Deep Image-based Illumination Harmonization

Zhongyun Bao[1], Chengjiang Long[2]*, Gang Fu[1], Daquan Liu[1], Yuanzhen Li[1], Jiaming Wu[1], Chunxia Xiao[1]*†
[1]School of Computer Science, Wuhan University, Wuhan, Hubei, China
[2] Meta Reality Labs, Burlingame, CA, USA
clong1@fb.com, xyzgfu@gmail.com, {zhongyunbao, daquanliu, cxxiao}@whu.edu.cn

Abstract

Integrating a foreground object into a background scene with illumination harmonization is an important but challenging task in computer vision and augmented reality community. Existing methods mainly focus on foreground and background appearance consistency or the foreground object shadow generation, which rarely consider global appearance and illumination harmonization. In this paper, we formulate seamless illumination harmonization as an illumination exchange and aggregation problem. Specifically, we firstly apply a physically-based rendering method to construct a large-scale, high-quality dataset (named IH) for our task, which contains various types of foreground objects and background scenes with different lighting conditions. Then, we propose a deep image-based illumination harmonization GAN framework named DIH-GAN, which makes full use of a multi-scale attention mechanism and illumination exchange strategy to directly infer mapping relationship between the inserted foreground object and the corresponding background scene. Meanwhile, we also use adversarial learning strategy to further refine the illumination harmonization result. Our method can not only achieve harmonious appearance and illumination for the foreground object but also can generate compelling shadow cast by the foreground object. Comprehensive experiments on both our IH dataset and real-world images show that our proposed DIH-GAN provides a practical and effective solution for image-based object illumination harmonization editing, and validate the superiority of our method against state-of-the-art methods. Our IH dataset is available at `https://github.com/zhongyunbao/Dataset`.

1. Introduction

As a part of scene editing, editing illumination for inserted object to achieve scene illumination harmonization is really important in computer vision [3, 9, 8, 39] and augmented reality (AR) as the unsatisfactory illumination harmonization greatly affect the user sense of reality. One example is illustrated in Figure 1. It is still difficulty to achieve a satisfactory result for illumination editing even by a experienced professional retoucher. Apparently, it is very challenging to automatically edit illumination harmonization without any human intervention.

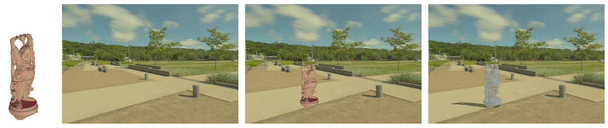
(a) (b) (c) (d)

Figure 1. Illumination editing for an inserted object in a single image. (a) Foreground object imagr with a illumination condition. (b) Background image with a new illumination condition. (c) Naive composite image. (d) Illumination harmonized image.

Some prior efforts have been made to solve this challenging task. In particular, Karsch *et al*. [22] presented an image editing system that supports drag-and-drop 3D object insertion, and Liao *et al*. [27, 28] proposed an approximate shading model for image-based object modeling and insertion. Although these methods produce perceptually convincing results, their performances highly depend on the quality of the estimated geometry, shading, albedo and material properties. However, in some cases, any errors or inaccurate estimation in either geometry, illumination, or materials may result in unappealing editing effects.

Such a shortcoming strongly motivates us to explore a deep learning-based method to directly learn mapping relationship between the inserted image-based foreground object and the real-world scene, and achieve scene illumination harmonization without any explicit inverse rendering (recovering 3D geometry, illumination, albedo and material). Obviously, a dataset with a lot of training image pairs of composite images without illumination harmonization and corresponding ground truth with illumination harmonization is strongly desired for the training purpose.

*This work was co-supervised by Chengjiang Long and Chunxia Xiao.
†Corresponding author.

However, existing datasets like iHarmony4 [4], shadow-AR dataset [30], HVIDIT dataset[13], *etc*., mainly focus on foreground object appearance or the foreground object shadow, what rarely consider global appearance and illumination harmonization. The dataset [41] considers both appearance and shadow of inserted foreground object, it contains only two types of foreground objects: car and person, which is not only unavailable, but also severely limits the generalization and robustness of the illumination harmonization task.

In this work, we first construct a large-scale, high-quality synthesized dataset named IH dataset for the object illumination editing task. To build our dataset, we first collect HDR panoramas to capture background images and illumination information from Laval's HDR dataset [11, 10] and the Internet, which are taken in various indoor and outdoor real-world scenes. Therefore, the scenes in our dataset are general and challenging. Besides, we also collect 60 3D object models with considerably different shapes and postures used as the foreground objects of our composite images. In general, our dataset finally contains 89,898 six-tuples in total, each with one input triplet (*i.e.*, a naive composite image, and the corresponding object mask and background mask), and another ground truth triplet (*i.e.*, a foreground object illumination map, a background illumination map, and a final illumination harmonization image). See Figure 2 for a six-tuples example.

Regarding to the deep learning model, we propose a novel learning-based scene illumination harmonization GAN framework named DIH-GAN, as shown in Figure 3, which incorporates both spatial attention learning [44, 17, 18, 15, 36] and adversarial learning [12, 1] to make the illumination of foreground compatible with background. Our DIH-GAN takes a naive composite image with shadow-free object as well as inserted object mask as input, and makes full use of multi-scale attention mechanisms and adversarial learning to directly infer mapping relationship between the inserted foreground object and the corresponding background scene. Besides, we propose an illumination exchange mechanism to edit object illumination and directly achieve seamless illumination integration between the foreground object and the background of composite image, which makes the synthesized image more harmonious and realistic. Note that, our proposed multi-scale attention mechanism and feature exchange mechanism play a key role, which can avoid the complicated inverse rendering process and directly generate reasonable illumination harmonized results.

Our main contributions are summarized as follows:

- We construct the first large-scale, high-quality image illumination harmonization dataset IH, which consists of 89,898 image six-tuples with a diversity of real-world background scenes and 3D object models.
- We propose a novel deep learning-based scene illumination harmonization GAN framework named DIH-GAN, which is a multi-task collaborative network and can directly perform illumination harmonization editing for the inserted object without explicit inverse rendering.
- Extensive experiments show that the proposed DIH-GAN can effectively achieve high-quality image illumination harmonization and significantly outperform existing state-of-the-art methods.

2. Related work

Object illumination editing. Traditional object illumination editing methods mainly concentrated on estimating the scene geometry, illumination and surface reflectance to edit the object. Previous methods [7, 21] have shown that coarse estimation of scene geometry, reflectance properties, illumination, and camera parameters work well for many image editing tasks. These methods require a user to model the scene geometry and illumination. The method [2] not only recovers shape, surface albedo and illumination for entire scenes, but also requires a coarse input depth map, while this method is not directly suitable for illuminating inserted object. Karsch *et al*. [22] presented a fully automatic method for recovering a comprehensive 3D scene model (geometry, illumination, diffuse albedo and camera parameters) from a single low dynamic range photograph. Liao *et al.* [28] presented an object relighting system that supports image-based relighting, although this method achieves impressive result, it still needs to reshape the object and model the scene.

These methods depend on the physical modeling of object and scene information, and inaccurate reconstruction results will lead to poor results. In contrast, our method automatically edits the object illumination, directly generates harmonized illumination results without complicated inverse rendering, and thus produces better visual effects.

Shadow generation. Recently, with the breakthrough in adversarial learning, generative adversarial network (GAN) [12, 1, 32] have been successfully applied to shadow detection, removal and generation [37, 6, 16, 42, 30]. For shadow generation, Liu *et al.* [30] proposed an ARShadowGAN model, which is able to directly model the mapping relation between the shadow of the foreground object and the corresponding real-world environment based on their constructed dataset. Similar to this method, our method also aims to generate the object shadow without explicit estimation of 3D geometric information. Besides that, our method considers the shading of the object itself. We not only realize reasonable object shadow generation with the similar effect as Liu *et al.*'s [30] method, but also edit the object illumination to achieve overall scene illumination harmonization.

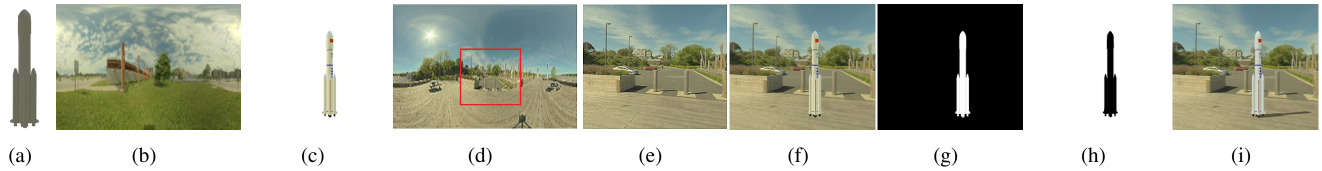

Figure 2. The illustration of a synthesized illumination harmonization image generation process. Given a 3D object with texture (a), we first apply a panorama illumination corresponding with (b) to render the 3D object and get an image-based object (c). Then we paste the image-based object into a background image (e) cropped from (d) directly without illumination adjustment. In this way, along with an object mask (g) and a background mask (h), we get a naive composite image (f) with illumination and shadow inconsistency between object and its surrounding. With the background illumination map (d), we then use Blender to synthesize the illumination harmonization image (i) and take it as the ground-truth for supervised learning. Note we consider (f), (g) and (h) as an input triplet, and (b), (d) and (i) as a ground truth triplet. The input triplet and corresponding ground truth triplet are treated as a six-tuple in our dataset. *Better view in electronic version.*

Image-to-image translation. Image-to-Image translation is to map an input image to a corresponding output image. It has been widely used in various tasks, including super-resolution [23, 25], image quality restoration [31, 38], image harmonization [19, 4, 29, 13], and so on. It is worth mentioning that Cong *et al.* [4] proposed a novel domain verification discriminator, with the insight that the foreground needs to be translated to the same domain as the background for image harmonization, but neglect to explicitly transform the foreground features in the generator. Recently, Ling *et al.* [29] treated image harmonization as a style transfer problem to explicitly formulates the visual style from the background and adaptively applies them to the foreground, Guo *et al.* [13] modeled image harmonization based on intrinsic image theory.

These methods all focus on the illumination of foreground object and do not consider object shadow generation task. Different from all existing methods, our task takes into account both illuminating the object and generating the cast shadow of the object, and achieves illumination harmonization for the whole scene.

3. Our IH Dataset

The construction process of the IH dataset includes three steps: (1) collecting images and 3D models, (2) filtering background images, and (3) rendering and composition. In the following, we will describe these steps in details.

Collecting images and 3D models. We first collect all images from the Laval's HDR panorama dataset[1] [11, 10], and capture 2,686 HDR panorama images from the Internet with a diversity of real-world scenes. For each panorama image, we extract 8 limited field-of-view crops to produce the background images of composite image, and also use the corresponding panorama images centering on the crops as the illumination to render ground truth results. We initially obtain 22,256 background images with the corresponding illumination map in total. Also, we collect 60 3D models from the website (`https://laozicloud.com`)[2] as inserted objects, which contain various types of objects, such as *bunny*, *person*, *lucy etc*.

Filtering background images. To ensure the quality of the dataset for our object illumination editing task, we further filter out the following three kinds of images: (1) without obvious or natural-looking illumination, (2) without a reasonable place to insert virtual object, and (3) with inconspicuous or no shadows. By this way, we finally obtain 12,253 remaining background images.

Rendering and compositing. With collected 3D models, background images, and the corresponding panorama maps, the ground truth object relighting images (see Figure 2 (i)) are rendered using Blender. Specifically, we first specify a plane at the bottom of the inserted object for casting shadows, then embed the 3D object into the cropped background image, and finally use the corresponding panorama map to render the illumination of the object to produce the final result. Note that, due to the corresponding backgrounds are real-world 2D scene images, we use Photoshop to manually annotate each foreground object in our dataset to obtain accurate masks.

In the construction process, we use 60 virtual models with different pose configures using the pipeline shown in Figure 2 to construct our dataset based on different background images, and we produce 169,672 synthesized ground truth illumination harmonization images in total. Moreover, to improve the training efficiency, we only use 89,898 six-tuples to train the our network in final. Each six-tuple consists of two triplets. One triplet as input data includes a naive composite image without illumination adjustment, and the corresponding object mask and background mask. The other one as ground-truth data includes a synthesized illumination harmonization image, one object illumination and one background illumination ground-truths. A visual six-tuple example is shown in Figure 2. Refer to the supplementary for more details of the dataset analysis.

[1] The author Zhongyun Bao signed the license and produced all the experimental results in this paper. Meta did not have access to the Laval's HDR panorama dataset.

[2] The author Zhongyun Bao purchased the 3D models for non-commercial research purpose only and produced the experimental results in this paper. Meta did not have access to these data.

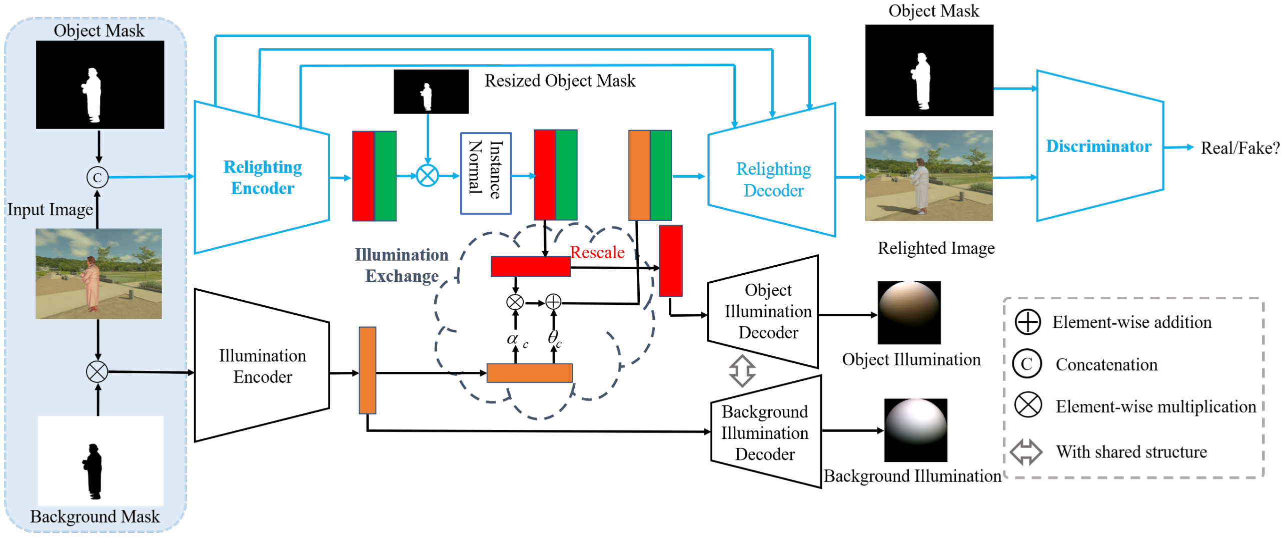

Figure 3. The overview of our proposed DIH-GAN. Given an input image with inserted object and the corresponding object mask and background mask, the generator of our DIH-GAN can generate the relighting image (R-Network marked in blue) and predict both object illumination and background illumination (I-Network marked in black), and the discriminator can distinguish whether the generated relighting image is real or fake. The Illumination exchange mechanism between the R-Network and the I-Network realizes the conversion of illumination information between the scene and the object.

4. Proposed Method

Our goal is to train a GAN that takes a naive composite image $\hat{Y}$ with inharmonious illumination, corresponding the object mask, background mask and corresponding target illumination as input, and directly generate the corresponding scene illumination harmonized image $\bar{Y}$. To achieve this goal, we propose a novel framework called DIH-GAN, of which the generator is a multi-task parallel network composed of two networks, *i.e.*, Relighting Network (R-Network) and Illumination Network (I-Network) to handle object and illumination separately. See Figure 3.

4.1. Generator

As shown in Figure 3, the generator of our DIH-GAN contains two parallel branch networks, *i.e.*, R-Network and I-Network. R-Network learns the overall features of the input image and I-Network predicts the object and background illumination. They work collaboratively to complete the task by illumination exchange mechanism.

Relighting Encoder. For the U-Net [33] like R-Network, there are five down-sampling blocks in encoder and each down-sampling block consists of a residual block with 3 consecutive convolutions, instance normalization and ReLU operation and halves the feature map with an average pooling operation. Each down-sampling block is followed by a multi-scale attention block which guides the network to infer the object shadow and generates the refinement feature maps. Note that we design such a multi-scale attention mechanism for two purposes: (1) to adaptively extract reliable multi-scale features and overcome the scale-variation across the image to assign larger weights to areas of interest for refinement; and (2) to guide the generation of shadows of the inserted objects by paying attention to real shadows and corresponding occluders in the scene.

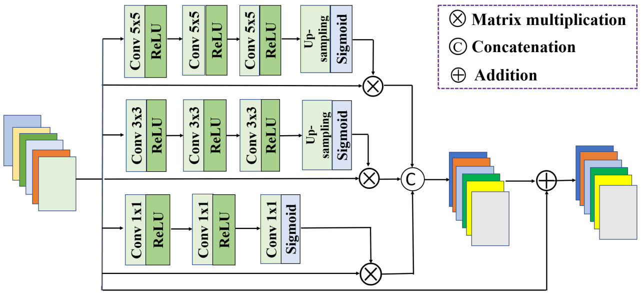

Figure 4. Illustration of our multi-scale attention mechanism.

As shown in Figure 4, the multi-scale attention block has three types convolution layers with three different kernel sizes, 1×1, 3×3, 5×5, to extract features in different scales. Specially, for the input feature map, the multi-scale attention block first extracts features using two 1×1 convolution layers with crossing channels and squeezing features, two 3×3 convolution layers and two 5×5 convolution layers to generate feature maps. Note that for 3×3 and 5×5 convolution, the feature map size of each channel has been changed and therefore we apply an up-sampling layer to recover the original size before feeding the feature map into the Sigmoid function to produce attention map. We separately conduct an element-wise multiplication on the input feature and the attention map at each scale to produce attended feature maps, which are then concatenated at channel-wise together and fed into a 1×1 convolution layer to recover the same channel number with the origin input feature. We apply a residual structure [14] to combine it with the origin input feature map together as final output. This residual mechanism not only accelerates the convergence speed but also correct image details such as border artifacts.

The final output features of the encoder include the illumination features F_{illu} and non-illumination features F_{noillu} of global image. This feature separation is enforced by the no-illumination loss $\mathcal{L}_{noillu}$ (see Eq. (6)).

Illumination Encoder. For the I-Network, the encoder has a similar structure to the one of the R-Network and takes the result of multiplying the input image and background mask as input to extract the illumination feature of background. The background illumination feature is then exchanged with the object illumination feature of R-Network by the illumination exchange mechanism. Note that the output of illumination encoder have two functions: the background illumination features are used in combination with object non-illumination features F_{noillu} of R-Network encoder to produce the illumination harmonized image in the decoder of R-Network; the background illumination features are also fed to the decoder of I-Network to predict the corresponding background illumination information through supervised learning.

Illumination Exchange Mechanism. After obtaining the features extracted by the two encoders, inspired by [29], we use the background illumination features in the I-Network to guide the foreground object illumination features in the R-Network, and exchange them for the input of the decoders.

To specify, the two sub-networks work together through the illumination exchange mechanism. It is worth mentioning that at the bottleneck feature of the R-Network, we perform multiplication operation on it with the object mask of the corresponding size, and get the normalized foreground object features F^{obj} by using IN [35]. This treatment is able to better realize the exchange of object illumination and background illumination, and achieve the illumination harmonization task.

The normalized foreground object feature F^{obj} can be divided into two parts: non-illumination features F^{obj}_{noillu} which is independent of illumination feature F^{obj}_{illu}. The illumination feature F^{obj}_{illu} have two functions: one is to be cropped by the resized object mask, rescaled to a larger size and then fed into the object illumination decoder of the I-Network to predict the object illumination. The other is to be affined by learned scale and bias from the background illumination features F^{bg}_{illu} extracted by the I-Network encoder, and then the affined features are concatenated with F^{obj}_{noillu} and fed into the R-Network decoder to generate the realistic Illumination harmonization image. The affine result is computed by:

$$\mathcal{F}_{affine} = \alpha_c F^{obj}_{illu} + \theta_c, \tag{1}$$

where θ_c and α_c are the mean and standard deviation of the activations of the background illumination features in channel c:

$$\theta_c = \frac{1}{N_{bg}} \sum_{h,w} F^{bg}_{illu}, \tag{2}$$

$$\alpha_c = \sqrt{\frac{1}{N_{bg}} \sum_{h,w} (F^{bg}_{illu} - \theta_c)^2}, \tag{3}$$

where N_{bg} is the total number of pixels of background illumination, h, w denote the height and width of features, respectively.

In the whole illumination harmonization task, we have supervised constraints on the relighting image, the non-illumination and illumination feature of the object, and the background illumination feature, respectively, which improves the harmonization accuracy of the relighting image.

Relighting Decoder. The decoder in the R-Network consists of five up-sampling layers. Each up-sampling layer doubles the feature map by nearest interpolation followed by consecutive dilated convolution, instance normalization and ReLU operations. The last feature map is activated by a sigmoid function. The R-network concatenates down-up sampling layers by skip connections.

Object/Background Illumination Decoders. Following [40], the decoder of the I-Network is to predict the illumination. In this paper, we use the shared structure for both object illumination and background illumination decoders.

4.2. Discriminator

The discriminator of DIH-GAN is designed to help the R-Network accelerate convergence and generate a plausible harmonized image. Following Patch-GAN [19], our discriminator consists of six consecutive convolutional layers. Each convolutional layer contains convolution, instance normalization and ReLU operations. We use Sigmoid function to activate last feature map produced by a convolution, and perform a global average pooling operation on the activated feature map to obtain the final output of the discriminator.

4.3. Loss functions

The total loss $\mathcal{L}_{total}$ is formulated with an illumination loss $\mathcal{L}_{illu}$, a non-illumination loss $\mathcal{L}_{noillu}$, a perceptual loss $\mathcal{L}_{per}$, an adversarial loss $\mathcal{L}_{adv}$, and a classical L_1-normal reconstruction loss $\mathcal{L}_{Recons}$ as follows:

$$\begin{aligned}\mathcal{L}_{total} =& \beta_1 \mathcal{L}_{illu} + \beta_2 \mathcal{L}_{noillu} + \beta_3 \mathcal{L}_{per} \\ &+ \beta_4 \mathcal{L}_{adv} + \mathcal{L}_{Recons},\end{aligned} \tag{4}$$

where β_1, β_2 , β_3, β_4 are weighting parameters controlling the influence of each term.

Illumination loss $\mathcal{L}_{illu}$ is the element-wise illumination loss between the generated illumination and the corresponding ground truth, *i.e.*,

$$\mathcal{L}_{illu} = \|Y_{OI} - \bar{Y}_{OI}\|_2^2 + \|Y_{BI} - \bar{Y}_{BI}\|_2^2, \tag{5}$$

where $\bar{Y}_{OI}$ and $\bar{Y}_{BI}$ represent the output foreground object illumination image, background illumination image, and relighting image respectively. Y_{OI} and $Y_{BI}$are their corresponding ground truth images.

Non-illumination feature loss $\mathcal{L}_{nonillu}$ is introduced to enforce non-illumination feature matching to improve the accuracy of the object relighting image. According to [43] and Retinex theory [24], reflectance is the inherent physical property of object, independent of illumination. Therefore, we expect that the same object under different illumination conditions have the same non-illumination (*i.e.*, reflectance) features,

$$\mathcal{L}_{nonillu} = (F^1_{nonillu} - F^2_{nonillu})^2 / N_{nonillu}, \tag{6}$$

where $F^1_{nonillu}$ and $F^2_{nonillu}$ are non-illumination features of the same insert object under the different illumination conditions, and $N_{nonillu}$ is the number of elements in $F_{nonillu}$.

Perceptual loss $\mathcal{L}_{per}$ [20] is used to measure the semantic difference between the generated image and the ground truth. Following [30], we use a VGG-16 model [34] pre-trained on ImageNet dataset [5] to extract feature and choose the first 10 VGG16 layers to compute feature map. $\mathcal{L}_{per}$ is defined as:

$$\begin{aligned}\mathcal{L}_{per} = &\text{MSE}(V_{Y_{FI}}, V_{\bar{Y}_{FI}}) + \text{MSE}(V_{Y_{BI}}, V_{\bar{Y}_{BI}}) \\ &+ \text{MSE}(V_{Y_R}, V_{\bar{Y}_R}),\end{aligned} \tag{7}$$

where MSE is the mean squared error, and $V_i = \text{VGG}(i)$ is the extracted feature map.

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

$$\mathcal{L}_{adv} = \log(\mathbf{D}(x, m, Y)) + \log(1 - \mathbf{D}(x, m, \bar{Y})), \tag{8}$$

where $\mathbf{D}(\cdot)$ is the probability that the image is "real". x is the input image and m is the corresponding mask, $\bar{Y}$ is the output of the generator of DIH-GAN, and Y is the ground-truth. The discriminator tries to maximize $\mathcal{L}_{adv}$ while the generator tries to minimize it.

4.4. Implementation details

Our DIH-GAN model is implemented by Tensorflow and runs with NVIDIA GeForce GTX 1080Ti GPU. We split the 89,898 six-tuples into 71,918 six-tuples for training and 17,980 six-tuples for testing. Note that, there is no crossover between the foreground objects in our training dataset and testing dataset. Our network is trained for 80 epochs, and the resolution of all images for training and testing is 256 × 256. The initial learning rate is 10^{-4}. We set β_1 = 25.0, β_2 = 6.0, β_3 = 0.04, β_4 = 0.5 and adopt Adam optimizer to optimize the DIH-GAN and discriminator.

5. Experiments

5.1. Evaluation Metrics and Experimental Settings

Evaluation metrics. To evaluate the performance of our DIH-GAN, we adopt two commonly-used evaluation metrics including RMSE and SSIM. In addition, we also introduce other two evaluation metrics including fMSE and fSSIM to evaluate the performance on foreground regions. These two metrics are to compute MSE and SSIM values between foreground regions of input and corresponding ground truth. Overall, smaller fMSE, RMSE, and larger fSSIM, SSIM indicate better results.

Compared methods. We choose one traditional relighting method ASI3D [22] with the similar task as ours, and other three deep learning-based methods from the related fields: one shadow generation method ARShadowGAN [30], and two image harmonization methods including DoveNet [4] and Intrinsic-Net [13]. For fair comparison, we re-train ARShadowGan, DoveNet and Intrinsic-Net on our training set, and test them on our testing set for our illumination harmonization task.

5.2. Comparison with Start-of-the-Art Methods

Quantitative comparison. Table 1 reports the quantitative comparison results on our testing set. As can be seen, our DIH-GAN achieves the best quantitative results on all these four evaluation metrics. This is mainly because the traditional methods ASI3D rely on the estimation accuracy of 3D information of objects and scenes. Inaccurate estimation of 3D information often leads to poor results. As a deep learning based method, our DIH-GAN does not require complicated 3D information estimation and instead it uses the attention mechanism to enhance the beneficial features for a better result. The best performance of DIH-GAN is mainly attributed to the multi-scale attention mechanism, feature exchange mechanism and adversarial learning, which can better guide the illumination editing of inserted object, refine the features and bridge the illumination gap between inserted object and background environment to obtain results closer to the ground truth.

Visual comparison. We provide some visual comparison results in Figure 5. As we can see, our DIH-GAN not only achieves the illumination transformation of different scenes, but also gains the best visual results with plausible object shadows and harmonious illumination. Among

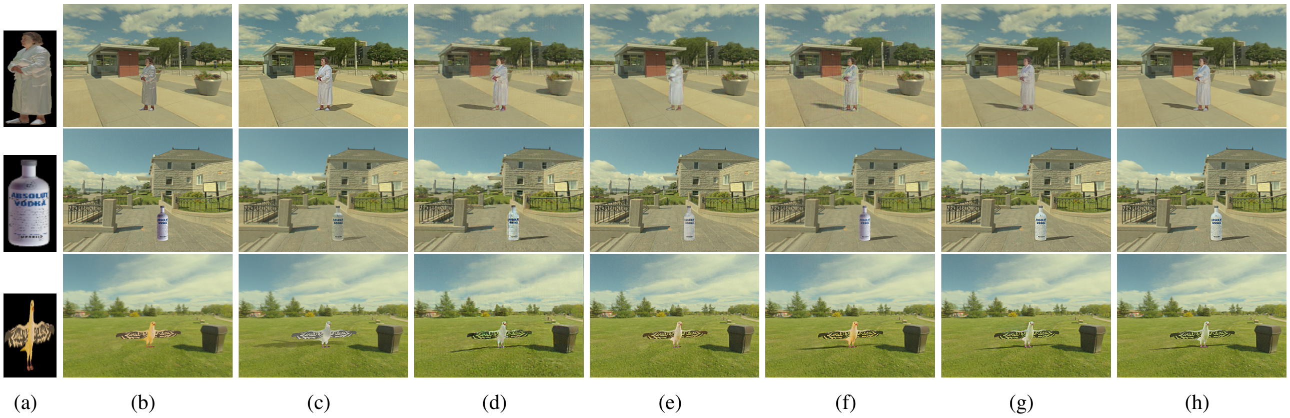

Figure 5. Visual comparison of our method against other start-of-the-art methods on our testing set. (a) 2D foreground object with a illumination condition. (b) Input composite image with a new illumination condition background. (c)-(g) Results produced by ASI3D DoveNet, Intrinsic-Net, ARShadowGAN, and our DIH-GAN respectively. (h) Ground truth.

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

Method	RMSE ↓	fMSE ↓	SSIM ↑	fSSIM ↑
ASI3D [22]	8.116	922.17	0.827	0.764
DoveNet [4]	6.825	698.71	0.934	0.876
Intrinsic-Net [13]	7.493	842.18	0.921	0.803
ARShadowGAN [30]	7.043	744.06	0.928	0.812
DIH-GAN	**6.421**	**618.44**	**0.957**	**0.882**

these competing methods, ASID3D may estimate inaccurate information of geometry and illumination of the object and scene. Although ARShadowGAN generates reasonable object shadows, it has weak illumination processing and therefore cannot reasonably edit the object illumination. For Intrinsic-Net and DoveNet, they target at the appearance harmonization of image and can not well address the object shadow (see the 2nd and 3rd rows of Figure 5 (b)(c)). In contrast, DIH-GAN is able to achieve the better results with plausible object shadow and harmonious illumination, which mainly because our network makes full use of the collaborative R-Network and I-Network parallel with the multi-scale attention mechanism, the illumination feature exchange mechanism, and the adversarial learning strategy to automatically infer the shadow and illumination generation of the object.

5.3. Ablation Study

We conduct ablation study to evaluate the performance of the proposed multi-scale attention mechanism (MSA), illumination exchange mechanism (IEM) and perceptual loss $\mathcal{L}_{per}$, non-illumination feature loss $\mathcal{L}_{nonillu}$ and adversarial loss $\mathcal{L}_{adv}$.

Table 2. Ablation study. "Basic" denotes our method without multi-scale attention mechanism (MSA), IEM and the used perceptual loss $\mathcal{L}_{per}$, non-illumination feature loss $\mathcal{L}_{nonillu}$, and adversarial loss $\mathcal{L}_{adv}$. The best results are marked in **bold**.

Method	RMSE ↓	SSIM ↑	fSSIM ↑
Basic	8.322	0.926	0.817
Basic + MSA + IEM	7.074	0.948	0.831
Basic + $\mathcal{L}_{adv}$ + IEM	7.116	0.944	0.828
Basic + $\mathcal{L}_{per}$ + IEM	7.148	0.935	0.822
Basic + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$ + $\mathcal{L}_{adv}$ + IEM	7.076	0.947	0.824
Basic + MSA + $\mathcal{L}_{adv}$ + $\mathcal{L}_{nonillu}$ + IEM	6.742	0.951	0.841
Basic + MSA + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$ + IEM	6.741	0.949	0.835
Basic + MSA + $\mathcal{L}_{adv}$ + $\mathcal{L}_{per}$ + IEM	6.522	0.953	0.858
Basic + MSA + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$ + $\mathcal{L}_{adv}$	7.092	0.945	0.826
DIH-GAN	**6.421**	**0.957**	**0.882**

The quantitative and visual comparison results are shown in Table 2 and Figure 6, respectively. As we can see in Table 2, our DIH-GAN with all components is able to obtain better results than other methods with one or two components in all three evaluation metrics. By comparing DIH-GAN with "Basic + MSA + $\mathcal{L}_{nonillu}$ + $\mathcal{L}_{adv}$ + IEM", "Basic + MSA + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$ + IEM" and "Basic + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$ + $\mathcal{L}_{adv}$ + IEM" respectively in Table 2, we find that our proposed multi-scale attention mechanism and the used perceptual loss ($\mathcal{L}_{per}$) and adversial loss ($\mathcal{L}_{adv}$) are all beneficial to our final results.

From Figure 6, we observe that DIH-GAN generates better object shadow and illumination than "Basic + MSA + IEM + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$" with unnatural result. The "Basic + MSA + IEM + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$" yields this worse results mainly because the network does not converge, which highlights the advantage of adversarial learning to accelerate network convergence in the task. Another observation is that "Basic + IEM + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$ + $\mathcal{L}_{adv}$" produces a relatively poor result with coarse object shadow and unnatural illumination compared to DIH-GAN, which demonstrates that our proposed multi-scale attention mechanism can make full use of important features to guide the shadow generation of inserted object and refine the extracted useful features of different scales. In addition, we find that the object with a poor illumination result from (f), mainly because there is no IEM to exchange illumination information. Although result (g) is closer to the best one produced by our

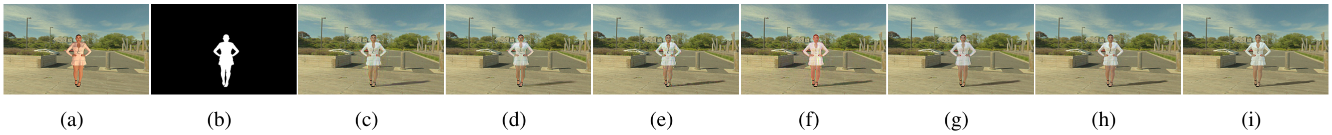

Figure 6. Ablation study for our DIH-GAN. (a) Input image without illumination harmonization. (b) mask. (c) Basic + MSA + IEM + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$. (d) Basic + IEM + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$ + $\mathcal{L}_{adv}$. (e) Basic + MSA + IEM + $\mathcal{L}_{nonillu}$ + $\mathcal{L}_{adv}$. (f) Basic + MSA + $\mathcal{L}_{per}$ + $\mathcal{L}_{nonillu}$ + $\mathcal{L}_{adv}$. (g) Basic + MSA + IEM + $\mathcal{L}_{adv}$ + $\mathcal{L}_{per}$. (h) DIH-GAN. (i) Ground truth. *Please zoom in to observe the detailed difference.*

DIH-GAN, the appearance of the object of DIH-GAN is more refined, which indicates that the non-illumination feature loss $\mathcal{L}_{nonillu}$ can encourage our network to generate a more accurate illumination harmonization image.

5.4. Perceptual Study

We conduct a perceptual study as done in [26] to further evaluate the performance of our method. We choose 100 testing images with various illumination conditions. Among them, 50 images are from the real-world, and the others are generated by our DIH-GAN.

Then we recruited 100 participants from a school campus for subject evaluation. We divide each image into three visual levels: (1) *Real*: realistic illumination harmonization, (2) *Fake*: unrealistic illumination harmonization with artifacts, (3) *Uncertain*: uncertain result which they can not make a decision. 52.4% of real images are judged to be real and the other real images are judged to be fake or uncertain. At the same time, 44.2% of the images produced by our DIH-GAN are judged to be real, which has almost the same evaluation result as the real images. This illustrates that our network can produce high-quality photorealistic results without artifacts.

5.5. Discussions

Robustness. To verify the robustness of our method, we test our DIH-GAN with new cases outside IH dataset. Specifically, our test cases contain four real-world scenes with different objects, two of which are virtual objects and the other two are real objects from the Internet. Note that the ground truths of all testing images are not available. The one visual result is shown in Figure 7. From Figure 7 (c) and (d), we can see that our DIH-GAN not only generates more plausible object shadows than ARShadowGAN, but also successfully achieves the illumination transformation thus to produce the scene illumination harmonization.

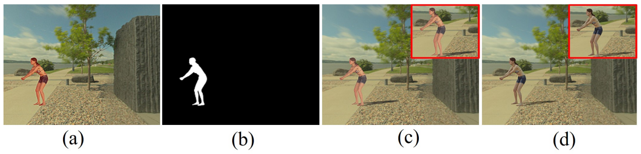

Figure 7. The robustness test. (a) Input image. (b) The corresponding input mask. (c)-(d) The illumination harmonization results from ARShadowGAN and the proposed DIH-GAN respectively.

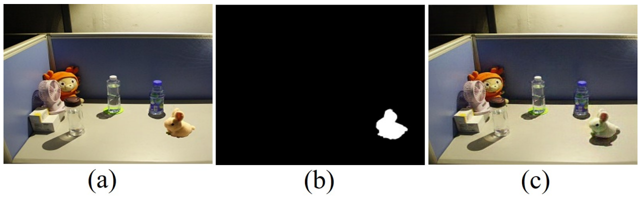

Figure 8. Illumination harmonization editing on a real image. From left to right: input image (a), the corresponding input mask (b) and the result (c).

Generalization. To verify the generalization ability, we test our DIH-GAN on 200 real-world images. Figure 8 presents one visual example. We also adopt the similar perceptual study as done in Section 5.4 to evaluate the performance of our method. The evaluation result of subjects is that 62.7%, 21.3% and 16.0% of all results produced by our method are real, fake and uncertain, respectively. It shows that DIH-GAN has a good generalization ability and is able to usually produce high-quality results without visually noticeable artifacts for unseen real-world images.

Limitations. Our work has two limitations. First, DIH-GAN mainly focus on the outdoor scenes, and is less able to produce satisfactory results for indoor scenes especially with dark lighting and multiple light sources. Second, our dataset only contains the planar shadows.

6. Conclusion and Future Work

In this work, we have presented a large-scale and high-quality illumination harmonization dataset IH and proposed a novel deep learning-based method DIH-GAN to edit the illumination of the inserted object and generate visually plausible illumination harmonized result without any intermediate inverse rendering process. In the future, we will extend our DIH-GAN and the dataset to address video illumination harmonization.

Acknowledgments

This work is partially supported by Prof. Chunxia Xiao's Key Technological Innovation Projects of Hubei Province (2018AAA062) and NSFC (No. 61972298).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. In *International conference on Machine Learning*, 2017. 2

[2] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 38(4), 2013. 2

[3] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Canet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4743–4752, 2021. 1

[4] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3, 6, 7

[5] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, and Fei Fei Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 6

[6] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *Proceedings of the IEEE International Conference on Computer Vision*, 2020. 2

[7] Jean Franxe, ois Lalonde, Derek W Hoiem, Alexei A. Efros, Carsten Rother, John M Winn, and Antonio Criminisi. Photo clip art. *ACM Transactions on Graphics*, 2007. 2

[8] Gang Fu, Qing Zhang, Qifeng Lin, Lei Zhu, and Chunxia Xiao. Learning to detect specular highlights from real-world images. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1873–1881, 2020. 1

[9] Gang Fu, Qing Zhang, Lei Zhu, Ping Li, and Chunxia Xiao. A multi-task network for joint specular highlight detection and removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7752–7761, 2021. 1

[10] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7175–7183, 2019. 2, 3

[11] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean Fran?Ois Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics*, 36(6), 2017. 2, 3

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[13] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16367–16376, 2021. 2, 3, 6, 7

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 5

[15] Tao Hu, Chengjiang Long, and Chunxia Xiao. A novel visual representation on text using diverse conditional gan for visual recognition. *IEEE Transactions on Image Processing*, 30:3499–3512, 2021. 2

[16] Xiaowei Hu, Yitong Jiang, Chi Wing Fu, and Pheng Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[17] Ashraful Islam, Chengjiang Long, Arslan Basharat, and Anthony Hoogs. Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[18] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 3, 5

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 6

[21] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):1–12, 2011. 2

[22] Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. Automatic scene inference for 3d object compositing. *ACM Transactions on Graphics*, 33(3):1–15, 2014. 1, 2, 6, 7

[23] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[24] Edwin H Land and John J McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, 1971. 6

[25] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3

[26] Chen Li, Kun Zhou, Hsiang Tao Wu, and Stephen Lin. Physically-based simulation of cosmetics via intrinsic image decomposition with facial priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1455–1469, 2018. 8

[27] Zicheng Liao, Kevin Karsch, and David Forsyth. An approximate shading model for object relighting. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015. 1

[28] Zicheng Liao, Kevin Karsch, Hongyi Zhang, and David Forsyth. An approximate shading model with detail decomposition for object relighting. *International Journal of Computer Vision*, 2018. 1, 2

[29] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9361–9370, 2021. 3, 5

[30] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 6, 7

[31] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in neural information processing systems*, pages 2802–2810, 2016. 3

[32] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *Computer ence*, pages 2672–2680, 2014. 2

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 6

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[36] Bhavan Vasu and Chengjiang Long. Iterative and adaptive sampling with spatial attention for black-box model explanations. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[37] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 2

[38] Li Xu, Jimmy SJ. Ren, Ce. Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. *Advances in neural information processing systems*, 2:1790–1798, 2014. 3

[39] Hanning Yu, Wentao Liu, Chengjiang Long, Bo Dong, Qin Zou, and Chunxia Xiao. Luminance attentive networks for hdr image and panorama reconstruction. In *Computer Graphics Forum*, volume 40, pages 181–192. Wiley Online Library, 2021. 1

[40] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2019. 5

[41] F. Zhan, S. Lu, C. Zhang, F. Ma, and X. Xie. *Adversarial Image Composition with Auxiliary Illumination*. Computer Vision – ACCV 2020, 2021. 2

[42] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computing Visual Media*, 5(1):105–115, 2019. 2

[43] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7194–7202, 2019. 6

[44] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai. An empirical study of spatial attention mechanisms in deep networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2020. 2