Computational Visual Media
https://doi.org/10.1007/s41095-0xx-xxxx-x

Research/Review Article

Illuminator: Image-based Illumination Editing for Indoor Scene Harmonization

Zhongyun Bao[1], Gang Fu[1], Zipei Chen[1], and Chunxia Xiao[1] (✉)

Abstract Illumination harmonization is an important but challenging task that aims to achieve illumination compatibility between the foreground and background under different illumination conditions. Most current studies mainly focus on achieving seamless integration between the appearance (illumination or visual style) of the foreground object itself and the background scene or producing the foreground shadow. They rarely considered global illumination consistency (i.e., the illumination and shadow of the foreground object). In our work, we introduce "Illuminator," an image-based illumination editing technique. This method aims to achieve more realistic global illumination harmonization, ensuring consistent illumination and plausible shadows in complex indoor environments. The Illuminator contains a shadow residual generation branch and an object illumination transfer branch. The shadow residual generation branch introduced a novel attention-aware graph convolutional mechanism to achieve reasonable foreground shadow generation. The object illumination transfer branch primarily transfers background illumination to the foreground region. In addition, we construct a real-world indoor illumination harmonization dataset called RIH, which consists of various foreground objects and background scenes captured under diverse illumination conditions for training and evaluating our Illuminator. Our comprehensive experiments, conducted on the RIH dataset and a collection of real-world everyday life photos, validate the effectiveness of our method.

Keywords indoor scene illumination harmonization, object illumination editing, seamless integration, shadow residual generation.

1 Introduction

Image Composition targets at producing a new composite image by cutting a foreground object from one image and pasting it on another background image. This is an important problem in computer vision [1–6] and graphics [7–15]. However, because the foreground and background are typically under different illumination conditions (e.g., illumination intensity, direction, and color temperature), composite images inevitably suffer from inharmonious illumination. Thus, illumination harmonization [16–19], which aims to achieve illumination compatibility between the foreground and the background, is a potential and challenging task.

Most previous traditional methods have been developed to address this challenging task by transferring statistical information between the foreground and background regions, such as color [20–25] and texture [26]. However, these approaches only work for simple cases. Recently, numerous deep learning-based methods [27–32] have been proposed for solving the image illumination harmonization problem from various perspectives, thereby producing more realistic illumination harmonization results. However, all these methods only consider appearance consistency and disregard the effects of latent shadows.

Liu et al. [33] and Yan et al. [30] mainly focused on the shadow generation of foreground objects to achieve the image composition task. However, they failed to ensure illumination consistency between the foreground and background and only targeted simple outdoor scenes with parallel illumination casting. Although these methods [34–36] consider both the global appearance and illumination consistency, they still suffer from the following difficulties. Song et al. [36] lacked control over the appearance preservation of the synthesized object and had a limited shadow generation space. Bao et al. [34] and Zhan et al. [35] primarily targeted simple outdoor scenes and failed to effectively generalize them to complex

[1] School of Computer Science, Wuhan University, Wuhan 430072, Hubei, China. E-mail: Z. Bao, zhongyunbao@whu.edu.cn; G. Fu, xyzgfu@gmail.com; Z. Chen, czpp19@whu.edu.cn; C. Xiao, cxxiao@whu.edu.cn

Manuscript received: 2022-XX-XX; accepted: 2022-XX-XX

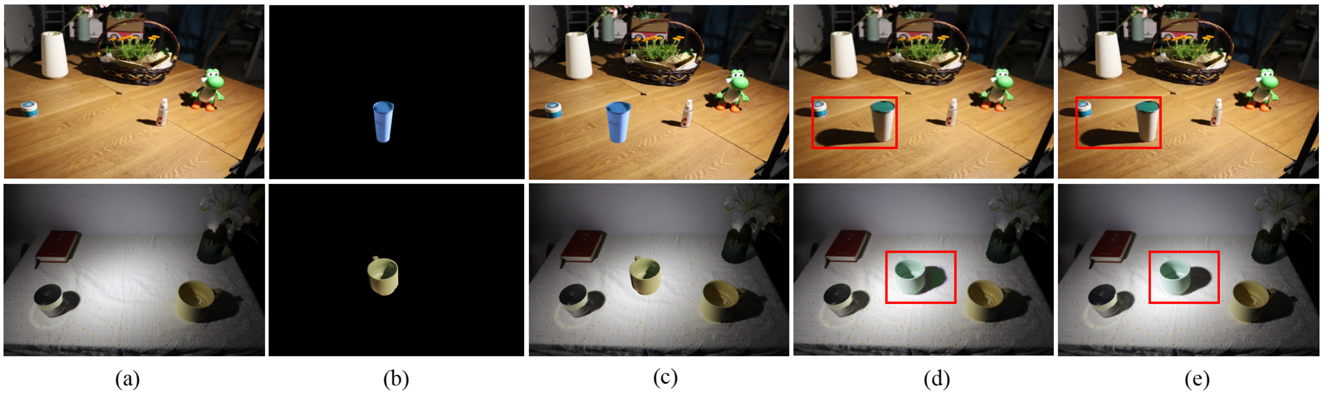

Fig. 1 Our visual illumination harmonization results on real-world scenes under various illumination conditions. From left to right are background scenes (a), foreground objects (b) captured under different illumination conditions from the corresponding background scenes, naive composite input images (c), our generated illumination harmonization results (d) and corresponding ground truths (e), respectively.

indoor environments.

These shortcomings strongly motivate us to seek an effective deep learning-based solution to explore more complex and challenging illumination harmonization tasks that consider global appearance and illumination harmonization, as shown in Figure 1. Deep learning-based techniques, particularly our task, generally require adequate paired training data consisting of composite images without global illumination harmonization and corresponding target images with global illumination harmonization.

However, existing image-based illumination harmonization datasets such as the ccHarmony dataset [19], iHarmony4 [27], shadow-AR dataset [33], HVIDIT dataset [29], and DESOBA dataset [30] only consider foreground object appearance or foreground object shadow, and rarely consider both. Although the datasets [35], IH [34] and VIDIT [37] consider both the appearance and shadow of a foreground object, they have limitations that make them ineffective for our task. For a synthetic dataset [35], limited foreground objects containing only two types of objects and simple outdoor scenes with parallel light projections severely limit their applications. The IH [34] and VIDIT datasets [37] are both synthesized using rendering software, which results in a considerable gap between the synthesized images and real-world images, and thus considerably limits the robustness of the algorithm. Therefore, we are encouraged to construct a complex and challenging real-world illumination harmonization dataset for global appearance and illumination harmonization tasks.

In this work, we construct a large-scale, high-quality, real-world illumination harmonization dataset named RIH in a controllable indoor environment. We first pre-determine over 800 real indoor scenes as backgrounds, and 600 common objects as foregrounds. Subsequently, as shown in Figure 5, we separately capture each background scene and the corresponding ground-truth image under different illumination conditions.

Based on background and ground truth images, a composite image is generated using a foreground object mask to cut the foreground object from one ground truth image and paste it into another background scene. To obtain more accurate foreground object masks, we employ professional photo editors to manually annotate the foreground objects. In addition, a light probe is used to record the illumination information in the scenes. In general, using the above composite method, our dataset finally contains 30000 seven-tuples in total, each with one input triplet (i.e., a naive composite image and the corresponding masks of the foreground object and background object-shadow) and another ground truth quadruplet (including foreground illumination, background illumination, object illumination consistency ground truth image, and global illumination harmonization ground truth image). The image composition of our dataset is shown in Figure 6.

With the paired training dataset, our goal is to achieve more realistic illumination harmonization results for naïve input composite images, focusing on global illumination consistency. Thus, inspired by the methods [29, 30], we propose a novel illumination harmonization method, Illuminator, as shown in Figure 2. It consists of a shadow residual generation branch and an object illumination transfer branch to produce global appearance and illumination harmonization.

In the shadow residual generation branch, due to the complexity of indoor illumination scenes, particularly the irregular cast shadows of background objects, previous methods [30, 33, 34] failed to generate plausible shadows for foreground objects by simply using background information cues. Thus, our key insight is to fully utilize the powerful modeling ability of graph convolution networks and propose a novel attention-aware graph convolutional mechanism to realize the reasonable generation of foreground object shadows. Specifically, we introduce a cross foreground-background attention (CFBA) module to effectively model the relative spatial interaction relationship between the foreground object and background scene, guiding the shadow residual generation of the foreground object. In contrast to previous works focusing only on shadow generation, our task considers both shadow generation and illumination harmonization for foreground objects. Therefore, the strategy of directly learning the object shadow cannot accurately produce light-transport effects (e.g., shadows), and we are encouraged to learn the foreground object shadow residuals.

In the object illumination transfer branch, we focus on achieving illumination seamless integration between the foreground object region and the background scene. Illumination accounts for shading effects [29, 38, 39] that are reflectance-independent. Thus, we first use intrinsic image properties, following the method in [29], to obtain foreground and background shading information from the input image, rather than physically based intrinsic images. Using the shading information, a shading transfer module is introduced to guide the foreground shading to be consistent with that of the background. In addition, we use a patch GAN [40] to push the generator to produce realistic illumination harmonization results. Our contributions are summarized as follows:

- We design a simple and effective illumination harmonization data capturing method, and present a real-world indoor dataset for global illumination harmonization tasks.
- We propose a novel global illumination harmonization framework, Illuminator, which redefines object illumination harmonization as a shading style consistency problem and introduces a novel cross foreground-background attention-aware graph convolutional mechanism, effectively achieving high-quality global illumination harmonization results.
- We evaluate our method and the state-of-the-arts on our RIH dataset and other real challenging images, and show the superiority of our method both quantitatively and qualitatively.

2 Related work

2.1 Image Harmonization

Image harmonization aims to achieve seamless illumination compatibility between the foreground and background under different illumination conditions. Previous image harmonization works mainly included traditional and deep learning-based methods. Specifically, traditional methods mainly focus on producing consistent visual appearances employing low-level statistics between the foreground and background, which include color [21, 41, 42], gradient information [43–45], and semantic information [25, 46]. However, they only match the appearance of the foreground with the background parts, while overlooking visual realism.

Recently, some deep-learning-based methods [47–49] have made further contributions to image harmonization tasks. Cong *et al.* [27, 50] regarded image harmonization as a domain translation that transforms the foreground domain into the background domain. Ling *et al.* [32] treated image harmonization as a style transfer problem and proposed a region-aware adaptive instance normalization module to address it. Guo *et al.* [29, 51] proposed image harmonization solutions using intrinsic image harmonization and transformers. In particular, intrinsic image harmonization seeks to achieve image harmonization via separable harmonization of reflectance and illumination. Jiang *et al.* [31] and Wang *et al.* [47] worked on the image harmonization problem from the perspectives of self-supervision and semi-supervision. In addition, Cong *et al.* [52] and Guerreiro *et al.* [48] began with full-resolution image harmonization to explore image harmonization tasks. Hang *et al.* [53] introduced contrastive learning to achieve image harmonization. Liu *et al.* [49] designed SwinIH, an image harmonization model based on the Swin Transformer architecture, to obtain impressive image harmonization results.

However, these methods mainly focus on achieving appearance and visual style consistency between the foreground and background. In contrast, we provide a novel perspective on image-shading style consistency and achieve global appearance and illumination harmonization.

2.2 Shadow Generation

The existing shadow generation methods can be divided into two categories: rendering-based and image-to-image translation methods. Some rendering-based shadow generation methods [54–56] require strong user interactions to obtain explicit illumination condition, reflectance, and scene geometry to generate plausible shadows. Although these methods [57–59] recover illumination information and scene geometry from a single image, inaccurate estimation typically produces unsatisfactory results. Worchel *et al.* [60] demonstrated efficient generation of shadows in the differentiable rendering of triangular meshes. Sheng *et al.* [61] introduced an interactive soft-shadow network (SSN) to generate controllable soft shadows for image composition. However, the assumption that the shadow receiver is only a ground plane limits its practical application. Sheng *et al.* [62] proposed a system PixHt-Lab for generating perceptually plausible light effects based on pixel height representation, and a method SSG++ guided by 3D-aware buffer channels to improve the soft shadow quality cast on general shadow receivers.

Image-to-image translation methods mainly focus on directly performing shadow-generation tasks in the 2D domain. ShadowGAN [63] proposed a local conditional discriminator and a global conditional discriminator to model the shape and direction of the shadow, respectively, and directly produced a plausible shadow for an inserted 3D foreground object. However, this method does not consider the occluders of real shadows. ARShadowGAN [33] is an attention mechanism that exploits background cues to guide foreground object shadow generation. Hong *et al.* [30] constructed an outdoor shadow-generation dataset and proposed a shadow-generation network SGRNet to generate shadows for foreground objects. However, they are suitable only for relatively simple scenes with parallel-illumination casting. Considering that our task targets both shadow generation and object illumination transfer, in contrast to the above methods, we introduce an attention-aware graph convolutional mechanism to effectively model the interaction relationship between the background and foreground to guide the foreground object shadow residual generation.

2.3 Graph Convolutional Networks

Graph Convolutional Network was first proposed in [64] for a semisupervised classification task. Recently, GCNs (GCNs) [65–69] have received considerable attention for computer vision and graphics tasks. Chen *et al.* [70] proposed a multilabel classification model based on a GCN to model object label dependencies to improve the recognition performance. Wan *et al.* [71] developed a HSI classification method based on a GCN to correctly discover contextual relations among pixels. Lin *et al.* [72] used GCNs to reconstruct detailed colors for mesh vertices to recover 3D facial shapes with high-fidelity textures from single-view images. Li *et al.* [68] proposed a GCN based mesh regression called IntagHand to demonstrate the effectiveness of GCN in a two-hand reconstruction task. Based on the above analysis, considering the ability of the graph to describe complex data relationships, particularly for our illumination harmonization task, we fully model the relative spatial position relationship between the foreground and background to effectively guide the shadow generation of the foreground object. Thus, without loss of generality, we leverage the powerful modeling capability of a GCN to design an attention-aware graph convolutional mechanism for illumination harmonization.

3 Proposed Method

Given a pair of naive composite image $\tilde{X}$ and the corresponding real ground truth image X, with a foreground mask M_f and a background object-shadow pair mask M_{os}, our goal is to learn a harmonization network G, which inputs $\tilde{X}$, M_f and M_{os}, and outputs the global illumination harmonization result as $\hat{X} = G\left(\tilde{X}, M_f, M_{os}\right)$ expected to be as harmonious as X.

To this end, we design an Illuminator to separately achieve foreground object shadow generation and illumination consistency between the foreground and background for global illumination harmonization. It primarily comprises two branches: a shadow residual generation branch and an object-illumination transfer branch, as shown in Figure 2. In the shadow residual generation branch, the key task is to generate a plausible foreground object shadow residual to achieve the shadow generation of the foreground object. The object-illumination transfer branch mainly focuses on producing consistent shading between the foreground and background. We also employ a generative adversarial network PatchGAN to force the generator to produce more realistic illumination harmonization results.

3.1 Generator

As shown in Figure 2, the generator mainly consists of two branches. The first branch targets at generating harmonious object illumination from a shading-style transfer perspective based on intrinsic image properties [29]. The second branch is the foreground object shadow residual generation module, which uses a novel attention-aware graph convolutional mechanism to solve the object shadow generation problems. The above two results, the final global illumination harmonization result is obtained by subtracting the foreground object shadow residual from the harmonious object illumination.

Object Illumination Transfer Architecture. In this work, we redefine object illumination harmonization as a shading style consistency problem. Thus, the key insight into shading style transfer is to achieve a consistent shading style between the foreground and background. Our object illumination transfer model uses a naive composite image $\tilde{X}$ and the corresponding foreground object mask M_f as input, and outputs a object illumination harmonization image $\hat{X}_a$.

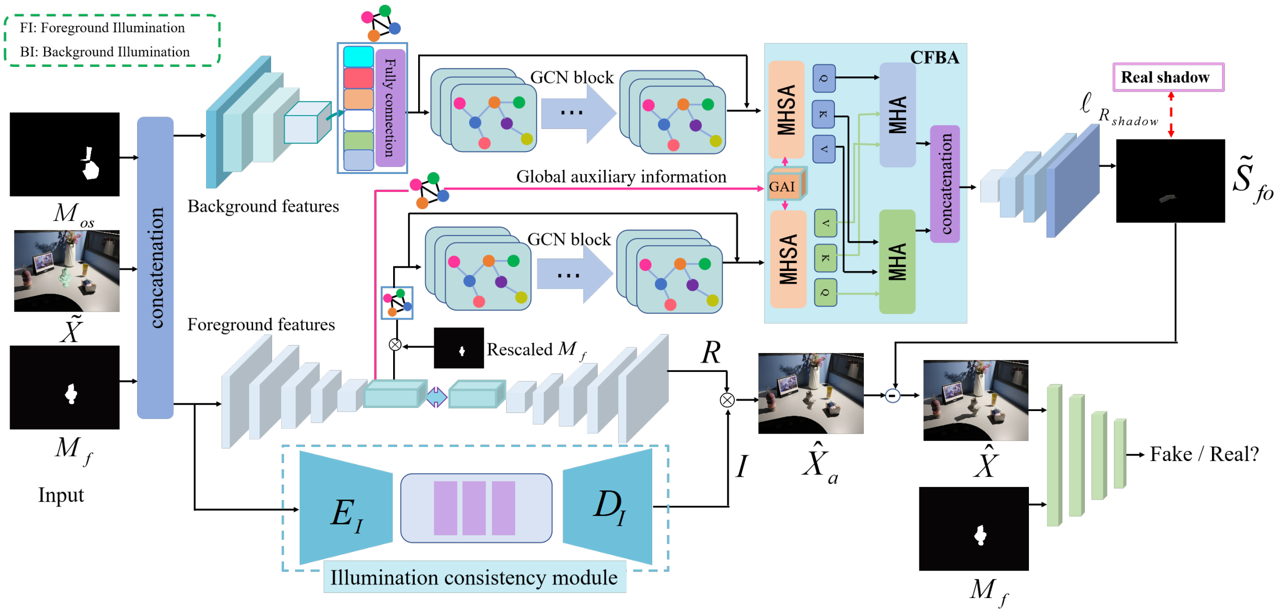

Fig. 2 Our pipeline consists of a shadow residual $\tilde{S}_{fo}$ generation branch and a object illumination transfer branch. The shadow residual generation branch takes the composite image $\tilde{X}$ and the corresponding background object-shadow mask M_{os} as input and outputs the foreground object shadow residual $\tilde{S}_{fo}$. The object illumination transfer branch takes the composite image $\tilde{X}$ and corresponding foreground object mask M_f as input and outputs the object illumination harmonization result $\hat{X}_a$ whose the foreground object illumination is consistent with the background. Our method finally outputs the realistic global appearance and illumination harmonization result.

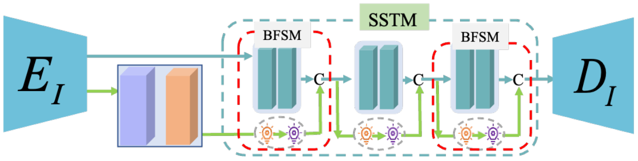

Fig. 3 The illumination consistency module. It mainly achieves the consistent illumination between the foreground and background.

We primarily adopt the encoder-decoder parts of the reflectance and shading of IntrinsicNet [29] to design our model. Specifically, for an input composite image and the corresponding foreground object mask, we first use the two encoders E to extract the shading and reflectance features, respectively. Subsequently, as shown in Figure 3, foreground shading and background shading were obtained by multiplying the corresponding masks and the composite image shading, respectively. Then, they are fed into the shading style transfer module (SSTM) following the shading encoder, achieving foreground shading style consistency with the background shading style.

The shading-style transfer module consists of three identical background-foreground shading mapping mechanisms (BFSM) in series. Each mechanism includes two branches: the first introduces the RAIN mechanism of [32], which guides the foreground object shading features to learn a style consistent with the background shading features; the second is a convolutional block consisting of three consecutive 3×3 convolution layers, which further extracts the composite image shading features. Their output features are then concatenated and fed into the next BFSM performing the above operation. After obtaining the output features of the shading-style transfer module, they were fed into the shading decoder D_I to produce a harmonious shading map. Finally, we conduct element-wise multiplication on the harmonious shading map I and reflectance map R obtained by decoding the reflectance features to achieve object illumination harmonization.

Object shadow residue generation architecture Our foreground object shadow residual generation network uses the composite image $\tilde{X}$ and the corresponding background object-shadow pair mask M_{os} as inputs, generating the foreground object shadow residual $\tilde{S_{fo}}$ directly. From the existing works, it is clear that background object-shadow pairs implicitly contain the illumination direction information [73] of the scene, which can provide rich cues for guiding foreground object shadow generation [30, 33, 34, 63]. However, they either focused on outdoor scenes with parallel lighting or ignored the occlusion problem of real shadows. Apparently, they fail to generate a plausible object shadow in complex real-world indoor scenes because the shadows cast by background objects have irregular directions and cannot effectively provide clear guidance. Therefore, a more effective foreground object shadow generation solution for indoor scenarios is required. We design a novel attention-aware graph convolutional mechanism that utilizes the cross foreground-background attention (CFBA) module to effectively associate background cues with the foreground object to learn the foreground object shadow residuals.

As shown in Figure 2, our architecture consists of three parts: a feature extraction module, a cross foreground-background attention-aware graph convolutional module, and a feature-decoding module. Specifically, we first adopt a CNN-based backbone to extract the foreground features FF_o and the background features BF_{os}. Furthermore, they are connected separately to same GCN block with four consecutive graph convolutional layers through a fully connected layer. We then construct the adjacency graph $\mathcal{G}$ as the input to the GCN, which can be defined as follows:

$$\mathcal{G} = \{V, E, A\}, \tag{1}$$

where V represents the set of nodes composed of all features, E is the set of edges, and A is the adjacency matrix describing the graph structure used to define the interconnections between all nodes. Specifically, for any two nodes V_i and V_j, if V_i is the nearest neighbor of V_j and there is an edge between them, the weight of the edge of the adjacency matrix is calculated as follows:

$$A_{ij} = \left\{ \begin{array}{ll} 1, & V_i \in N_k(V_j) \\ 0, & otherwise \end{array} \right. ,$$

where $N_k(V_j)$ represents the K(K=10) nearest neighbor of V_j. The graph convolution of each layer is defined as follows.

$$X^{(k+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\hat{A}\tilde{D}^{-\frac{1}{2}}X^{(k)}W^{(k)}\right), \tag{2}$$

where σ denotes ReLU activation function for fast convergence and the node representations $X^{(k)}$ and $X^{(k+1)}$ are the input and output of the k-th layer, respectively. $\hat{A}$ is the self-connection adjacency matrix, that is, $\hat{A} = A + I_n$, used to improve the stability of the model training, where I_n denotes the unit matrix. $W^{(k)}$ is a trainable weight matrix and $\tilde{D}_{ii}$ can be calculated as

$$\tilde{D}_{ii} = \sum\nolimits_j A_{ij}. \tag{3}$$

Considering that background information can provide key clues for the foreground, such as lighting information and relative position relationships, it is particularly important to effectively model the correlation between the background and foreground. However, because of the complex geometric structure and spatial layout of indoor scenes as well as the irregular casting of background object shadows, simply representing the interaction [30, 33, 34, 63] between the foreground and background cannot achieve satisfactory results. Thus, inspired by the methods [68, 74], we design a cross foreground-background attention (CFBA) module, as shown in Figure 2, to implicitly express the spatial correlation between the foreground and the background.

We utilize the global features of the input image as global auxiliary information (GAI) to further guide the generation of foreground object shadows in a reasonable direction. First, we concatenated the GAI with the foreground and background graph features, respectively. They are then fed into the multi head self-attention module (MHSAM), obtaining the corresponding query (Q_F/Q_B), key (K_F/K_B), and value (V_F/V_B) features of the foreground and background, respectively. We exchange K and V values of the foreground and background, and combine the multi head attention mechanism to achieve the cross foreground-background attention operator for modeling the spatial correlation between the foreground and background. Specifically, we input the Q_F value of the foreground and the K_B and V_B values of the background, as well as the Q_B value of the background and the K_F and V_F values of the foreground into the multi head attention mechanism as follows:

$$\mathcal{R}_{F\to B} = \mathit{Softmax}(\frac{(Q_B)(K_F)^T}{\sqrt{D}})V_F, \tag{4}$$

$$\mathcal{R}_{B\to F} = \mathit{Softmax}(\frac{(Q_F)(K_B)^T}{\sqrt{D}})V_B, \tag{5}$$

where D is a normalization constant and $\mathcal{R}_{F\to B}$ and $\mathcal{R}_{B\to F}$ are the corresponding constructed correlation features between the foreground and background. Finally, we concatenate $\mathcal{R}_{F\to B}$ and $\mathcal{R}_{B\to F}$, and pass them through the feature decoding module to obtain the foreground object shadow residual, which is further supervised by $\mathcal{L}_{R_{shadow}}$.

$$\mathcal{L}_{R_{shadow}} = \|(\tilde{S_{fo}} - (\hat{X} - X))M_{fos}\|_2^2, \tag{6}$$

where M_{fos} is the shadow mask of the foreground object.

3.2 Discriminator

To ensure that the generated illumination harmonization results is more realistic, it should be closer to a real ground-truth image. We design the discriminator following PatchGAN [40], which concatenates the input image $\tilde{X}$ and foreground object mask M_f as the input. Our discriminator consists of four consecutive convolutions with valid padding, instance normalization, and Leaky ReLU operations. Subsequently, a sigmoid function is used to activate the last feature map produced by the convolution, and the activated feature map is further passed through the global average pooling to produce the final output.

3.3 Training Losses and Details

Our illumination harmonization task includes two key components: transforming the illumination information of foreground objects and generating reasonable shadows for foreground objects. Thus, our total loss $\mathcal{L}_{total}$ consists of five sub-losses. The reconstruction loss $\mathcal{L}_{recons}$ is used to reconstruct the global appearance and illumination information of the output image. The reflectance loss $\mathcal{L}_{reflec}$ is used to constrain the appearance (content information) of the image to remain unchanged, whereas the illumination loss $\mathcal{L}_{illu}$ and shadow residual loss $\mathcal{L}_{R_{shadow}}$ are used to ensure that the illumination and shadow information of the generated foreground object are closer to the target image. Furthermore, an adversarial loss $\mathcal{L}_{adv}$ is further used to refine the generated result. The total loss is expressed as follows:

$$\begin{aligned}\mathcal{L}_{total} =& \beta_1\mathcal{L}_{recons} + \beta_2\mathcal{L}_{reflec} + \beta_3\mathcal{L}_{illu} \\ &+ \beta_4\mathcal{L}_{R_{shadow}} + \beta_5\mathcal{L}_{adv},\end{aligned} \tag{7}$$

where β_1, β_2, β_3, β_4 and β_5 are the hyperparameters controlling the influence of each term.

Reconstruction Loss. This is a classical L_1 loss between the output image $\hat{X}$ and the corresponding real ground-truth image X, which is expressed as

$$\mathcal{L}_{recons} = \|\hat{X} - X\|_1. \tag{8}$$

Reflectance and Illumination Losses. Following [29], we regard $\nabla\hat{R} \approx \nabla X_a$ as a constraint to harmonize reflectance and reflectance loss:

$$\mathcal{L}_{reflec} = \mathrm{E}_{(\triangledown\hat{R},\triangledown X_a)}[\| \triangledown \hat{R} - \triangledown X_a\|_1], \tag{9}$$

where $\triangledown\hat{R}$ and $\triangledown X_a$ are the predicted reflectance gradient and real object appearance illumination harmonization ground-truth image gradient, respectively. E denotes the expectation. The object illumination harmonization loss is defined as follows [29]:

$$\mathcal{L}_{illu} = \mathrm{E}_{(I,X_a)}[\|I - X_a\|_2], \tag{10}$$

where I denotes shading predicted by the generator.

Adversarial losses. $\mathcal{L}_{adv}$ is utilized to describe the competition between the generator and the discriminator as:

$$\mathcal{L}_{adv} = \log(\mathbf{D}(\tilde{X}, M_f, X)) + \log(1 - \mathbf{D}(\tilde{X}, M_f, \hat{X})), \tag{11}$$

where $\mathbf{D}(\cdot)$ is the probability that the image is “real”. $\tilde{X}$ is the input image, M_f is the corresponding mask, $\hat{X}$ is the output of the generator of Illuminator, and X is the ground truth. We set β_1 = 60.0, β_2 = 10.0, β_3 = 1.0, β_4 = 40, β_5 = 1.0, and adopt the Adam optimizer to optimize the entire network.

Our Illuminator is implemented using PyTorch and run on the NVIDIA GeForce GTX 3090Ti GPU. We divide the 30000 seven-tuples of RIH dataset into two parts: 25000 for training, and 5000 for testing. There are no overlapping background scenes or foreground objects between the test and training sets. The Illuminator is trained for 100 epochs with batch size of 1, and the resolution of all images is 512×512. The initial learning rate is 10^{-4}, and the number of layers in the GCN is set to 3.

4 Our RIH Dataset

Our dataset contains over 800 indoor scenes covering various common daily life and office environments as backgrounds, and 600 foreground objects with different shapes, types, and materials as foregrounds for illumination editing. Each scene with a foreground object was captured under various illumination conditions using several predetermined controllable intelligent spotlights with different color temperatures. In addition, to enrich the complexity and diversity of the captured backgrounds, we use various reference objects with different materials and shapes to fill the background during the capture process. In the following section, we introduce our dataset in detail from two perspectives: image capture, image composition and pairing.

4.1 Image Capture

To capture a real-world dataset for our illumination harmonization task, we design a simple and effective data-capturing method and build an intelligent photographic device that only requires a set of simple photography and lighting equipment. Specifically, our capture device consists of several intelligent spotlights with different illumination intensities and color temperature, additional disturbing light sources to enhance the data, a light probe that records the illumination information of scenes, a digital camera (Canon camera with 6D Mark II and 4608×3456 resolution), and the corresponding fixture, that is, a magnetic circular light track. Figure 4 (a) and (b) show the specific construction of the data acquisition device and a live case, respectively.

We first fix the spotlights and camera on the light track at equal intervals and the camera bracket, respectively. The camera and spotlights were maintained at the same height to form a closed loop. A captured scene is then built under the closed loop, which is located in the field of view where the beam of light intersects the field of view of the camera. Additionally, a light probe was placed at the scene to record the corresponding illumination information. To ensure the quality of the captured image, the following three aspects must be emphasized. (1) The pitch angle of the camera and the direction of the spotlights can be adjusted appropriately

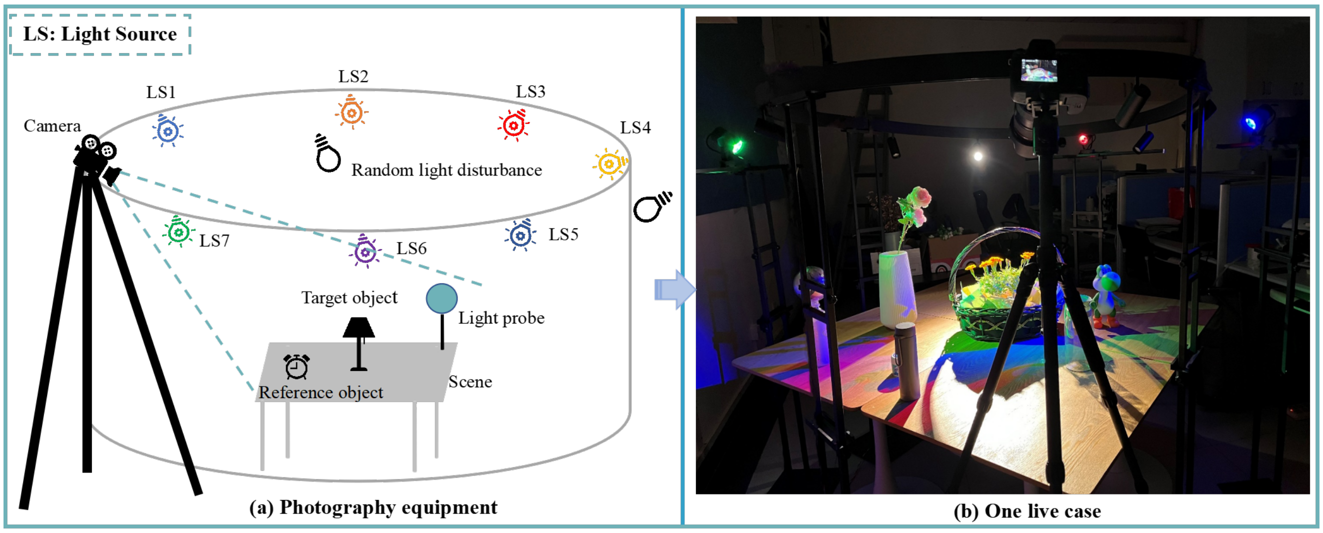

Fig. 4 (a) and (b) are the data acquisition device and one live case, respectively. The data acquisition device (a) consists of light sources, a digital camera (Canon camera with 6D Mark II and 4608 × 3456 resolution), and a magnetic circular light track.

according to the photography requirements. (2) We use a remote-control switch and trigger to control spotlight state (closed or open) and camera shutter, respectively. (3) The entire photography process is in a controllable environment that switches off the room lights and shuts windows.

For ease of description, we consider the dataset capture process based on a scene as an example. For a target background scene, we first capture a series of background scene images under different lighting conditions as the backgrounds of the dataset. To this end, under each lighting condition, we only turned on the corresponding light and turned off the other lights to ensure data quality. During the entire process of capturing background scenes under different lighting conditions, the camera pose and position of the light probe remain fixed. The background capturing process is illustrated in the first row of Figure 5.

Subsequently, we keep the above conditions constant and place the foreground objects in the background scene. After placing each foreground object, we repeat the same photography steps to capture the background to record the entire scene as the ground truth image. The process of capturing a ground truth image is shown in the last row of Figure 5. To increase the diversity of the data and simulate indoor light sources, we also adjust different light sources directions, and use extra light sources with different intensities to perturb the foreground objects in random directions.

4.2 Image Composition and Pairing

With the illumination harmonization of the ground truth images and background scene images, the non-illumination harmonization composite image is produced by cutting the foreground object from one ground truth image and pasting it into another background scene image. Therefore, more accurate masks are required to obtain high-quality composite images. First, we test state-of-the-art algorithms [73, 75] to automatically obtain masks, including foreground object masks, background object-shadow pair masks, and the light probe mask. However, due to the complexity of indoor scenes, their effectiveness is far from satisfactory, which significantly affects the composite quality of the dataset. Therefore, we employ professional photo editors to annotate the masks manually. Using these masks, we can effectively achieve image composition and pairing, and the complete process is shown in Figure 6.

As shown in Figure 6, for convenience, we introduce a dataset composition process based on one scene and a foreground object. We first use the foreground object mask to multiply the corresponding global appearance and illumination harmonization ground-truth image to obtain a foreground object under one illumination condition (Figure 6 (a)). Subsequently, a background mask is used to perform the same operation with the same background scene image under different illumination conditions to obtain the corresponding insertion position of the foreground object (Figure 6 (c)). Using the foreground object and corresponding position scenes, composite images are produced by pasting the foreground object into the position scenes (Figure 6 (b)). In this manner, we achieve all image composites for other foreground objects and background scenes. In general, our dataset finally contains 30000 seven-tuples in total, of which each seven-tuple ((Figure 6 (d)) consists of one input triplet and a ground

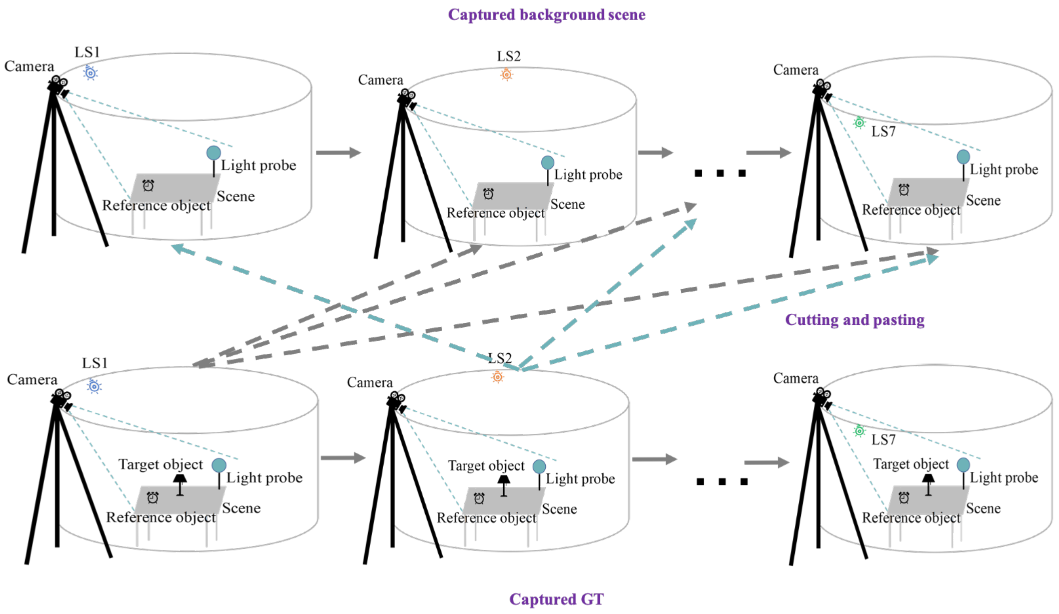

Fig. 5 Acquisition of foreground objects and background scenes under different lighting conditions and their specific synthesis.

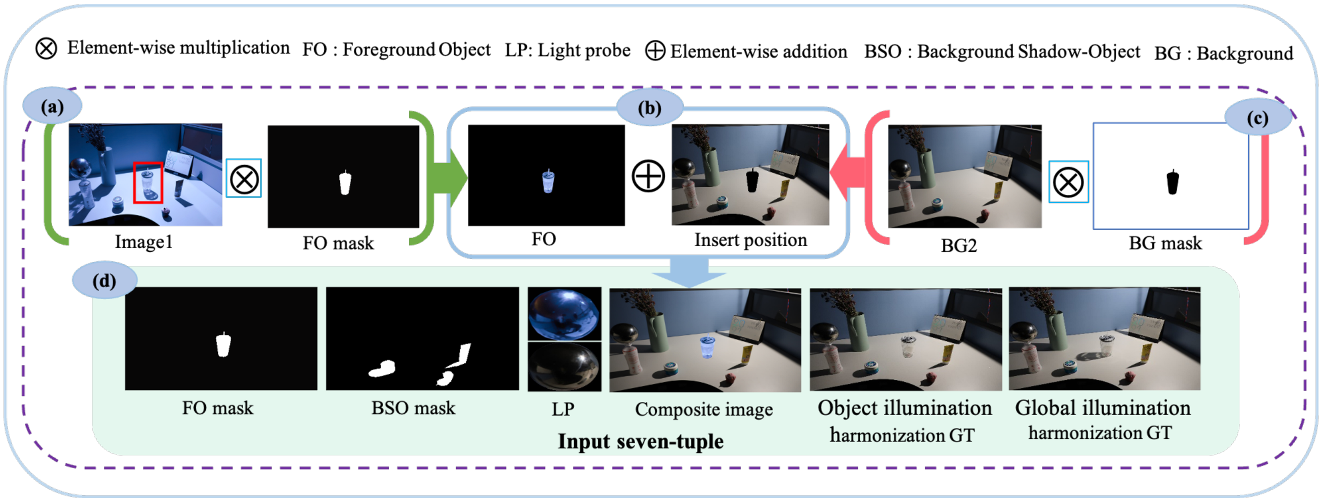

Fig. 6 Illustration of the image composition and pairing. (a) and (c) represent the process of obtaining the foreground object and the corresponding insertion position in the background, respectively. (b) represents a simple composite process between foreground and background. The bottom row (d) shows the paired result forming an input seven-tuple, which includes a foreground object mask, a background shadow-object pair mask, the illumination images of the foreground and background scene, a naive composite image, a foreground object illumination harmonization ground truth image and the global illumination consistency ground truth image.

truth quadruplet. The input triplet contains a naive composite image, a corresponding foreground mask, and a background object-shadow pair mask. The ground truth quadruplet consists of the foreground illumination information, background illumination information, object illumination harmonization ground truth, and the global illumination harmonization ground truth. One visual seven-tuple example is shown in Figure 6 (d).

5 Experiments

To verify the superiority of our proposed Illuminator, we compare our Illuminator with four state-of-the-art illumination harmonization methods on the real-world RIH dataset, and provide quantitative and qualitative assessments.

5.1 Experimental Settings and Evaluation Metrics.

Compared methods . We compared our illuminator with four related baseline methods, where AICNet [35] and DIH [34] performed the same task as in this work, and IntrinsicNet [29]

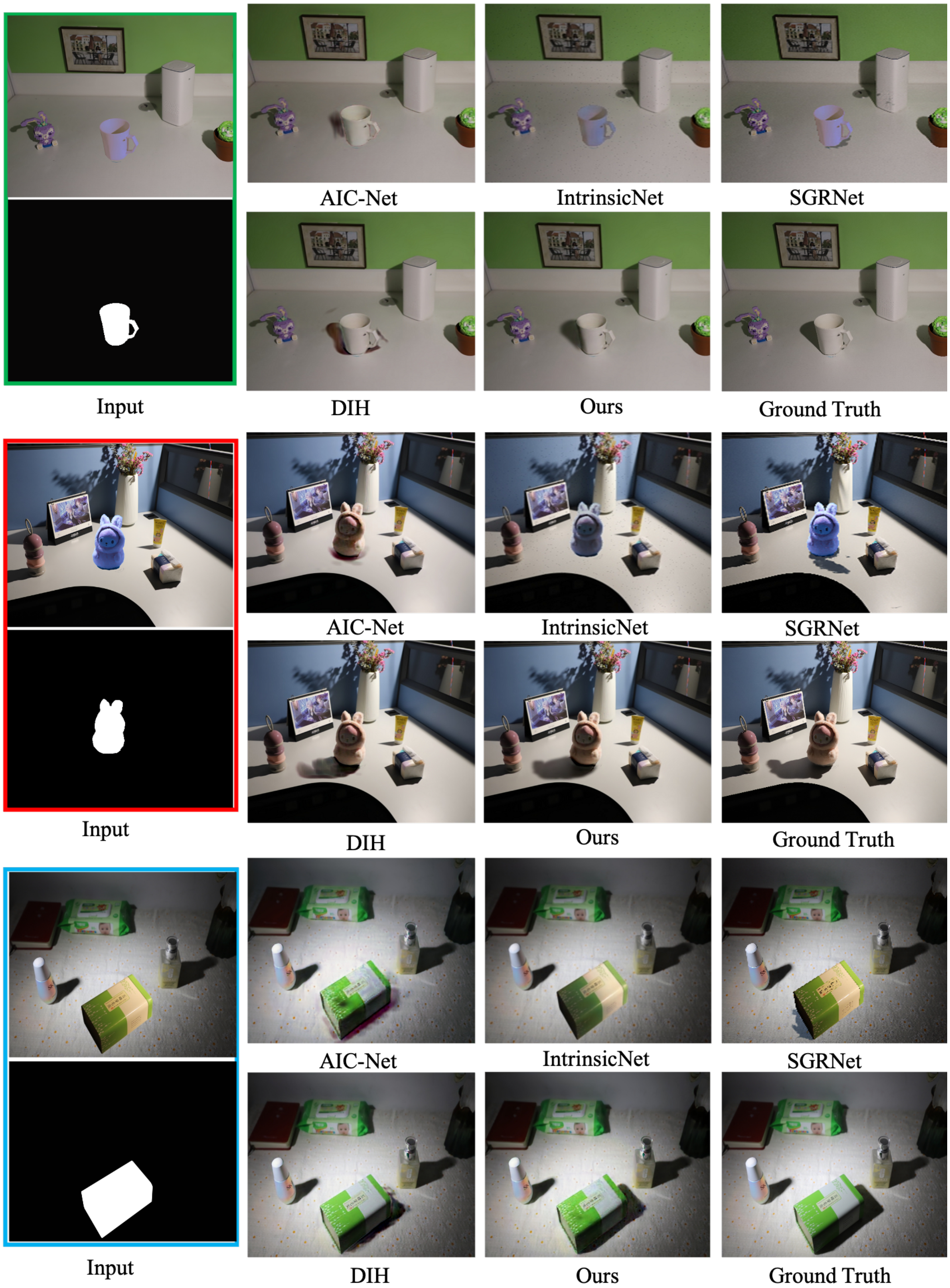

Fig. 7 Visual comparison of our method with state-of-the-art methods on three real-world scenes with various materials and foreground objects.

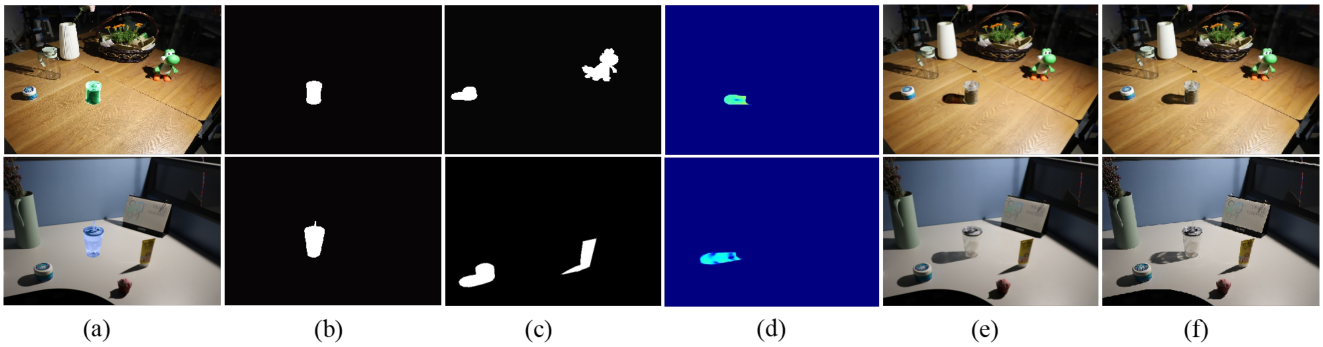

Fig. 8 Visual results of our method. From (a) to (f) are input naive composite images, foreground object masks, background object-shadow masks, shadow residual images, output global illumination harmonization results and the ground truth images, respectively.

Table 1 Results of quantitative comparison on our testing set. "↑" indicates the higher the better, and "↓" indicates the lower the better. The best results are marked in bold.

Method	RMSE ↓	SSIM ↑	fMSE ↓	fSSIM ↑
AICNet [35]	5.904	0.884	579.186	0.901
IntrinsicNet [29]	7.113	0.814	1024.256	0.891
SGRNet [30]	9.712	0.773	1417.211	0.862
DIH [34]	4.725	0.897	478.254	0.923
Ours	**4.372**	**0.912**	**342.239**	**0.947**

and SGRNet [30] were mainly focused on image harmony and object shadow generation, respectively.We train and test all these methods based on our real-world RIH dataset.

Evaluation Metrics. We used four metrics to evaluate the image illumination harmonization results: relative mean square error (RMSE), structural similarity index measure (SSIM), foreground mean square error (fMSE), and foreground structural similarity index measure (fSSIM). Generally, smaller RMSE and fMSE and larger SSIM and fSSIM indicate better image illumination harmonization results.

5.2 Comparison with State-of-the-Arts

All methods are trained and tested on the training and testing sets of RIH dataset, respectively. The quantitative comparison results of different methods on the testing set are reported in Table 1. Our Illuminator achieves better quantitative results than other state-of-the-art methods on all four metrics. We also observe that IntrinsicNet [29] is better than SGRNet [30] in terms of the fMSE and fSSIM metrics. This is primarily because IntrinsicNet [29] focuses on the image harmonization task and fully accounts for the illumination of the foreground object, whereas SGRNet [30] only focuses on foreground object shadow generation and ignores object illumination consistency. Besides, for the SSIM and RMSE metrics, DIH [34] achieves the better results than AICNet [35], which is mainly attributed to the fact that DIH [34] fully utilizes illumination exchange module and multi-scale attention mechanism to achieve the generation of foreground object shadows and illumination. Our Illuminator obtains the best performance, mainly because Illuminator considers global illumination consistency and constructs an effective GCN with the CFBA mechanism to fully model the spatial interaction relation between the foreground and the background.

Figure 7 shows some visual comparison results. Among these competing methods, we observe that IntrinsicNet [29] mainly targets image appearance harmonization and fails to address foreground object shadow generation. Although AICNet [35] considers image appearance and illumination harmonization, it uses spherical harmonic parameters to represent illumination, and cannot model detailed high-frequency illumination information (see the visual results of AICNet [35] in Figure 7). In addition, neither ACINet [35], SGRNet [30] nor DIH [34] can effectively handle foreground object shadow generation in complex indoor scenarios. In contrast, our method achieve the best visual results with plausible foreground object shadows and harmonious object illumination. Figure 8 presents some image illumination harmonization results produced by Illuminator. The fourth column (d) shows the generated object shadow residual marked with a striking colour. Our Illuminator effectively achieves image illumination harmonization in indoor scenes.

5.3 Ablation Study

To verify the effectiveness of each design choice of our Illuminator, we conduct an ablation study by modifying the Illuminator architecture to evaluate the performance of the

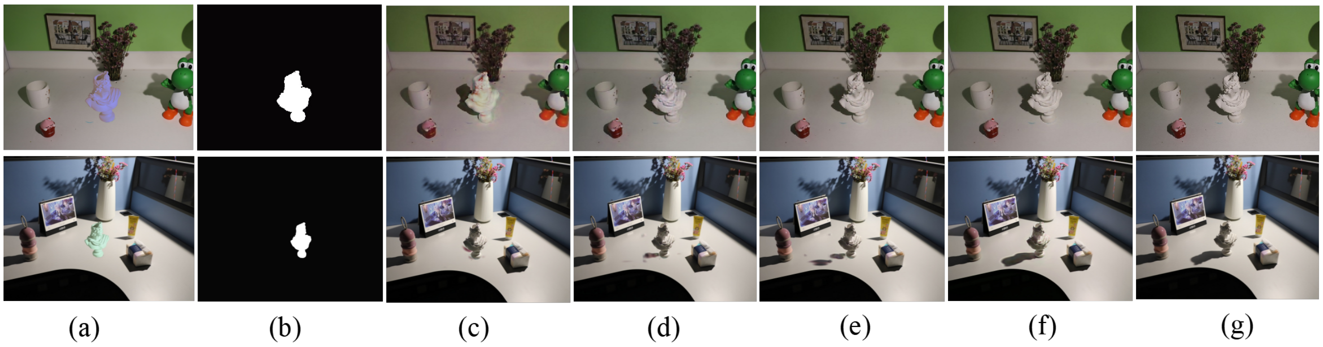

Fig. 9 The first row and the second row are visualization results of ablated versions of the SSTM and CFBA, respectively. For the top row, from (a) to (g) are the input composite image (a), foreground object mask (b), the corresponding different output results: baseline (c), BFSM×2 (d), BFSM×3 (e), BFSM×4 (f), and the corresponding ground truth images (g). In the bottom row, from left to right are the input composite image (a), foreground object mask (b), *baseline* (c), w/o CFBA (d), w/o GAI (e), Ours (f) and the corresponding ground truth images (g), respectively.

proposed cross foreground-background attention-aware graph convolutional mechanism (CFBA) and shading style transfer module (SSTM).

Specifically, to investigate that the shading style transfer module is crucial for object illumination harmonization tasks, we conduct four experiments including Illuminator without SSTM as *baseline*, Illuminator with SSTM consisting of 2, 3, 4 BFSMs, that is, BFSM×2, BFSM×3, BFSM×4, respectively.

From Table 2, we can observe that our method (Illuminator / (BFSM×3)) achieves the best performance on all four evaluation metrics. Comparing Illuminator / (BFSM×3) with the *baseline*, our method shows the great superiority. This strongly demonstrates the importance of BFSM×3 for illumination harmonization in our method. However, when we use BFSM×2, although the performance is better than that of the baseline, it is not as good as BFSM×3. The results for the BFSM×4 show that more parts do not guarantee that better performance but consumes more time and memory for training and testing. To demonstrate the roles of these components intuitively, the first row in Figure 9 presents different visual illumination harmonization results. We observe that the Illuminator with the BFSM×3 and BFSM×4 generate the better illumination harmonization result, while BFSM ×4 requires more parameters and time consumption.

Table 2 Ablation study of the shading style transfer module (SSTM). "Baseline" denotes our method without BFSM module, and BFSM×2, BFSM×3, BFSM×4 indicates using 2, 3, and 4 BFSMs, respectively. The best results are marked in **bold**.

Method	RMSE ↓	SSIM ↑	fMSE ↓	fSSIM ↑
Baseline	6.019	0.821	898.417	0.703
Illuminator / (BFSM×2)	5.891	0.860	566.712	0.928
Illuminator / (BFSM×3)	**4.372**	**0.912**	**342.239**	**0.947**
Illuminator / (BFSM×4)	4.403	0.891	369.702	0.941

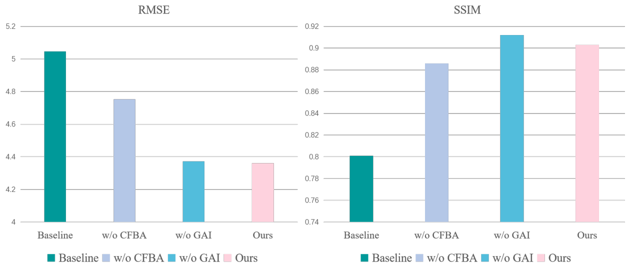

Fig. 10 Quantitative results of ablation study of the shadow residual generation module (CFBA).

Section 3.1 introduces a cross foreground-background attention-aware graph convolutional mechanism (CFBA) for foreground object shadow residual generation. We also analyze the performance of different settings. To verify the effectiveness of CFBA, we compare our Illuminator with its three ablated versions: 1) baseline, that is, removing CFBA and global auxiliary information (GAI); 2) w/o GAI, that is, indicating the CFBA module without GAI; 3) w/o CFBA, that is, retaining GAI while replacing the CFBA with two simple CNN modules. From Figure 10, we can see that our Illuminator achieves better performance than the other components in RMSE and SSIM metrics. Compared to the *baseline*, our method improves 0.684 and 0.102 in RMSE and SSIM, respectively. The second row in Figure 9 shows the results of different components, and we can see that *baseline* fails to produce a foreground object shadow, mainly because of the lack of effective modeling of the relative position relationship between the foreground and the background. Although both without CFBA and without GAI generated object shadows, their results are poor. Contrastly, our Illuminator achieves reasonable and satisfactory results.

5.4 User Study on Real Composite Images

To further verify the generalization capability of our illuminator, we collect 100 additional real composite images outside the RIH dataset and compare the illuminator with the four baseline methods. These real composite images do not have corresponding ground-truth images and contain background scenes and foreground objects that differ significantly from those in the RIH dataset. Following [25, 27, 29], we conduct a user study on the collected images. Specifically, given an input composite image, we can obtain five different illumination harmonization results by using different methods (four baselines and our method). Then, we create image pairs according to randomly choosing two images from five images. Therefore, we can obtain 1000 image pairs based on 100 composite images. We then recruit 100 participants and require them to select the visually more realistic image for each pair. Finally, we collect 100000 pairwise results in total and calculate the global ranking of all methods using the Bradley-Terry model (B-T model) [76, 77] .

Figure 11 shows the B-T scores of Illuminator and the four baseline methods. The proposed Illuminator obtains the highest score. This demonstrates that our method still generalizes well to unseen images outside the dataset. Figure 12 shows the visual illumination harmonization results obtained by our method for two other real-word images randomly selected from the collected images.

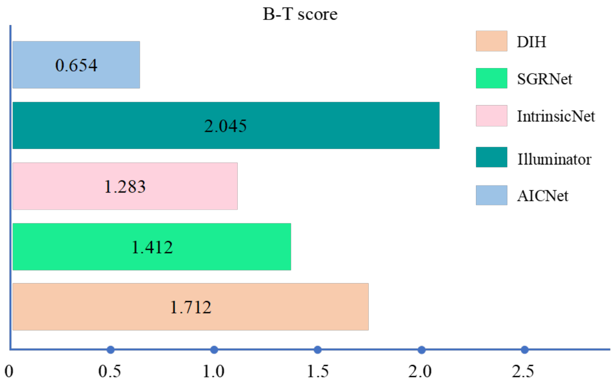

Fig. 11 B-T scores of different methods on our collected images.

5.5 Limitations

Our Illuminator has the following limitations. (1) Illuminator fails to achieve satisfactory illumination harmonization for multiple target objects in a scene. (2) Illuminator may produce unrealistic illumination harmonization results for natural scenes illuminated by multiple light sources.

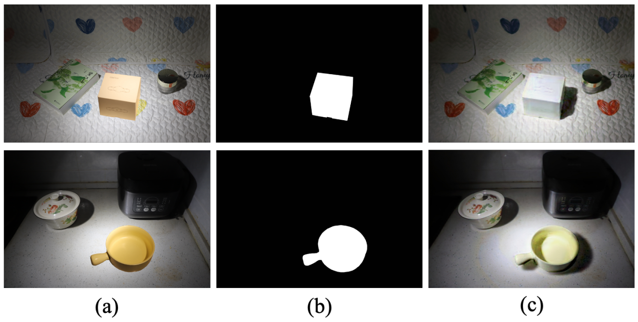

Fig. 12 Illumination harmonization results of our method on other real composite images. From (a) to (c) are input naive composite images, the corresponding foreground object masks, and our results, respectively.

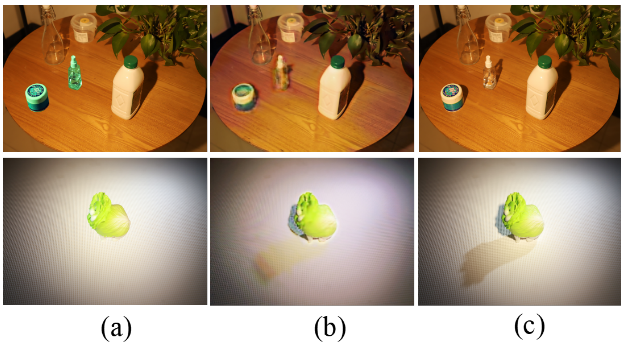

Fig. 13 Limitations of our method. The first row is the multi-object illumination harmonization result, and the second row is the illumination harmonization result under the multi-light sources condition. (a), (b) and (c) are the input composite images, the outputs of our method and the ground truth images, respectively.

Figure 13 (first row) shows an example of limitation (1). Although we can perform an illumination harmonization operation for each object to achieve multi-object illumination harmonization using our method, this is not an ideal way. Our method fails to achieve satisfactory illumination harmonization results when editing multi-object at the same time.

From the second row of Figure 13, Illuminator fails to address multi-light source image illumination harmonization, that is, Limitation (2). Because our method focuses on image-based illumination harmonization without explicitly estimating any 3D information, solving this limitation based on 2D scenes is interesting but challenging and is left as our future work.

6 Conclusion and Future Work

In this work, we have presented an image-based object illumination editing method called Illuminator for indoor scene illumination harmonization, which focuses on producing more realistic illumination harmonization results for challenging indoor scenes. First, we construct a large-scale, high-quality

RIH dataset for real-world indoor illumination harmonization task, and propose a simple yet effective approach to obtain real datasets for other related tasks like shadow generation and removal, image inpainting, and scene relighting.

We then propose a novel illumination harmonization method named Illuminator, which consists of a shadow residual generation branch and an object illumination transfer branch, achieving physically more realistic global illumination harmonization results. The shadow residual generation branch introduces a novel cross foreground-background attention-aware graph convolutional mechanism to model the spatial interaction relationship between the foreground and background, producing plausible shadows for foreground objects. The object illumination transfer branch mainly focuses on achieving illumination consistency between the foreground and background from a shading style perspective. Using these branches, our Illuminator can produce realistic illumination harmonization results and achieve the best performance in terms of both quantitative metrics and qualitative effects. In the future, we plan to extend our Illuminator to address illumination harmonization in complex natural scenes with multiple objects and light sources.

Declaration of competing interest

The authors have no competing interests to declare. Content of this article.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No. 61972298 and No. 62372336), CAAI-Huawei MindSpore Open Fund and Wuhan University-Huawei GeoInformatices Innovation Lab.

References

[1] Chen Z, Long C, Zhang L, Xiao C. CANet: A Context-Aware Network for Shadow Removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 4743–4752.

[2] Yu H, Liu W, Long C, Dong B, Zou Q, Xiao C. Luminance attentive networks for HDR image and panorama reconstruction. *Computer Graphics Forum*, 2021, 40(7): 181–192.

[3] Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Mu TJ, Zhang SH, Martin RR, Cheng MM, Hu SM. Attention mechanisms in computer vision:A survey. *Computational Visual Media*, 2022, 8(3): 38.

[4] Lin Z, Zhang Z, Zhu ZY, Fan DP, Liu XL. Sequential interactive image segmentation. *Computational Visual Media*, 2023, 9(4): 753–765.

[5] Fu G, Zhang Q, Zhu L, Li P, Xiao C. A Multi-Task Network for Joint Specular Highlight Detection and Removal. *IEEE*, 2021.

[6] Fang F, Luo F, Zhang HP, Zhou HJ, Xiao CX. A Comprehensive Pipeline for Complex Text-to-Image Synthesis. , 2020, 35(3): 16.

[7] Eisemann E, Durand F. Flash photography enhancement via intrinsic relighting. *ACM transactions on graphics (TOG)*, 2004, 23(3): 673–678.

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

[9] Guo K, Lincoln P, Davidson P, Busch J, Yu X, Whalen M, Harvey G, Orts-Escolano S, Pandey R, Dourgarian J, et al.. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 2019, 38(6): 1–19.

[10] Fu K, Jiang Y, Ji GP, Zhou T, Zhao Q, Fan DP. Light field salient object detection: A review and benchmark. *Computational Visual Media*, 2022, 8(4): 26.

[11] Lan Y, Duan Y, Liu C, Zhu C, Xiong Y, Huang H, Xu K. ARM3D: Attention-based relation module for indoor 3D object detection. *Computational Visual Media*, 2022, 8(3): 20.

[12] Cao T, Luo F, Fu Y, Zhang W, Zheng S, Xiao C. DGECN: A Depth-Guided Edge Convolutional Network for End-to-End 6D Pose Estimation, 2022.

[13] Fu Y, Yan Q, Liao J, Xiao C. Joint Texture and Geometry Optimization for RGB-D Reconstruction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Li Y, Luo F, Xiao C. Self-supervised coarse-to-fine monocular depth estimation using a lightweight attention module. *:*, 2022, 8(4): 17.

[15] Li YZ, Zheng SJ, Tan ZX, Cao T, Luo F, Xiao CX. Self-Supervised Monocular Depth Estimation by Digging into Uncertainty Quantification. *Journal of Computer Science and Technology*, 2023, 38(3): 510–525.

[16] Huang HZ, Xu SZ, Cai JX, Liu W, Hu SM. Temporally coherent video harmonization using adversarial networks. *IEEE Transactions on Image Processing*, 2019, 29: 214–224.

[17] 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, 300–316.

[18] Ke Z, Sun C, Zhu L, Xu K, Lau RW. Harmonizer: Learning to perform white-box image and video harmonization. In *European Conference on Computer Vision*, 2022, 690–706.

[19] Niu L, Tan L, Tao X, Cao J, Guo F, Long T, Zhang L. Deep Image Harmonization with Globally Guided Feature Transformation and Relation Distillation. *arXiv preprint arXiv:2308.00356*, 2023.

[20] Lalonde JF, Efros AA. Using color compatibility for assessing

image realism. In *2007 IEEE 11th International Conference on Computer Vision*, 2007, 1–8.

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

[22] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015, 234–241.

[23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[24] Zakharov E, Shysheya A, Burkov E, Lempitsky V. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, 9459–9468.

[25] Tsai YH, Shen X, Lin Z, Sunkavalli K, Lu X, Yang MH. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 3789–3797.

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

[27] 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, 8394–8403.

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

[29] 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*, 2021, 16367–16376.

[30] Hong Y, Niu L, Zhang J, Zhang L. Shadow generation for composite image in real-world scenes. *arXiv preprint arXiv:2104.10338*, 2021.

[31] 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*, 2021, 4832–4841.

[32] 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*, 2021, 9361–9370.

[33] Liu D, Long C, Zhang H, Yu H, Dong X, Xiao C. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 8139–8148.

[34] 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 (CVPR)*, 2022, 18542–18551.

[35] Zhan F, Lu S, Zhang C, Ma F, Xie X. Adversarial image composition with auxiliary illumination. In *Proceedings of the Asian Conference on Computer Vision*, 2020, 1–17.

[36] Song Y, Zhang Z, Lin Z, Cohen S, Price B, Zhang J, Kim SY, Aliaga D. ObjectStitch: Object Compositing With Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 18310–18319.

[37] Helou ME, Zhou R, Barthas J, Süsstrunk S. VIDIT: Virtual image dataset for illumination transfer. *arXiv preprint arXiv:2005.05460*, 2020.

[38] Grosse R, Johnson MK, Adelson EH, Freeman WT. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, 2009, 2335–2342.

[39] Barron JT, Malik J. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 2014, 37(8): 1670–1687.

[40] Demir U, Unal G. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

[41] Pitie F, Kokaram AC, 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*, volume 2, 2005, 1434–1439.

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

[43] Pérez P, Gangnet M, Blake A. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, 2003, 313–318.

[44] Tao MW, Johnson MK, Paris S. Error-tolerant image compositing. *International journal of computer vision*, 2013, 103(2): 178–189.

[45] Jia J, Sun J, Tang CK, Shum HY. Drag-and-drop pasting. *ACM Transactions on graphics (TOG)*, 2006, 25(3): 631–637.

[46] Tsai YH, Shen X, Lin Z, Sunkavalli K, Yang MH. Sky is not the limit: semantic-aware sky replacement. *ACM Trans. Graph.*, 2016, 35(4): 149–1.

[47] 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 (CVPR)*, 2023, 5927–5936.

[48] Guerreiro JJA, 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 (CVPR)*, 2023, 5917–5926.

[49] Liu S, Huynh CP, 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 (CVPR)*, 2023, 18290–18299.

[50] 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)*, 2021, 1–6.

[51] 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*, 2021, 14870–14879.

[52] 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 (CVPR)*, 2022, 18470–18479.

[53] 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 (CVPR)*, 2022, 19710–19719.

[54] Karsch K, Sunkavalli K, Hadap S, Carr N, Jin H, Fonte R, Sittig M, Forsyth D. Automatic Scene Inference for 3D Object Compositing. *ACM Transactions on Graphics*, 2014, 33(3): 1–15.

[55] Kee E, O'brien JF, Farid H. Exposing Photo Manipulation from Shading and Shadows. *ACM Trans. Graph.*, 2014, 33(5): 165–1.

[56] Liu B, Xu K, Martin RR. Static scene illumination estimation from videos with applications. *Journal of Computer Science and Technology*, 2017, 32(3): 430–442.

[57] Liao B, Zhu Y, Liang C, Luo F, Xiao C. Illumination animating and editing in a single picture using scene structure estimation. *Computers & Graphics*, 2019, 82: 53–64.

[58] Zhang J, Sunkavalli K, Hold-Geoffroy Y, Hadap S, Eisenman J, Lalonde JF. All-weather deep outdoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 10158–10166.

[59] Arief I, McCallum S, Hardeberg JY. Realtime estimation of illumination direction for augmented reality on mobile devices. *Color and Imaging Conference*, 2012, 2012(1): 111–116.

[60] Worchel M, Alexa M. Differentiable Shadow Mapping for Efficient Inverse Graphics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 142–153.

[61] Sheng Y, Zhang J, Benes B. SSN: Soft Shadow Network for Image Compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 4380–4390.

[62] Sheng Y, Zhang J, Philip J, Hold-Geoffroy Y, Sun X, Zhang H, Ling L, Benes B. PixHt-Lab: Pixel Height Based Light Effect Generation for Image Compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 16643–16653.

[63] Zhang S, Liang R, Wang M. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 2019, 5(1): 105–115.

[64] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[65] Monti F, Boscaini D, Masci J, Rodola E, Svoboda J, Bronstein MM. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 5115–5124.

[66] Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[67] Tse THE, Kim KI, Leonardis A, Chang HJ. Collaborative Learning for Hand and Object Reconstruction With Attention-Guided Graph Convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 1664–1674.

[68] Li M, An L, Zhang H, Wu L, Chen F, Yu T, Liu Y. Interacting Attention Graph for Single Image Two-Hand Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 2761–2770.

[69] Wu SC, Tateno K, Navab N, Tombari F. Incremental 3D Semantic Scene Graph Prediction From RGB Sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 5064–5074.

[70] Chen ZM, Wei XS, Wang P, Guo Y. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, 5177–5186.

[71] Wan S, Gong C, Zhong P, Pan S, Li G, Yang J. Hyperspectral image classification with context-aware dynamic graph convolutional network. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 59(1): 597–612.

[72] Lin J, Yuan Y, Shao T, Zhou K. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 5891–5900.

[73] Wang T, Hu X, Wang Q, Heng PA, Fu CW. Instance shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 1880–1889.

[74] Jiaqi YU, Nie Y, Long C, Wenju XU, Zhang Q, Guiqing LI. Monte Carlo Denoising via Auxiliary Feature Guided Self-Attention. *ACM Transactions on Graphics*, 2021.

[75] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2017, 2961–2969.

[76] Bradley RA, Terry ME. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 1952, 39(3/4): 324–345.

[77] Lai WS, Huang JB, Hu Z, Ahuja N, Yang MH. A comparative study for single image blind deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 1701–1709.