Spatially-Varying Illumination-Aware Indoor Harmonization

Zhongyun Hu[1], Jiahao Li[1], Xue Wang[1], Qing Wang[1*]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China.

*Corresponding author(s). E-mail(s): qwang@nwpu.edu.cn;
Contributing authors: zy_h@mail.nwpu.edu.cn; swafim2@mail.nwpu.edu.cn;
xwang@nwpu.edu.cn;

Abstract

In this paper, we address the problem of spatially-varying illumination-aware indoor harmonization. Existing image harmonization works either focus on extracting no more than 2D information (e.g., low-level statistics or image filters) from the background image or rely on the non-linear representations of deep neural networks to adjust the foreground appearance. However, from a physical point of view, realistic image harmonization requires the perception of illumination at the foreground position in the scene (i.e., Spatially-Varying (SV) illumination), especially for indoor scenes. To solve indoor harmonization, we present a novel learning-based framework, which attempts to mimic the physical model of image formation. The proposed framework consists of a new neural harmonization architecture with four compact neural modules, which jointly learn SV illumination, shading, albedo, and rendering. In particular, a multilayer perceptron-based neural illumination field is designed to recover the illumination with finer details. Besides, we construct the first large-scale synthetic indoor harmonization benchmark dataset in which the foreground focuses on humans and is rendered and perturbed by SV illuminations. An object placement formula is also derived to ensure that the foreground object is placed in the background at a reasonable size. Extensive experiments on synthetic and real data demonstrate that our proposed approach achieves better results than prior works.

Keywords: Image harmonization, Spatially-varying illumination, Shading, Deep learning

1 Introduction

Image harmonization aims at adjusting the appearance of the foreground image so that it is perfectly merged into the environment of the background image. However, some existing image harmonization methods only extract no more than 2D information (such as image filters [1, 2], low-level statistics [3, 4], or semantics [5, 6]) from the background image. As shown in Fig. 1(c), this kind of information can only be used to adjust foreground brightness and color, not *shading*. The other methods [7–11] rely on the non-linear representations of deep neural networks to adjust the foreground appearance. However, since they do not explicitly consider certain factors such as shading and illumination, their models provide limited representations for indoor scenes with spatially-varying illumination. This causes these methods to sometimes produce wrong results that do not match the illumination distribution of the background in 3D space, such as transferring the irrelevant illuminant in the background to the foreground. In fact, from the perspective of physical image formation [12], illumination is the crucial factor that affects

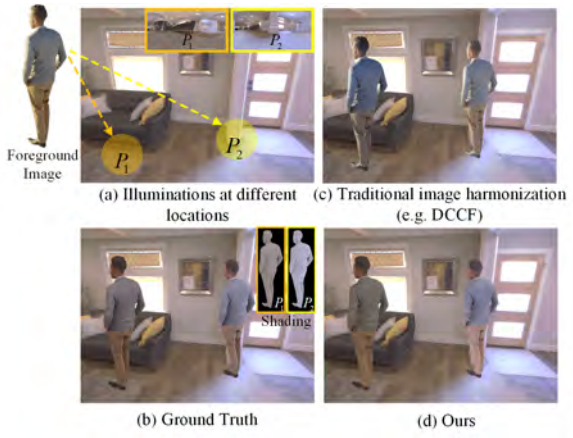


Fig. 1 Different from the directional light in outdoor scenes (i.e., the sun), the illumination in indoor scenes is spatially varying (a). As a result, the appearance of the foreground object at different locations may be different (b). However, some existing image harmonization works only extract no more than 2D information (e.g., image filters [1, 2] and low-level statistics [3, 4]) from the background image, which greatly reduces the realism of the indoor composite image (c). In contrast, our proposed harmonization framework can not only produce more accurate shading results consistent with the background illuminations, but also effectively remove the original illumination effects (d). Here P_1 and P_2 refer to different foreground placement locations.

the foreground's brightness, color, and shading. Thus background illumination should be perceived if we want the foreground to blend seamlessly into the background image.

Recently, Bao et al. [13] first proposed to extract background illumination to adjust the foreground appearance. But they assume that the illumination is directional, which only applies to outdoor scenes. As shown in Fig. 1(a), the spatially-varying illumination of indoor scenes poses a greater challenge to existing image harmonization works. Therefore, the goal of this paper is to solve the problem of spatially-varying illumination-aware indoor harmonization.

Before addressing indoor harmonization, the first obstacle we face is the lack of a large-scale dataset for indoor harmonization. In existing large-scale image harmonization datasets (e.g., iHarmony4 [7], IH [13], and HLIP [6]), either the foreground perturbation only contains brightness and color variations [6, 7], or the illumination is assumed to be directional [13]. A dataset containing both foreground illumination perturbations and spatially-varying illumination does not yet exist. In this work, we construct the first large-scale synthetic indoor harmonization dataset where the foreground focuses on humans and is perturbed and rendered by spatially varying illuminations. In dataset construction, an object placement formula is derived so that the size of the foreground object changes with the placement position, in order to comply with the principle of "near and big, far and small in space". In addition, our carefully collected illumination maps and 3D models are both reconstructed from the real world in order to achieve photo-realistic renderings.

To solve indoor harmonization, in this paper, we propose a novel physically-inspired, learning-based framework. Specifically, it is composed of four compact neural modules to simulate the process of physical image formation, namely an illumination estimation module, a shading module, an albedo estimation module, and a rendering module. In particular, the albedo estimation module is carefully designed to be connected to the rendering module by physically-meaningful deep features rather than the albedo itself, which effectively avoids the accumulated error caused by inaccurate estimation of albedo. Our experiments show that this module can effectively remove the effects of original illumination. In addition, unlike the existing illumination estimation methods based on parametric representation [14–17], we introduce a neural illumination field into the illumination estimation module. The Neural Illumination Field (NIF) is designed as a Multi-Layer Perceptron (MLP) with Fourier features-based position encoding. By introducing a large number of Fourier features, our NIF is able to more accurately characterize illumination with the same number of parameters compared to [17]. More importantly, the NIF-based illumination estimation module enables our entire image harmonization framework to achieve spatially-varying illumination-aware indoor harmonization in 3D space.

Extensive experiments on this large-scale benchmark dataset demonstrate that our proposed method outperforms three state-of-the-art methods in terms of brightness, color, and shading. In addition, we carefully collect and construct a small indoor harmonization evaluation

dataset of real composite images where the foreground is placed in multiple positions. A user study of five state-of-the-art methods and our proposed method on this evaluation dataset shows that our results are not only visually pleasing but also, more importantly, consistent with the background illumination distribution in 3D space. Finally, experiments on the HVIDIT dataset [18] also demonstrate that not only can our framework be extended to general object types, but the use of illumination information helps improve the harmonization performance.

In summary, our main contributions are as follows:

1. We construct a large-scale synthetic indoor harmonization benchmark dataset, in which the foreground is perturbed and rendered by spatially varying illuminations. In dataset construction, we also derive an object placement formula to make the foreground object match the background at a reasonable size. To the best of our knowledge, this is the first image harmonization dataset to include the spatially-varying illumination property.
2. We propose a novel physically-inspired, learning-based framework for spatially-varying illumination-aware indoor harmonization, the core of which is an illumination estimation module equipped with an MLP-based NIF with Fourier features-based positional encoding to recover the illumination accurately.
3. Our proposed framework achieves state-of-the-art performance on both the indoor harmonization benchmark and HVIDIT. A user study on real composite images is also conducted to verify the superiority of our framework, especially in dealing with spatially-varying harmonization.

2 Related Work

In this section, we briefly review image harmonization works. We also discuss illumination estimation, intrinsic image decomposition, and image relighting works related to this paper.

2.1 Image Harmonization

For image harmonization, we divide it into statistic-based methods and learning-based methods. Below, we will discuss them separately.

Statistic-based methods: The early image harmonization works [3, 4, 19–25] mainly concentrated on matching low-level statistics consistency between different images, such as mean and variance [3], contrast and noise [24]. In particular, Xue et al. [25] identified key statistics that influence the realism of composite images through human visual perception experiments. Lalonde and Efros [4] used global color statistics that are calculated on a large real image dataset to improve the realism of composite images. However, in terms of photo-realism, the results of matching the hand-crafted statistics are far from satisfactory.

Learning-based methods: Deep learning has recently achieved state-of-the-art results on a lot of computer vision tasks, including image harmonization. Tsai et al. [5] proposed the first end-to-end neural network for image harmonization. These learning-based methods usually formulate image harmonization as an image-to-image translation task while ensuring visual consistency in different aspects, such as semantics consistency [5, 6, 26], style consistency [27, 28], domain consistency [7, 29, 30], appearance consistency [10], and reflectance consistency [18]. It is worth noting that a learning-based illumination harmonization framework is proposed to ensure illumination consistency by Bao et al. [13], but they just assume that the illumination is directional, which only works for outdoor scenes. In addition, some latest deep learning techniques, such as attention mechanisms [8, 31] or Transformer [11, 32, 33], have also been exploited to boost the performance of image harmonization. High-resolution image harmonization [2, 9, 34, 35] are also considered. More recently, Xue et al. [2] and Ke et al. [1] simultaneously integrated comprehensible image filters into the learning framework, which facilitates user editing. Valanarasu et al. [36] proposed an interactive portrait harmonization method. Xu et al. [37] performed color harmonization in an intermediate high dynamic range color space instead of the standard color space to correct color discrepancy effectively. However, none of these methods take into account spatially-varying illumination-aware indoor harmonization, nor even alter the foreground shadings except for [13, 38]. In contrast, in this paper, we propose a physically-inspired learning framework to solve the problem of indoor harmonization.

2.2 Illumination Estimation

Early illumination estimation works [16, 39–43] usually estimate a single, global scene illumination from a limited field-of-view image. However, a single global illumination does not take into account the spatially-varying lighting effects, which may result in unrealistic renderings, especially for indoor scenes. Recently, a growing number of works [15, 17, 44–48] have focused on spatially-varying illumination estimation. Garon et al. [15] first proposed a two-stream global and local neural network to estimate a spherical harmonics illumination for each local image patch. Li et al. [17] proposed to use spherical Gaussians rather than spherical harmonics to better recover high-frequency lighting effects. They also proposed to estimate parametric models of both visible and invisible light sources for editing indoor scene lighting [48]. In addition, 3D volumetric representations [44, 45] are also used to represent spatially-varying illuminations, but accompanied by a high computational cost. Recently, neural fields have shown great success in the field of view synthesis and extended to many other fields [49, 50]. In this work, we propose a novel neural fields-based illumination estimation module to reconstruct the illumination with finer details.

2.3 Intrinsic Image Decomposition

Intrinsic Image Decomposition (IID) [51, 52] is the task that aims to decompose a single image into an albedo image and a shading image from the perspective of physical image formation. Similarly, our framework is also inspired by the physical image formation and incorporates shading and albedo estimation for better harmonizing images. However, compared to these IID works [53–57] where albedo and shading must be explicitly estimated, we can estimate them implicitly during the inference phase. Therefore, we have more freedom in the design of network architecture. For example, instead of running the entire albedo estimation module to estimate the final albedo, we only need to obtain the intermediate albedo feature as input to the rendering module. In addition, IID works focus on decoupling the shading from the image, but our goal is to render a new shading of the foreground image with the estimated background illumination. The closest work to ours is [58], where they proposed a global-local spherical harmonics lighting model to improve the results of IID. However, as discussed in Sec. 2.2, spherical harmonics lighting often loses some high-frequency details.

2.4 Image Relighting

Early image relighting works [59–61] focus on capturing multiple images under different illumination conditions to reconstruct the light transport function for relighting the objects. Note that the target illumination here is given directly without estimating it when relighting. Recently, several deep relighting networks [62–66] with illumination estimation have been proposed to relight portraits or human bodies only using a single RGB image. However, illumination estimation only targets portraits or human bodies rather than complex natural scenes. In other words, these relighting methods are not specifically designed for image harmonization and can not be applied to it directly. More importantly, they lack the ability to perceive spatially varying illumination, which is the core topic of this paper.

3 Dataset Construction

In this section, we introduce a novel large-scale synthetic indoor harmonization dataset in which the foreground is rendered and perturbed by *spatially-varying illumination*. Below we describe the dataset construction process in detail, which covers data collection, spatially-varying illumination generation, background rendering, foreground rendering, and object placement.

3.1 Data Collection

To construct our dataset, we collect High Dynamic Range (HDR) illumination maps with depth annotations, high-quality 3D interior models, and 3D human models. Specifically, the Laval Indoor HDR dataset [39] and the Replica dataset [67] are collected to generate SV HDR illumination maps. They cover various types of indoor scenes, such as shopping malls, bedrooms, offices, and corridors. We also collect a total of 135 high-quality 3D human models from 3D People [68], of which 117 are used for training and 18 for testing. A rich variety of humans are included, with diversity across genders (male, female), ages, poses,

and clothing (colors, accessories). Note that both the Replica dataset and the 3D human models are reconstructed from the real world to achieve photo-realistic renderings.

3.2 Spatially-varying Illumination Generation

The spatially-varying illumination maps are generated from the Laval Indoor HDR dataset and the Replica dataset. The overall generation process is shown in Fig. 2. For the Laval Indoor HDR dataset, we first filtered the HDR illumination maps with no depth annotations, inaccurate geometries, or low brightness, resulting in the remaining 65 illumination maps. In particular, for filtering geometries, we first transformed each illumination map into a point cloud based on its depth, and then visualized it one by one in MeshLab software [69] to determine whether the geometry reflected in the scene content is consistent with the annotated geometry. For example, we removed those geometries whose surfaces have depth variations but were mis-annotated as flat surfaces. 42 are randomly selected for training and 23 for testing. For each illumination map, we then manually annotated the planar surface within the illumination map to ensure the foreground object was in a suitable location. Specifically, we used Labelme software [70] to label the floor in the illumination map. We further transform the illumination map and its depth into the corresponding point cloud. Specifically, for an illumination map with a resolution of $H \times W$, assuming that the coordinates of a certain pixel are (u, v), its corresponding longitude and latitude coordinates are $(a, b) = (2u\pi/(W-1), -0.5\pi + v\pi/(H-1))$. Given its depth d, the spatial coordinates $[x, y, z]^T$ can be calculated as follows:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = d \begin{bmatrix} \cos a \cos b \\ \sin a \cos b \\ -\sin b \end{bmatrix}. \quad (1)$$

Finally, given a target pixel sampled from the annotated planar surface, the point cloud is translated and projected into the target illumination map via the Z-buffering algorithm [71].

For the Replicate dataset, we first manually selected the camera placement locations using MeshLab software. Then, for each location, a panoramic camera is placed to render an HDR illumination map and its depth map using the Habitat-Sim platform [72]. We further filtered the illumination maps with large black holes caused by missing textures. A total of 720 illumination maps are retained, 561 for training and 159 for testing. Note that the illumination maps in the same scene are either all for training or all for testing. Finally, the object placement locations are directly based on the camera placement locations. But we removed those locations where foreground objects could not be placed due to occlusion, too small space, or being close to the black hole.

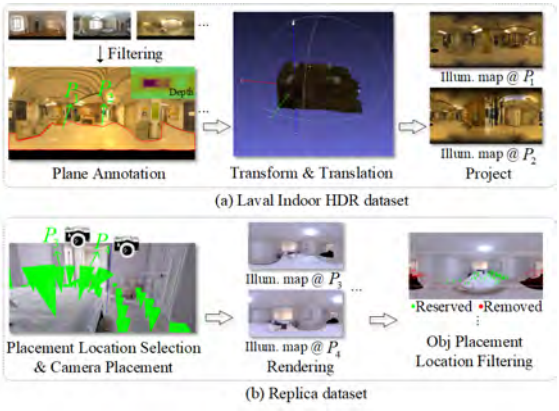


Fig. 2 The process of spatially-varying illumination generation from the Laval Indoor HDR dataset and Replica dataset. P_1 and P_2 represent the locations where illumination maps are to be rendered. P_3 and P_4 represent the locations where panoramic cameras are placed.

3.3 Background Rendering

As shown in Fig. 3, the background image is cropped from the illumination map using a virtual perspective camera with a resolution of 640×480 pixels. In addition, for each background image from the Replica dataset, we also render per-pixel illuminations with a resolution of 16×32 pixels.

3.4 Foreground Rendering

For each pair of the illumination map and 3D model, we use Blender [73] with the Cycle Renderer to render the foreground image along with its foreground mask, shading, and albedo as shown in Fig. 4. Each image is rendered with a resolution of 640×480 pixels. Samplings Per Pixel (SPP) is set to around 200∼400. To increase the richness of object poses, each object is rotated at one angle

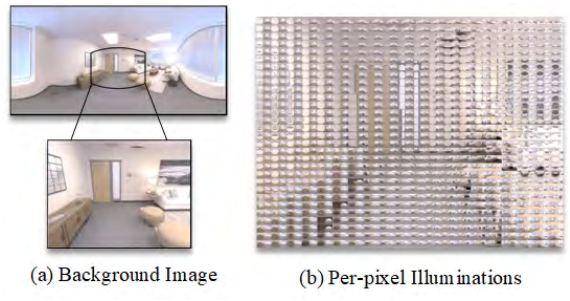


Fig. 3 The background image is cropped from the illumination map, and per-pixel illuminations are rendered only from the Replica dataset.

that is randomly sampled from a pre-defined set of 8 angles, ranging from 0 to 360 degrees with an increment of 45 degrees.

Fig. 4 For each pair of the illumination map and 3D scan model, a quad-tuple (i.e., image, shading, albedo, and mask) is rendered.

3.5 Object Placement

Once the foreground image is rendered, we place it in the background image. To match the background image size, as shown in Fig. 5, a foreground image scale factor s is inferred,

$$\tan\beta = \frac{H_b/2 - H_p}{H_b/2} \cdot \tan\left(\alpha/2\right), \tag{2}$$

$$s = \frac{H_p}{H_f} \cdot \frac{h_o \tan\beta \tan\left(\pi/2 - \alpha/2\right)}{h_c\left(1 - \tan\beta \tan\left(\pi/2 - \alpha/2\right)\right)}, \tag{3}$$

where α and h_c denote the Vertical Field of View (VFOV) and the height of the camera, respectively. h_o is the foreground object height, and H_p is the vertical pixel distance from the placement pixel p to the bottom of the background image. H_b is the background image height, and H_f is the original foreground image height. Note that the image plane here is perpendicular to the ground. See supplementary material for a more general formula with camera pitch angle and object support height. The scale factor S is then used to resize the foreground image and its mask. Finally, the composite image is obtained by alpha-blending the resized foreground image and the background image.

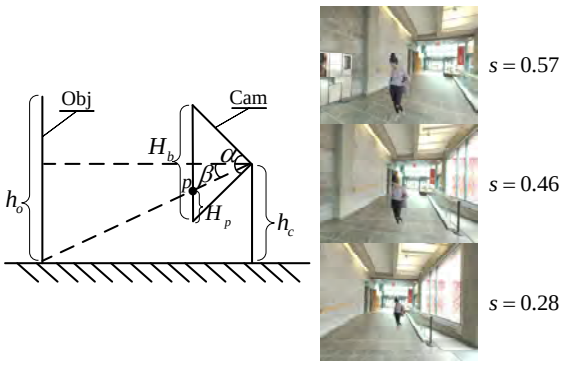


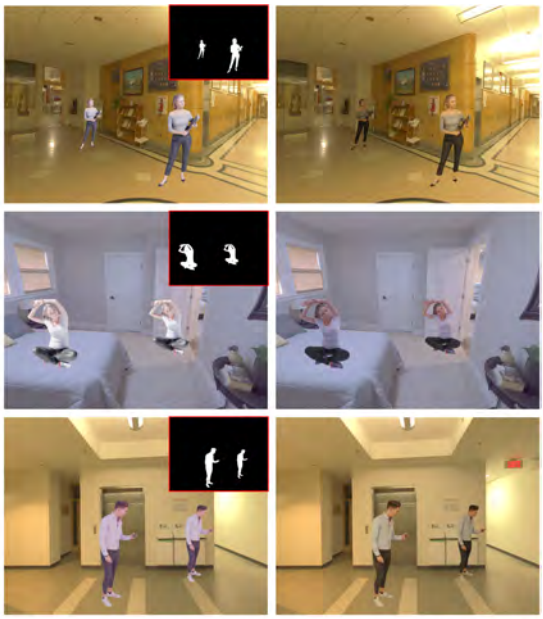
Fig. 5 The estimate of S in foreground object placement. Given the different placement positions of the foreground object, the corresponding scales are inferred to match the geometry of the background.

The above procedure is used to create both unharmonious and harmonious composite images. But the foreground of the unharmonious composite image is rendered by a randomly selected illumination. In addition, to increase the diversity of the constructed dataset, the outdoor HDR illumination maps, collected from Poly Haven [74] and HDR MAPS [75], are also used to render the foreground image but only for the unharmonious composite image. In Fig. 6, we show some high-quality examples from our constructed dataset.

Data source	Laval	Replica	Total
#Train	16,141	55,944	72,085
#Test	2192	4,570	6,762

Table 1 The number of training and test images on each data source.

In summary, we produce a total of 78,847 images, 72,085 for training, and 6,762 for testing. In Tab. 1, we also show the statistics on each data source in detail.



(a) Unharmonious Composite Image (b) Harmonious Composite Image (Ground Truth)

Fig. 6 High-quality examples from our constructed large-scale image harmonization dataset. Note that the foreground is rendered and perturbed by *spatially varying illumination*. Compared to the widely used iHarmony4 dataset [7], our challenging setting is in line with real-world image harmonization.

4 Method

4.1 Problem Formulation

Unlike existing image harmonization methods [7, 27, 28] that directly map an unharmonious image to a harmonious one, we seek to incorporate the physical principles of image formation into our method to improve the realism of the composite image. The rendering equation [12], which consists of several input physical terms, is often used to simulate the physical process of image formation in the real world. However, it is hardly possible to directly use this equation to render a new foreground image because we cannot accurately estimate all input physical terms from limited observations (i.e., a foreground image, a background image, and a foreground mask), which is a severely ill-posed problem.

Although not all physical terms can be or need to be estimated, we believe that the three physical terms are essential for image harmonization: illumination, shading, and albedo. First, the rendering equation indicates that illumination is the only variable that makes the same object look different. Moreover, our experiment also shows that the performance improvement of image harmonization also benefits from the use of illumination information. Therefore, illumination estimation is performed to ensure illumination consistency between the foreground and the background. In addition, Bao et al. [13] also proposed to adjust the foreground appearance by perceiving the illumination of the background scene. However, on one hand, they mainly target outdoor scenes and do not involve spatially-varying illumination estimation; on the other hand, the spherical harmonics they used clearly have limited ability to represent high-frequency details [17]. Second, how to make good use of the estimated illumination is equally important for indoor harmonization. Because, for the goal of image harmonization, what we ultimately need is the adjusted foreground appearance, and the illumination is only the cause of appearance changes, not the appearance itself. From the perspective of physical rendering, the foreground appearance adjustment requires both removing the original illumination of the foreground and rendering the new illumination effects in the foreground during the image harmonization process. Considering the complexity of this harmonization process, we therefore break it down into two simpler parts: *re-shading* and *albedo estimation*. For re-shading, it requires rendering a new shading from two inputs: the background illumination and the input foreground image with the original illumination. Compared to directly rendering the final appearance, re-shading is more conducive to rendering since it does not involve foreground textures. For albedo estimation, it is introduced with the aim of removing the effects of the original illumination. Finally, an albedo image that removes the original illumination and a shading image under the background illumination, which correspond to different components of the harmonized foreground appearance, are combined to finish the image harmonization process.

As a result, in this paper, we propose to explicitly estimate the three key physical terms (i.e., illumination, shading, and albedo), and formulate image harmonization in a physically meaningful form:

$$\hat{L}_M = f\left(I_b, M\right), \hat{A} = g\left(\tilde{I}_f\right), \tag{4}$$

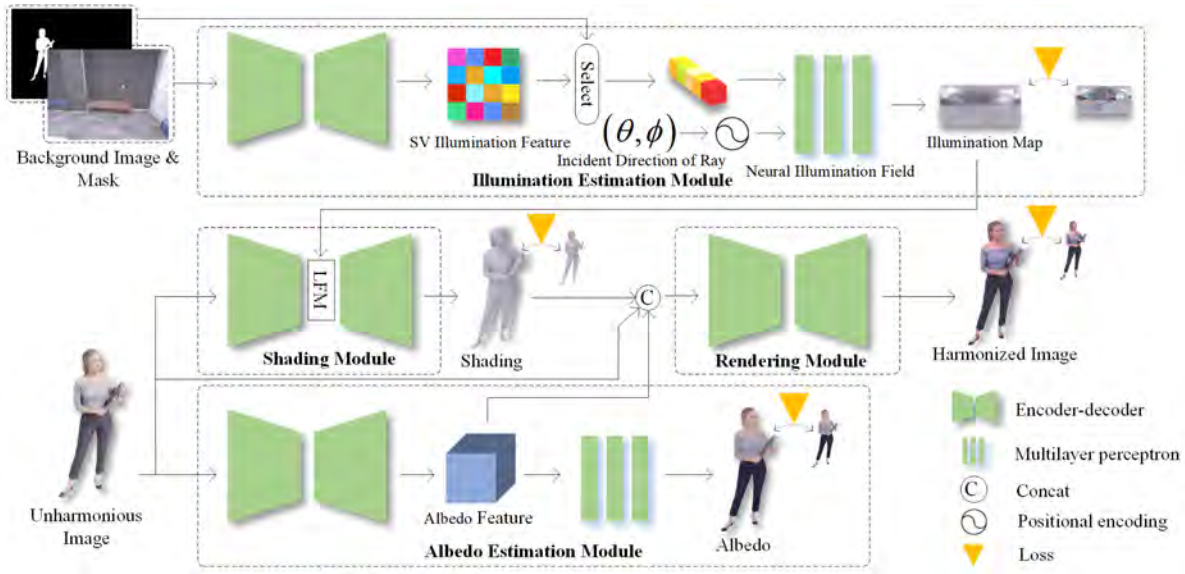


Fig. 7 An overview of the proposed framework for indoor harmonization. Inspired by the physical principle of image formation, it consists of four neural modules, which jointly learn SV illumination, shading, albedo, and rendering.

$$\hat{S} = h\left(\tilde{I}_f, \hat{L}_M\right), \hat{I}_f = r\left(\hat{A}, \hat{S}\right), \tag{5}$$

where f is an illumination estimation function that takes a background image I_b and a foreground mask M as input, g is an albedo estimation function whose input is an unharmonious foreground image $\tilde{I}_f$, h is a shading function whose input is $\{\hat{L}_M, \tilde{I}_f\}$, and r is a rendering function that takes the estimated albedo $\hat{A}$ and shading $\hat{S}$ as input and outputs the harmonized foreground image $\hat{I}_f$. The subscript M of $\hat{L}_M$ indicates that illumination L is located in the foreground region M. We model all these functions $\{f, g, h, r\}$ using neural networks. The details of the network architecture are described in Sec. 4.2.

4.2 Network Architecture

The overview of our proposed image harmonization network architecture is shown in Fig. 7. It is composed of four neural network modules, namely an illumination estimation module, a shading module, an albedo estimation module, and a rendering module. We elaborate on the details of each module below.

Illumination Estimation Module. As mentioned above, the Illumination Estimation Module (IEM) f takes a background image I_b and a foreground mask M as input, and estimates the illumination $\hat{L}_M$. However, it is very challenging to directly estimate a general illumination representation (i.e., illumination map) via an encoder-decoder-based neural network because a large number of parameters of illumination map (e.g., 16×32×3=1536 parameters) lead to a high-dimensional output space. Recently, Li et al. [17] sought to estimate a parametric representation of illumination, namely Spherical Gaussian (SG) based illumination (typically with 84 parameters). But, in practice, some problems still remain: 1) Spherical Gaussian-based illumination representation can only recover high-frequency details due to the limited amount of parameters, and some low-frequency details are lost; 2) Due to the various changes in the shape, size, and number of indoor light sources, there may be multiple solutions for the same illumination map, which leads to the instability of the model convergence.

Benefiting from the powerful complex signals modeling capability of neural fields [49, 50], we propose to construct a neural illumination field via a multilayer perceptron to avoid the above two problems. Specifically, we first construct an encoder-decoder network to encode the background image into a low-dimensional spatially-varying illumination feature F^{sv}_{illum}. Then, according to the placement position of the foreground object in the mask, we select the corresponding illumination feature F^{i}_{illum} from F^{sv}_{illum}. Finally,

the illumination feature F^i_{illum} combined with the incident direction (θ, ϕ) of the ray are fed to an MLP-based neural field (i.e., neural illumination field) to reconstruct a complete illumination map $\hat{L}_M$. Before (θ, ϕ) are fed into the neural illumination field, the positional encoding function $\gamma(p)$ is applied separately to each of (θ, ϕ) by introducing Fourier features:

$$\gamma(p) = \left(\cdots, \cos\left(2^K \pi p\right), \sin\left(2^K \pi p\right), \cdots\right), \quad (6)$$

where $K \in [0, L-1]$. We set $L = 128$ in our experiments and (θ, ϕ) are normalized to $(-1, 1)$. Note that once F^{sv}_{illum} is estimated for one background image, only the neural illumination field is used in IEM when the foreground is placed in different locations, which reduces the computational complexity of the model.

NIF-based illumination v.s. SG-based illumination. We will analyze and compare them in detail from the perspective of the basis and its projection coefficient. First of all, although the SG function itself can represent high-frequency and low-frequency signals (by changing its bandwidth parameter), only a few SG functions are used to mainly represent high-frequency illumination (i.e., those light sources) to avoid a high-dimensional output space in practice [17]. In addition, since the SG functions do not satisfy orthogonality, the solution to the same illumination is not unique. In contrast, the proposed NIF, which is designed as an MLP with Fourier features-based positional encoding, can alleviate these problems. The key here is to introduce numerous *Fourier features* [49, 76], which is actually equivalent to introducing a set of orthogonal bases. The illumination feature fed to the NIF can essentially be interpreted as the coefficients projected onto these orthogonal bases. The illumination feature (i.e., the coefficients) and the Fourier features (i.e., the bases) are passed through the NIF to capture the complex interactions between them, thereby producing an illumination map. Note that the number of illumination feature here does not need to be exactly the same as the number of Fourier features, and can even be less than the number of Fourier features. In other words, these Fourier features allow our NIF to model both high and low-frequency signals from a *low-dim* illumination feature. In addition, for a given illumination map, the projection coefficients for these orthogonal bases (i.e., the Fourier features) are also unique, which may allow the model to converge more stably. Finally, our experiments also show that NIF-based IEM outperforms SG-based IEM both quantitatively and qualitatively.

The structure of the IEM is shown in Fig. 7. It consists of an encoder-decoder network and an MLP. The MLP is composed of four fully-connected layers and each layer is followed by a rectified linear activation function.

Shading Module. Once the estimated illumination $\hat{L}_M$ is obtained, it is fed into the Shading Module (SM) h together with the input unharmonious image $\tilde{I}_f$ to achieve the shading result $\hat{S}$. Specifically, the encoder first extracts the object feature from $\tilde{I}_f$, as shown in Fig. 7. Then, we use a Lighting guided Feature Modulation (LFM) block [65] to modulate the object feature with $\hat{L}_M$. Finally, we feed the modulated object feature containing the background illumination to a decoder, resulting in a harmonized shading image.

Albedo Estimation Module. The Albedo Estimation Module (AEM) takes an unharmonious foreground image $\tilde{I}_f$ as input to estimate albedo $\hat{A}$. It is used to remove the original illumination of the foreground object. In principle, the estimated albedo should be used as input to the rendering module. However, due to the decrease in the number of feature channels, decoding albedo feature F_{albedo} to albedo $\hat{A}$ is actually a feature compression process, which may result in information loss (especially the loss and distortion of textures in albedo). This inaccurate estimate is ultimately transmitted to our harmonized result through the rendering module, causing a degradation in model performance. So instead of exporting the estimated albedo $\hat{A}$ to the rendering module, we feed the albedo feature F_{albedo} into the rendering module to alleviate this problem. We refer the reader to Sec. 5.5 for more details. In addition, due to the possible distortion of the texture in albedo estimation, the input unharmonious foreground image $\tilde{I}_f$ is combined with the albedo feature F_{albedo} and fed into the rendering module.

The structure of the AEM is shown in Fig. 7. It is composed of an encoder-decoder network and an MLP-based albedo decoder. They are

used to estimate the albedo feature and albedo, respectively.

Rendering Module. As mentioned above, the inputs to the Rendering Module (RM) are now the shading image $\hat{S}$, the albedo feature F_{albedo}, and the input unharmonious foreground image $\tilde{I}_f$. We feed them into an encoder-decoder network to obtain the final harmonized image $\hat{I}_f$ as shown in Fig. 7.

In our proposed harmonization framework, all encoder-decoder networks adopt a U-Net-like structure with skip connections. It mainly consists of two parts: down-sampling blocks and up-sampling blocks. See the supplementary material for the detailed network parameters of the IEM, SM, AEM, and RM.

4.3 Loss Functions

Our loss function consists of two parts: illumination estimation loss and reconstruction loss. The illumination estimation loss l_{illum} is defined as a logL_2 loss:

$$l_{illum} = \left\| \log(L_M + 1) - \log\left(\hat{L}_M + 1\right) \right\|_2^2, \quad (7)$$

where L_M denotes the illumination map Ground Truth (GT). In addition, since per-pixel illumination GT is available in the Replica dataset, the l_{illum} is also used for other estimated illumination at locations where the foreground is not placed.

The $\mathcal{L}_2$ loss is adopted for shading reconstruction. Besides, inspired by [77], the SSIM metric is also used to supervise the learning of the AEM. Thus, the reconstruction loss for albedo is defined as,

$$l_{rec}^{albedo} = \left\| A - \hat{A} \right\|_2 + \lambda(1 - \text{SSIM}(A, \hat{A})), \quad (8)$$

where A denotes the albedo GT. We set $\lambda = 1$ in our experiment. Similarly, the l_{rec}^{albedo} is also used as the shading reconstruction loss $l_{rec}^{shading}$ and the foreground reconstruction loss l_{rec}^{render}. As a result, the final total loss l_{total} is defined as follows:

$$l_{total} = l_{illum} + l_{rec}^{albedo} + l_{rec}^{shading} + l_{rec}^{render}. \quad (9)$$

5 Experiments

In this section, we first introduce experimental setups, including evaluation metrics and training details. We next compare our method with these state-of-the-art methods both qualitatively and quantitatively. Then, a user study on real data is conducted to validate the effectiveness of our method. We also compare our neural illumination field with parametric illumination. Finally, we perform the ablation study to demonstrate the contribution of each component of our framework in isolation.

5.1 Experimental Setups

Evaluation metrics. For evaluation of image harmonization, fSSIM[78], LPIPS[79], and fPSNR are selected. Note that the prefix f indicates that the metric is only calculated on the foreground region. Besides, we also choose the MPS [80] (Mean Perceptual Score), which is a normalized average of the fSSIM and LPIPS, as the determinant metric to rank these methods. Since LPIPS is calculated over the entire image, its value is relatively small. To balance the weight between fSSIM and LPIPS, we modified the original MPS by multiplying LPIPS by 10:

$$\text{MPS} = 0.5 * (\text{fSSIM} + (1 - 10 * \text{LPIPS})). \quad (10)$$

Training details. We implement the model via the PyTorch framework [81] and train the model on 1 Intel Xeon Gold 6246 CPU and 2 NVIDIA TITAN RTX GPUs. The parameters of our networks are initialized with Kaiming Uniform Initialization [82]. We optimize the model parameters by the Adam optimizer for 80 epochs, with a learning rate = 1e-4, betas = (0.9, 0.99). The batch size is set to 8 to maximize GPU memory utilization. Note that due to the directionality of the illumination, we only applied data augmentation techniques (including random rotation and flipping) to AEM and RM.

5.2 Comparison with State-of-the-Art Methods

To validate the superiority of the proposed method, we compare it with four representative open-source image harmonization methods,

Fig. 8 Qualitative comparison of different methods on the test set. We show representative examples with close-up details focusing on brightness, color, shading, and artifacts. Our method outperforms all other approaches with more accurate and sharper details. Zoom in for a better view.

namely Lalonde and Efros [4], Sg-MMH [6], Harmonizer [1] and DCCF [2]. Note that Sg-MMH, Harmonizer, and DCCF are all learning-based state-of-the-art methods. To make a fair comparison, we re-train the three methods on our training set according to the training configurations given by the authors. See the supplementary for their training details. When their losses converge, we report their results on the test set. In addition, we also report the results of their released pre-trained models on the test set.

Method	MPS↑	fSSIM↑	LPIPS(×10)↓	fPSNR↑
Lalonde and Efros [4]	0.775	0.737	0.187	14.84
Sg-MMH*[6]	0.823	0.784	0.138	17.57
DCCF*[2]	0.836	0.796	0.124	18.16
Harmonizer*[1]	0.857	0.812	0.099	18.30
Sg-MMH [6]	0.885	0.854	0.084	22.27
Harmonizer [1]	0.890	0.855	0.075	21.28
DCCF [2]	0.897	0.858	0.064	22.72
Ours	**0.921**	**0.894**	**0.052**	**23.58**

Table 2 Quantitative results of different methods on the test set. All the methods are ranked by MPS in ascending order. The best results are marked in bold. The second-best results are underlined. $*$ denotes that the pre-trained model released by the authors is used. ↑ denotes the higher the better, and ↓ denotes the lower the better. It can be seen that our method achieves the best MPS result.

Quantitative comparison. The quantitative comparison on the test set is shown in Tab. 2. First, it can be observed that our method achieves the best MPS of 0.921, a 0.024 improvement over that of the second-best method (DCCF). Specifically, for LPIPS and fSSIM, our results achieve approximately 18.8% and 4.2% improvements over DCCF, respectively. The MPS of Lalonde and Efros's method is the worst at only 0.775.

Second, benefiting from our dataset, all the re-trained Sg-MMH, Harmonizer, and DCCF outperform their own pre-trained models by a large margin, with improvements of 7.5%, 3.9%, and 7.3% in MPS, respectively. This is mainly due to the fact that each input and GT pair in their original training datasets only contains color and brightness variations. In contrast, each pair in our dataset also contains additional shading variations caused by varying illuminations. It implies that the new setting introduced in our constructed dataset poses a greater challenge to existing image harmonization methods.

Method	Sg-MMH [6]	Harmonizer [1]	DCCF [2]	Ours
# Parameters (M)	39.502	4.727	18.092	9.109
Inference time (ms)	13	9	40	36

Table 3 Comparison in terms of the number of model parameters and inference time.

In Tab. 3, we also compare the number of model parameters and inference time of these learning-based methods. All methods are running on the test set using a single TITAN RTX GPU. The average inference time (excluding image load and write time) for a single image with an original resolution of 640×480 is reported, except that Sg-MMH runs on 512×512 images since their method only supports fixed resolutions. Although DCCF achieves the second-best MPS result, its number of model parameters is also the second-largest, reaching 18.092M. In contrast, our model not only achieves the best MPS result but also has the second-smallest number of parameters, approximately half that of DCCF. However, our model has the second-longest inference time of 36 ms. This is mainly because our model is composed of four modules, which need to be calculated sequentially, resulting in more time overhead. The Harmonizer has the smallest number of parameters and the fastest inference time, which are 4.727M and 9 ms respectively. Sg-MMH has the largest number of parameters (39.502M), but its inference time is only around 13 ms, mainly because it consists of only one simple U-Net.

Qualitative comparison. The qualitative comparison on the test set is shown in Fig. 8. We show the results with different illumination conditions on typical indoor scenes, including the classroom, shopping mall, and bedroom. It can be observed that our method produces more realistic results, where the foreground is compatible with the background image. Taking Fig. 8 (a) as an example, the foreground object of the input composite image appears to be illuminated from the right of the image, as is evident in the shadows of the man's face. Given that the background scene consists of multiple light sources evenly located on the ceiling, the man should be under smooth illumination at P_1 and P_2. The Sg-MMH, Harmonizer, and DCCF fully preserve the original illumination on the foreground object. Both DCCF and Harmonizer are struggling to remove them, but some shadows still remain on the white cloth. Lalonde and Efros' method mistakenly transfers the color of the glass walls to the man and thus produces greenish results. In contrast, our result is closer to the ground truth in terms of brightness, color, and shading. The same can be observed in Fig. 8 (b). For the lady at P_1, these competing methods preserve some shadows from the original illumination. In addition, their results are also overall darker compared to the ground truth due to the presence of the original shadow. Our result is closer to ground truth in terms of brightness and shading, but it is slightly less yellow. Fig. 8

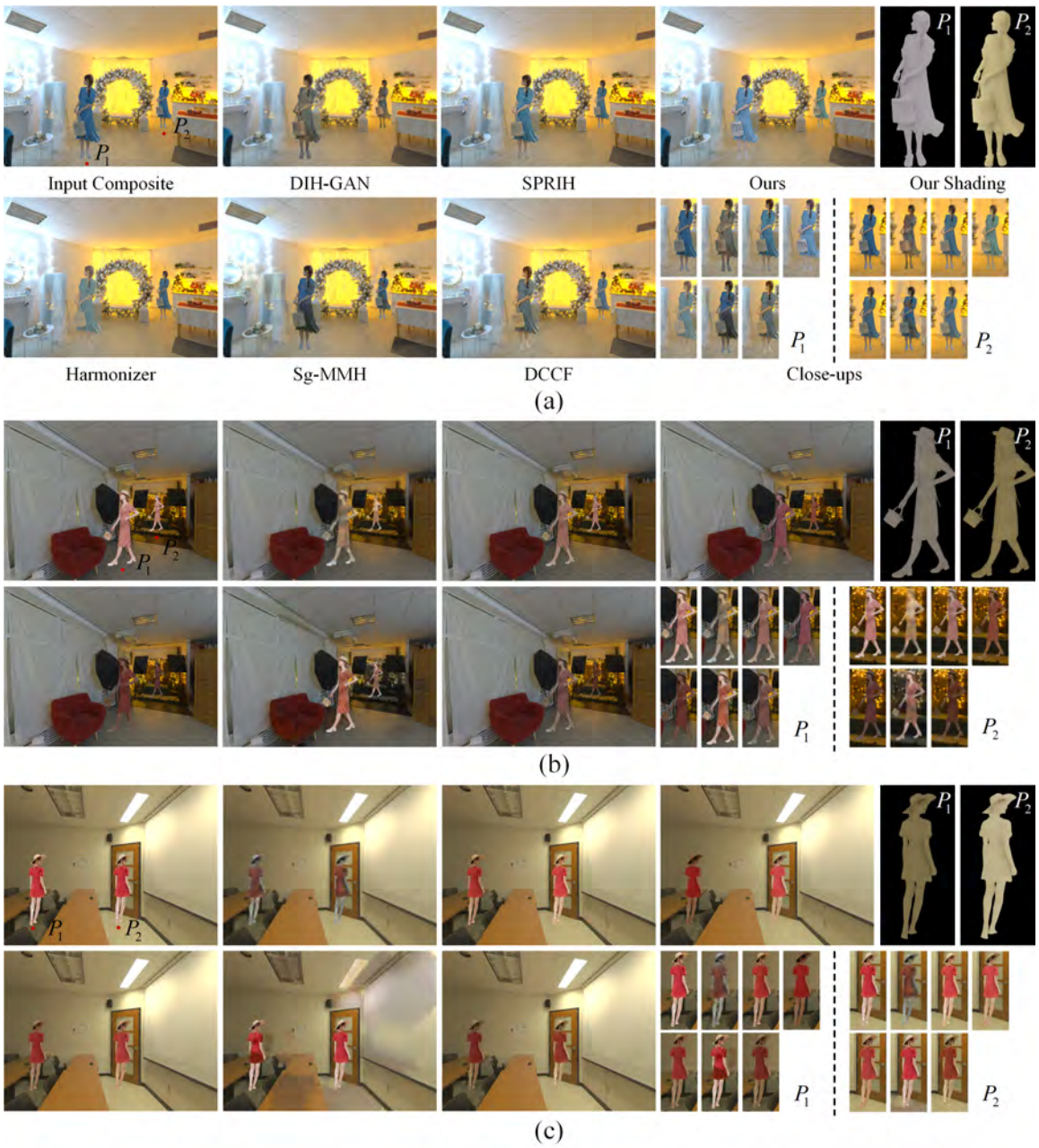


Fig. 9 Qualitative comparison of five state-of-the-art methods and our proposed method on real composite images. We show representative examples with different illumination conditions. Due to the lack of ability to perceive 3D information (that is, spatially-varying illumination), these competing methods may produce obviously erroneous results that do not match the lighting distribution of the real physical world. In contrast, our method produces more accurate and sharper results.

(c) shows a challenging example where the light source is not directly visible in the background image. However, it can be seen from the bedroom floor that the intensity of light gradually decreases from P_2 to P_1. The results produced by these competing methods for the two locations are not much different in brightness. In contrast, our results are in line with the light intensity distribution of the bedroom.

Method	Score (Q1)↑	Score (Q2)↑	Score (Q3)↑	Total Score↑
DIH-GAN [13]	0.118	0.217	0.103	0.146
Sg-MMH [6]	0.395	0.574	0.800	0.590
SPRIH [38]	1.444	1.202	1.506	1.384
DCCF [2]	1.435	1.231	1.574	1.413
Harmonizer [1]	1.734	1.390	1.140	1.421
Ours	**1.968**	**1.867**	**1.669**	**1.835**

Table 4 User study on real data. All the methods are ranked by the total score in ascending order.

5.3 User Study on Real Data

A user study on real data is conducted to verify the superiority of our proposed method for indoor harmonization. First, we made a total of 52 real composite images of which both the foreground images and the background images are collected from the Internet. Specifically, 18 real foreground humans for clothing display are collected from Taobao [83] and cover different genders (male, female), ages, poses, and clothing. 15 indoor background images are collected from Poly Haven [74], HDR MAPS [75], and Laval [39], and cover various indoor scenes, such as classrooms, workshops, and offices. Note that the background images collected from Laval are not used in our constructed indoor harmonization dataset. All the foreground images and background images are captured by real professional cameras. Next, we select three open-source state-of-the-art methods (i.e., DCCF [2], Harmonizer [1] and Sg-MMH [6]) as the competing methods. We used their publicly released pre-trained models to process these composite images. In addition, we also compare two methods that are close to our work, namely SPRIH [38] and DIH-GAN [13]. Their results are provided by the authors.

Then, for each composite image processed by these six methods, we ask 31 individuals to score the visual quality. As inspired by [10], the following three questions are considered for scoring: (1) Are the brightness and color of the foreground and background consistent; (2) Are the shadings/shadows 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).

Finally, we report the results in Tab. 4. It can be seen that we achieve the biggest relative improvement in Q2. This is mainly because our proposed harmonization framework effectively removes the effects of original illumination, and numerous illumination variations are also included in our training data.

In Fig. 9, we also show qualitative results of different methods. Fig. 9 (a) shows a very challenging example. The lady of the input composite image appears to be illuminated by a dark and smooth illumination in her original environment. However, the background scene is illuminated by two bright light sources, a cold white light source on the left side of the image and a warm yellow light source in the middle of the image. We placed the lady near each of the two light sources, respectively. For the lady at P_1 who is close to the white light source, both Harmonizer and DCCF mistakenly transfer the color of the yellow light source to the lady. SPRIH still shows a weak yellow color. The results of Sg-MMH and DIH-GAN also contain some yellow artifacts, which are visually undesirable. In fact, due to the lack of ability to perceive 3D information (i.e. spatially-varying illumination), these methods are easily misled by the illusion that the lady is closer to the yellow light source in the view of 2D image space. In contrast, our method produces bright white results that appear to be illuminated by the cold white light source. In addition, for the lady at P_2 who is close to the yellow light source, the results produced by these competing methods are too dark, and the yellow color is not obvious enough. Our results show a more pronounced yellow appearance. The same can also be observed in Fig. 9 (b). Fig. 9 (c) shows an example where the background only contains a yellow light source. In the real physical world, the energy of light decreases rapidly with the square of the distance. However, the results of these competing methods show that there is not much difference in the brightness of the lady at the two positions. In addition, since Sg-MMH is a mask-free harmonization method, they seem to identify the background as an inharmonious region and adjust it, resulting in many unpleasant artifacts. In contrast, our harmonized results are not only visually pleasing but consistent with the light intensity distribution of the background scene.

5.4 Comparison with Parametric Illumination

The efficacy of our neural illumination field (NIF) based IEM is compared with the parametric

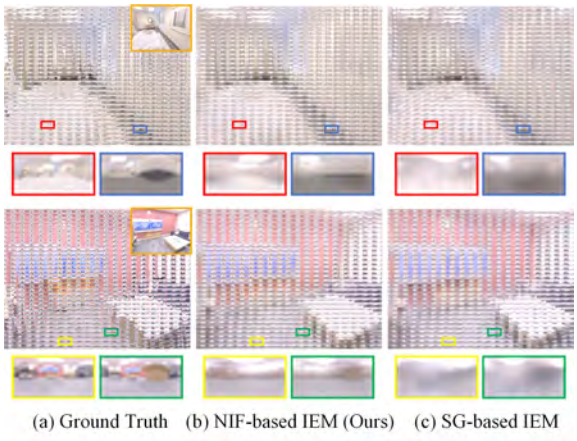


Fig. 10 Qualitative comparisons of SG-based IEM and NIF-based IEM. As shown in the close-ups, our NIF-based IEM produces more accurate illumination results in terms of the number, shape and position of light sources and background texture.

illumination-based IEM for spatially-varying illumination estimation. Here, we adopt spherical Gaussian (SG) [17] as the parametric illumination representation for comparison. NIF-based IEM and SG-based IEM are both trained and tested on our per-pixel illumination dataset, which is constructed from the Replica dataset. For a fair comparison, the channel dimension of the illumination feature produced by the encoder-decoder network in both IEMs is set to 84. The quantitative results are reported in Tab. 5. Compared with the SG-based IEM, the MAE and RMSE of our NIF-based IEM decreased by 7.9% and 3.0%, respectively. In Fig. 10, we show the qualitative results. In addition, we also combine the two IEMs with other modules to compare their impacts on indoor harmonization, as shown in Fig. 11. For example, on the left side of the illumination map in Fig. 11 (a), as indicated by the red arrow, the result of SG-based IEM loses the light source. Correspondingly, at the red dased box in the foreground (mainly affected by the left half of the illumination map), their rendering appears dark. In contrast, our result based on NIF-based IEM is slightly brighter, which is closer to GT. The same can also be seen in Fig. 11 (b).

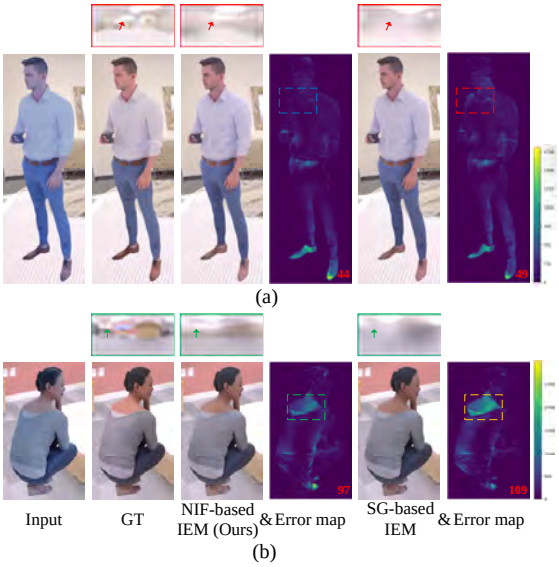


Fig. 11 The impact of SG-based IEM and NIF-based IEM on image harmonization. Note that due to the inaccurate illumination estimation of SG-based IEM (e.g., the light sources are missing at the red arrow and green arrow), this will inevitably cause inaccurate renderings at the corresponding position in the foreground (such as the red and orange dashed boxes). In contrast, our renderings based-on NIF-based IEM show more accurate results as shown in the error maps. The error map is calculated by squared error, and the red number indicates the mean of the error map.

	MAE ↓	RMSE↓
SG-based IEM	0.164	0.405
NIF-based IEM (Ours)	**0.151**	**0.393**

Table 5 Quantitative comparison of using SG-based IEM and NIF-based IEM for spatially-varying illumination estimation.

5.5 Model Analysis

Ablation on our framework. In Tab. 6, we perform an ablation study to demonstrate the effect of each component in our framework.

Configuration	MPS↑	fSSIM↑	LPIPS(×10)↓
Baseline model (=Model 0)	0.887	0.851	0.078
Model 0 + SSIM Loss (=Model 1)	0.891	0.870	0.089
Model 1 + IEM (=Model 2)	0.897	0.866	0.072
Model 2 + SM (=Model 3)	0.908	0.880	0.064
Model 3 + AEM (=Ours)	0.921	0.894	0.052

Table 6 Ablation study on the proposed framework.

We start with an end-to-end model commonly used in the field of image harmonization as our

baseline model (i.e., Model 0). Specifically, the baseline model consists of only a rendering module whose input is a composite image and a foreground mask, and the output is a harmonized image. Only the basic $\mathcal{L}_2$ loss is included to train the baseline model. It achieves an MPS of 0.887. When the SSIM loss (Model 1) is added to the baseline model, we observe a 0.004 MPS improvement where the gain comes from SSIM.

We next add the illumination estimation module (Model 2) to Model 1. Here, the LFM is also added to the bottleneck of the rendering module to make use of the estimated illumination. Experimental results show a 0.006 improvement in MPS due to the utilization of illumination information.

We then add the shading module (Model 3) to Model 2. At this point, the LFM is moved to the shading module. The generated foreground shading is later combined with the unharmonious foreground image as input to the rendering module. Compared with directly using illumination information, the way of converting illumination information to the shading, which is approaching the final harmonious image, can bring a significant improvement of 0.011 in MPS.

Finally, We add an albedo estimation module (Ours) to Model 3. The current model shows the largest improvement of 0.013 in MPS. This is mainly because the model can effectively remove the effect of original illumination, as shown in Fig. 12. For example, the blue box in the input image shows that the foreground man is originally illuminated by the light source on the right, as is evident in the shading variation of the man's clothes. However, the results of ours without AEM more or less preserve the original illumination effects. In contrast, ours with AEM not only effectively removes the original illumination effects of the man, but also re-render it under a smooth background illumination.

Effect of albedo feature. We feed the albedo feature into RM instead of the estimated albedo itself, because the estimated albedo sometimes suffers from texture distortion problems as shown in Fig. 13(b), especially at the blue dashed box. This erroneous estimate may be passed to the harmonized result through RM, and ultimately lead to the degradation of model performance, as shown in Tab. 7 and Fig. 13(c). In contrast, the albedo feature contains more rich information and can effectively alleviate this problem because there is no feature reduction process (that is, compressing from albedo feature to albedo). As shown in Fig. 13(d), even if the estimated albedo has some problems, our proposed framework (i.e., albedo feature to RM) still shows more accurate results.

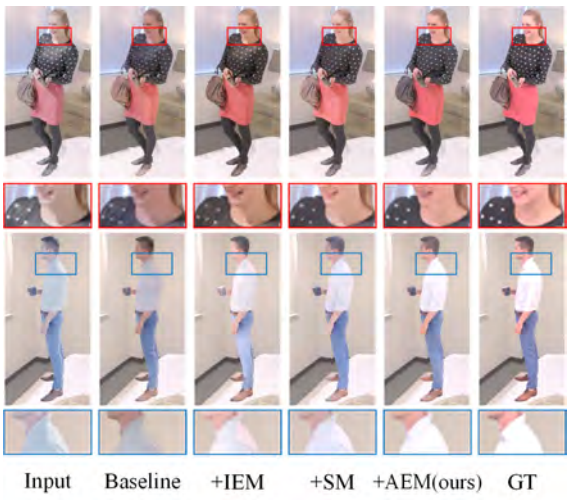


Fig. 12 Qualitative results of ablation study. Ours without AEM (i.e., Baseline model, Model 1 + IEM, Model 2 + SM) retain the effect of the original illumination, as shown in the red and blue boxes. They also produce some distorted textures. In contrast, ours with AEM is closer to ground truth.

Configuration	MPS↑	fSSIM↑	LPIPS(×10)↓
Ours w. a shared backbone	0.913	0.884	0.058
Ours w.o. albedo feat.	0.914	0.887	0.060
Ours	0.921	0.894	0.052

Table 7 Quantitative results of ours without albedo feature and ours with a shared backbone in SM and AEM.

Effect of a shared backbone in AEM and SM. Since AEM and SM share the same input, we study the effect of a shared backbone in AEM and SM. Specifically, we first use a shared encoder network to extract features from the input image, and then send them to their respective decoder networks to obtain shading and albedo features respectively. The other modules of our model remain unchanged. Although the model with a shared backbone is more concise and has fewer parameters, its MPS is 0.008 lower than our MPS as shown in Tab. 7.

The performance of our framework on a general object harmonization benchmark.

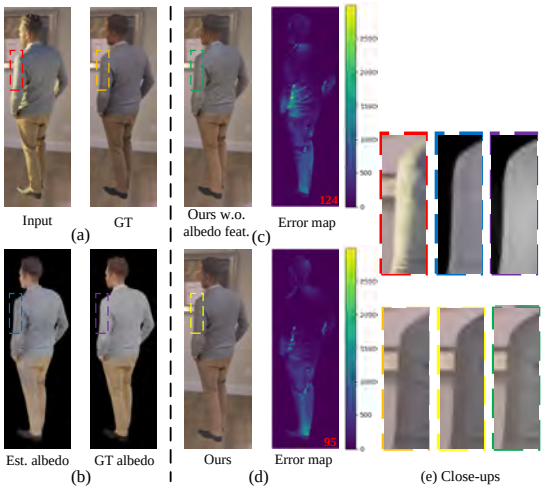


Fig. 13 Qualitative comparisons w.r.t. albedo feature. Note the man's highlights on the left in the red dashed box of the input image. Ours w.o. albedo feature still retains part of the highlights, as shown in the green dashed box. In contrast, ours shows more accurate results in the yellow dashed box compared to GT. The error map is calculated by squared error, and the red number indicates the mean of the error map.

To further investigate the harmonization performance of our framework on general objects, we select a general object harmonization dataset HVIDIT [18] as the benchmark dataset because it contains shading variations caused by color temperature and direction changes of a single global point light source. We train different versions of our framework on the HVIDIT and report the test results in Tab. 8.

Method	PSNR↑	SSIM↑	fMSE↓
Retinex-Net [84]	36.32	0.9321	1603.21
DIH [5]	36.62	0.9310	1207.03
S^2AM [6]	36.24	0.9206	1230.92
DoveNet [7]	36.80	0.9585	1186.19
IIH [18]	41.55	0.9914	800.92
Ours (Model 1)	41.71	0.9945	796.80
Ours (Model 1 + Data aug.)	42.04	0.9950	762.71
Ours (Model 2)	**42.40**	**0.9954**	**695.37**

Table 8 Quantitative results of our framework on the test set of HVIDIT. The results of all other competing methods are directly copied from [18]. It can be seen that our method achieves the best results.

Specifically, we first trained our baseline model with SSIM loss (i.e., Model 1) on the HVIDIT dataset, and we achieved a test result of 796.80 in fMSE, which is 4.12 lower than that of IIH. Then, some data augmentation techniques (i.e., random flipping and rotation) are added to increase the diversity of training data, bringing the fMSE down to 762.71. Finally, we further added the IEM to Model 1 (i.e., Model 2) and used the ground truth color temperature of background illumination to supervise the learning of the IEM. With the use of illumination information, our method achieves the best fMSE of 695.37, a 13.2% decrease compared to IIH. Note that the ground truth color temperatures are obtained from the VIDIT dataset [80] that HVIDIT is based on, and the HVIDIT itself does not provide them. Besides, since the background scene from VIDIT is only lit by a global point light source, our current IEM removes the NIF used for spatially varying illumination estimation and only consists of an encoder network for estimating a global color temperature. The HVIDIT dataset contains 5 pre-defined color temperature values [2500K, 3500K, 4500K, 5500K, 6500K], and each background image corresponds to a specific color temperature. Therefore, here we use the cross-entropy loss function as l_{illum} instead of the logL_2 loss in Eq. 7. In Fig. 14, we also show qualitative results.

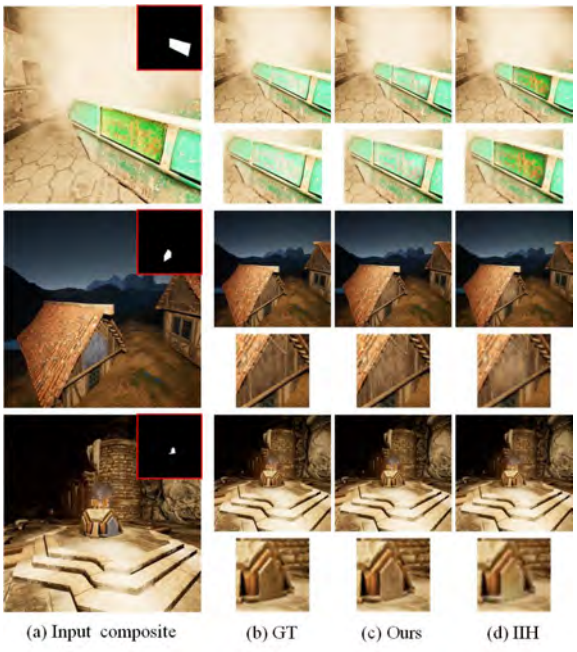


Fig. 14 Qualitative results of our proposed method and IIH on the HVIDIT. Compared with IIH, our results are closer to GT.

Generalization to non-human objects. To evaluate the generalization performance of our

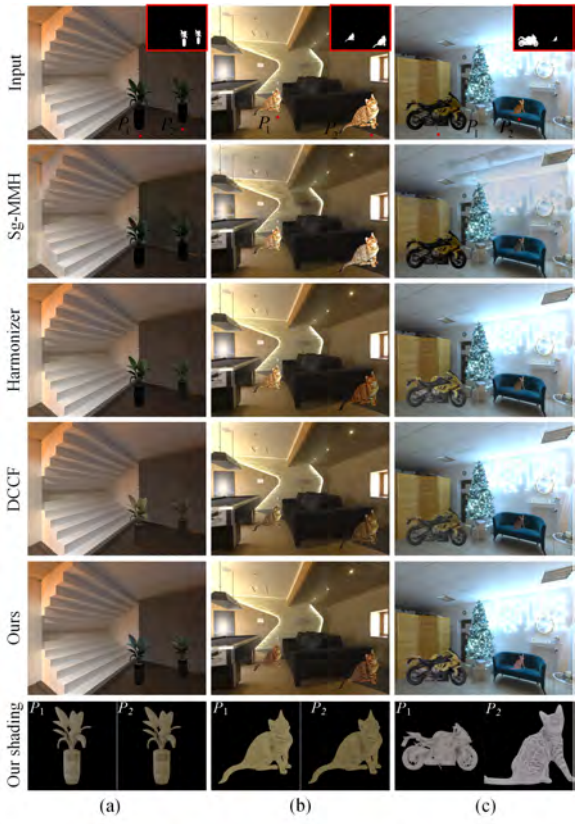


Fig. 15 Generalization to non-human objects.

method (trained on our indoor harmonization dataset) to non-human objects, Fig.15 shows the results of our method on some common objects, such as plants, cats, and motorcycles. The results of other methods, obtained by their released pre-trained models, are also given for comparison. It can be observed that our method produces more reasonable results in terms of brightness, color and shading, which are consistent with the illumination distribution of indoor scenes. Taking Fig.15(a) as an example, the potted plant in the input image was originally in a dark environment and was mainly illuminated by a light source coming from its right side, which can be seen from the shadows on the leaves. However, the main light source in the background image is from the left side and the light intensity of the background decreases from left to right. The results of Sg-MMH at P_1 and P_2 have almost no difference in brightness. The results of Harmonizer are interesting. Although the leaves at P_1 is brighter than that at P_2, the pot at P_1 is darker than that at P_2, which is self-contradictory. The results produced by DCCF are reasonable in terms of brightness, but the color of the leaves appears too yellow. In contrast, the results produced by our method appear to be consistent with the background illumination in terms of brightness and color, and our shadings also reflect that the main light source of the background is from the left side. The same can be observed in Fig.15(b). Fig.15(c) shows an example where the background scene is illuminated by some cold white light sources. The results of these competing methods are almost too dark. In contrast, our method produces results that appear to be illuminated by those cold white light sources.

5.6 Discussion

Discussion about our dataset. 1) Our current dataset only focuses on the human body. There are two main reasons: first, in actual scenarios, whether for professional or amateur Photoshop users, the main object of photo editing is often people [36]; second, the construction of large-scale datasets includes several very time-consuming steps such as data collection, cleaning, and processing, which also require a lot of human labor. Anyway, through the human body, this work investigates a new problem (i.e., spatially-varying illumination-aware indoor harmonization). 2) We also note that the latest work [85] has photometrically calibrated the Laval indoor HDR dataset to obtain more accurate HDR panoramas. Based on the latest dataset, more panoramas can be exploited to further increase the diversity of our indoor harmonization dataset. 3) The background images sometimes exhibit some lens distortion. It is mainly introduced in the process of capturing images using the camera with a fisheye lens, especially for the Laval indoor HDR dataset. In addition, some distortion could also be introduced during the image-stitching process for an HDR panorama.

Discussion about our proposed framework. 1) Our method sometimes produced a texture-less appearance as shown in Fig. 16 (especially in the blue box). This is mainly because our SM does not effectively perceive the geometry of the input foreground image, thus rendering too smooth shading. Note that the illumination map is also one of

two inputs of the SM, but even a completely wrong illumination estimate will cause shading variations due to changes in the object's surface geometry. In other words, the estimated illumination is not the main cause of smooth shading. We also give our albedo result for reference. It can be seen that our estimated albedo completely removes the original illumination effects. Finally, the smooth shading and the albedo feature are combined to produce a texture-less appearance through the RM. In the future, we will focus on improving the model's ability to perceive geometric changes in foreground objects. For example, we could consider introducing depth information or using off-the-shelf depth estimators in our SM to improve the rendered shading result. 2) Besides, the inaccurate illumination estimation causes our final harmonized result to look slightly warmer as shown in Fig. 16. In addition to directly improving the IEM, another interesting and simple alternative is to incorporate image color correction into our framework as a post-processing step. For example, we can train a lightweight network [86] to predict a nonlinear mapping function from the initial harmonized image, and then combine it with the polynomial kernel function to perform global color correction on the initial result. 3) Regarding the FOV mismatch between foreground and background, at the methodology level, our proposed framework first extracts illumination from the background and then uses the illumination to adjust the foreground appearance. In other words, we do not directly process the composite image that contains both the foreground and background but process the foreground and background separately. In contrast, most existing image harmonization methods [2, 5–7, 27] directly process the composite image, which may be sensitive to the FOV mismatch. 4) Few works consider both image harmonization and shadow generation tasks. In the future, one of the promising extensions of our framework is to integrate the shadow generation module that can be guided by our estimated spatially-varying illumination. 5) Introducing more complex reflectance properties into our framework, such as the specular BRDFs [87], could further improve the realism of image harmonization.

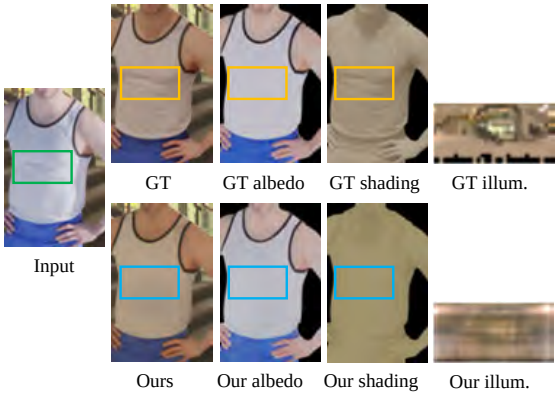


Fig. 16 The limitation of our proposed method. Since the SM does not effectively perceive some subtle surface geometric changes (such as the wrinkles of clothes in the green box), it cannot render accurate shading that reflects the geometric changes of the object. This ultimately causes our harmonized result to look texture-less, as shown by the blue box. In addition, the estimated inaccurate illumination causes our harmonized result to look slightly warmer than GT.

6 Conclusion

In this paper, we have contributed a large-scale photo-realistic indoor harmonization benchmark dataset. In the dataset construction, we use an object placement formula to place the foreground object in the background at a reasonable size. We also present a novel physically-inspired, learning-based indoor harmonization framework, which allows using perceived spatially-varying illumination to adjust the foreground appearance. It consists of four compact neural modules, among which the albedo module can effectively remove the effects of original illumination. An MLP-based neural field with positional encoding is also included in the illumination estimation module to recover the illumination more accurately. We qualitatively and quantitatively demonstrate the efficacy of our proposed framework at better matching the lighting distribution of the background compared to the other competing methods.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 62031023.

Data Availability. Both our large-scale indoor harmonization benchmark dataset and real evaluation dataset are available at https://github.com/waldenlakes/IndoorHarmony-Dataset. The raw

data used to construct our datasets are available from the Laval Indoor HDR dataset [39], the Replica dataset [67], Poly Haven [74], HDR MAPS [75], 3D People [68], and Taobao [83].

References

[1] Ke, Z., Sun, C., Zhu, L., Xu, K., Lau, R.W.: Harmonizer: Learning to perform white-box image and video harmonization. In: European Conference on Computer Vision (2022). Springer

[2] Xue, B., Ran, S., Chen, Q., Jia, R., Zhao, B., Tang, X.: DCCF: Deep comprehensible color filter learning framework for high-resolution image harmonization. In: European Conference on Computer Vision (2022). Springer

[3] Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer graphics and applications **21**(5), 34–41 (2001)

[4] Lalonde, J.-F., Efros, A.A.: Using color compatibility for assessing image realism. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8 (2007). IEEE

[5] Tsai, Y.-H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.-H.: Deep image harmonization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3789–3797 (2017)

[6] Ren, X., Liu, Y.: Semantic-guided multi-mask image harmonization. In: European Conference on Computer Vision (2022). Springer

[7] Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: DoveNet: Deep image harmonization via domain verification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)

[8] Cun, X., Pun, C.-M.: Improving the harmony of the composite image by spatial-separated attention module. IEEE Transactions on Image Processing **29**, 4759–4771 (2020)

[9] Cong, W., Tao, X., Niu, L., Liang, J., Gao, X., Sun, Q., Zhang, L.: High-resolution image harmonization via collaborative dual transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18470–18479 (2022)

[10] Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z.: Ssh: A self-supervised framework for image harmonization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4832–4841 (2021)

[11] Guo, Z., Gu, Z., Zheng, B., Dong, J., Zheng, H.: Transformer for image harmonization and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

[12] Kajiya, J.T.: The rendering equation. In: Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques, pp. 143–150 (1986)

[13] Bao, Z., Long, C., Fu, G., Liu, D., Li, Y., Wu, J., Xiao, C.: Deep image-based illumination harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18542–18551 (2022)

[14] Gardner, M.-A., Hold-Geoffroy, Y., Sunkavalli, K., Gagné, C., Lalonde, J.-F.: Deep parametric indoor lighting estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7175–7183 (2019)

[15] Garon, M., Sunkavalli, K., Hadap, S., Carr, N., Lalonde, J.-F.: Fast spatially-varying indoor lighting estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6908–6917 (2019)

[16] Zhan, F., Zhang, C., Hu, W., Lu, S., Ma, F., Xie, X., Shao, L.: Sparse needlets for lighting estimation with spherical transport loss. In: Proceedings of the IEEE/CVF International

Conference on Computer Vision, pp. 12830–12839 (2021)

[17] Li, Z., Shafiei, M., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2475–2484 (2020)

[18] Guo, Z., Zheng, H., Jiang, Y., Gu, Z., Zheng, B.: Intrinsic image harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16367–16376 (2021)

[19] Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003 Papers, pp. 313–318 (2003)

[20] Pitie, F., Kokaram, A.C., Dahyot, R.: 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, pp. 1434–1439 (2005). IEEE

[21] Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., Xu, Y.-Q.: Color harmonization. In: ACM SIGGRAPH 2006 Papers, pp. 624–630 (2006)

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

[23] Pitié, F., Kokaram, A.C., Dahyot, R.: Automated colour grading using colour distribution transfer. Computer Vision and Image Understanding **107**(1-2), 123–137 (2007)

[24] Sunkavalli, K., Johnson, M.K., Matusik, W., Pfister, H.: Multi-scale image harmonization. ACM Transactions on Graphics **29**(4), 1–10 (2010)

[25] Xue, S., Agarwala, A., Dorsey, J., Rushmeier, H.: Understanding and improving the realism of image composites. ACM Transactions on Graphics **31**(4), 1–10 (2012)

[26] Sofiiuk, K., Popenova, P., Konushin, A.: Foreground-aware semantic representations for image harmonization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1620–1629 (2021)

[27] Ling, J., Xue, H., Song, L., Xie, R., Gu, X.: Region-aware adaptive instance normalization for image harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9361–9370 (2021)

[28] Hang, Y., Xia, B., Yang, W., Liao, Q.: Scs-co: Self-consistent style contrastive learning for image harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19710–19719 (2022)

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

[30] Cao, J., Cong, W., Niu, L., Zhang, J., Zhang, L.: Deep image harmonization by bridging the reality gap. (2022). BMVC

[31] Hao, G., Iizuka, S., Fukui, K.: Image harmonization with attention-based deep feature modulation. In: BMVC (2020)

[32] Guo, Z., Guo, D., Zheng, H., Gu, Z., Zheng, B., Dong, J.: Image harmonization with transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14870–14879 (2021)

[33] Liu, S., Huynh, C.P., Chen, C., Arap, M., Hamid, R.: Lemart: Label-efficient masked region transform for image harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18290–18299 (2023)

[34] Liang, J., Cun, X., Pun, C.-M.: Spatial-separated curve rendering network for efficient and high-resolution image harmonization. In: European Conference on Computer Vision (2022). Springer

[35] Guerreiro, J.J.A., Nakazawa, M., Stenger, B.: Pct-net: Full resolution image harmonization using pixel-wise color transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5917–5926 (2023)

[36] Valanarasu, J.M.J., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Echevarria, J., Ma, Y., Wei, Z., Sunkavalli, K., Patel, V.M.: Interactive portrait harmonization. In: International Conference on Learning Representations (2023)

[37] Xu, K., Hancke, G.P., Lau, R.W.H.: Learning image harmonization in the linear color space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12570–12579 (2023)

[38] Wang, K., Gharbi, M., Zhang, H., Xia, Z., Shechtman, E.: Semi-supervised parametric real-world image harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5927–5936 (2023)

[39] Gardner, M.-A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., Lalonde, J.-F.: Learning to predict indoor illumination from a single image. ACM Transactions on Graphics **36**(6), 1–14 (2017)

[40] Zhang, J., Lalonde, J.-F.: Learning high dynamic range from outdoor panoramas. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4519–4528 (2017)

[41] LeGendre, C., Ma, W.-C., Fyffe, G., Flynn, J., Charbonnel, L., Busch, J., Debevec, P.: Deeplight: Learning illumination for unconstrained mobile mixed reality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5918–5928 (2019)

[42] Somanath, G., Kurz, D.: Hdr environment map estimation for real-time augmented reality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11298–11306 (2021)

[43] Weber, H., Garon, M., Lalonde, J.-F.: Editable indoor lighting estimation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, pp. 677–692 (2022). Springer

[44] Srinivasan, P.P., Mildenhall, B., Tancik, M., Barron, J.T., Tucker, R., Snavely, N.: Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8080–8089 (2020)

[45] Wang, Z., Philion, J., Fidler, S., Kautz, J.: Learning indoor inverse rendering with 3d spatially-varying lighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12538–12547 (2021)

[46] Zhu, Y., Zhang, Y., Li, S., Shi, B.: Spatially-varying outdoor lighting estimation from intrinsics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12834–12842 (2021)

[47] Tang, J., Zhu, Y., Wang, H., Chan, J.H., Li, S., Shi, B.: Estimating spatially-varying lighting in urban scenes with disentangled representation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, pp. 454–469 (2022). Springer

[48] Li, Z., Shi, J., Bi, S., Zhu, R., Sunkavalli, K., Hašan, M., Xu, Z., Ramamoorthi, R., Chandraker, M.: Physically-based editing of indoor scene lighting from a single image. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, pp. 555–572 (2022). Springer

[49] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.:

Nerf: Representing scenes as neural radiance fields for view synthesis. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pp. 405–421 (2020). Springer

[50] Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., Sridhar, S.: Neural fields in visual computing and beyond. In: Computer Graphics Forum, vol. 41, pp. 641–676 (2022). Wiley Online Library

[51] Narihira, T., Maire, M., Yu, S.X.: Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2992–2992 (2015)

[52] Sato, S., Yao, Y., Yoshida, T., Kaneko, T., Ando, S., Shimamura, J.: Unsupervised intrinsic image decomposition with lidar intensity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13466–13475 (2023)

[53] Li, Z., Snavely, N.: Learning intrinsic image decomposition from watching the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9039–9048 (2018)

[54] Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: Revisiting deep intrinsic image decompositions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8944–8952 (2018)

[55] Liu, Y., Li, Y., You, S., Lu, F.: Unsupervised learning for intrinsic image decomposition from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3248–3257 (2020)

[56] Zhu, Y., Tang, J., Li, S., Shi, B.: Derendernet: Intrinsic image decomposition of urban scenes with shape-(in) dependent shading rendering. In: 2021 IEEE International Conference on Computational Photography (ICCP), pp. 1–11 (2021). IEEE

[57] Das, P., Karaoglu, S., Gevers, T.: Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19790–19799 (2022)

[58] Zhou, H., Yu, X., Jacobs, D.W.: Glosh: Global-local spherical harmonics for intrinsic image decomposition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7820–7829 (2019)

[59] Debevec, P., Hawkins, T., Tchou, C., Duiker, H.-P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 145–156 (2000)

[60] Xu, Z., Sunkavalli, K., Hadap, S., Ramamoorthi, R.: Deep image-based relighting from optimal sparse samples. ACM Transactions on Graphics **37**(4), 1–13 (2018)

[61] Meka, A., Haene, C., Pandey, R., Zollhöfer, M., Fanello, S., Fyffe, G., Kowdle, A., Yu, X., Busch, J., Dourgarian, J., *et al.*: Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. ACM Transactions on Graphics **38**(4), 1–12 (2019)

[62] Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7194–7202 (2019)

[63] Sun, T., Barron, J.T., Tsai, Y.-T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P.E., Ramamoorthi, R.: Single image portrait relighting. ACM Transactions on Graphics **38**(4), 79–1 (2019)

[64] Kanamori, Y., Endo, Y.: Relighting humans: occlusion-aware inverse rendering for full-body human images. ACM Transactions on Graphics **37**(6), 1–11 (2018)

[65] Wang, Z., Yu, X., Lu, M., Wang, Q., Qian, C., Xu, F.: Single image portrait relighting via explicit multiple reflectance channel modeling. ACM Transactions on Graphics **39**(6), 1–13 (2020)

[66] Lagunas, M., Sun, X., Yang, J., Villegas, R., Zhang, J., Shu, Z., Masia, B., Gutierrez, D.: Single-image full-body human relighting. arXiv preprint arXiv:2107.07259 (2021)

[67] Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., Nardi, R.D., Goesele, M., Lovegrove, S., Newcombe, R.: The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)

[68] 3D People. https://3dpeople.com

[69] Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G., *et al.*: Meshlab: an open-source mesh processing tool. In: Eurographics Italian Chapter Conference, vol. 2008, pp. 129–136 (2008). Salerno, Italy

[70] Labelme. https://github.com/wkentaro/labelme

[71] Catmull, E.E.: A subdivision algorithm for computer display of curved surfaces. PhD thesis (1974)

[72] Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., Batra, D.: Habitat 2.0: Training home assistants to rearrange their habitat. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)

[73] Community, B.O.: Blender - a 3D Modelling and Rendering Package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018). Blender Foundation. http://www.blender.org

[74] Poly Haven. https://polyhaven.com/hdris

[75] HDR MAPS. https://hdrmaps.com/

[76] Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in Neural Information Processing Systems **33**, 7537–7547 (2020)

[77] Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on computational imaging **3**(1), 47–57 (2016)

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

[79] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)

[80] El Helou, M., Zhou, R., Süsstrunk, S., Timofte, R., Afifi, M., Brown, M.S., Xu, K., Cai, H., Liu, Y., Wang, L.-W., Liu, Z.-S., Li, C.-T., Dipta Das, S., Shah, N.A., Jassal, A., Zhao, T., Zhao, S., Nathan, S., Beham, M.P., Suganya, R., Wang, Q., Hu, Z., Huang, X., Li, Y., Suin, M., Purohit, K., Rajagopalan, A.N., Puthussery, D., Hrishikesh, P.S., Kuriakose, M., Jiji, C.V., Zhu, Y., Dong, L., Jiang, Z., Li, C., Leng, C., Cheng, J.: Aim 2020: Scene relighting and illumination estimation challenge. In: Bartoli, A., Fusiello, A. (eds.) Computer Vision – ECCV 2020 Workshops, pp. 499–518. Springer, Cham (2020)

[81] Paszke, A., Gross, S., Massa, F., Lerer, A.,

Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[82] He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)

[83] Taobao. https://www.taobao.com/

[84] Chen, W., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference (2018)

[85] Bolduc, C., Giroux, J., Hébert, M., Demers, C., Lalonde, J.-F.: Beyond the pixel: a photometrically calibrated hdr dataset for luminance and color prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8071–8081 (2023)

[86] Hu, Z., Nsampi, N.E., Wang, X., Wang, Q.: PNRNet: Physically-inspired neural rendering for any-to-any relighting. IEEE Transactions on Image Processing **31**, 3935–3948 (2022)

[87] Ashikmin, M., Premože, S., Shirley, P.: A microfacet-based BRDF generator. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 65–74 (2000)