PNRNet: Physically-Inspired Neural Rendering for Any-to-Any Relighting

Zhongyun Hu, Ntumba Elie Nsampi, Xue Wang, and Qing Wang, *Senior Member, IEEE*

***Abstract*— Existing any-to-any relighting methods suffer from the task-aliasing effects and the loss of local details in the image generation process, such as shading and attached-shadow. In this paper, we present PNRNet, a novel neural architecture that decomposes the any-to-any relighting task into three simpler sub-tasks, i.e. lighting estimation, color temperature transfer, and lighting direction transfer, to avoid the task-aliasing effects. These sub-tasks are easy to learn and can be trained with direct supervisions independently. To better preserve local shading and attached-shadow details, we propose a parallel multi-scale network that incorporates multiple physical attributes to model local illuminations for lighting direction transfer. We also introduce a simple yet effective color temperature transfer network to learn a pixel-level non-linear function which allows color temperature adjustment beyond the predefined color temperatures and generalizes well to real images. Extensive experiments demonstrate that our proposed approach achieves better results quantitatively and qualitatively than prior works.**

***Index Terms*— Any-to-any relighting, physical image formation, neural rendering.**

I. Introduction

THE goal of this paper is to generate a relit image from the original RGB-D image to match the illumination setting of the given guide RGB-D image. As shown in Fig. 1, the inputs consist of a source image of a complex scene (Fig. 1a) and a guide image under novel lighting (Fig. 1b), and the output is the relit source image under the novel lighting provided by the guide image. Different from traditional image-based relighting where the target illumination is given explicitly [1]–[3], the illumination in any-to-any relighting is implicitly contained in the guide image. This has attracted much attention because it can benefit many applications when casual photographers do not have the expertise in lighting, but the images containing desired illumination settings can be easily obtained.

Inverse rendering-based relighting methods [4]–[6] explicitly recover illumination, geometry and material properties of the scene, then forward rendering acts on these scene factors for relighting. However, this is an ill-posed problem, considering these factors interact in complex ways to form images and different combinations of these factors could produce the same image [7]. In contrast, some learning-based approaches [2], [8], [9] do not have an explicit inverse rendering step for relighting. Instead, a single relighting network is trained to generate relit images from one or more input images. In particular, Zhou *et al.* [9] and Sun *et al.* [8] propose to directly relight the input portrait from an implicit neural representation in the latent space without explicitly reconstructing the intrinsic properties. However, these methods do not apply to any-to-any relighting.

Recently, several learning-based methods [10]–[12] have been proposed to solve any-to-any relighting successfully. But there still exist two challenges that need to be addressed. First, the visual effects of the lighting direction transfer task (e.g. the shadows in the second row of Fig. 1h) are introduced in the image generation process where only the temperature transfer task is involved, which is called the task-aliasing effect in this paper. In fact, the end-to-end model tends to learn a generalisable representation consisting of both on-task and off-task features by integrating relighting of the color temperature and direction of the light source into one single model. Second, existing any-to-any relighting methods fail at preserving the local shading and attached-shadow details. The main reason is that the commonly adopted encoder-decoder structure performs down-sampling operations in the encoder to obtain a larger receptive field to account for global illumination effects, such as cast shadows and inter-reflections. However, from the perspective of the physical image formation process, the down-sampling operations could break the pixel-level local illumination modeling.

To that end, we propose a novel neural architecture that decomposes the any-to-any relighting task into three independent sub-tasks: lighting estimation, color temperature transfer, and lighting direction transfer. Thus the task-aliasing effects in the relit image can be suppressed efficiently rather than combining both on-task and off-task features into one single end-to-end model. First, to relight the source image with the illumination setting of the guide image, we train a network for estimating the illumination setting which will be fed into the following sub-tasks. Second, to adjust the color temperature of the source image, we train a fully-connected neural network that learns the pixel-level non-linear color mapping from estimated color temperature. Finally, to preserve the local shading and attached-shadow details, we train a parallel multi-scale network that maintains a high-resolution representation throughout the whole process to model the pixel-level local illumination using multiple physical attributes.

Manuscript received August 18, 2021; revised March 1, 2022 and April 22, 2022; accepted May 15, 2022. Date of publication May 30, 2022; date of current version June 9, 2022. This work was supported by NSFC under Grant 62031023 and Grant 61801396. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ran He. *(Corresponding author: Qing Wang.)*

The authors are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: qwang@nwpu.edu.cn).

Code is available at: https://github.com/waldenlakes/PNRNet

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2022.3177311, provided by the authors.

Digital Object Identifier 10.1109/TIP.2022.3177311

1941-0042 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

Our contributions can be summarized as follows:

1) We analyze the physical process of image formation under near-field point light sources and derive its relighting formulation. Based on the relighting formulation, we propose to solve any-to-any relighting by learning three functions sequentially. Each function is modeled as a neural network that takes into account the characteristics of the respective physical formation process.

2) We propose two efficient sub-networks that can be separately plugged into different neural networks for specific usage. One is a fully-connected neural network for estimating pixel-level non-linear color temperature mapping. The other one is a parallel multi-scale lighting direction transfer network for modeling pixel-level local illumination using multiple physical attributes.

3) We demonstrate the state-of-the-art performance of the proposed method on the VIDIT benchmark dataset. More importantly, our method shows better generalization results on real data. The code and models are released in order to inspire more research in this direction.

II. Related Work

A. Image-Based Relighting

Image-based relighting methods directly reconstruct the light transport function to relight the objects without an explicit estimate of the physical attributes of the objects. Debevec *et al.* [1] first proposed to relight the objects by densely sampling the light transport function using thousands of images. Furthermore, the coherence of the light transport function [13]–[15] is utilized to relight the objects using fewer samples. However, these approaches still require hundreds of images, and this acquisition process is very time-consuming. Recently, driven by the success of deep learning, Xu *et al.* [2] used a non-linear CNN-based representation that exploited correlations in light transport across scenes to relight the objects with only five images. Meka *et al.* [3] proposed a learning-based solution to reconstruct the light transport function from two spherical gradient images. Such acquisition systems need to be specially designed to simulate the desired illumination. Zhou *et al.* [9] and Sun *et al.* [8] argued that the utility was usually limited due to the requirements of multiple images of the scene under controlled or known illuminations, deep neural networks with encoder-decoder structures are proposed to relight the portrait using a single RGB image. Despite all that, such approaches have often focused on objects of a specific class (e.g. portraits or human bodies), more complex scenes should be further considered. What's more, they usually represent the incident lighting using environment maps or spherical harmonics (SH) that are assumed as distant lighting. As a result, these methods are incapable of rendering scenes with near-field lighting effects.

B. Inverse Rendering

Inverse rendering is to estimate physical attributes (e.g., geometry, reflectance, and illumination) of a scene from observed appearance. Once the reflectance and illumination are estimated, the any-to-any relighting problem can be viewed as a natural extension of inverse rendering, which is then performed by an additional physically based rendering pipeline. Traditional inverse rendering [16]–[21] jointly optimize the scene properties by extensive priors to achieve the set of values that best explain the observed image. However, directly optimizing all scene parameters is often a highly under-constrained problem, which in turn results in severe artifacts in relighting.

In the past few years, researchers have concentrated on data-driven approaches for learning priors instead of hand-crafted priors [22]–[24]. Sengupta *et al.* [22] proposed a residual appearance renderer by employing a self-supervised reconstruction loss to learn inverse rendering on images. Yu and Smith [23] used multi-view stereo supervision to train an hourglass-based neural network with skip connections to predict normal and albedo from a single image. Li *et al.* [24] proposed an inverse rendering network to estimate shape, spatially-varying lighting, and SVBRDF from a single image. Still they are limited to what is expressible by their physically-based rendering model, whether for inverse rendering or relighting.

Some other learning-based relighting methods [4]–[6], [25]–[27] have taken neural rendering into the relighting task. Based on the inverse rendering network [23], Yu *et al.* [6] further proposed a self-supervised neural rendering framework for outdoor scene relighting. Bi *et al.* [27] proposed a neural rendering framework with a scene appearance representation to enable relighting from several unstructured mobile phone flash images. Nestmeyer *et al.* [26] and Wang *et al.* [25] proposed to decompose the input image into intrinsic components using neural networks for single image portrait relighting. Sang and Chandraker [5] proposed a joint learning approach to estimate shape and SVBRDF, as well as relight the object from a single image under point light or environment illumination. However, they require accurate reflectance or multi-view data for supervised training, which is difficult to obtain in practice. In addition, they often focus on single objects rather than complex scenes. In contrast, we propose to solve any-to-any scene relighting by directly learning the relighting functions from RGB-D images without an explicit reflectance estimate.

C. Any-to-Any Relighting

Any-to-any relighting is first proposed in [28], [29], which aims to relight a source image with the illumination settings of a guide image. Hu *et al.* [10] proposed a encoder-decoder structure augmented with a self-attention scheme to improve global illumination effects. Yang *et al.* [11] proposed to solve any-to-any relighting as an image-to-image translation task, where the goal was to directly map the source images and the guides images to the relit images using a single stream structure network. Yazdani *et al.* [12] proposed to fuse the intrinsic image-based relighting results and direct relighting results by a learned weight map. These methods all integrate relighting of the color temperature and direction of the light source into one single end-to-end model, which would suffer from task-aliasing effects. More importantly, their generated relit images often contain blur and artifacts, and do not well

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

(a) Source Images (b) Guide Images (c) Ground Truth (d) Ours (e) AMDIR-Net [12] (f) SA-AE [10] (g) DPR [9] (h) Insets

Fig. 1. Given one RGB-D image of a source scene (a) and one RGB-D image of a guide scene (b), our proposed method is able to relight the source scene with the illumination of the guide scene (d), while maintaining the local shading and attached-shadow details and accurately recovering the color temperature of the relit image (h).

maintain the local shading and attached-shadow details, particularly for the shadow regions in source images. Inspired by the physical principles of image formation, we try to propose a novel neural architecture to mitigate these problems. Last but not least, some image restoration works [30], [31] also adopted the disentanglement idea to design different sub-networks with physically meaningful functions, which is proven to achieve promising performance.

III. Physical Image Formation

Unlike existing any-to-any relighting methods that fail at maintaining the local details, we seek to incorporate the physical principles of image formation into our neural rendering design to improve local illumination effects. The physical process of image formation in the real world, where light sources emit photons that interact with the objects in the scene, is often formulated as the rendering equation [32]:

$$I\left(\mathbf{x}, \mathbf{w_o}\right) = \int_{\Omega^+} f_r\left(\mathbf{x}, \mathbf{w_i}, \mathbf{w_o}\right) L\left(\mathbf{x}, \mathbf{w_i}\right) \left(\mathbf{w_i} \cdot \mathbf{n_x}\right) \mathrm{d}\mathbf{w_i}, \quad (1)$$

where $L\left(\mathbf{x}, \mathbf{w_i}\right)$ and $I\left(\mathbf{x}, \mathbf{w_o}\right)$ are the incoming and outgoing radiance respectively along $\mathbf{w_i}$ and $\mathbf{w_o}$ from a particular surface point $\mathbf{x}$. $f_r\left(\mathbf{x}, \mathbf{w_i}, \mathbf{w_o}\right)$ is the bidirectional reflectance distribution function (BRDF) describing the optical properties of materials, and $\left(\mathbf{w_i} \cdot \mathbf{n_x}\right)$ represents the weakening factor of outgoing radiance due to incident angle.

Assuming that the scene is illuminated with a point light source with the direction $\mathbf{l}$ and the color temperature T, Eq. 1 can be reformulated as:

$$I\left(\mathbf{x}\right) = \frac{f_r\left(\mathbf{x}, \mathbf{l}\right) L\left(\mathbf{x}, \mathbf{l}, T\right) \left(\mathbf{n_x} \cdot \mathbf{l}\right)}{1 + d_{\mathbf{x}}^2\left(\mathbf{l}\right)}, \quad (2)$$

where d is the distance between the point light source and the surface point. L is attenuated with the distance by a factor $1/\left(1 + d^2\right)$, and also determined by its color temperature T. In the case of a fixed viewpoint, we discard the viewing direction $\mathbf{w_o}$ for simplicity. Note that the point light source is conceptually simple, but it leads to challenging relighting problems, such as shadow synthesis and removal. Compared to the directional light source, it also needs to account for near-field illumination effects.

When the scene is illuminated with a point light source with different lighting directions and color temperatures, the appearance of the scene will change accordingly. Suppose the point light source is rotated from $\mathbf{l}_{src}$ to $\mathbf{l}_{tgt}$ and its color temperature changes from T_{src} to T_{tgt}, f_r is a Lambertian BRDF, we can obtain the new outgoing radiance $I_{tgt}(\mathbf{x})$ using Eq. 2 as the following relighting formulation:

$$I_{tgt}(\mathbf{x}) = \overbrace{\underbrace{I_{src}(\mathbf{x}) \frac{L\left(\mathbf{x}, T_{tgt}\right)}{L\left(\mathbf{x}, T_{src}\right)}}_{g} \cdot \frac{\left(1 + d_{\mathbf{x}}^2\left(\mathbf{l}_{src}\right)\right)\left(\mathbf{n_x} \cdot \mathbf{l}_{tgt}\right)}{\left(1 + d_{\mathbf{x}}^2\left(\mathbf{l}_{tgt}\right)\right)\left(\mathbf{n_x} \cdot \mathbf{l}_{src}\right)} V}^{f} \quad (3)$$

local term O

Here a visibility term (also called a global term later) V that takes into account the occlusion between the point light source and the surface point is added to produce shadows. Since the change in lighting directions doesn't affect L in this case, we can discard $\mathbf{l}$ of L for simplicity.

Motivation for neural rendering design. Based on the above relighting formulation, we propose to solve any-to-any relighting by learning three functions, a lighting estimation function h, a color temperature transfer function g, and a lighting direction transfer function f. For the function h, because both T and $\mathbf{l}$ in Eq. 3 are unknown quantities, we first need to use h to estimate T and $\mathbf{l}$ from the known quantities I and D. Once T and $\mathbf{l}$ are obtained, the function f and the function g are proposed to model the two parts of Eq. 3 respectively. Specifically, g is learned from the surface point's appearance and color temperatures. f is learned from the output produced by g, lighting directions, as well as additional physical properties. Thus we formulate our any-to-any relighting learning as:

$$T, \mathbf{l} = h\left(I, D\right), \quad (4)$$

$$I_{tgt}\left(\mathbf{x}\right) = f\left(g\left(I_{src}\left(\mathbf{x}\right), T_{src}, T_{tgt}\right), \mathbf{x}, \mathbf{n_x}, \mathbf{l}_{src}, \mathbf{l}_{tgt}, D_{src}\right), \quad (5)$$

where D denotes the scene depth. We model the functions f, g and h using neural networks.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

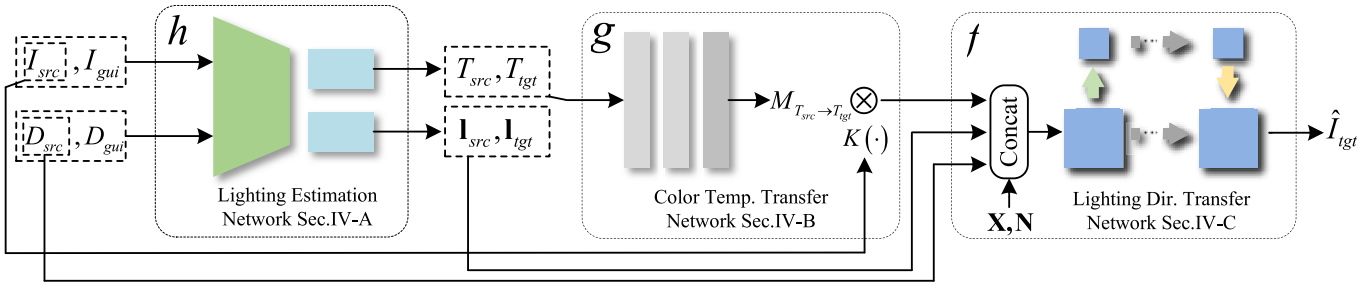


Fig. 2. An overview of our proposed pipeline. The whole pipeline decomposes any-to-any relighting into three sub-networks: 1) lighting estimation network, 2) color temperature transfer network and 3) lighting direction transfer network. The lighting estimation network takes two RGB-D images captured from the source scene and the guide scene respectively as input and predicts their illumination settings including the color temperature and lighting direction. The estimated illumination settings and the source RGB-D image are then fed into the color temperature transfer network and the lighting direction transfer network to generate the target relit image. **X** and **N** denote the position map and the normal map respectively.

We further observe that, the relighting formulation in Eq. 3 can be viewed as the product of a local term O and a global term V. Based on the observation, the generation of the local term needs only to account for the local properties and illumination settings of the surface point. Conversely, the generation of the global term needs to account for the interaction between the current surface point and other surface points. In other words, the global term needs to account for the presence of occluders between the light source and the surface points.

In summary, the function g only describes the local pixel-to-pixel mapping relation, whereas the function f further considers the relation between distant surface points via occlusions in space, while preserving the local pixel-to-pixel mapping relation. These important observations will serve as design guidelines for our networks.

IV. Method

As mentioned in Sec. III, we formulate any-to-any relighting as a regression problem modeled by three functions: f, g and h. In this section, these three functions are further modeled as corresponding feedforward neural networks. Fig. 2 shows the proposed pipeline. By decomposing any-to-any relighting into three sub-tasks, we allow each network to focus on a relatively easier task with direct supervision, which effectively avoids task-aliasing effects. Each network is trained individually with its own input and output pairs derived from ground truth. The details of these three neural networks are described as follows.

A. Lighting Estimation

The lighting estimation network (LE-Net) takes a single RGB-D image as input and outputs estimated illumination settings including a lighting direction and a color temperature. As shown in Fig. 3, we adopt a hard parameter sharing scheme [33] to implement the lighting estimation network. Specifically, the lighting estimation network is composed of a parameter sharing module and two task-specific output modules. The parameter sharing module consists of four residual convolution blocks (RCBs) and a global average pooling (GAP) layer. Each RCB is followed by a max-pooling layer and the GAP layer is used to yield a shared feature. The shared feature is then fed into the task-specific output modules to predict their respective illumination setting.

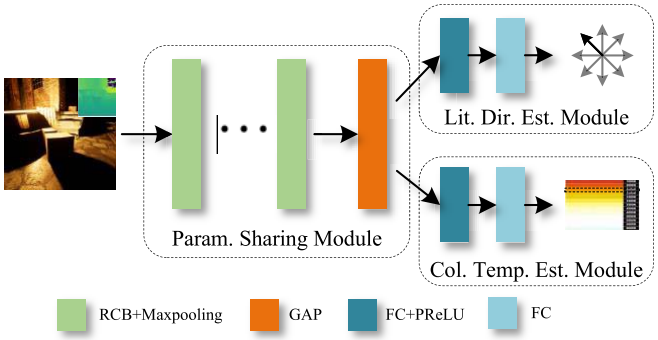


Fig. 3. Lighting Estimation Network. A parameter sharing module is applied to learn a general representation from multi-tasks, while a lighting direction estimation module and a color temperature estimation module are used to estimate their respective illumination settings from shared features.

B. Color Temperature Transfer

The color temperature transfer network (CTT-Net) uses the illumination settings predicted by the lighting estimation network to relight the source image. As discussed in Sec. III, g only considers the local mapping relation. Thus a pixel-to-pixel nonlinear color mapping function $M_{T_{src}\to T_{tgt}}$ is designed to map the source image I_{src} with a color temperature T_{src} to the target image $\hat{I}_{tgt}$ with a color temperature T_{tgt}:

$$\hat{I}_{tgt} = g\left(I_{src}, T_{src}, T_{tgt}\right) = M_{T_{src}\to T_{tgt}}\ \mathrm{K}\left(I_{src}\right), \qquad (6)$$

where $\mathrm{K}\left(I_{src}\right) : R^3 \to R^n$ is a kernel function that transforms the 3-dimensional RGB to a n-dimensional space. Followed by [34], [35], a polynomial kernel function with a 3^{rd} degree expansion is adopted as below:

$$\mathrm{K} : [\mathrm{R}, \mathrm{G}, \mathrm{B}]^T \to [\mathrm{R}, \mathrm{G}, \mathrm{B}, \mathrm{R}^2, \mathrm{G}^2, \mathrm{B}^2, \mathrm{RG}, \mathrm{GB}, \mathrm{RB}, \mathrm{R}^3, \mathrm{G}^3, \mathrm{B}^3, \mathrm{RG}^2, \mathrm{GB}^2, \mathrm{RB}^2, \mathrm{GR}^2, \mathrm{BG}^2, \mathrm{BR}^2, \mathrm{RGB}]^T. \qquad (7)$$

Higher degree polynomials are also chosen as the kernel function, however we did not observe any noticeable improvements in our experiment.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

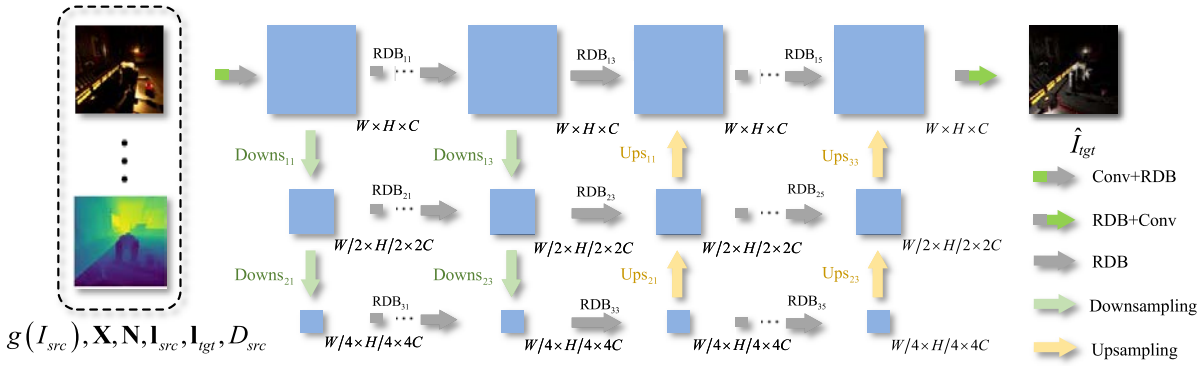


Fig. 4. Lighting Direction Transfer Network. Inspired by the physical image formation process, multiple physical attributes of the surface point are fed into the network to assist in pixel-level illumination modeling. A parallel multi-scale architecture is also adopted to maintain high-resolution representations through high-resolution RDBs and strengthen the representations with parallel low-resolution RDBs, which well transfers the local details under one illumination condition to that under another illumination condition.

In Eq. 6, we estimate the nonlinear mapping function $M_{T_{src}\to T_{tgt}}$ using a fully-connected network, whose inputs are the color temperature T_{src} of the source image and the color temperature T_{tgt} of the target image. Fig. 2 shows the network architecture. In contrast to the U-Net [36], we directly utilize fully connected layers to model the local mapping relation. Specifically, the color temperatures T_{src} and T_{tgt}, which are both represented as one-hot vectors, are concatenated together and fed into two fully-connected layers. The fully connected layers output a 128-dimensional feature vector and a 256-dimensional feature vector respectively. Each fully-connected layer is followed by a rectified linear activation function. The 256-dimensional feature vector is finally passed to the last fully-connected layer to output $M_{T_{src}\to T_{tgt}}$.

C. Lighting Direction Transfer

The lighting direction transfer network (LDT-Net) aims to re-render the intermediate relit image under the estimated lighting directions. As discussed in Sec. III, to maintain local shading and attached-shadow details in the image synthesis process, the neural network is designed with two important features. First, multiple physical attributes to guide: not only the depth map but also the position and normal of the surface point are fed into the network to assist in pixel-level illumination modeling. For the position map, we use the depth map and camera intrinsic matrix to calculate 3D spatial positions. We then use the plane fitting method to estimate the normal of each 3D point to obtain the normal map. See more details in the supplementary. Second, a parallel multi-scale architecture [37]–[39] to render: we propose to maintain high-resolution representations through high-resolution convolutions (modeling the local mapping relation) and strengthen the representations with parallel low-resolution convolutions (modeling the interactions between distant surface points). In contrast, existing relighting methods usually recover high-resolution representations from low-resolution representations outputted by an encoder, which breaks the pixel-level illumination modeling.

As a result, a parallel multi-scale architecture incorporating multiple physical attributes is designed to model the function f. Fig. 4 shows the network architecture. Specifically, the lighting direction transfer network mainly consists of parallel three-stream sub-networks. Each sub-network is composed of five residual dense blocks [40] (RDBs) and keeps the feature map size constant. Besides, the top sub-network includes one input convolutional layer at the start and one output convolutional layer at the end. The input convolutional layer is used to extract rich feature maps from the inputs, and feature maps are then fed to the subsequent RDB. On the contrary, the output convolutional layer generates the target relit image from the final feature maps. The information flow between sub-networks is implemented via several down-sampling blocks and up-sampling blocks. The down-sampling block reduces the feature map size by a factor of 2 and multiplies the feature map number by 2, while the up-sampling block performs the reverse operation. Note that the top sub-network ($RDB_{11}\cdots RDB_{15}$) can maintain the original resolution representations to model the pixel-level relationship between the input and the output without losing local information. See the appendix for more details of the network architectures including LE-Net, CTT-Net and LDT-Net.

D. Training the Model

1) Loss Functions: In our work, lighting estimation is seen as a classification task to obtain the direction and color temperature of the point light source. Therefore, we apply the cross-entropy loss function H to train the LE-Net:

$$L_c = H\left(p_{temp}, q_{temp}\right) + H\left(p_{dir}, q_{dir}\right), \tag{8}$$

where p_{temp} and p_{dir} are the true color temperature and lighting direction respectively, q_{temp} and q_{dir} are the predicted color temperature and lighting direction respectively.

For the CTT-Net, we adopt the $\mathcal{L}_1$ loss to train the network. For the LDT-Net, the $\mathcal{L}_1$ loss is also used to minimize the errors. In addition, inspired by [41], SSIM is utilized to make the network learn to produce visually pleasing images. Thus, the loss L_d for the LDT-Net is defined as:

$$L_d = \|\hat{I}_{tgt} - I\|_1 + \lambda(1 - \mathrm{SSIM}(\hat{I}_{tgt}, I)), \tag{9}$$

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

where $\hat{I}$ and I are the generated relit image and ground truth respectively. λ is set to 0.1 in our experiment. Note that we train these three networks separately. Theoretically, based on the pre-trained networks, we can continue to train the PNRNet in an end-to-end fashion, but we did not observe any performance improvement in our experiment.

2) Data Augmentation and Training Details: Data augmentation is a critical step for training neural networks to reduce over-fitting. To this end, we increase the diversity of the color temperatures by interpolating between the predefined color temperatures. We conduct our experiments using Pytorch on 2 NVIDIA Titan RTX GPUs. The parameters of the network are initialized with Kaiming Uniform Initialization [42]. We use Adam optimizer [43] with parameters: learning rate $= 5e-5$, betas $= (0.9, 0.999)$. Consequently, the batch size is set to be 16 to maximize GPU memory utilization.

V. Results and Analysis

A. Datasets and Evaluation Metric

Our method is trained on a novel Virtual Image Dataset for Illumination Transfer (VIDIT) dataset [44], which contains 390 different scenes and is split into three different mutually-exclusive sets: training set (300 scenes), validation set (45 scenes) and test set (45 scenes). Each scene is captured with 40 predetermined illumination settings, which is a combination of 5 color temperatures (2500K, 3500K, 4500K, 5500K, and 6500K) and 8 light directions (N, NE, E, SE, S, SW, W, NW). Note that the test set is kept private for the challenging benchmarking purpose, while the train and validation sets are made public for academic evaluation. We also qualitatively compare relighting results on the DiLiGenT-MV dataset [45] and evaluate the proposed color temperature transfer network on our own captured data.

The evaluation of lighting estimation, introduced by [28], is based on the prediction accuracy:

$$\underbrace{\frac{1}{N}\sum_{i=0}^{N-1}\left(\frac{\left|\hat{A}_i - A_i\right| \bmod 180}{180}\right)^2}_{\text{AngLoss}} + \underbrace{\frac{1}{N}\sum_{i=0}^{N-1}\left(\hat{T}_i - T_i\right)^2}_{\text{TempLoss}} \quad (10)$$

where $\hat{A}_i$ is the predicted angle (0-360) for test sample i and A_i is the corresponding ground-truth value. $\hat{T}_i$ is the temperature prediction for test sample i and T_i is the corresponding ground-truth value. T_i takes values equal to [0, 0.25, 0.5, 0.75, 1], which correspond to the color temperature values [2500K, 3500K, 4500K, 5500K, 6500K] respectively.

In addition to the common image quality metrics, such as PSNR and SSIM, the Mean Perceptual Score (MPS) [28], specially designed for any-to-any relighting performance evaluation, is also employed as follows:

$$MPS = 0.5 \cdot (S + (1 - L)) \quad (11)$$

where S is the SSIM score, and L is the LPIPS score. MPS is one of the most important quantitative metrics responsible for the human perception of digital image quality.

TABLE I
Effects of Relighting Decomposition and the Order

	MPS ↑	SSIM ↑	LPIPS ↓	PSNR ↑
Ours w/o decom.	0.6917	0.6806	0.2973	18.243
LDT-Net → CTT-Net	0.7063	0.6949	0.2823	18.521
Ours	**0.7090**	**0.6986**	**0.2806**	**18.593**

In order to independently evaluate the color temperature transfer performance, we define a cycle-consistent loss:

$$\frac{1}{N_T}\sum_{i=0}^{N_T}\text{Metric}\left(I_{src}, M_{T_i \to T_{src}}\left(M_{T_{srs} \to T_i}\left(I_{src}\right)\right)\right) \quad (12)$$

where N_T indicates the number of color temperatures.

B. Ablation Study

1) Any-to-Any Relighting Decomposition: To validate the effect of any-to-any relighting decomposition, we additionally train a single network for simultaneous relighting under novel color temperatures and lighting directions. Note that the single network is based on our lighting direction transfer network, and it also takes the estimated source and target temperature as extra inputs. Tab. I shows the quantitative comparison. The performance gap between this network (ours w/o decom.) and ours demonstrates that naively integrating the relighting of color temperatures and lighting directions into one single model without task decomposition or intermediate supervision does not work as well as our decomposition approach. In addition, we exchange the order of the color temperature transfer and the lighting direction transfer for comparison. As shown in Tab. I, the performance of the model "LDT-Net → CTT-Net" is slightly lower than that of our proposed model ("CTT-Net → LDT-Net"). Actually, compared to the lighting direction transfer task, the color temperature transfer task is much easier and has less impact on the subsequent task.

2) Effects of Physical Attributes: In order to prove the gain that the network obtains from integrating the physical attributes, we retrain the lighting direction transfer network without physical attributes **X** and **N**. As shown in Tab. II, all evaluation indicators have declined to varying degrees. In particular, the SSIM value is decreased by about 3.1%. We also retrain the LDT-Net without **X** or **N**. The quantitative results in Tab. II show that the LDT-Net benefits from each of the physical attributes. In fact, from the perspective of the physical imaging process, these physical attributes determine the appearance of the scene, especially the local illumination effects. For the normal map **N**, taking the first row in Fig. 5(g) as an example, we can see that ours or ours w/o **X** is able to generate visually plausible local shading and attached-shadow details. For the position map **X**, since the radiance of a point light source attenuates with distance, the incident radiance will vary with the position of the surface point, which accounts for near-field illumination. It demonstrates from the second row of Fig. 5(g) that ours or ours w/o **N** can produce darker details than ours w/o **X** and ours w/o **X**, **N**. Note that the light source is on the left of the image and the close-up details are located at the right of the image.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

TABLE II
EFFECTS OF PHYSICAL ATTRIBUTES

	MPS ↑	SSIM ↑	LPIPS ↓	PSNR ↑
Ours w/o $\mathbf{X}, \mathbf{N}$	0.6812	0.6702	0.3078	18.093
Ours w/o $\mathbf{X}$	0.6925	0.6795	0.2946	18.189
Ours w/o $\mathbf{N}$	0.7006	0.6911	0.2900	18.314
Ours	**0.7090**	**0.6986**	**0.2806**	**18.593**

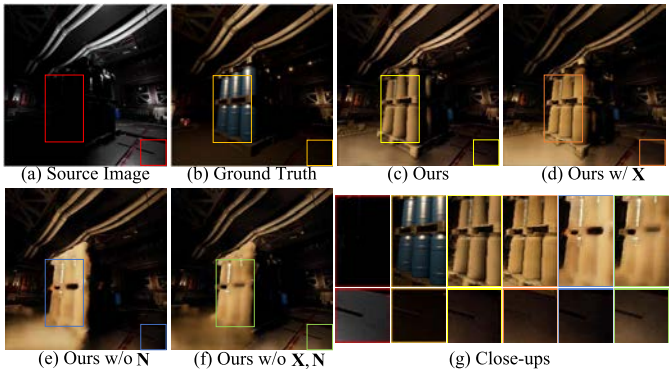


Fig. 5. Qualitative comparisons w.r.t. physical attributes.

3) Effects of Fusion Network: We also add another network, which is called fusion network (F-Net), to generate the final optimized relit image from the LDT-Net output and the CTT-Net output using a parallel structure or a serial structure. The difference between the parallel structure and the serial structure is that the LDT-Net input of the former is the source image, while the LDT-Net input of the latter is the output of the CTT-Net. Note that we fix the parameters of the three sub-networks (i.e., LE-Net, CTT-Net and LDT-Net) and only update the parameters of the fusion network. The fusion network is designed with 6 convolution layers and each is followed by a rectified linear activation function. Tab. III reports the results. It shows that ours w/ F-Net (Serial) achieves the best MPS result, which is a 0.37% improvement over ours. In fact, as shown in Fig. 7, the LDT-Net could result in a slight degradation of color temperature quality, and the F-Net (Serial) uses the color temperature information in the CTT-Net output to re-correct the color temperature of the LDT-Net output (namely our PNRNet result), which leads to better results. In contrast, the F-Net (Parallel) also needs to perform color temperature transfer on the LDT-Net output, but the color temperature gap between the LDT-Net output and the CTT-Net output is larger than that of the F-Net (Serial), which may lead to a slight decrease in performance. As shown in the close-up details in Fig. 6, compared to our PNRNet result (yellow box), both ours w/ Serial F-Net (green box) and ours w/ Parallel F-Net (blue box) achieve higher color temperature results, which is close to the ground truth. Finally, it is worth noting that the F-Net will introduce additional computational overhead and memory overhead.

4) Different Settings of LDT-Net: We also conduct ablation studies on different settings (i.e., different combinations of the height and width) of the LDT-Net. Note that the height represents the number of parallel sub-networks, and the width represents the number of blocks contained in the sub-network. Tab. IV reports the results. It shows that increasing the height and width of the LDT-Net leads to higher MPS.

TABLE III
EFFECTS OF FUSION NETWORK

	MPS ↑	SSIM ↑	LPIPS ↓	PSNR ↑
Ours	0.7090	0.6986	0.2806	18.593
Ours w/ F-Net (Parallel)	0.7113	0.7030	0.2804	**18.727**
Ours w/ F-Net (Serial)	**0.7116**	**0.7031**	**0.2800**	18.721

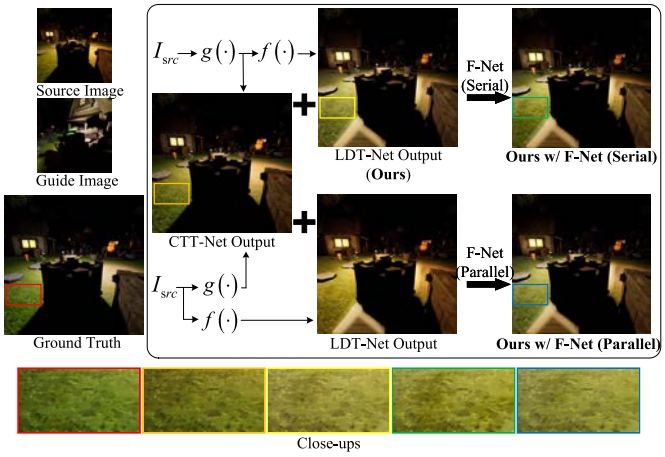


Fig. 6. Qualitative comparisons of the F-Net with different structures.

TABLE IV
RESULTS OF DIFFERENT SETTINGS OF LDT-NET ON THE NTIRE 2021 TRACK2 VALIDATION SET

Height	Width	MPS	SSIM	LPIPS	PSNR
1	2	0.7004	0.6905	0.2898	18.082
	4	0.7029	0.6922	0.2865	18.163
	6	0.7055	0.6949	0.2839	18.351
2	2	0.7037	0.6937	0.2864	18.344
	4	0.7061	0.6947	0.2825	18.502
	6	0.7087	0.6974	0.2800	18.561
3	2	0.7054	0.6949	0.2841	18.352
	4	0.7073	0.6966	0.2821	18.572
	6	0.7090	0.6986	0.2806	18.593

C. Comparative Study

We compare with prior works in three aspects, color temperature transfer, any-to-any relighting and model complexity. In addition to existing any-to-any relighting methods [10], [12], we also compare the deep portrait relighting method (DPR) [9] that does not explicitly reconstruct the surface reflectance. Since the DPR focuses on directional lighting which is represented as spherical harmonics, for a fair comparison, we modify the outputs of their original lighting estimation network to the combination of lighting directions and color temperatures in order to train the model on the VIDIT dataset, while all others remain unchanged. Owing to our unsuccessful attempt to reproduce the results of S3Net [11] as reported in their paper, we do not compare our method against theirs.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

Fig. 7. Qualitative comparison of color temperature transfer on the AIM 2020 track2 validation set. We show representative examples under five predefined color temperatures (2500K - 6500K). The zoom-in details indicate that existing end-to-end relighting models tend to produce undesired shadows and artifacts. In contrast, our CTT-Net can generate more accurate and shadow-free results.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

TABLE V
QUANTITATIVE COMPARISON OF COLOR TEMPERATURE TRANSFER ON THE AIM 2020 TRACK2 VALIDATION SET

Method	MAE ↓	PSNR ↑	MPS ↑	SSIM ↑
DPR[9]	0.0396	23.570	0.8659	0.8584
SA-AE[10]	0.0308	25.497	0.8873	0.8774
AMIDR-Net[12]	0.0342	24.630	0.9351	0.8702
CTT-Net + LDT-Net	0.0113	34.943	0.9659	0.9498
CTT-Net	**0.0022**	**52.039**	**0.9942**	**0.9885**

TABLE VI
QUANTITATIVE COMPARISON OF LIGHTING ESTIMATION ON THE AIM 2020 TRACK2 VALIDATION SET

Method	TotalLoss ↓	AngLoss ↓	TempLoss ↓
DPR[9]	0.1930	0.1625	0.0305
SA-AE[10]	0.1556	0.1195	0.0361
AMIDR-Net[12]	0.2402	0.1986	0.0416
Ours ($\mathcal{L}_1$ loss)	0.1319	0.0833	0.0486
Ours (Eq. 8)	**0.1083**	**0.0792**	**0.0292**

TABLE VII
QUANTITATIVE COMPARISON OF ANY-TO-ANY RELIGHTING ON THE NTIRE 2021 TRACK2 VALIDATION SET

Method	MPS ↑	SSIM ↑	LPIPS ↓	PSNR ↑
DPR[9]	0.6697	0.6557	0.3163	17.586
SA-AE[10]	0.6814	0.6699	0.3071	18.209
AMIDR-Net[12]	0.6779	0.6940	0.3381	**19.830**
Ours	**0.7090**	**0.6986**	**0.2806**	18.593

1) Color Temperature Transfer: We compared the performance of color temperature transfer with previous relighting methods on the AIM 2020 track2 validation set [28]. The AIM 2020 track2 validation set contains 45 images. Each image provides corresponding illumination setting ground-truth which can be fed into our CTT-Net for independent performance evaluation without introducing errors caused by inaccurate lighting estimation. Following the calculation steps of Eq. 12, we transform each image from the original temperature to five predefined color temperatures, and then transform them backward to compute the differences between the input image and the predicted results. The metrics in Eq. 12 are chosen as MAE, PSNR, MPS and SSIM respectively. Tab.V shows the quantitative results. It can be seen that both CTT-Net and CTT-Net + LDT-Net outperform previous methods by a large margin. In fact, CTT-Net + LDT-Net reduces the performance of color temperature transfer, which implies interference between two sub-tasks. Fig. 7 shows the qualitative results. Compared to our method, other relighting methods [9], [10], [12] often introduce unnecessary shadows and artifacts. For example, the insets of Fig. 7(b) show that the shadows are removed incorrectly by [12] and [9]. The highlights on the blue buckets in Fig. 7(c) are either suppressed or saturated by [10], [12] and [9]. The insets of Fig. 7(d) show that the shading details are also not well preserved in their results. In contrast, our CTT-Net does not alter the original shading and shadow details which are only caused by the changes in lighting direction. It implies that the decomposition of any-to-any relighting can effectively avoid the task-aliasing effects. In fact, the key to solving the task-aliasing effect is to recognize deep features from different tasks. Compared to existing methods where different tasks share common deep features, the idea of decomposition makes the deep features inherently distinguishable. In other words, it makes the deep features from different tasks independent of each other and not interfere with each other. At the same time, our CTT-Net is able to recover images with more accurate color temperature and sharp details. This suggests that simply modeling the pixel-to-pixel mapping relationship with fully-connected layers is sufficient for the color temperature transfer task. It can be confirmed from Tab.V that the performance of our method is much better than that of these relighting methods [10], [12] based on the multi-scale modeling with a U-Net structure.

2) Any-to-Any Relighting: We also compare the performance of any-to-any relighting with prior works on the NTIRE 2021 track2 validation set [29]. The NTIRE 2021 track2 validation set contains 90 input image and guide image pairs. Tab.VII shows the quantitative results. We achieve the best MPS over all other methods due to the preservation of local fine details. AMIDR-Net achieves the highest PSNR value due to the ensemble of multiple models, but it, in turn, produces blurry results. Considering the AIM 2020 track2 validation set provides the illumination setting ground-truth, we also evaluate the performance of lighting estimation on it. The comparative results are reported in Tab.VI. Our LE-Net achieves the lowest loss in both lighting direction prediction and color temperature prediction, which also contributes to the improvement of any-to-any relighting performance. We also train the LE-Net using the $\mathcal{L}_1$ loss function. It shows that the $\mathcal{L}_1$ loss function achieves comparable results to the cross-entropy loss function in the light direction estimation accuracy, but its color temperature estimation accuracy is slightly worse. Fig. 8 shows the qualitative results of any-to-any relighting. Fig. 8(a) shows that if the source image and the guide image share the same illumination setting, the relit image produced by our method is almost the same as the source image. On the contrary, DPR introduces shadows and AMIDR-Net generates blurry results with a lower color temperature. The insets of Fig. 8(b)(c) indicate that our method produces visually accurate shading and shadow details, which are consistent with the source geometry structure. It implies that our LDT-Net, which incorporates multiple physical attributes, is capable of generating local details. We also note that the guide image in Fig. 8(f) is almost completely dark, and only a little information in the upper left corner area can be utilized to infer the illumination setting. The corresponding insets indicate that our method can accurately estimate the illumination setting and produce the results without breaking the shading variation. We also provide qualitative relit results on the NTIRE 2021 track2 test set in Fig. 9. Note that the test set does not contain ground truth.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

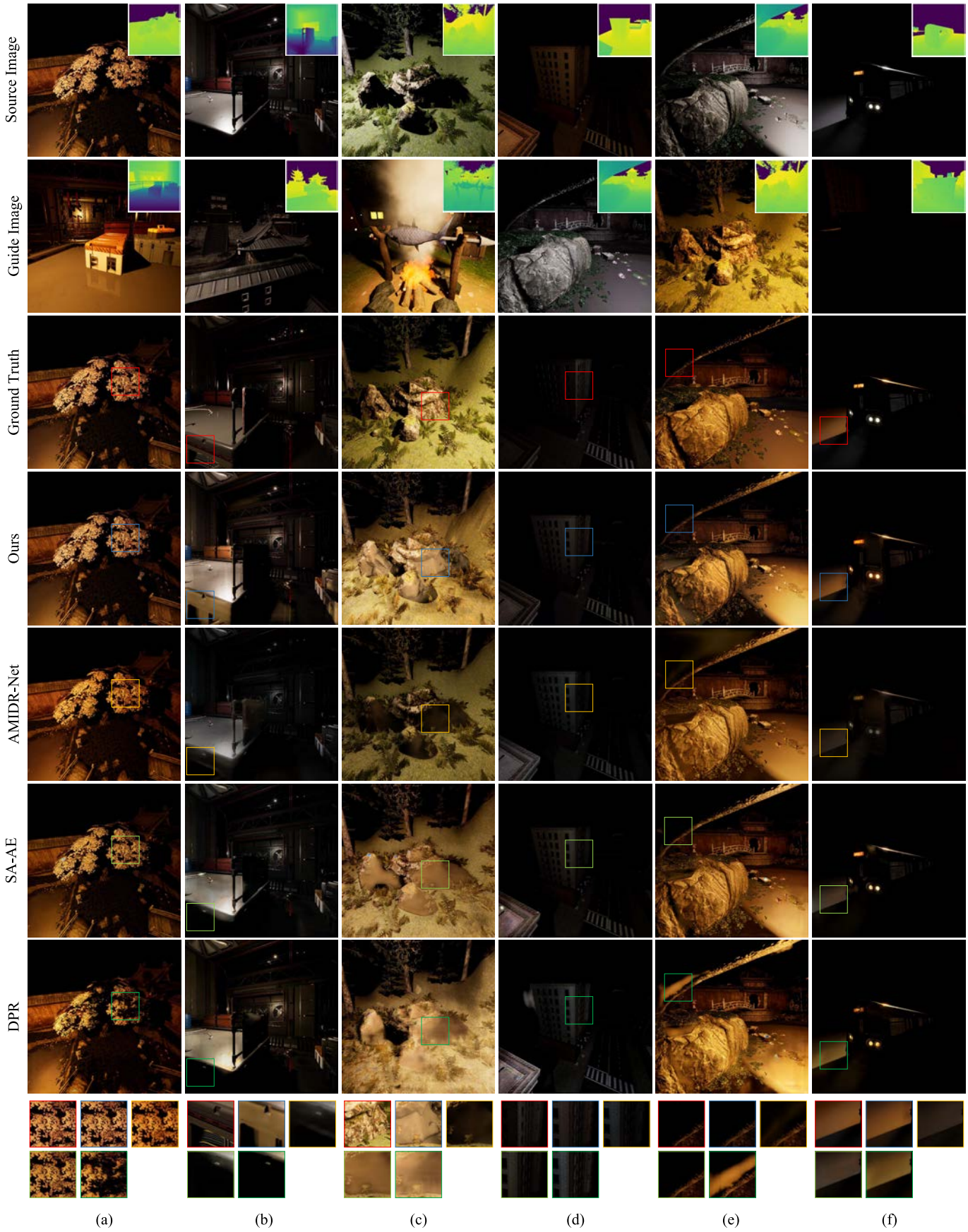


Fig. 8. Qualitative comparison of any-to-any relighting on the NTIRE 2021 track2 validation set. We show representative examples with zoom-in details focusing on color temperatures (a)(f), shading variation (c)(d), and shadows (b)(e). Note that our method outperforms all other approaches with sharper and more accurate relit results.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

Fig. 9. Qualitative comparison of any-to-any relighting on the NTIRE 2021 track2 test set. Note that our method is able to well preserve the shading and attached-shadow details, while other methods suffer from the artifacts.

TABLE VIII
COMPARISON IN TERMS OF MODEL COMPLEXITY AND INFERENCE TIME

Method	DPR[9]	SA-AE[10]	AMIDR-Net[12]	Ours
# Parameters (M)	**2.4**	20.3	190.8	17.0
Inference time (S)	**0.886**	0.990	1.383	0.960

3) Model Complexity and Inference Time: Tab.VIII reports the comparison in terms of model complexity and inference time among different methods. Note that all methods are tested on the NTIRE 2021 track2 validation set with a single RTX Titan GPU. The average inference time for a single image with a resolution of 1024*1024 is reported. Although the DPR with the fewest network parameters can handle portrait relighting well, it cannot be easily extended to challenging scene relighting. The AMIDR-Net achieves the highest PSNR due to the ensemble of three multiple models but accordingly leads to the most network parameters. In contrast, the SA-AE and our method have almost the same amount of network parameters, but our method achieves the best visual effects among all the methods.

D. Generalization to Real Data

We further use real images to demonstrate the extensibility of our method to real data in Fig. 10 and Fig. 11. As shown in Fig. 10, the input real image is captured using an iPhone 8 plus cellphone at 1:00 PM, and we manually set its depth map to 0. The image is first identified by the LE-Net as a color temperature of 6500K and then fed into the CTT-Net to generate the images under other predefined color temperatures (2500K, 3500K, 4500K and 5500K). We also interpolate color mapping functions (see the appendix for implementation details) to generate the images under user-defined color temperatures (3000K and 5000K), while existing end-to-end relighting models cannot handle this case. Compared to the reference images, it is obvious that our method produces more accurate relighting outputs under novel color temperatures.

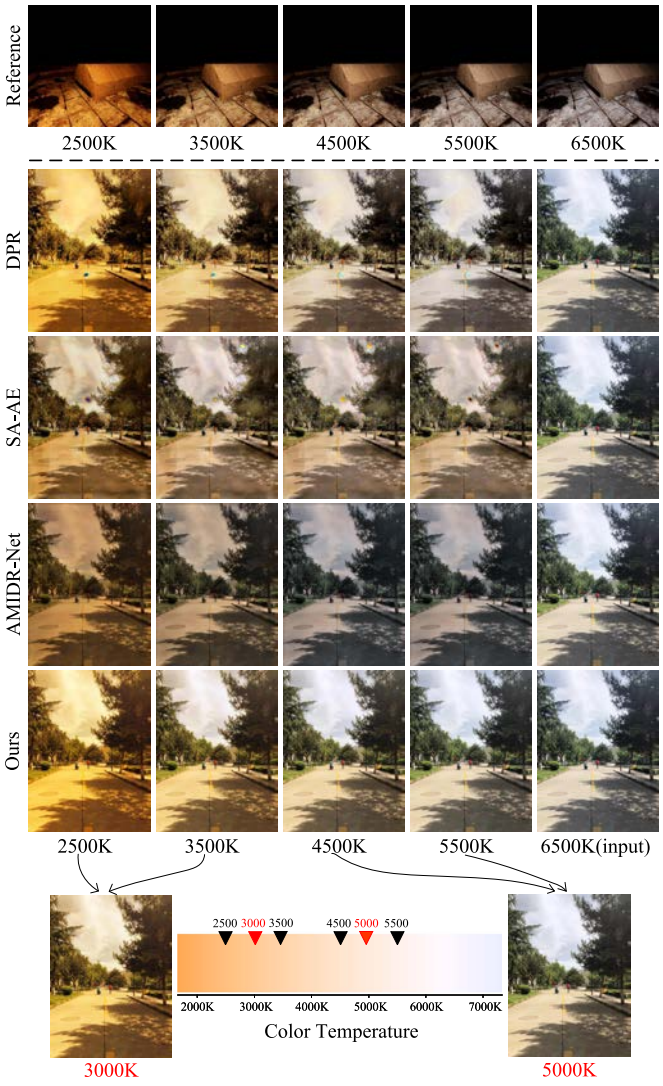


Fig. 10. Results of the CTT-Net on the real data.

We also provide qualitative comparison using two examples in the DiLiGenT-MV dataset [45]. Note that the positions of nearby point light sources are in front of the object, which are not exactly the same as the positions of our point light sources. Therefore, we select two images whose light source positions are approximately in the south as the source images. As shown in Fig. 11, compared with other methods, our method is able to generate plausible shading and attached-shadow details which are consistent with the target lighting directions.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

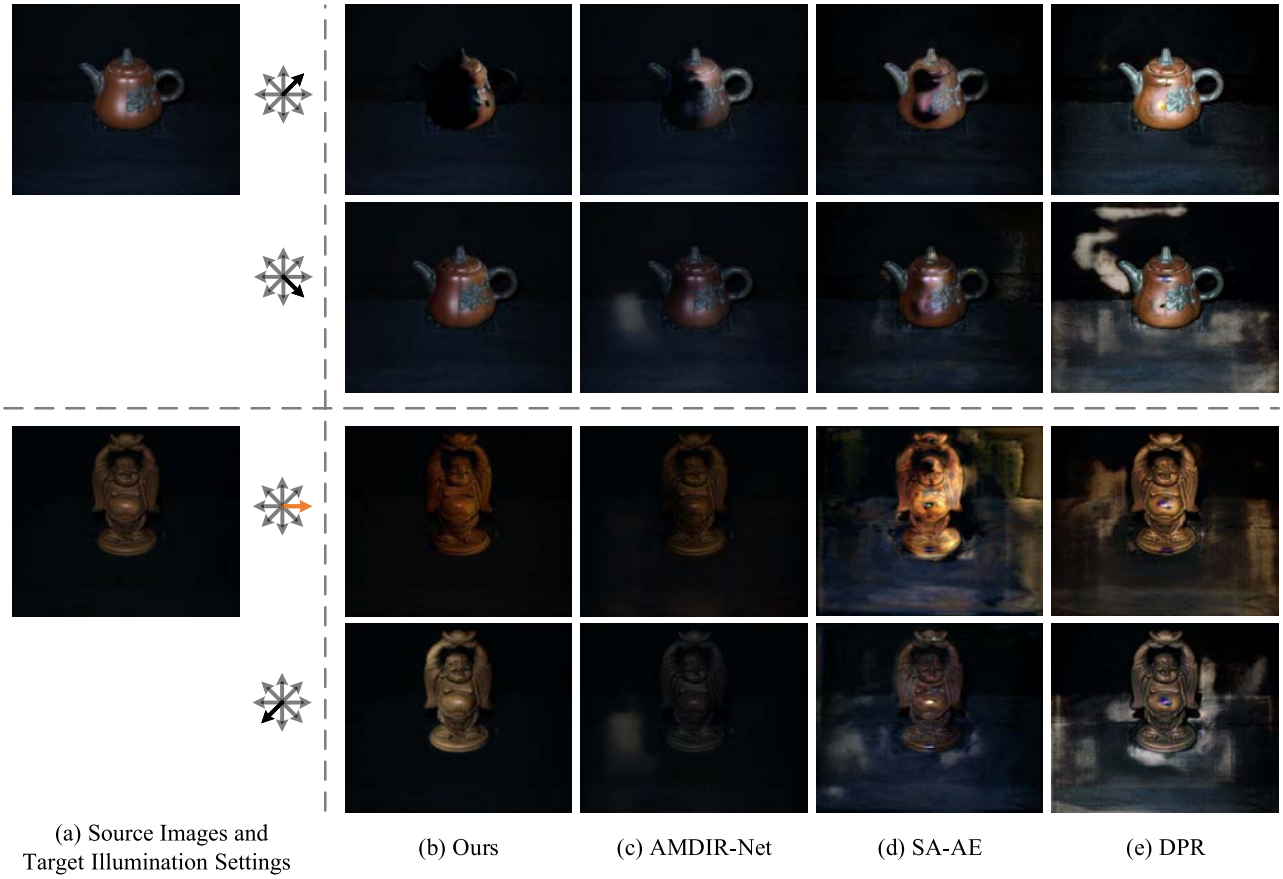


Fig. 11. Qualitative comparison of relighting results on the DiLiGenT-MV dataset consisting of real images. Our method is able to generate plausible shading and attached-shadow details which are consistent with the target lighting directions.

VI. Conclusion

In this paper, we present a novel network architecture for relighting a source RGB-D image with the illumination setting provided by a guide RGB-D image. Based on the relighting formulation, we propose to decompose the task into three simpler sub-tasks, which are lighting estimation, color temperature transfer and lighting direction transfer, and train corresponding sub-network separately. Inspired by the physical image formation process, we propose to solve lighting direction transfer with a parallel multi-scale network that incorporates multiple physical attributes to model the local illumination without missing the fine details. A simple yet effective fully-connected neural network is designed to estimate the non-linear color mapping function to transfer images from one color temperature to another color temperature in a pixel-level manner. Extensive experiments show that the proposed decomposition solution can produce relit images with better local shading and attached-shadow details than prior works.

This work also has a few limitations that can be the subject of future work. First, our model can be extended to handle relighting of multiple near-field point light sources by redesigning the lighting estimation network and collecting more training data under multiple light sources. Second, we just consider Lambertain BRDFs. It would be interesting to extend our model to handle objects with specular reflection by introducing specular BRDFs [46]. Finally, the regions relit from hard-shadow regions can only recover the shading details without generating texture details. A flow vector [47] can be learned to select similar textures from adjacent regions to fill in the regions.

References

[1] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 2000, pp. 145–156.

[2] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi, "Deep image-based relighting from optimal sparse samples," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–13, Aug. 2018.

[3] A. Meka *et al.*, "Deep reflectance fields: High-quality facial reflectance field inference from color gradient illumination," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Aug. 2019.

[4] Y. Kanamori and Y. Endo, "Relighting humans: Occlusion-aware inverse rendering for full-body human images," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–11, Dec. 2018.

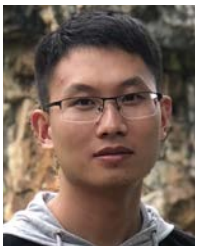
[5] S. Sang and M. Chandraker, "Single-shot neural relighting and SVBRDF estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 85–101.

[6] Y. Yu, A. Meka, M. Elgharib, H.-P. Seidel, C. Theobalt, and W. A. Smith, "Self-supervised outdoor scene relighting," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 84–101.

[7] D. C. Knill and W. Richards, *Perception as Bayesian Inference. Chapter The Perception of Shading and Reflectance*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

[8] T. Sun *et al.*, "Single image portrait relighting," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Aug. 2019..
[9] H. Zhou, S. Hadap, K. Sunkavalli, and D. Jacobs, "Deep single-image portrait relighting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7194–7202.
[10] Z. Hu, X. Huang, Y. Li, and Q. Wang, "SA-AE for any-to-any relighting," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*. Cham, Switzerland: Springer, 2020, pp. 535–549.
[11] H.-H. Yang, W.-T. Chen, and S.-Y. Kuo, "S3Net: A single stream structure for depth guided image relighting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 276–283.
[12] A. Yazdani, T. Guo, and V. Monga, "Physically inspired dense fusion networks for relighting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 497–506.
[13] W. Matusik, M. Loper, and H. Pfister, "Progressively-refined reflectance functions from natural illumination," *Rendering Techn.*, vol. 1, no. 2, pp. 299–308, 2004.
[14] P. Peers *et al.*, "Compressive light transport sensing," *ACM Trans. Graph.*, vol. 28, no. 1, pp. 1–18, Jan. 2009.
[15] D. Reddy, R. Ramamoorthi, and B. Curless, "Frequency-space decomposition and acquisition of light transport under spatially varying illumination," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 596–610.
[16] J. T. Barron and J. Malik, "Shape, albedo, and illumination from a single image of an unknown object," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 334–341.
[17] J. T. Barron and J. Malik, "Color constancy, intrinsic images, and shape estimation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 57–70.
[18] J. T. Barron and J. Malik, "Intrinsic scene properties from a single RGB-D image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 17–24.
[19] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1670–1687, Aug. 2015.
[20] K. Karsch *et al.*, "Automatic scene inference for 3D object compositing," *ACM Trans. Graph.*, vol. 33, no. 3, pp. 1–15, 2014.
[21] S. Duchêne *et al.*, "Multi-view intrinsic images of outdoors scenes with an application to relighting," *ACM Trans. Graph.*, vol. 34, no. 5, pp. 1–16, Oct. 2015, doi: 10.1145/2756549.
[22] S. Sengupta, J. Gu, K. Kim, G. Liu, D. Jacobs, and J. Kautz, "Neural inverse rendering of an indoor scene from a single image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8598–8607.
[23] Y. Yu and W. A. P. Smith, "InverseRenderNet: Learning single image inverse rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3155–3164.
[24] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2475–2484.
[25] L.-W. Wang, W.-C. Siu, Z.-S. Liu, C.-T. Li, and D. Lun, "Deep relighting networks for image light source manipulation," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2020, pp. 550–567.
[26] T. Nestmeyer, J.-F. Lalonde, I. Matthews, and A. Lehrmann, "Learning physics-guided face relighting under directional light," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5124–5133.
[27] S. Bi *et al.*, "Deep reflectance volumes: Relightable reconstructions from multi-view photometric images," 2020, *arXiv:2007.09892*.
[28] M. E. Helou *et al.*, "AIM 2020: Scene relighting and illumination estimation challenge," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2020, pp. 499–518.
[29] M. E. Helou *et al.*, "NTIRE 2021 depth guided image relighting challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 566–577.
[30] B. Li, Y. Gou, J. Z. Liu, H. Zhu, J. T. Zhou, and X. Peng, "Zero-shot image dehazing," *IEEE Trans. Image Process.*, vol. 29, pp. 8457–8466, 2020.
[31] B. Li, Y. Gou, S. Gu, J. Z. Liu, J. T. Zhou, and X. Peng, "You only look yourself: Unsupervised and untrained single image dehazing neural network," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1754–1767, May 2021.
[32] J. T. Kajiya, "The rendering equation," in *Proc. 13th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 1986, pp. 143–150.
[33] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*.
[34] G. D. Finlayson, M. Mackiewicz, and A. Hurlbert, "Color correction using root-polynomial regression," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1460–1470, May 2015.
[35] M. Afifi, A. Punnappurath, A. Abdelhamed, H. C. Karaimer, A. Abuolaim, and M. S. Brown, "Color temperature tuning: Allowing accurate post-capture white-balance editing," in *Proc. Color Imag. Conf.* Springfield, VI, USA: Soc. Imag. Sci. Technol., 2019, pp. 1–6.
[36] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[37] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.
[38] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7314–7323.
[39] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[40] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.
[41] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[44] M. El Helou, R. Zhou, J. Barthas, and S. Süsstrunk, "VIDIT: Virtual image dataset for illumination transfer," 2020, *arXiv:2005.05460*.
[45] M. Li, Z. Zhou, Z. Wu, B. Shi, C. Diao, and P. Tan, "Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials," *IEEE Trans. Image Process.*, vol. 29, pp. 4159–4173, 2020.
[46] M. Ashikmin, S. Premože, and P. Shirley, "A microfacet-based BRDF generator," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 2000, pp. 65–74.
[47] P. P. Srinivasan, R. Tucker, J. T. Barron, R. Ramamoorthi, R. Ng, and N. Snavely, "Pushing the boundaries of view extrapolation with multiplane images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 175–184.



Zhongyun Hu received the B.E. degree in electronic and information engineering from Changzhou University in 2016 and the master's degree in circuits and systems from Northwestern Polytechnical University in 2019, where he is currently pursuing the Ph.D. degree with the School of Computer Science. His research interests include computer vision and computer graphics, with particular interests in relighting, light estimation, and image harmonization.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.

Ntumba Elie Nsampi received the B.E. degree in software engineering from the School of Mathematics and Computer Science, Zhejiang Normal University, China, in 2019, and the M.E. degree in computer science and technology from Northwestern Polytechnical University, China, in 2022. His research interests include computer vision and computer graphics.

Xue Wang received the B.S. and Ph.D. degrees from Northwestern Polytechnical University in 2007 and 2017, respectively. From 2012 to 2014, she studied at the University of Pennsylvania as a Visiting Ph.D. Student financed by the China Scholarship Council. She is currently an Associate Research Fellow with the School of Computer Science, Northwestern Polytechnical University. She focuses on building machines that understand the social signals and events that multiview/light field videos portray. Her research interests include computer vision, computational photography, and machine learning.

Qing Wang (Senior Member, IEEE) graduated from the Department of Mathematics, Peking University, in 1991. He received the master's and Ph.D. degrees from the Department of Computer Science, Northwestern Polytechnical University, in 1997 and 2000, respectively. Then, he joined Northwestern Polytechnical University, where he is currently a Professor with the School of Computer Science. He worked as a Research Scientist at the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, from 1999 to 2002. He also worked as a Visiting Scholar at the School of Information Engineering, The University of Sydney, Australia, in 2003 and 2004. In 2009 and 2012, he visited the Human Computer Interaction Institute, Carnegie Mellon University, for six months, and the Department of Computer Science, University of Delaware, for one month. He has published more than 100 papers in the international journals and conferences. His research interests include computer vision and computational photography, such as 3D vision, light field imaging and processing, and novel view synthesis. He is a member of ACM. In 2006, he was awarded as the Outstanding Talent Program of New Century by Ministry of Education, China.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on June 11,2022 at 00:01:56 UTC from IEEE Xplore. Restrictions apply.