# SIDNet: Learning Shading-aware Illumination Descriptor for Image Harmonization

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

***Abstract*—Image harmonization aims at adjusting the appearance of the foreground to make it more compatible with the background. Without exploring background illumination and its effects on the foreground elements, existing works are incapable of generating a realistic foreground shading. In this paper, we decompose the image harmonization task into two sub-problems: 1) illumination estimation of the background image and 2) re-rendering of foreground objects under background illumination. Before solving these two sub-problems, we first learn a shading-aware illumination descriptor via a well-designed neural rendering framework, of which the key is a shading bases module that generates multiple shading bases from the foreground image. Then we design a background illumination estimation module to extract the illumination descriptor from the background. Finally, the Shading-aware Illumination Descriptor is used in conjunction with the neural rendering framework (SIDNet) to produce the harmonized foreground image containing a novel harmonized shading. Moreover, we construct a photo-realistic synthetic image harmonization dataset that contains numerous shading variations with image-based lighting. Extensive experiments on both synthetic and real data demonstrate the superiority of the proposed method, especially in dealing with foreground shadings.**

***Index Terms*—Image Harmonization, Illumination, Shading Field, Neural Rendering.**

## I. Introduction

GIVEN a composite image of which the foreground and background taken from different images, image harmonization aims to adjust the appearance of the foreground to make it compatible with the background. A lot of works [1]–[8] have been proposed to solve the inharmony problem in the composite image. As shown in Fig. 1, these image harmonization methods, however, tend to focus on adjusting the low-level statistics (i.e., color and brightness) of the foreground rather than its shading.

The failure to model shadings can be attributed to the lack of a comprehensive understanding of the background illumination and its effects on the foreground elements, especially the direction and distribution of illumination cannot be perceived by these methods. In particular, learning-based approaches, such as [1], [2], [6], [8], are generally formulated as the image-to-image translation task where the illumination is implicitly transferred from the background to the foreground. Moreover, existing large-scale image harmonization datasets [1], [6] are devoid of perceivable shading variations. It is questionable whether or not the networks trained with these datasets could deal with shading variations.

Collecting, processing, and distributing real-world datasets is often associated with data gathering costs, quality problems, and privacy concerns. Recently, researchers have turned to synthetic data as an adequate solution in the face of the numerous data-related challenges posed by real-world data and the requirements of recent AI technologies [1]. Several color transfer algorithms [9]–[12] have been used to generate visually acceptable images with varying colors and brightness. However, the shading variation, which is also important for image harmonization, has not been taken into consideration. To this end, we construct a large-scale photo-realistic image harmonization dataset that contains color, brightness and shading variations with image-based lighting [13]. Unlike existing synthetic datasets [7], [14] of which the foreground objects or the illumination maps are created by CG software, we refer to real models/illumination captured from the real world, with the aim of achieving photo-realistic renderings.

In this paper, we propose to decompose the image harmonization task into two sub-problems: (1) illumination estimation of the background image, and (2) re-rendering of foreground objects under background illumination. The general lighting representation (i.e. illumination maps [15]) that is able to record the complete illumination (including the directional information) can be used to solve sub-problem (1). However, using such a representation brings considerable challenges due to its large number of parameters. Our key to solving (1) lies in the proposal of an efficient illumination representation with fewer parameters that also can retain directional information. For sub-problem (2), in contrast to the spherical harmonics lighting model [16], we intend to render complex global illumination effects such as cast shadows to further improve the realism of the composite image.

To achieve these ends, we propose a novel Neural Rendering Framework that accounts for global illumination effects while learning a shading-aware illumination descriptor from the illumination maps. Its key component is a neural Shading Bases Module, which is utilized to generate multiple shading bases from the foreground image. Each shading basis corresponds to a specific illumination distribution. It then combines with the illumination descriptor, which is encoded by an Illumination Encoder Module, to render a shading image. We propose to reconstruct the shading image as a pretext task in order to simultaneously supervise the learning of both the shading bases and the illumination descriptor. Note that the GT shading image is automatically generated by a path tracing algorithm, and also contains global illumination effects.

Z. Hu, N. Nsampi, X. Wang and Q. Wang (corresponding author) are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (E-mail: qwang@nwpu.edu.cn).

Project website: https://waldenlakes.github.io/IllumHarmony/

Z. Hu and N. Nsampi contributed equally to this work.

Manuscript received April 19, 2021; revised August 16, 2021.


Fig. 1. Given a composite image (a) of which the foreground and background taken from different images, our proposed method is able to produce the harmonized image (e) containing a novel foreground shading (f) that conforms to the background illumination. In contrast, existing methods [1], [2] only adjust the brightness and color of the foreground. Here, for both methods, we use their publicly released pre-trained models.

Once we pre-define the shading-aware illumination descriptor, the illumination of the background image could be estimated via the proposed Background Illumination Estimation Module and then is used in conjunction with our Neural Rendering Framework to generate the harmonized foreground image which contains a harmonized shading. We name this novel Shading-aware Illumination Descriptor-based image harmonization Network SIDNet.

Our contributions can be summarized as follows:

(1) We propose a neural Shading Bases Module, which decomposes the shading field into multiple shading components, to generate a novel foreground shading using the estimated illumination descriptor. To the best of our knowledge, this is the first of its kind to explicitly model shadings in image harmonization.

(2) We design a novel Neural Rendering Framework to learn the shading-aware illumination descriptor from the illumination maps in a self-supervised manner.

(3) We provide a large-scale photo-realistic synthesized image harmonization dataset containing challenging shading variations.

## II. Related Work

In this section, we briefly discuss existing works related to image harmonization. In addition, image relighting methods relevant to the proposed work are also included.

### A. Image Harmonization

Traditional image harmonization methods [3], [4], [9], [11], [17]–[21] focus on matching low-level statistics between images. The pioneer work [9] matched the means and variances of the color histograms between images in a decorrelated color space. Lalonde and Efros [3] then combined global color statistics obtained over a large natural image set and local color statistics to improve the realism of the composite images. Sunkavalli et al. [21] proposed to match contrast, texture, noise, and blur of visual appearance using multi-scale pyramid representations to produce realistic composites. Xue et al. [4] identified key statistical measures that most affected the realism of a composite image. However, the adjustment of low-level statistics can not handle shading variations.

In the past few years, researchers have concentrated on deep neural network-based approaches for image harmonization [1], [2], [5], [6], [8], [22]–[25]. Zhu et al. [5] proposed a CNN-based classifier model for the perception of realism to guide a traditional color adjustment method to produce more realistic outputs. The first end-to-end CNN model for image harmonization was proposed by Tsai et al. [6]. These methods usually formulate image harmonization as an image-to-image translation task by ensuring visual consistency between the foreground and the background in different aspects, such as the domain consistency [1], [26], the visual style consistency [2], [27], and the reflectance/illumination consistency [8], [14]. Furthermore, the attention [28], [29] or self-attention [30] mechanism is applied to improve the realism of the composite images. However, without considering the physical principles of image formation, learning-based methods lack the perception of illumination information in the background image, especially the direction of illumination. This inevitably leads to their inability to generate a realistic foreground shading, which severely degrades the realism of the composite images.

### B. Image Relighting

Traditional image-based relighting methods [31]–[33] directly reconstruct the light transport function to relight the objects using multiple images under different illumination conditions. Note that the illumination here is always explicitly provided. Recently, several deep neural networks with illumination estimation modules [34]–[39] are proposed to relight objects of a specific class (e.g., portraits and human bodies) using a single RGB image. Still, the illumination estimation is only considered for specific objects rather than the natural scenes. Given multi-view images, Yu et al. [40] proposed the first single image-based outdoor scene relighting method along with lighting estimation for the scene. They used spherical harmonics lighting model [16] to generate the shading. However, it could not handle global illumination. Moreover, it is worth noting that although these relighting methods [34]–[40] with illumination estimation can be applied

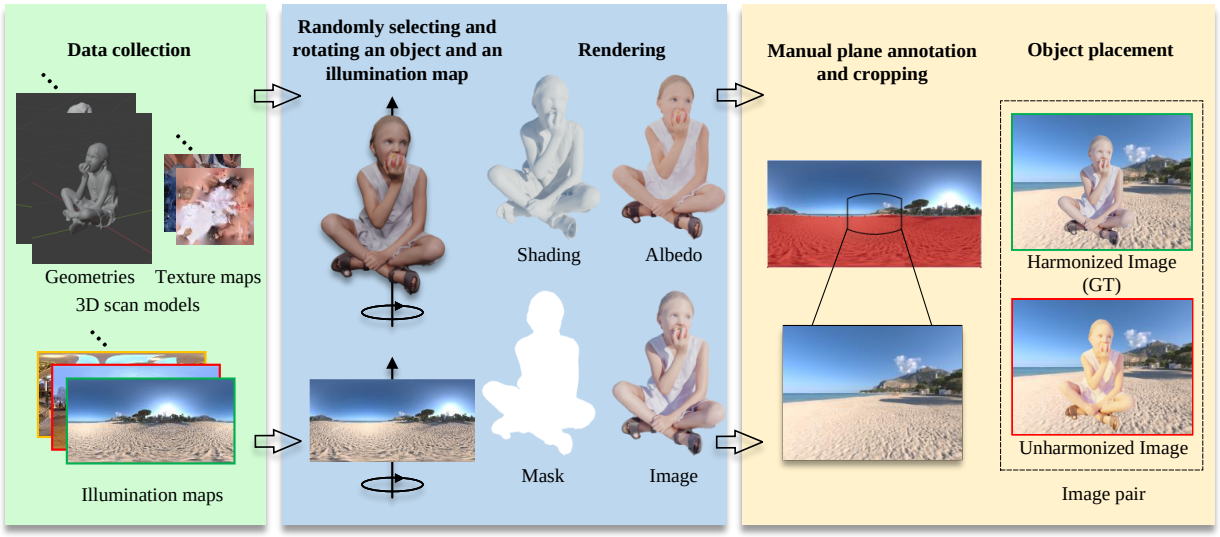


Fig. 2. The pipeline of constructing the pair of harmonized image and unharmonized image. It mainly covers data collection, rendering, and object placement. To train the proposed model, shading and albedo are also rendered.

to image harmonization, additional computational overhead would be also introduced, since illumination estimation for the background image is often accompanied by meanwhile estimating other physical attributes in the background image. In other words, these relighting methods are not specifically designed for image harmonization.

## III. Image Harmonization Dataset

The goal of our image harmonization dataset is to introduce a new challenging benchmark with photo-realistic synthesized images and plentiful variations (color, brightness, and shading) to the community of image harmonization. The pipeline of constructing the pair of harmonized image (ground truth) and unharmonized image is illustrated in Fig. 2. Below we introduce the construction process, which covers data collection, rendering and object placement.

### A. Data Collection

To construct our dataset, we collect both high-quality 3D human models and high dynamic range (HDR) illumination maps. The collected 3D models are acquired from [41] using photogrammetric 3D scanning methods. A rich variety of humans are included, with the diversity across genders (male, female), ages, poses, and clothing (colors, accessories). We collect a total of 138 high-quality 3D humans, of which 120 are used for training and 18 for testing.

Our illumination maps are collected from the internet source Poly Haven [42] and HDR MAPS [43], which offer diverse high dynamic range panoramic images. Fig. 3 shows the t-SNE visualization of our illumination maps. We mainly select outdoor illumination maps, resulting in a total of 318 high dynamic range panoramic images. In order to generate images with different kinds of variations (color, brightness, and shading), we ensure that the selected illumination maps are diverse across weather conditions (sunny, cloudy, and overcast), illuminant colors, time of the day, and locations.

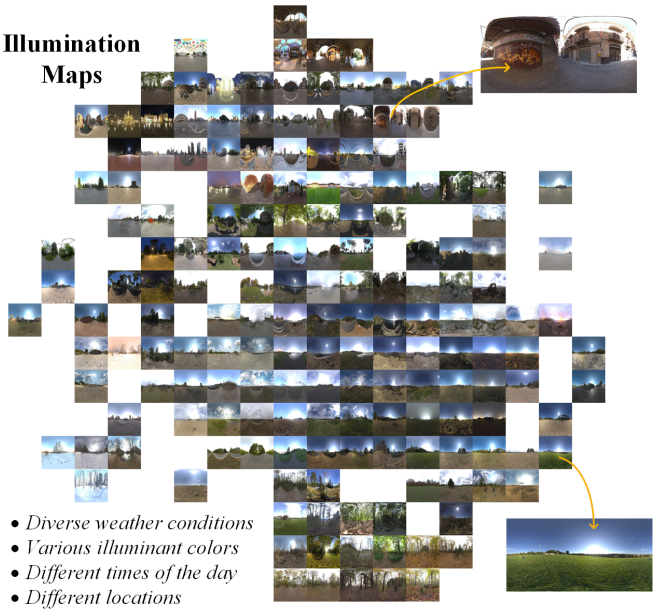


Fig. 3. The t-SNE visualization of our collected illumination maps, which contain diverse weather conditions, illuminant colors, time of the day, and locations.

From the 318 images, 191 images are used for training and the remaining 127 ones are reserved for testing. All the selected illumination maps come with the resolution of 8k, which are resized to 2k before rendering.

### B. Rendering

To generate training and test images, we use Blender [44] with Cycle Renderer. Each object is first placed on a planar surface within Blender's environment. We then randomly sample (without replacement) half of the images as illumination maps. For each possible pair of object and illumination map,

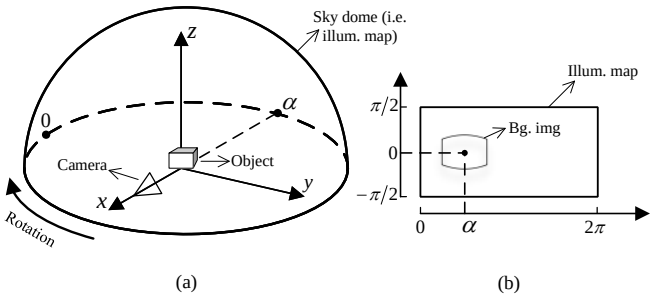


Fig. 4. The Blender rendering system (a) and the cropping horizontal center $\alpha$ of the illumination map for the background image (b).

we randomly sample 4 rotation angles from a pre-defined set of 8 angles, ranging from 0 to 360 degrees with an increment of 45 degrees. The sampled angles are used to rotate both the object and the illumination map, resulting in a total of 16 path-traced images per object. This process increases the richness of object poses and provides sufficient shading variations for our model to learn. Specifically, as shown in Fig. 4(a), a Blender rendering system includes three parts: a sky dome, a camera, and an object. We randomly select one illumination map as the sky dome. In order to generate various shading variations on the foreground object, we need to first rotate the illumination map by a randomly sampled angle, and then render a foreground image. Assume that the horizontal coordinate of the rotated illumination map pointed by the camera is $\alpha$ at this time. As shown in Fig. 4(b), the illumination map must be cropped with $\alpha$ as the horizontal center to obtain the background image to ensure the illumination consistency between the rendered foreground image and the background image. In the next subsection, this illumination consistency enables the foreground image to be placed within the background image.

For each object, we generate path-traced images, shading images, albedo images, and foreground masks, which are all rendered with 480 $\times$ 640 resolution. About 200$\sim$300 samples per pixel are used for generating path-traced images.

### C. Object Placement

The location of an object within an image conveys important clues for image harmonization. Here, object placement and tuple building for training and test sets will be elaborated. We assume that all objects are placed on planar surfaces. To distinguish between planar and non-planar surfaces within an image, we manually annotate the planar surface in the illumination map from which the background image will be extracted.

For a given background image extracted from an illumination map using a virtual perspective camera, we randomly select a pixel belonging to the annotated planar surface. Note that we discard the background images that do not contain the annotated plane surface. We then crop the rendered object as the foreground image and select one image corner as the reference point. Finally, we randomly resize the cropped object and compose it with the background image, so that the reference point uses the randomly selected pixel.

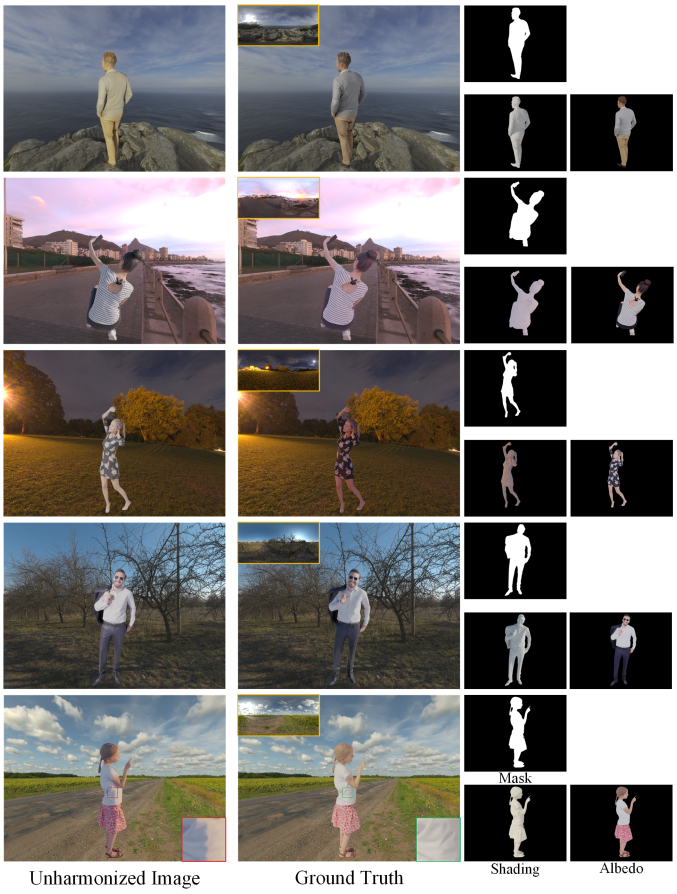


Fig. 5. High-quality examples from our constructed dataset. Red and green insets in the bottom row indicate that our dataset contains challenging *shading variations*.

To create the training and test tuple, we first select a rendered image of an object and its corresponding illumination map. We then rotate the illumination map based on the angle used for rendering and extract an image crop (background image) using a virtual perspective camera. Here, a standard gamma tone mapping ($\gamma = 2.2$ ) is also applied to the illumination map before extracting the background image. Lastly, we perform object placement and compose foreground/background images as described above. The same procedure is used to create unharmonized and harmonized images, only the unharmonized image contains the same object rendered under a different illumination. Fig. 5 shows some representative examples from our constructed dataset.

### D. Dataset Summary

Our dataset has a total of 143,390 training images and 22,048 test images, which cover a wide range of scenes and illumination conditions. We further split the training set and the test set into four categories based on illumination conditions as reported in Tab. I. The binary foreground (object) mask is also provided for each image.

## IV. Method

We decompose the image harmonization task into two sub-problems: (1) illumination estimation of background im-

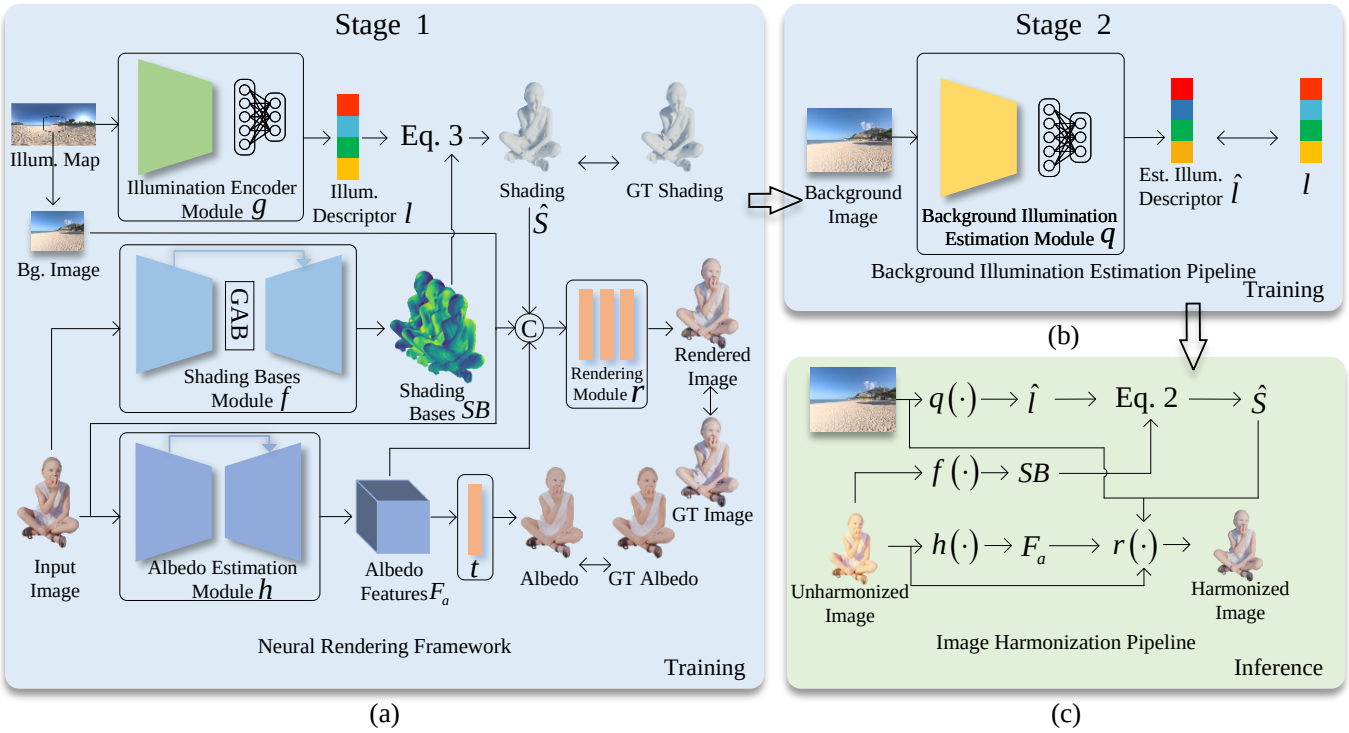


Fig. 6. An overview of our proposed image harmonization method. Our method has two training stages: training the Neural Rendering Framework (NRF) and training the Background Illumination Estimation Module (BIEM). The key to the first stage is to learn a shading-aware illumination descriptor, which is then estimated from the background image in the second stage. During inference, our image harmonization pipeline combines partial modules of the NRF $\{f, h, r\}$ and the BIEM $q$ to adjust the foreground appearance using the estimated background illumination $\hat{l}$.

TABLE I
THE NUMBER OF TRAINING AND TEST IMAGES ON EACH SCENE.

| Scene | Sunny | Sunrise/Sunset | Cloudy | Night | All |
|---|---|---|---|---|---|
| #Train | 62,074 | 27,471 | 48,051 | 5,794 | 143,390 |
| #Test | 10,472 | 3,608 | 7,508 | 460 | 22,048 |

ages, and (2) re-rendering of foreground objects. The overall pipeline of the proposed image harmonization algorithm is illustrated in Fig. 6. We first train a Neural Rendering Framework to learn the shading-aware illumination descriptor in a self-supervised manner (Sec. IV-A). Then we train a Background Illumination Estimation Module to estimate the shading-aware illumination descriptor from the background image (Sec. IV-B). The inference pipeline of image harmonization is briefly introduced in Sec. IV-C. Finally, we elaborate on the training and implementation details in Sec. IV-D.

### A. Neural Rendering Framework

As shown in Fig. 6(a), the Neural Rendering Framework is composed of three neural network modules and one rendering module. First, the Shading Bases Module and the Illumination Encoder Module generate the shading using the input foreground image and the illumination map. Then, the Albedo Estimation Module makes an estimate of the albedo from the input image. Finally, the Rendering Module combines the albedo feature, the shading, the background image and the input image to re-render the input image under a novel illumination. Below we describe these modules in detail.

**Shading Bases Module.** Inspired by the illumination cone theory [45], the Shading Bases Module $f$, parameterized by $\theta_f$, is designed to generate a set of $K$ shading bases $SB \in \mathbb{R}^{K\times H\times W}$, given the input foreground image $\tilde{I} \in \mathbb{R}^{3\times H\times W}$,

$$SB = f(\tilde{I}; \theta_f). \tag{1}$$

The Shading Bases Module, based on a U-Net architecture [46], consists of a downsampling sub-module and a upsampling sub-module. The downsampling sub-module is mainly composed of a series of Residual Dense Blocks [47] (RDBs) followed by max-pooling layers. The upsampling sub-module is composed of several convolution layers and upsampling layers. In addition, we utilize the Global Attention Block (GAB) at the bottleneck of the downsampling sub-module to imitate long-range interactions between distant pixels in global illumination. The GAB is composed of 6 transformer layers [48].

**Illumination Encoder Module.** The purpose of the Illumination Encoder Module $g$ is to encode the illumination map $L \in \mathbb{R}^{3\times H'\times W'}$ as a low dimensional illumination descriptor $l \in \mathbb{R}^{3\times K}$,

$$l = g(L; \theta_g), \tag{2}$$

where $K \ll H' \times W'$. In addition, in order to simultaneously perceive shading and illumination distribution, our illumination descriptor combines different shading bases to generate the final shading $\hat{S} \in \mathbb{R}^{3\times H\times W}$ which contains global illumination effects,

$$\hat{S}_{cij} = \sum_{k=1}^{K} l_{ck} \times SB_{kij}. \tag{3}$$

The network architecture of the Illumination Encoder Module is similar to the downsampling sub-module of the Shading Bases Module. The only difference is that the transformer layers are replaced by three fully-connected layers for outputting the illumination descriptor, which further reduces the amount of network parameters. Note that the first two fully-connected layers are both followed by a rectified linear activation function.

**Albedo Estimation Module.** The Albedo Estimation Module $h$ is designed to extract the albedo feature $F_a \in \mathbb{R}^{C\times H\times W}$ from the input foreground image $\tilde{I}$,

$$F_a = h(\tilde{I}; \theta_h). \tag{4}$$

Then, one convolution layer $t$ that takes $F_a$ as input is adopted to estimate the albedo $\hat{A} \in \mathbb{R}^{3\times H\times W}$: $\hat{A} = t\left(F_a; \theta_t\right)$.

The network architecture of the Albedo Estimation Module is the same as that of the Shading Bases Module without GAB. In addition, the channel number of RDBs is reduced by half.

**Rendering Module.** After obtaining the albedo feature and the shading, the Rendering Module $r$ performs the final rendering,

$$\hat{I} = r(F_a, \hat{S}, \tilde{I}, B; \theta_r), \tag{5}$$

where $B$ denotes the background image. In order to preserve details in the foreground, the input image $\tilde{I}$ is also fed to the Rendering Module. Note that the output image $\hat{I}$ shares the same content with the input image $\tilde{I}$ but under a different illumination condition $L$.

The network architecture of the Rendering Module is the same as that of the Albedo Estimation Module.

### B. Background Illumination Estimation Module

Once we have obtained the shading-aware illumination descriptor via the Neural Rendering Framework, the goal of the Background Illumination Estimation Module $q$, which is shown in Fig. 6 (b), is to estimate the illumination descriptor given the input background image,

$$\hat{l} = q(B; \theta_q). \tag{6}$$

The Background Illumination Estimation Module (BIEM) shares the same network architecture with the Illumination Encoder Module (IEM). However, there are two main differences between BIEM and IEM. First, the inputs of IEM and BIEM are different. As shown in Eq. 2 and Eq. 6, the input of IEM is the illumination map $L$, while the input of BIEM is the background image $B$. Note that the illumination map is not available in the inference stage of image harmonization and only the background image is used. Second, the key to the first training stage is to train an IEM to compress a high-dimensional illumination map into a low-dimensional shading-aware illumination descriptor. Once the first training stage is finished, the pre-trained IEM will be later used in the second training stage to supervise the training of the BIEM. In other words, we use the BIEM to estimate an illumination descriptor from a background image, where the ground-truth illumination descriptor is provided by the pre-trained IEM.

Refer to the supplementary materials for more implementation details of all network structures.

### C. Image Harmonization Pipeline

As shown in Fig. 6(c), our image harmonization pipeline consists of two modules, namely Background Illumination Estimation Module $q$ and Foreground Rendering Module $\{f, h, r\}$.

The Foreground Rendering Module leverages partial modules of the existing Neural Rendering Framework to re-render the input unharmonized foreground image to make it more compatible with the background image.

### D. Training Details

We train the model on our image harmonization dataset with ground truth $\{I, A, S, L\}$, where $I, A, S$ denote the harmonized image, albedo and shading respectively. Our training is divided into two stages: training the Neural Rendering Framework and training the Background Illumination Estimation Module.

At the first stage, we train the Neural Rendering Framework. The $\mathcal{L}_1$ loss is applied for shading, albedo and the output image. In addition, inspired by [49], the SSIM metric is utilized to encourage the neural network to produce visually pleasing images. Thus, the loss $L_{NR}$ for the Neural Rendering Framework is defined as,

$$\begin{aligned} L_{NR} = \|S - \hat{S}\|_1 + \|A - \hat{A}\|_1 + \|I - \hat{I}\|_1 + \lambda(1 - \mathrm{SSIM}(S, \hat{S})) \\ +\lambda(1 - \mathrm{SSIM}(A, \hat{A})) + \lambda(1 - \mathrm{SSIM}(I, \hat{I})), \end{aligned} \tag{7}$$

where the weight $\lambda$ is set to 1 in our experiments.

At the second stage, we train the Background Illumination Estimation Module. The $\mathcal{L}_1$ loss is used for the illumination descriptor. Also, the predicted illumination descriptor and the shading bases are utilized to render the shading and then minimize the error between the rendered shading and the ground truth shading. The loss $L_{BIE}$ for the Background Illumination Estimation Module is defined as,

$$L_{BIE} = \|l - \hat{l}\|_1 + \left\| S - \sum_k \hat{l}_{ck} \times SB_{kij} \right\|_1. \tag{8}$$

## V. Experiments

To validate the effectiveness of our image harmonization pipeline, we first compare our method with several state-of-the-art methods. Then we compare our neural illumination descriptor against the common illumination representation (i.e., HDR illumination maps) and our neural shading bases against the spherical harmonic bases to demonstrate their advantages in terms of rendering quality. A user study on real data is also conducted to confirm the effectiveness of our method. Finally, we perform extensive ablation studies to illustrate the contribution of each component of our framework in isolation.

TABLE II
QUANTITATIVE EVALUATION OF DIFFERENT METHODS ON THE TEST SET. THE BEST RESULTS ARE MARKED IN BOLD. THE SECOND BEST RESULTS ARE UNDERLINED. OUR METHOD ACHIEVES THE BEST RESULTS ON THE ENTIRE TEST SET WITH THE FEWEST PARAMETERS.

| Sub-dataset | Evaluation Metric | Input Composite | Lalonde and Efros [3] | DoveNet [1] | Guo et al. [8] | RainNet [2] | D-HT [30] | CDTNet [25] | SCS-Co [27] | Ours | Ours w/ Illum. map |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sunny | fMAE↓ | 0.099 | 0.110 | 0.073 | 0.076 | 0.067 | 0.074 | 0.091 | 0.119 | 0.062 | **0.055** |
| | fPSNR↑ | 19.52 | 18.33 | 21.38 | 21.15 | 22.02 | 21.47 | 20.21 | 17.90 | 22.61 | **23.77** |
| | fSSIM↑ | 0.804 | 0.787 | 0.821 | 0.851 | 0.831 | 0.810 | 0.811 | 0.757 | 0.869 | **0.893** |
| | LPIPS↓ (×e-2) | 1.089 | 2.353 | 1.111 | 1.304 | 0.815 | 1.114 | 1.096 | 3.182 | 0.791 | **0.678** |
| Sunrise/Sunset | fMAE↓ | 0.083 | 0.119 | 0.071 | 0.072 | 0.060 | 0.070 | 0.080 | 0.091 | 0.061 | **0.058** |
| | fPSNR↑ | 21.48 | 18.07 | 22.35 | 22.49 | 23.60 | 22.30 | 21.63 | 20.34 | 23.38 | **23.92** |
| | fSSIM↑ | 0.873 | 0.820 | 0.888 | 0.892 | 0.899 | 0.866 | 0.882 | 0.843 | 0.926 | **0.941** |
| | LPIPS↓ (×e-2) | 0.993 | 2.249 | 0.915 | 1.207 | 0.685 | 1.008 | 1.030 | 4.316 | 0.634 | **0.517** |
| Cloudy | fMAE↓ | 0.084 | 0.101 | 0.070 | 0.074 | 0.063 | 0.071 | 0.082 | 0.089 | 0.057 | **0.056** |
| | fPSNR↑ | 21.75 | 19.38 | 22.64 | 22.24 | 23.67 | 22.60 | 21.81 | 20.64 | 24.14 | **24.21** |
| | fSSIM↑ | 0.881 | 0.843 | 0.899 | 0.905 | 0.908 | 0.873 | 0.893 | 0.852 | 0.935 | **0.947** |
| | LPIPS↓ (×e-2) | 0.897 | 1.997 | 0.772 | 1.142 | 0.641 | 0.918 | 0.879 | 4.680 | 0.558 | **0.487** |
| Night | fMAE↓ | 0.171 | 0.122 | 0.085 | 0.093 | **0.080** | 0.098 | 0.182 | 0.136 | 0.088 | 0.094 |
| | fPSNR↑ | 16.07 | 17.65 | 20.81 | 20.21 | **21.41** | 20.20 | 15.15 | 17.34 | 20.16 | 19.80 |
| | fSSIM↑ | 0.701 | 0.736 | 0.819 | 0.821 | 0.818 | 0.791 | 0.690 | 0.746 | 0.840 | **0.849** |
| | LPIPS↓ (×e-2) | 2.078 | 2.264 | 1.291 | 1.777 | 1.127 | 1.501 | 2.154 | 4.252 | 1.146 | **1.105** |
| All | fMAE↓ | 0.093 | 0.109 | 0.072 | 0.075 | 0.065 | 0.073 | 0.088 | 0.105 | 0.061 | **0.056** |
| | fPSNR↑ | 20.53 | 18.63 | 21.95 | 21.72 | 22.83 | 21.96 | 20.88 | 19.22 | 23.21 | **23.86** |
| | fSSIM↑ | 0.840 | 0.810 | 0.859 | 0.876 | 0.868 | 0.841 | 0.848 | 0.803 | 0.900 | **0.918** |
| | LPIPS↓ (×e-2) | 1.028 | 2.213 | 0.967 | 1.243 | 0.741 | 1.038 | 1.033 | 3.900 | 0.693 | **0.596** |
| | Parameters↓ | - | - | 54.756M | 40.863M | 54.763M | 34.299M | **2.744M** | 44.900M | 10.403M | |

### A. Experimental Setup

**Evaluation metrics.** We evaluate the realism of harmonized images using fMAE, fPSNR, fSSIM [50] and LPIPS [51], where the prefix f indicates that the metric measurement is calculated only using the foreground region.

**Baselines.** We compare with one traditional method [3] and six deep learning-based methods [1], [2], [8], [25], [27], [30]. For deep learning-based methods, we select recent open-source methods [1], [2], [8], [30] achieving state-of-the-art performance. In addition, Cong et al. [25] provided us with their code and pre-trained model. For a fair comparison, we re-train their models on our image harmonization dataset according to the experiment settings given by the authors. We report their results when the training losses converge. Refer to the supplementary materials for more experimental details. The results of SCS-Co [27] are provided by the authors.

### B. Comparison with State-of-the-art

**Quantitative results.** Tab. II summarizes the quantitative results obtained by our method as well as the competing methods. Our method achieves the best results on the *sunny* and *cloudy* scenes, which can be attributed to its ability to generate realistic shadings. However, our method gets lower scores on the *night* scene compared to previous works. This is primarily due to the fact that the night images lack noticeable shading variations. Overall, our method achieves the best performance in all metrics when using the entire test set for evaluation. In addition, compared with other learning-based methods, CDTNet specially integrates with the color mapping module. However, this module cannot handle shading variations and may result in limited performance. Since SCS-Co does not consider the perception of illumination and its training data only contains variations in brightness and color, its performance is severely degraded on our test data which also contains shading variations.

We demonstrate the effect of using the illumination maps as inputs to extract the illumination descriptors. As can be observed from Tab. II, using the illumination maps as inputs (ours w/ illum. maps) leads to a significant increase in the rendering performance.

We also compare our method against the baselines using the number of parameters. Despite its complexity, our entire framework has a total of 10.403M parameters, which is approximately one-fifth of the amount of the second-best baseline with 50.763M parameters.

**Qualitative results.** Harmonized images produced by different methods are compared in Fig. 7. We display the qualitative results with different lighting conditions on several scenes, including *sunny*, *cloudy*, *sunrise/sunset*, and *night*. Our method produces compelling results that are closer to the ground truth in terms of photo-realism. For instance, in the first column of Fig. 7, there is an observable illumination inconsistency between the foreground and the background in the input composite image. Specifically, the background suggests that the main illumination source is located at the rear right, whereas, the foreground appears to be illuminated from the left. The result of Lalonde et al. [3] shows greenish colors, and all the other comparative methods [1], [2], [8], [25], [27], [30] basically retain the original illumination (e.g., the boy neck in close-ups). In contrast, our method consistently relights the foreground object, making it more consistent with the background illumination.

In the fifth column of Fig. 7, the foreground object in the input image appears to be illuminated from the right, whereas the background is a cloudy image. Ideally, under such background illumination, the foreground object should appear smooth lighting. The result of Lalonde and Efros [3] is inconsistent in terms of both color and illumination. The results of CDTNet [25] and SCS-Co [27] almost completely preserve the effect of the original lighting. Although RainNet [2], DoveNet [1], Guo et al. [8] and D-HT [30] produce the results that are a step closer to the ground truth, the highlights


Fig. 7. Qualitative comparison of different methods on our test set. We show representative examples with close-up details focusing on shading variations. Our method outperforms all other approaches with more accurate and sharper results.

on the woman's left arm are improperly preserved. Our method not only effectively delights the foreground object, but also re-renders it under a smooth illumination.

**Effects of the inferred shading and albedo.** For the single image harmonization task, there are two challenges: (1) removing the original illumination on the foreground and (2) generating the shadings under the background illumination. In this paper, we design the Albedo Estimation Module and the Shading Bases Module to solve these two problems respectively. As shown in Fig. 8, our inferred albedos effectively remove the original illumination effects, and our inferred shadings correctly contain the effects of the background illuminations. As a result, our harmonized images are more realistic and physically correct. In contrast, those image-to-image harmonization methods perform poorly on these two aspects. As shown in Fig. 7(a)(e), for example, the original light on the boy's left nose and on the woman's clothes are not well eliminated, and even artifacts are introduced in these areas. Moreover, since neither explicit shading modeling nor light perception is conducted, these methods fail to generate

TABLE III
QUANTITATIVE COMPARISON OF DIFFERENT ILLUMINATION REPRESENTATIONS FOR IMAGE HARMONIZATION ON THE TEST SET.

| | fMAE↓ | fPSNR↑ | fSSIM↑ | LPIPS↓ |
|---|---|---|---|---|
| HDR illum. map | 0.097 | 19.91 | 0.858 | 0.017 |
| Our illum. descriptor | **0.061** | **23.21** | **0.900** | **0.007** |

plausible shadings.

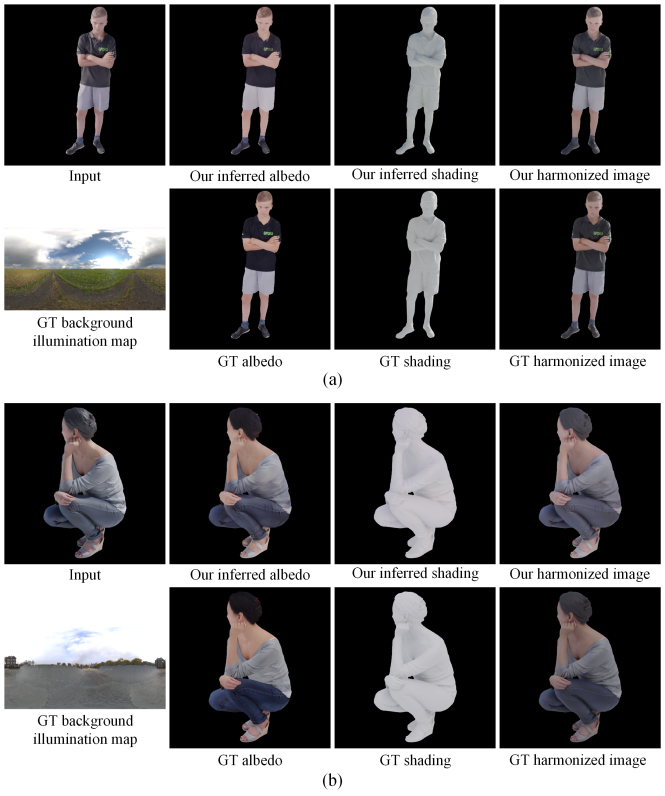


Fig. 8. Visualization of our intermediate results. Note that our inferred albedos effectively remove the original illumination effects, and our inferred shadings correctly contain the effects of the background illuminations.

### C. Comparison with HDR Illumination Map

The efficacy of our learned illumination descriptor is compared with the HDR illumination map for light estimation. Specifically, we train an encoder-decoder based neural network to map the background image to its corresponding panoramic HDR illumination map. We train this network until its loss converges. Then, we use it to estimate the HDR panoramic image from the background image. The estimated HDR image is further used as part of our Neural Rendering Framework. The rendered images are compared against those generated using our learned illumination descriptor, which is reported in Tab. III. Note that our learned illumination descriptors achieve obviously better performance compared to the estimated HDR images. In fact, it is very difficult to accurately estimate the HDR illumination map from the background image due to its huge amount of parameters.

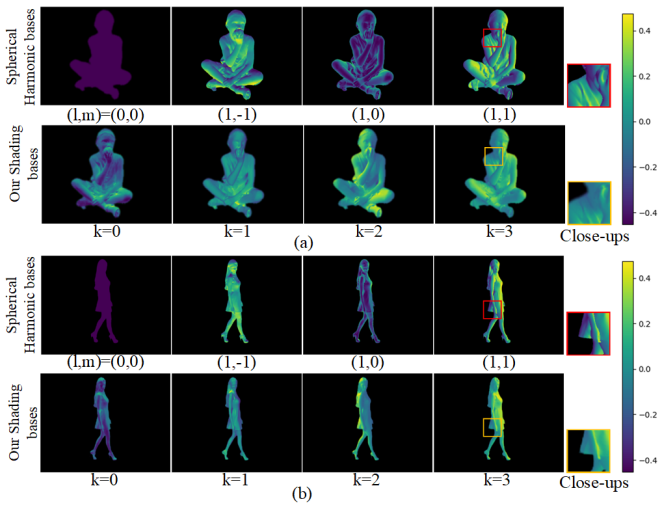


Fig. 9. Comparison with Spherical Harmonic bases.

### D. Comparison with Spherical Harmonic Bases

In Fig. 9, we compare our shading bases against the Spherical Harmonics (SH) bases [16]. Our main objective is to emphasize the advantage of our shading bases in comparison to the SH bases in terms of generating cast shadows.

We generate the first 4 SH bases $Y_{lm}$ (with (l,m) = {(0,0), (1,-1), (1,0), (1,1)}, where $l \geq 0$ and $-l \leq m \leq l$) for comparison. $(0, 0)$ indicates ambient illumination and has no specific illumination direction. $(1, -1)$, $(1, 0)$ and $(1, 1)$ show that the light source is located below, behind and to the left of the little girl, respectively.

For this experiment, we set $K$ of the Shading Bases Module to 4. We visualize the shading bases learned by our Shading Bases Module in Fig. 9, where $k$ is used to denote the $k_{th}$ shading basis. The values $k = 0, 1, 2, 3$ indicate that the light source is located behind, in front, to the right, and to the left of the little girl, respectively.

In comparison to the SH bases, our shading bases contain the cast shadow effects which are explicitly omitted by the spherical harmonic bases. Taking the last column of Fig. 9 as an example, both our shading basis and the spherical harmonics basis are illuminated from the right. From the close-ups in Fig. 9(a), it can be observed that the shoulder region is occluded by the head. Spherical harmonics however produce bright intensities without cast shadows. While our shading bases contain cast shadows that are congruent with the illumination direction, as shown in the close-ups.

### E. User Study on Real Data

We also conduct user study on real data to validate the performance of our proposed method. We made 58 composite images of which both the foreground images and the background images are collected from the Internet. Specifically, foreground humans are collected from Taobao [52] and captured by real cameras for clothes display. The background images are collected from Poly Haven [42] and HDR MAPS [43], which are all captured by professional digital cameras. We will make this benchmark dataset publicly available. For

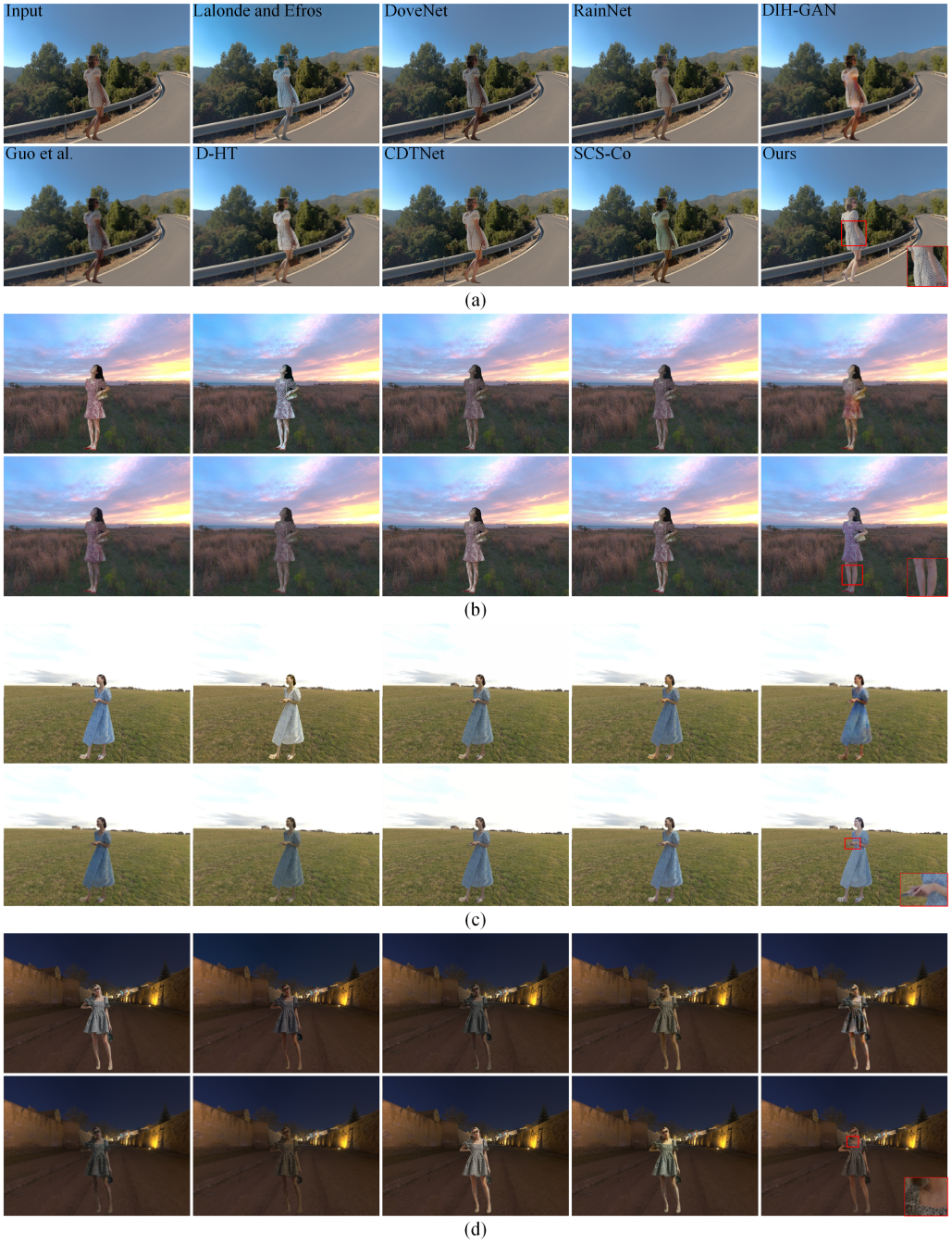


Fig. 10. Qualitative comparison of different methods on real data across different weather conditions. The details in red boxes show that our method is capable of generating plausible shadings which are consistent with the background illumination, while other methods can only adjust the color and brightness of the foreground.

deep learning-based methods, we compare against the state-of-the-art methods, namely DoveNet [1], RainNet [2], DIH-GAN [14], CDTNet [25], SCS-Co [27], Guo et al. [8] and D-HT [30]. Note that for DoveNet, RainNet, CDTNet, Guo et al. and D-HT, we use their released pre-trained models to process the input composite images. The results of DIH-GAN and SCS-Co are provided by the authors. We also compare against the traditional method proposed by Lalonde and Efros [3]. For each composite image processed by these nine methods, we ask 27 individuals to score the visual quality. As inspired by [24], the following three questions are considered for scoring: (1) Are the brightness and color of the foreground and background consistent; (2) Are the illumination directions of the foreground and background consistent; and (3) Are the texture distortions/artifacts of the foreground serious. The visual quality score ranges from 0 to 3 (worst to best quality). Tab. IV reports the results. It shows that we achieve a large advantage on question 2, which is mainly due to the fact that neither previous methods [1], [2], [8], [25], [30] nor their corresponding training data have yet considered shading variations. Besides, the recently proposed DIH-GAN [14] does not explicitly model foreground shading, and many of their

3D models are created by CG software, which results in poor generalization to real data. As shown in Fig. 10, the details in red boxes show that our method is able to generate foreground shadings that are consistent with the background illumination. In contrast, other methods fail to generate plausible shading and even erroneously transfer the color of the background objects (e.g., the grasses) to the foreground due to the lack of perception of the illumination in the background.

TABLE IV
USER STUDY ON REAL DATA.

| | Score (Q1) | Score (Q2) | Score (Q3) | Overall Score |
|---|---|---|---|---|
| Lalonde and Efros [3] | 0.987 | 1.084 | 1.893 | 1.322 |
| DIH-GAN [14] | 1.293 | 1.466 | 1.017 | 1.259 |
| DoveNet [1] | 1.670 | 1.480 | 1.883 | 1.678 |
| RainNet [2] | 1.636 | 1.469 | 1.720 | 1.608 |
| Guo et al. [8] | 1.512 | 1.477 | 1.633 | 1.541 |
| D-HT [30] | 1.656 | 1.494 | 1.901 | 1.684 |
| SCS-Co [27] | 1.645 | 1.473 | 1.897 | 1.672 |
| CDTNet [25] | 1.749 | 1.497 | 1.912 | 1.719 |
| Ours | **2.051** | **1.915** | **1.981** | **1.982** |

### *F. Generalization to Indoor Scenes and Non-Human Objects*

Fig. 11 shows the generalization of our method to indoor scenes. First, our method is able to perceive illumination for indoor scenes, especially the illumination direction. For example, in Fig. 11 (b), the background image indicates that the primary light source in the scene comes from the right side (i.e., the windows), and our generated shading (e.g., the details in the red box) is consistent with the direction of the primary light source in the scene. Second, as shown in Fig. 11 (a), our result, especially the details in the blue box, is more realistic owing to the appropriate illumination brightness and color. This can also be observed in Fig. 11 (c).

However, our method does not yet account for spatially varying illumination estimation which can further improve the realism for indoor scene harmonization. In the future, one of the potential solutions is to estimate an illumination descriptor for individual background image pixel.

Fig. 12 shows the generalization of our method to non-human objects. Although the constructed training set only covers human objects, our approach generalizes well to non-human objects, specifically generating reasonable shadings that are consistent with the target background lighting. For example, in the second row of Fig. 12, it can be observed from the background image that the sun is located behind the right side of the toy car. Not only is our harmonized toy car mostly backlit, but its right side is partially illuminated by the sun, as shown in the blue box. In addition, adding more different types of objects to the training set could further improve the generalization performance of the proposed method.

### *G. Ablation Study*

The ablation study is conducted to demonstrate the effectiveness of each component on the Neural Rendering Framework (NRF).

**Neural rendering framework Ablation.** We demonstrate how the use of the albedo features $F_a$, the input unharmonized image $\hat{I}$ and the background image $B$ as additional inputs to

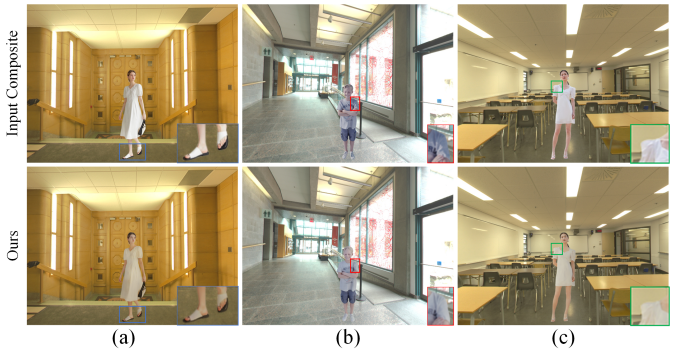


Fig. 11. Generalization to indoor scenes. Zoom in for more details.


Fig. 12. Generalization to non-human objects.

the Rendering Module can improve the overall performance. We also perform an ablation on the loss function of our Neural Rendering Framework. Quantitative results are reported in Tab. V.

We start with the baseline NRF which uses the concatenated shading and albedo as inputs to the Rendering Module. Replacing the albedo image with the albedo feature $F_a$ results in better performance. We attribute this improvement to the albedo feature $F_a$ which contains richer information.

We then proceed to add the input unharmonized image $\hat{I}$ in conjunction with the albedo feature $F_a$ and the shading as inputs to the Rendering Module, and observe a slight improvement in performance. The delighting, which occurs as a consequence of albedo estimation, can result in a loss of information in the final rendered image. Therefore, using

TABLE V
ABLATION STUDY ON NEURAL RENDERING FRAMEWORK.

| | fPSNR↑ | fSSIM↑ | LPIPS↓ (×e-2) |
|---|---|---|---|
| Baseline NRF | 22.63 | 0.893 | 0.775 |
| Baseline NRF + $F_a$ | 23.25 | 0.897 | 0.735 |
| Baseline NRF + $F_a$ + $\hat{I}$ | 23.18 | 0.901 | 0.704 |
| Baseline NRF + $F_a$ + $\hat{I}$ + $B$ | 23.82 | 0.906 | 0.647 |
| Baseline NRF + $F_a$ + $\hat{I}$ + $B$ + SSIM loss | **23.86** | **0.918** | **0.596** |

TABLE VI
EFFECTS OF THE NUMBER OF SHADING BASES ($K$).

| $K$ | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|
| fPSNR↑ | 29.44 | 30.81 | 31.21 | 31.35 | **31.42** | 31.41 |
| fSSIM↑ | 0.919 | 0.931 | 0.935 | 0.936 | **0.937** | **0.937** |

the unharmonized image $\hat{I}$ as additional input can make up for the lost information. When the background image is also added to the Rendering Module, there is a moderate increase in performance. In fact, by exploiting the brightness and color information of the background image, the Rendering Module is able to generate the foreground appearance more accurately.

Finally, in addition to the $\mathcal{L}_1$ loss of the baseline NRF, the SSIM loss function is added. Experimental results show that adding the SSIM loss significantly improves the performance of our framework, especially in terms of the foreground SSIM (fSSIM) and LPIPS metric measurements.

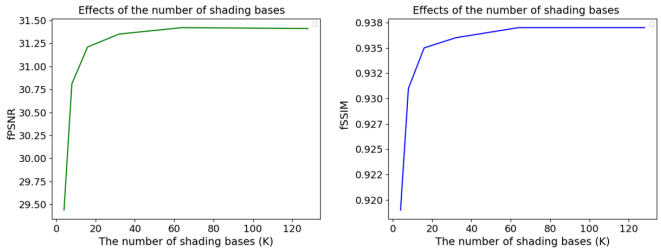


Fig. 13. Effects of the number of shading bases.

**Effects of the number of shading bases ($K$).**
To measure the influence of the number of shading bases on the performance of our Shading Bases Module, we train the Shading Bases Module with 6 different values of $K$ (4, 8, 16, 32, 64, 128). The results are reported in Tab. VI. We also visualize them in Fig. 13. Note that the reported metrics (fPSNR, fSSIM) are obtained by comparing the results generated by our Shading Bases Module against the corresponding ground truth shading images. There exists a significant increase in the performance when the value of $K$ increases from 4 to 32, then the increase becomes less obvious when $K > 32$. To balance between the rendering performance and the computational complexity, we resolve to use $K = 32$ as the optimal number of shading bases.

### H. Discussions

**Impact of object placement.** Object placement aims to place the foreground within the background image with a suitable location and size. From the view of physical image formation, the most important factor affecting image harmonization is the location of objects. Because different locations may have different lighting, depending on the type of scene. Especially in indoor scenes, the lighting at different locations may vary greatly. Therefore, the acquisition of illumination at different locations in scenes and its accurate estimation pose a greater challenge to image harmonization. In the near future, we will continue to focus on image harmonization with spatially-varying lighting estimation.

## VI. CONCLUSIONS

In this paper, we have contributed a large-scale photo-realistic image harmonization dataset involving variations in color, brightness, and shading. In addition, a novel Neural Rendering Framework is designed to learn a shading-aware illumination descriptor from the illumination maps. A neural Shading Bases Module is proposed to generate the foreground shading using the shading-aware illumination descriptor estimated from the background. Extensive experiments on the self-constructed dataset and real data demonstrate the effectiveness of our proposed method.

**Limitations.** This work has several limitations that can be further improved. At now, we focus on one specific object type (i.e. human body), which limits the application scope of our method. Extending to different types could improve the ability to generalize across a wide spectrum of objects. Additionally, only Lambertian objects are considered. When introducing the specular BRDFs [53], our model could be applied to the objects with specular reflection.

## REFERENCES

[1] W. Cong, J. Zhang, L. Niu, L. Liu, Z. Ling, W. Li, and L. Zhang, "DoveNet: Deep image harmonization via domain verification," in *CVPR*, 2020.
[2] J. Ling, H. Xue, L. Song, R. Xie, and X. Gu, "Region-aware adaptive instance normalization for image harmonization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9361–9370.
[3] J.-F. Lalonde and A. A. Efros, "Using color compatibility for assessing image realism," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.
[4] S. Xue, A. Agarwala, J. Dorsey, and H. Rushmeier, "Understanding and improving the realism of image composites," *ACM Transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–10, 2012.
[5] J.-Y. Zhu, P. Krahenbuhl, E. Shechtman, and A. A. Efros, "Learning a discriminative model for the perception of realism in composite images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3943–3951.
[6] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang, "Deep image harmonization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3789–3797.
[7] J. Cao, W. Cong, L. Niu, J. Zhang, and L. Zhang, "Deep image harmonization by bridging the reality gap," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.
[8] Z. Guo, H. Zheng, Y. Jiang, Z. Gu, and B. Zheng, "Intrinsic image harmonization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16 367–16 376.
[9] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.
[10] X. Xiao and L. Ma, "Color transfer in correlated color space," in *Proceedings of the 2006 ACM international conference on Virtual reality continuum and its applications*, 2006, pp. 305–309.
[11] F. Pitié, A. C. Kokaram, and R. Dahyot, "Automated colour grading using colour distribution transfer," *Computer Vision and Image Understanding*, vol. 107, no. 1-2, pp. 123–137, 2007.
[12] U. Fecker, M. Barkowsky, and A. Kaup, "Histogram-based prefiltering for luminance and chrominance compensation of multiview video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 9, pp. 1258–1267, 2008.
[13] P. Debevec, "Image-based lighting," in *ACM SIGGRAPH 2006 Courses*, 2006, pp. 4–es.
[14] Z. Bao, C. Long, G. Fu, D. Liu, Y. Li, J. Wu, and C. Xiao, "Deep image-based illumination harmonization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 542–18 551.

[15] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '97. USA: ACM Press/Addison-Wesley Publishing Co., 1997, p. 369–378. [Online]. Available: https://doi.org/10.1145/258734.258884

[16] R. Ramamoorthi and P. Hanrahan, "An efficient representation for irradiance environment maps," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 497–500.

[17] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*, 2003, pp. 313–318.

[18] F. Pitie, A. C. Kokaram, and R. Dahyot, "N-dimensional probability density function transfer and its application to color transfer," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1434–1439.

[19] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu, "Color harmonization," in *ACM SIGGRAPH 2006 Papers*, 2006, pp. 624–630.

[20] J. Jia, J. Sun, C.-K. Tang, and H.-Y. Shum, "Drag-and-drop pasting," *ACM Transactions on Graphics (SIGGRAPH)*, 2006.

[21] K. Sunkavalli, M. K. Johnson, W. Matusik, and H. Pfister, "Multi-scale image harmonization," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, pp. 1–10, 2010.

[22] S. Song, F. Zhong, X. Qin, and C. Tu, "Illumination harmonization with gray mean scale," in *Computer Graphics International Conference*. Springer, 2020, pp. 193–205.

[23] K. Sofiiuk, P. Popenova, and A. Konushin, "Foreground-aware semantic representations for image harmonization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1620–1629.

[24] Y. Jiang, H. Zhang, J. Zhang, Y. Wang, Z. Lin, K. Sunkavalli, S. Chen, S. Amirghodsi, S. Kong, and Z. Wang, "SSH: A self-supervised framework for image harmonization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4832–4841.

[25] W. Cong, X. Tao, L. Niu, J. Liang, X. Gao, Q. Sun, and L. Zhang, "High-resolution image harmonization via collaborative dual transformations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 470–18 479.

[26] W. Cong, L. Niu, J. Zhang, J. Liang, and L. Zhang, "BargainNet: Background-guided domain translation for image harmonization," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[27] Y. Hang, B. Xia, W. Yang, and Q. Liao, "SCS-Co: Self-consistent style contrastive learning for image harmonization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 710–19 719.

[28] X. Cun and C.-M. Pun, "Improving the harmony of the composite image by spatial-separated attention module," *IEEE Transactions on Image Processing*, vol. 29, pp. 4759–4771, 2020.

[29] G. Hao, S. Iizuka, and K. Fukui, "Image harmonization with attention-based deep feature modulation." in *BMVC*, 2020.

[30] Z. Guo, D. Guo, H. Zheng, Z. Gu, B. Zheng, and J. Dong, "Image harmonization with transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 870–14 879.

[31] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000, pp. 145–156.

[32] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi, "Deep image-based relighting from optimal sparse samples," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–13, 2018.

[33] A. Meka, C. Haene, R. Pandey, M. Zollhöfer, S. Fanello, G. Fyffe, A. Kowdle, X. Yu, J. Busch, J. Dourgarian *et al.*, "Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[34] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs, "Deep single-image portrait relighting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7194–7202.

[35] T. Sun, J. T. Barron, Y.-T. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. E. Debevec, and R. Ramamoorthi, "Single image portrait relighting." *ACM Trans. Graph.*, vol. 38, no. 4, pp. 79–1, 2019.

[36] Y. Kanamori and Y. Endo, "Relighting humans: occlusion-aware inverse rendering for full-body human images," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–11, 2018.

[37] Z. Wang, X. Yu, M. Lu, Q. Wang, C. Qian, and F. Xu, "Single image portrait relighting via explicit multiple reflectance channel modeling," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–13, 2020.

[38] S. Sang and M. Chandraker, "Single-shot neural relighting and SVBRDF estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 85–101.

[39] M. Lagunas, X. Sun, J. Yang, R. Villegas, J. Zhang, Z. Shu, B. Masia, and D. Gutierrez, "Single-image full-body human relighting," in *Eurographics Symposium on Rendering (EGSR)*. The Eurographics Association, 2021.

[40] Y. Yu, A. Meka, M. Elgharib, H.-P. Seidel, C. Theobalt, and W. A. Smith, "Self-supervised outdoor scene relighting," in *European Conference on Computer Vision*. Springer, 2020, pp. 84–101.

[41] "3D People," https://3dpeople.com.

[42] "Poly Haven," https://polyhaven.com/hdris.

[43] "HDR MAPS," https://hdrmaps.com/.

[44] B. O. Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[45] P. N. Belhumeur and D. J. Kriegman, "What is the set of images of an object under all possible illumination conditions?" *International journal of computer vision*, vol. 28, no. 3, pp. 245–260, 1998.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[47] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[49] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[52] "Taobao," https://www.taobao.com/.

[53] M. Ashikmin, S. Premože, and P. Shirley, "A microfacet-based BRDF generator," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000, pp. 65–74.