, then the decomposition can be expressed as:

$$S = R \odot L + N, \quad (2)$$

where R represents the reflectance, L stands for the illumination, N denotes the noise, and $\odot$ denotes the element-wise multiplication. The reflectance component describes the intrinsic color of a scene and is invariant under different lighting conditions.

4. Method

Targeting at real-world images captured in a low-light environment with varying brightness, we aim to learn a neural representation from the images and enable to synthesize and enhance novel views. According to the robust Retinex model [23], we propose to learn a neural scene representation consisting of decomposed reflectance, illumination, and noise. In addition, we design an illumination adjustment module to allow intuitive editing of the brightness for novel views.

Fig. 2 shows an overview of our method, which consists of two stages. The first stage is the unsupervised decomposition process. In Sec. 4.1, we introduce the modules to estimate the reflectance, illumination, and noise components. Then, we present the illumination adjustment module in Sec. 4.2 and the training losses in Sec. 4.3, followed by the implementation details in Sec. 4.4.

4.1. Unsupervised Intrinsic Decomposition

Density and reflectance estimation. NeRF forms a volumetric scene representation, in which volume density only depends a spatial location. Similar to the density, reflectance is only related to the spatial location, based on the property that the reflectance at a location stays invariant across different views. While separate MLPs can be trained to learn the mapping from spatial locations to the density and the reflectance, we propose to reduce neural network parameters by weight sharing. Specifically, we leverage a

single MLP_{θ_1} to learn the shared feature w, and send w to two shallow neural networks ϕ_1, ϕ_2 for predicting the density σ and the reflectance R, respectively (Fig. 2). Since the density is non-negative, Softplus (f_p) activation function is used. Meanwhile, Sigmoid (f_g) is employed for the reflectance to adapt to the value range prior from 0 to 1. In summary, the density and the reflectance of a spatial location can be formulated as:

$$w = MLP_{\theta_1}\left(\boldsymbol{r}\left(x\right)\right),$$
$$\sigma\left(\boldsymbol{r}\left(x\right)\right) = f_p\left(\phi_1\left(w\right)\right), R\left(\boldsymbol{r}\left(x\right)\right) = f_g\left(\phi_2\left(w\right)\right) \quad (3)$$

where $\boldsymbol{r}(x)$ stands for a point on the ray $\boldsymbol{r}$, $\sigma(\cdot)$ is the density, $R(\cdot)$ refers to the reflectance.

Illumination estimation. Different from the multi-view invariant reflectance, the illumination may vary across multiple views. Therefore, we learn an illumination embedding μ for each input view to encode the varying illumination by leveraging *Generative Latent Optimization* [2]. The embedding vector μ has the length η (η is set to 48).

To minimize the coupling between the illumination and the reflectance, we leverage the feature from the hidden layer of ϕ_1 in the density branch as one input to MLP_{θ_2}. Then, μ and the position-encoded direction $\tilde{\boldsymbol{d}}$ are concatenated with the feature z to form the input for MLP_{θ_2}. Since the Retinex theory assumes all color channels share the same illumination, MLP_{θ_2} predicts a one-channel illumination component L_h in high dynamic range

$$L_h\left(\boldsymbol{r}\left(x\right)\right) = MLP_{\theta_2}\left(z, \tilde{\boldsymbol{d}}, \mu\right). \quad (4)$$

Note that the feature z of the position $\boldsymbol{r}\left(x\right)$ also make MLP_{θ_2} aware of the 3D location information. Then we use a neural network-based tone mapping module $\mathcal{T}$ to convert the high dynamic range illumination to low dynamic range illumination $L_l\left(\boldsymbol{r}\left(x\right)\right) = \mathcal{T}\left(L_h\left(\boldsymbol{r}\left(x\right)\right)\right)$. The training of the decomposition stage only produces illumination embeddings for training images. The illumination embeddings of test images are initially unavailable. For test images, we first optimize its illumination embeddings such that the reconstructed images match the input.

Noise estimation. Vanilla NeRF assumes that the input images are noise-free and share the same brightness level. However, images captured under a low-light condition suffer from the camera sensor noise. The noise generally varies across different views. RawNeRF [31] learns a radiance field in the linear RGB space with multi-view RAW images to remove noise, whereas our method targets at the commonly used sRGB images. Note that the quantization operation of sRGB will alter the distribution of noise, such that the multi-view setting of NeRF cannot eliminate the impact of noise thoroughly.

To this end, we propose to guide the neural networks to learn clean reflectance and illumination components, and

optimize a 2D noise map N individually for each view

$$N = \lambda_N \cdot \mathrm{Tanh}\left(\Pi\right), \quad (5)$$

where Π is a trainable image with the same resolution as the corresponding view, $\mathrm{Tanh}(\cdot)$ is the tangent hyperbolic function, λ_N is a hyperparameter to constrain the magnitude of noise. Since the noise is camera sensor related, we do not learn a noise value for every 3D point sample.

Rendering. Given the illumination L and reflectance R for a point sample $\boldsymbol{r}(t_k)$, we compute its color c by taking a product as $c(\boldsymbol{r}\left(t_k\right)) = L\left(\boldsymbol{r}\left(t_k\right)\right) \odot R\left(\boldsymbol{r}\left(t_k\right)\right)$. Finally, the pixel color of the ray $\boldsymbol{r}$ can be calculated by

$$\widehat{C}\left(\boldsymbol{r}\right) = N\left(\boldsymbol{r}\right) + \sum_{k=1}^{K} T\left(t_k\right)\alpha\left(\sigma\left(t_k\right)\delta_k\right)c\left(t_k\right), \quad (6)$$

where $N\left(r\right)$ is the corresponding noise for the ray $\boldsymbol{r}$ and t_k indicates the parametric distance.

4.2. Illumination Adjustment

With the decomposed illumination, a trivial adjustment method is to apply linear transformation $\kappa \cdot L_l$ or gamma transformation $\kappa \cdot \left(L_l\right)^{\gamma}$ to the illumination, where κ and γ are parameters. However, these transformations require manual parameter tuning and tends to introduce color distortion artifacts [22].

To this end, we propose an illumination adjustment stage to transform the input illumination into the adjusted illumination (Fig. 2). The adjusted image can be composited from the product of the adjusted illumination and the reflectance. Because the brightness of images does not have a quantitative characterization, we introduce a variable ϵ to describe the ratio between the output brightness and the input brightness. We leverage images of the same input view with both higher and lower brightness than the input view as the supervisory signals for the stage 2. Images of the high brightness level use two times longer exposure time than the input view, meanwhile images in the low brightness level use only half the exposure time of the input view. Based on the supervisory data, we train the stage 2 to learn the illumination adjustment operations by using the ϵ as a conditional input.

4.3. Training Losses

To enable the robust Retinex decomposition for stage 1, we design the loss $\mathcal{L}_{s1}$ which can be expressed as:

$$\mathcal{L}_{s1} = \mathcal{L}_{recon} + \mathcal{L}_R + \mathcal{L}_L + \mathcal{L}_N, \quad (7)$$

where $\mathcal{L}_{recon}$, $\mathcal{L}_R$, $\mathcal{L}_L$, $\mathcal{L}_N$ denotes the constraint on the reconstructed image, the illumination, the reflectance, and the noise, respectively. These loss terms are detailed as below.

Reconstruction loss. The decomposed reflectance R, illumination L, and noise N should be able to reconstruct the input view S. Therefore, the reconstruction loss is

$$\mathcal{L}_{recon} = \lambda_c \cdot \| S - (R \cdot L + N) \|_1. \tag{8}$$

Illumination loss. We first use an illumination consistency term $\| L - L_0 \|_1$ to guide the estimated illumination L to be similar to the initial illumination estimation L_0. In the Retinex theory, L_0 is computed from the maximum value of R, G, B channels as $L_0(p) = \max_{c \in \{R,G,B\}} S(p)$, where p is a pixel of the input view. Also, to encourage the illumination map be smooth in textural regions and preserve structural boundary, we apply a weighted total variation term $\| w_h \cdot (\nabla_h L) \|_1 + \| w_v \cdot (\nabla_v L) \|_1$, where $\nabla_{h,v}$ refers to the horizontal and vertical gradients. The weights w_h, w_v are defined as $w_h = 1/(\nabla_h L_0)$, $w_v = 1/(\nabla_v L_0)$, respectively. Here, the weights ensure that structural boundaries are preserved. Therefore, the illumination loss is

$$\mathcal{L}_L = \lambda_i \cdot \| L - L_0 \|_1 + \lambda_g \cdot (\| w_h \cdot (\nabla_h L) \|_1 + \| w_v \cdot (\nabla_v L) \|_1). \tag{9}$$

Reflectance loss. Based on the estimated illumination, we can compute the reflectance via a pixel-wise division as S/L. Therefore, we can form the constraint to guide the learning of the reflectance as

$$\mathcal{L}_R = \lambda_r \cdot \| R - S/\text{sg}(L) \|_1, \tag{10}$$

where $\text{sg}(\cdot)$ is the stop gradient operation, which helps to stabilize the training process.

Noise regularization loss. Based on the observation that the noise level in the dark regions is perceptually higher, we introduce the constraint $\| S \cdot N \|_F$, where S provides the information of dark regions and $\| \cdot \|_F$ is the Frobenius norm. In addition, the noise distribution of small patches on a noise map should be the same. Thus, we introduce a standard deviation term $V [M(\zeta_{\{K\}})]$ to encourage that noise patches share a uniform mean value, where $V(\cdot)$ and $M(\cdot)$ denote the operators to compute standard deviation and mean, respectively, $\zeta_{\{K\}}$ is a set of K noise patches. The noise regularization loss is formulated as

$$\mathcal{L}_N = \lambda_n \cdot \| S \cdot N \|_F + \lambda_s \cdot V [M(\zeta_{\{K\}})]. \tag{11}$$

The above $\lambda_c, \lambda_i, \lambda_g, \lambda_r, \lambda_n, \lambda_s$ are balancing weights. We find the values of the balancing weights with the grid search method.

To train the illumination adjustment module of stage 2, we use the $\mathcal{L}^2$ reconstruction loss:

$$\mathcal{L}_{s2} = \| C_d(\boldsymbol{r}) - \widehat{C_d}(\boldsymbol{r}) \|_2^2 + \| C_e(\boldsymbol{r}) - \widehat{C_e}(\boldsymbol{r}) \|_2^2. \tag{12}$$

Here, C_d and C_e are the reference darkened and enhanced pixel color, respectively, and $\widehat{C_d}$ and $\widehat{C_e}$ are the predicted colors.

4.4. Implementation Details

Architecture. The highest positional encoding frequency is 2^{15} for locations and 2^4 for directions. In the decomposition stage, MLP_{θ_1} takes in the position encoded location and has eight fully-connected hidden layers with 256 neurons in each layer. MLP_{θ_2} receives the incoming features and processes them using a fully-connected hidden layer with 128 neurons. At the illumination adjustment stage, the adjustment neural network has a 128-dimensional fully-connected hidden layer. All hidden layers are configured with ReLU activations.

Training details. We implement our approach with PyTorch. We first train the decomposition stage to minimize the loss (Eq. (7)) by utilizing the images from the middle brightness level. Afterwards, we keep the trained weights of the decomposition stage. For the illumination adjustment stage, we leverage the images from the high brightness level and the low brightness level as supervisory data. To learn the enhancement operation, we construct training pairs that have a two-times relation between the exposure time (Sec. 4.2) and set the ratio ϵ to 2. Similarly, we construct training pairs that have a one-half relation between the exposure time to learn the darkening operation and set the darkening ratio ϵ to $1/2$.

Both stages are trained by the Adam optimizer with its default hyperparameters. the initial learning rate 5×10^{-4} is exponentially decayed to 5×10^{-5} with a scheduler. λ_N of Eq. (5) is set to 0.2. The values of balancing weights λ_c, $\lambda_i, \lambda_g, \lambda_r, \lambda_n, \lambda_s$ for loss terms are presented in the supplementary. The decomposition stage is trained for $45,000$ iterations. The mini-batch size of rays is $10,000$. The rays are collected from 100 sampled 10×10 small patches from input views. The illumination adjustment stage is trained for $6,000$ iterations with a mini-batch size $30,000$. The training of stage 1 takes about 20 hours on a Nvidia A100 GPU, and stage 2 takes 2 hours. We refer readers to the supplementary for other details.

5. Experiments

In the experiments, we capture five datasets of real-world scenes. Four of them are indoor scenes and the other one is an outdoor scene. In this section, we present results of four forward-facing scenes. The other $360°$ scene is in the supplementary material. All scenes are captured with a smartphone under low-light conditions. Each dataset has three different scales with the brightness ranging from low to high. Each scale has 30 views and captures two brightness levels for a view. The image resolution is 1000×750 and camera poses are estimated from the captures in the high scale using COLMAP [18]. For the stage 1, we use 25 views of the middle scale for training and the remaining five views for evaluation. To train the stage 2, we utilize the

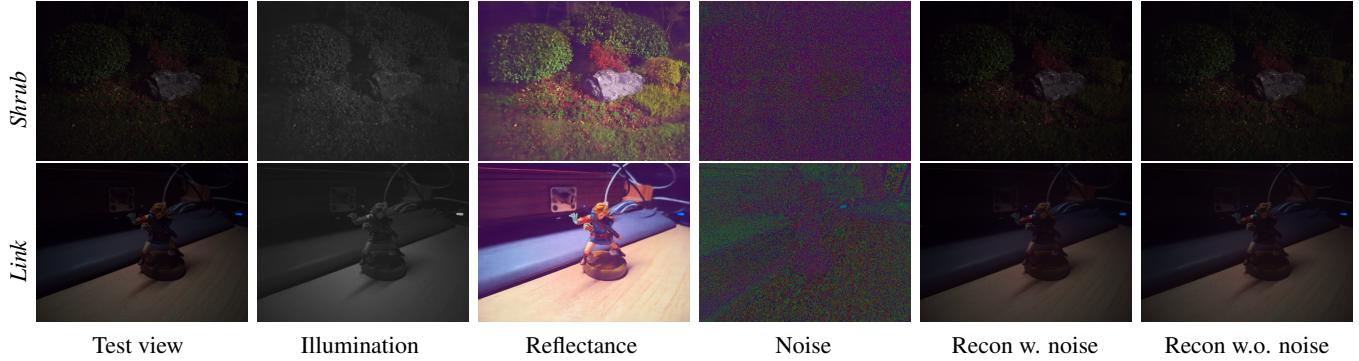| Test view | Illumination | Reflectance | Noise | Recon w. noise | Recon w.o. noise |

Figure 3. The decomposition results of our approach on the test views of the *Shrub* and *Link* scenes. The reconstructions with and without the noise component are also presented. The noise maps are normalized to $[0, 0.4]$ for the visualization.

corresponding 25 views from the low and the high scales as the supervision. Also, the remaining views are used for test.

We first evaluate the decomposition stage of our approach on the test images of two scenes. Then, we compare our approach on the novel view synthesis and enhancement task against the state-of-the-art NeRF-W [29] method. Note that Aleth-NeRF [7] does not deal with camera sensor noise and cannot process low-light images with heterogeneous brightness. Therefore, it is not included in the comparison.

5.1. Evaluation of the Decomposition

Since ground-truth decomposition is unavailable, we present the qualitative decomposition results of a test view from the *Shrub* scene and the *Link* scene in Fig. 3. Our approach achieves a plausible decomposition including illumination, reflectance, and noise. The decomposed results are consistent with the assumption of the Retinex theory, that the illumination component is smooth in textual details and the reflectance component is free of noise since the noise has been separated out. Please refer to the supplementary for the decomposition results on other scenes.

5.2. Novel View Synthesis and Enhancement

In the experiment, we compare the novel view synthesis and enhancement results between our approach and the NeRF-W [29] method on five scenes. Note that NeRF-W has no designs to process the camera sensor noise of low-light images and cannot synthesize novel views with an enhanced brightness level beyond the training views. Thus, for the NeRF-W, we incorporate the non-local mean method of OpenCV as the pre-denoiser and leverage three state-of-the-art 2D low-light enhancement methods SCI [28], DCE [10], and EnGAN [13] for pre-enhancing the training images. Our approach decomposes a test view into three components and enhances the illumination component in stage 2. The qualitative comparisons on the enhanced test views of four scenes are presented in Fig. 4. Tab. 1 presents the cor-

responding quantitative metrics in terms of PSNR, SSIM, and LPIPS. Our approach produces enhanced novel views with both higher visual quality and better numerical metrics than the compared methods. Note that the results of the NeRF-W suffer from severe contour artifacts and loss of details, whereas our approach successfully reconstructs more geometric and textural details.

5.3. Generalization of Illumination Adjustment

In Fig. 5, we demonstrate the generalization ability of our illumination adjustment module to achieve new brightness levels that are not observed from training images. Along with the adjusted images, we present their illumination components. Our approach can robustly darken or enhance the input illumination component of the stage 2 (4th column). If the ratio $\epsilon > 1$ is set, the illumination is enhanced; If ϵ is within $(0, 1)$, the illumination is darkened. As the ratio increases, the image becomes brighter. It is worth noting that the ratio values 0.125, 0.25, 4, and 8 are not observed during training, but our approach still yields reasonable adjustment results without undesired noise and artifacts.

5.4. Ablation Study

Homogeneous and heterogeneous brightness. In Fig. 6, we investigate training our approach with images of the same brightness level. The count of training views is the same as our approach. While the ablation method successfully obtains a decomposition, our approach using images of heterogeneous brightness provides a higher decomposition quality. As shown in the closeup views of the reflectance (Fig. 6), the ablated method misses the textural details on the wall and the face of the toy. Also, its illumination presents inconsistency in local regions. Due to its decomposition error, visible geometric and textural details remain as residuals in the noise map.

Noise module. We conduct an ablation on the noise module of the Retinex decomposition. The noise map and its related losses are removed. We compare the decomposed
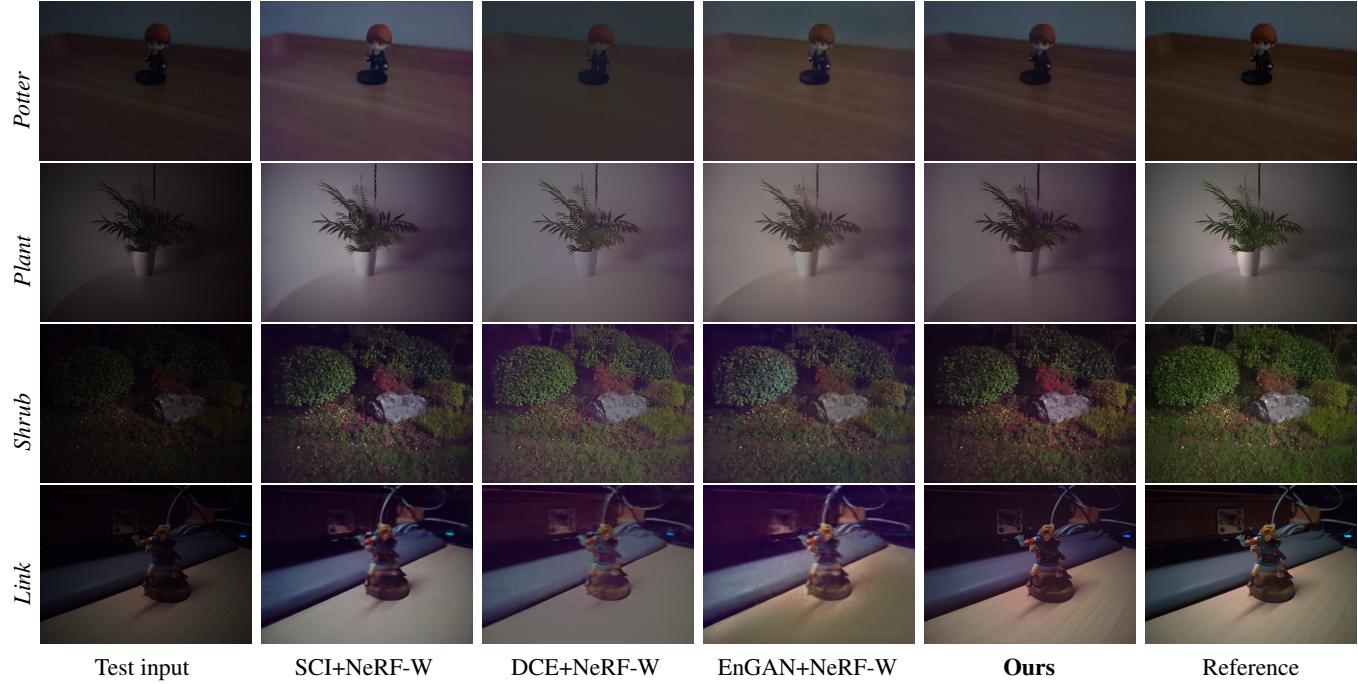
Figure 4. View synthesis and enhancement comparisons between our approach and the SCI+NeRF-W, DCE+NeRF-W, and EnGAN+NeRF-W methods on the same test views of four scenes. The reference images are from the high scale and they are pre-processed by a non-local mean denoiser.

Table 1. Comparisons of the average quantitative metrics on the enhanced test views of four scenes. The best metrics are in bold.

| Method | Potter | | | Plant | | | Shrub | | | Link | | |
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SCI+NeRF-W	20.539	0.868	0.188	22.299	0.902	0.227	22.121	0.641	0.568	20.659	0.831	0.382
DCE+NeRF-W	25.063	0.938	0.164	18.978	0.848	0.221	18.808	0.570	0.513	19.091	0.774	0.427
EnGAN+NeRF-W	21.568	0.883	0.166	18.354	0.841	0.208	19.931	0.591	0.566	17.243	0.741	0.457
Ours	**27.393**	**0.952**	**0.152**	**22.946**	**0.917**	**0.216**	**23.436**	**0.701**	**0.443**	**26.442**	**0.901**	**0.184**

Table 2. Average quantitative results on the test images of the *Potter* scene for the ablations of loss terms. The best metrics (including PSNR, SSIM, and LPIPS) are shown in bold.

Ablation	PSNR↑	SSIM↑	LPIPS↓
w.o. $\|L - L_0\|_1$	35.769	0.973	0.087
w.o. $\|w \cdot (\nabla L)\|_1$	35.890	0.967	0.070
w.o. $\|R - S/L\|_1$	36.099	0.971	0.071
w.o. $\|S \cdot N\|_F$	35.627	0.967	0.071
w.o. $V\left[M\left(\zeta_{\{K\}}\right)\right]$	35.624	0.963	0.074
Ours	**36.864**	**0.975**	**0.065**

reflectance and illumination components of a test view of the *Plant* scene in Fig. 7. Without decomposing the noise, the noise causes distraction for learning the details of the scene. As shown in the closeup views on the top row, structural details of the plant are destroyed. By contrast, our

approach manages to reconstruct these fine details.

Loss functions. We further conduct the ablation of each loss function individually on the *Potter* scene. The average quantitative metrics on the test images are tabulated in Tab. 2, which indicates that the noise regularization losses and the illumination losses play significant roles. The qualitative visual comparisons on the decomposed results and other ablation studies are presented in the supplementary.

6. Limitations and Future Work

Generalization across scenes. Our novel view synthesis and enhancement approach is trained on a per scene basis. While it realizes plausible decomposition and enhancement results after the training on a real-world scene, it does not generalize directly to unseen scenes. A future avenue of this work is to enable the generalization across different scenes. **Pose estimation for low-light views.** We estimate camera
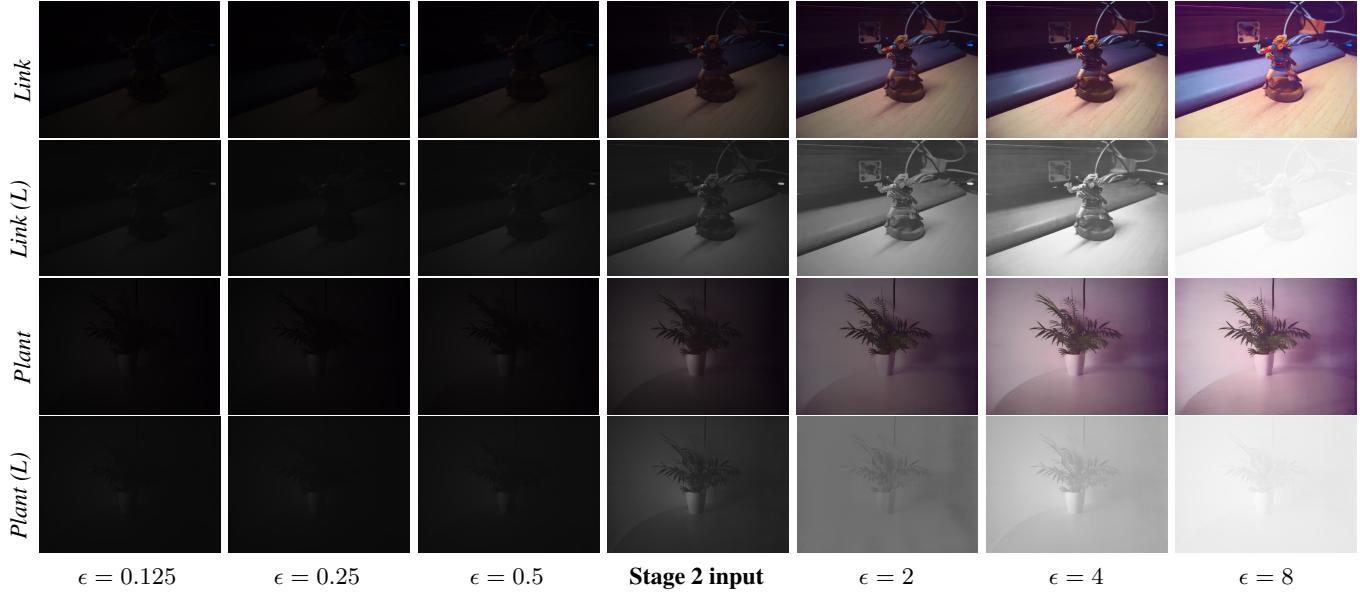
Figure 5. Demonstrations on the generalization ability of our illumination adjustment module on the *Link* and *Plant* scenes. Note that the ϵ values 0.125, 0.25, 4, and 8 are not observed during the training. The second and fourth rows present the adjusted illumination components. The first and third rows are the product of the reflectance and the adjusted illumination.
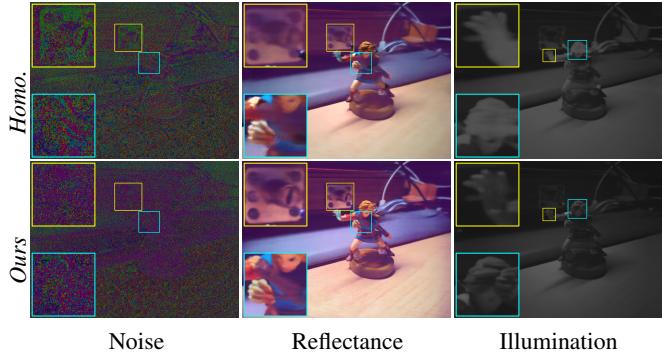

Figure 6. Brightness homogeneity ablation by the comparisons on the decomposition results between training using images with homogeneous brightness and ours for the *Link* scenes. The noise maps are normalized to $[0, 0.4]$ for visualization.
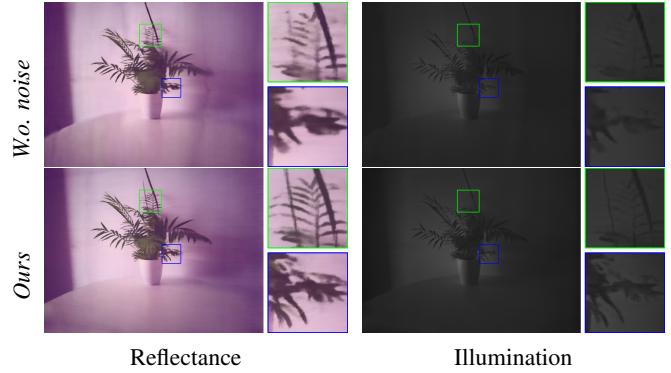

Figure 7. Noise module ablation by the comparison on the decomposed reflectance and illumination for the *Plant* scene. The ablated method cannot reconstruct the details of the foliage as shown in the closeup views.

poses using the images of the high scale. In the case of extremely low-light conditions, the weak signals in the captured images will cause difficulty for the camera pose estimation. We investigate the camera pose estimation using the images from the low scale and present the results in the supplementary. Tackling the pose estimation under extreme low illumination is an important future research avenue.

7. Conclusion

We have demonstrated a novel approach to learn neural representations from multi-view low-light sRGB images with heterogeneous brightness. The tough low-light conditions lead to low pixel values and significant camera sensor noise. Our core idea is to decompose the multi-view low-light images into the invariant reflectance, varying illumination, and individual noise map in an unsupervised manner, according to the robust Retinex theory. Based on the decomposition, we have introduced an effective and intuitive illumination adjustment module for editing the brightness of novel views, without altering the intrinsic reflectance. This work achieves a crucial step towards novel view synthesis from real-world heterogeneous low-light captures and improves the controllability of editing the brightness of novel views.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022. 2

[2] Piotr Bojanowski, Armand Joulin, David Lopez-Pas, and Arthur Szlam. Optimizing the latent space of generative networks. In *International Conference on Machine Learning*, pages 600–609. PMLR, 2018. 4

[3] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. NeRD: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 2

[4] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan Barron, Hendrik Lensch, and Varun Jampani. Samurai: Shape and material from unconstrained real-world arbitrary image collections. *Advances in Neural Information Processing Systems*, 35:26389–26403, 2022. 2

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 2

[6] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12952, 2022. 2

[7] Ziteng Cui, Lin Gu, Xiao Sun, Yu Qiao, and Tatsuya Harada. Aleth-NeRF: Low-light condition view synthesis with concealing fields. *arXiv preprint arXiv:2303.05807*, 2023. 1, 6

[8] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022. 2

[9] Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22252–22261, 2023. 2

[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. 2, 6

[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, 2016. 2

[12] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. HDR-NeRF: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, 2022. 2

[13] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 2, 6

[14] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. TensoIR: Tensorial inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[15] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image processing*, 6(7):965–976, 1997. 2

[16] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. Properties and performance of a center/surround retinex. *IEEE transactions on image processing*, 6(3):451–462, 1997. 2

[17] Kim Jun-Seong, Kim Yu-Ji, Moon Ye-Bin, and Tae-Hyun Oh. HDR-Plenoxels: Self-calibrating high dynamic range radiance fields. In *ECCV*, 2022. 2

[18] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 5

[19] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 1

[20] SeokYeong Lee, JunYong Choi, Seungryong Kim, Ig-Jae Kim, and Junghyun Cho. ExtremeNeRF: Few-shot neural radiance fields under unconstrained illumination. *arXiv preprint arXiv:2303.11728*, 2023. 1

[21] Chongyi Li, Chunle Guo, and Change Loy Chen. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[22] Chongyi Li, Chunle Guo, Ling-Hao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2021. 2, 4

[23] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6):2828–2841, 2018. 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

[25] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. NeRO: Neural geometry and BRDF reconstruction of reflective objects from multiview images. In *SIGGRAPH*, 2023. 2

[26] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. LL-Net: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017. 2

[27] Feifan Lv, Feng Lu, Jianhua Wu, and Chong Soon Lim. MBLLEN: Low-light image/video enhancement using cnns. In *British Machine Vision Conference*, 2018. 2

[28] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022. 2, 6

[29] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 1, 2, 6

[30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2

[31] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. NeRF in the Dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022. 2, 4

[32] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41 (4):102:1–102:15, 2022. 2

[33] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 2

[34] Naama Pearl, Tali Treibitz, and Simon Korman. NAN: Noise-aware NeRFs for burst-denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022. 2

[35] Stephen M Pizer, R Eugene Johnston, James P Ericksen, Bonnie C Yankaskas, and Keith E Muller. Contrast-limited adaptive histogram equalization: Speed and effectiveness. In *Proceedings of the first conference on visualization in biomedical computing, Atlanta, Georgia*, page 1, 1990. 2

[36] Xutong Ren, Wenhan Yang, Wen-Huang Cheng, and Jiaying Liu. LR3M: Robust low-light enhancement via low-rank regularized retinex model. *IEEE Transactions on Image Processing*, 29:5862–5876, 2020. 2

[37] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2

[38] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 2

[39] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, pages 703–735. Wiley Online Library, 2022. 2

[40] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5481–5490. IEEE, 2022. 2

[41] Haoyuan Wang, Xiaogang Xu, Ke Xu, and Rynson WH Lau. Lighting up NeRF via unsupervised decomposition and enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12632–12641, 2023. 2

[42] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6849–6857, 2019. 2

[43] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 2

[44] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. URetinex-Net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2022. 2

[45] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality. *IEEE Transactions on Image Processing*, 30:3461–3473, 2021. 2

[46] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. 2

[47] Weicai Ye, Shuo Chen, Chong Bao, Hujun Bao, Marc Pollefeys, Zhaopeng Cui, and Guofeng Zhang. IntrinsicNeRF: Learning intrinsic neural radiance fields for editable novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[48] Runsheng Yu, Wenyu Liu, Yasen Zhang, Zhi Qu, Deli Zhao, and Bo Zhang. DeepExposure: Learning to expose photos with asynchronously reinforced adversarial learning. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[49] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 2

[50] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. NeRfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics*, 40 (6):1–18, 2021. 1, 2

[51] 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

[52] Zunjin Zhao, Bangshu Xiong, Lei Wang, Qiaofeng Ou, Lei Yu, and Fa Kuang. RetinexDIP: A unified deep framework for low-light image enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1076–1088, 2021. 2

[53] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13106–13113, 2020. 2