Foreground Harmonization and Shadow Generation for Composite Image

Jing Zhou
School of Computer Science
Wuhan University, China
zhoujing@whu.edu.cn

Ziqi Yu
School of Computer Science
Wuhan University, China
ziqiyu@whu.edu.cn

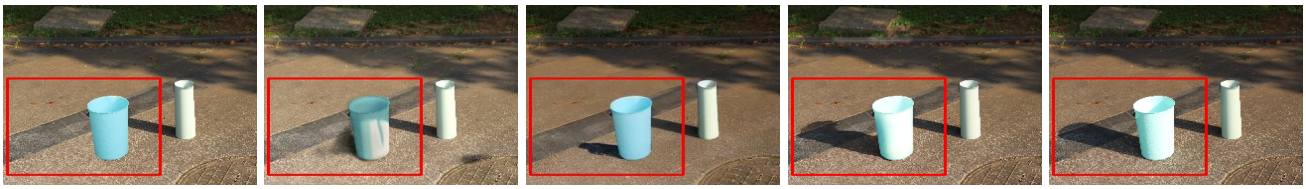
Zhongyun Bao
School of Computer Science
Wuhan University, China
tantouxy@163.com

Gang Fu
Department of Computing
The Hong Kong Polytechnic University, China
xyzgfu@gmail.com

Weilei He
School of Computer Science
Wuhan University, China
weileihe090@whu.edu.cn

Chao Liang
School of Computer Science
Wuhan University, China
cliang@whu.edu.cn

Chunxia Xiao*
School of Computer Science
Wuhan University, China
cxxiao@whu.edu.cn



(a) Composite image (b) DIH-GAN (c) SGDiffusion (d) Ours (e) Ground Truth

Figure 1: Illumination editing effects. Given an composite image, our goal is to harmonize the foreground object and generate its cast shadow. From left to right are composite image, DIH-GAN, SGDiffusion and our method, Ground Truth, respectively.

ABSTRACT

We propose a method for lighting and shadow editing of outdoor disharmonious composite images, including foreground harmonization and cast shadow generation. Most existing works can only perform foreground appearance editing task or only focus on shadow generation. In fact, lighting not only affects the brightness and color of objects, but also produces corresponding cast shadows. In recent years, diffusion models have demonstrated their strong generative capabilities, and due to their iterative denoising properties, they have a significant advantage in image restoration task. But it fails to preserve content structure of image. To this end, we propose an effective model to tackle the problem of foreground lighting-shadow editing. Specifically, we use a coarse shadow prediction module (SP) to generate coarse shadows for foreground objects. Then, we use the predicted results as prior knowledge to guide the generation of harmony diffusion model. In this process, the primary task is to learn lighting variation to harmonize foreground regions, the secondary task is to generate high-quality cast shadow containing more details. Considering that existing datasets do not support the dual tasks of image harmonization and shadow generation, we construct a real outdoor dataset, named IH-SG, covering various lighting conditions. Extensive experiments conducted on existing benchmark datasets and the IH-SG dataset demonstrate the superiority of our method.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Image harmonization, shadow generation, diffusion model

ACM Reference Format:
Jing Zhou, Ziqi Yu, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao. 2024. Foreground Harmonization and Shadow Generation for Composite Image. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681355

*Corresponding authors

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681355

1 INTRODUCTION

Image coposite refers to the process of combining images from different sources to create new images, which has a variety of applications like advertisement propaganda and digital entertainment. However, due to variations in lighting conditions, camera parameters, and other factors, composite images often have inconsistent lighting statistics compared to real images. This necessitates image harmonization to adjust the appearance of the foreground for visual consistency. Additionally, most existing image harmonization methods focus solely on the lighting effects of foreground. However, lighting also produces corresponding cast shadows, which provide important clues about object shape, position, and relative depth. Therefore, shadow generation is equally essential for achieving lighting-shadow consistency.

For image harmonization, most traditional methods [5, 21, 33, 34, 36, 46, 48, 54] focus on matching low-level appearance statistics. They fail to solve the significant appearance differences between foreground and background images. Deep learning-based methods provide powerful capabilities for modeling regional appearances to facilitate harmonization. Some methods [9, 50] explore the semantic information to reconstruct coordinated images. Several methods [6, 8, 17, 26] explore domain adaptation to bring the predicted foreground closer to the original background domain. These methods treat the background image as a whole and may ignore changes in spatial lighting. Guo et al. [16] introduced the Retinex theory into the image harmonization task. But intrinsic decomposition itself is a difficult problem. The methods [14, 15] explore Transformer-based method for image harmonization. Considering the limitations of existing datasets, self-supervised or semi-supervised methods are proposed [22, 30, 32, 52]. The methods [7, 12] adjust the image through color transformation, but ignore lighting effects. Tan et al. [47] used unrelated L, a, b features to guide image reconstruction, which can better adjust brightness.

For shadow generation, rendering-based methods [42–44] require explicit knowledge of lighting, reflectance, material properties, and scene geometry to generate shadows for inserted virtual objects using rendering techniques. However, obtaining such knowledge is often impractical in real-world scenarios. The estimated results are influenced by the accuracy of the input information [11, 20]. Deep learing-based methods [19, 28], on the other hand, directly learn the mapping from input images to output images with foreground shadows, without requiring explicit knowledge of lighting, reflectance, etc. Bao et al. [2] considered both harmonizing the foreground objects and generating reasonable shadows for the foreground objects. But this method only focuses on indoor images.

To address these issues, we propose a novel method for both image harmonization and shadow generation in this paper. As our task requires generating plausible cast shadows for the foreground objects, we develop a coarse shadow prediction module to effectively utilize background information to generate coarse shadows for foreground objects. Considering the powerful generation capability of diffusion models, inspired by [13], we exploit a conditional diffusion model as the backbone network. Compared to textual information, images provide rich structural and semantic features to assist in image reconstruction, so we use composite image with coarse shadows as a condition to guide the diffusion model. The harmonization diffusion model can better guide the lighting editing of inserted objects, bridge the lighting gap between inserted objects and background environments. Additionally, the diffusion model can iteratively refine the shadow regions, and achieves more realistic shadow effects closer to real images.

The existing dataset are not well-suited for our task. IHarmony4 [8] provides different color conversions but lacks attention to lighting. RealHM [22] and RdHarmony [3] require a significant amount of manpower and technical resources. CcHarmony [32] focuses on realistic lighting changes, but has a complex filming process. ShadowAR dataset [28] is collected through rendering models. However, the attributes of shadows may also not match those of real images. DESOBA and DESOBAv2 [19, 29] use real images as target images to remove shadows from the foreground to generate composite images. Bao et al. [2] proposed an indoor dataset for foreground harmonization and shadow generation, but only focusing on indoor scenes. In this paper, we construct a new outdoor real-world dataset (IH-SG) for image harmonization and shadow generation tasks.

Our contributions can be summarized as follows:

- We construct a new outdoor real-world dataset (IH-SG) for image harmonization and shadow generation task.
- We propose a new image lighting-shadow editing method based on conditional diffusion model, which can achieve controllable harmonization of foreground regions and reasonable generation of cast shadows.

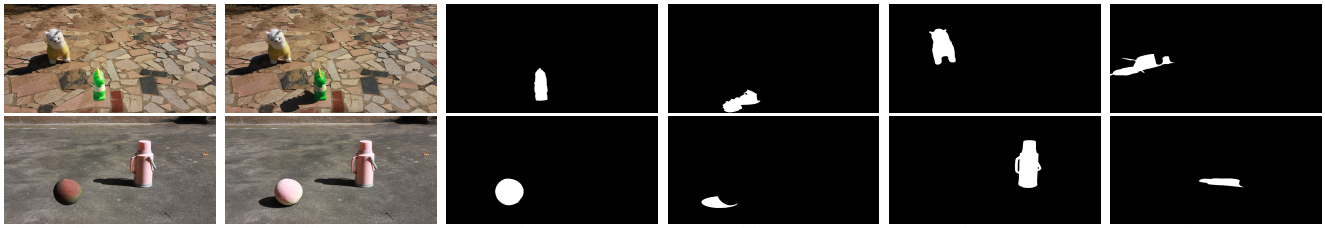
Extensive experiments conducted on public datasets and our IH-SG dataset demonstrate the effectiveness of our method.

2 RELATED WORK

2.1 Image Harmonization

Traditional image harmonization methods primarily focus on adjusting the low-level appearance statistics between foreground objects and the background, such as color statistics [5, 34, 36, 54], and gradient information [21, 33, 48]. The limited representation capability of low-level features can negatively impact their performance. Especially when there are significant differences between the foreground and background regions.

Recent research has built reasonably sized datasets [8, 22, 32] to advance learning-based approaches. CNN-based methods analyze semantic information [9, 50]. Since image harmonization adjusts the foreground lighting or style to match the background, domain adaptation methods [6, 8, 26] have also been proposed to explore the idea of domain harmonization. Guo et al. [16] introduced Retinex theory into the image harmonization task and decomposed the synthetic image into reflectance and illumination. With the rise of Transformers, Guo et al. [14, 15] applied the Transformer framework to image harmonization task. But intrinsic decomposition is a difficult problem. Some methods treat image harmonization as a style transfer problem. These methods have achieved advanced research results through contrastive learning [17], high resolution [23] or color space adjustment [7, 12, 47]. Shen et al. [41] trained Global Perception Adaptive Coordination Kernel. Bao et al. [2] generated harmonious objects and shadows on a synthetic dataset. Bao et al. [1] fouses on indoor scenes. In the concurrent work, Yu et al. [55] uses stable-diffusion model to handle image harmonization and shadow generation tasks. Unlike existing methods, we learn



(a) Input (b) GT (c) M_{fo} (d) M_{fs} (e) M_{bo} (f) M_{bs}

Figure 2: IH-SG dataset. From left to right, they are composite images, real images, foreground object masks, foreground shadow masks, background object masks and background shadow masks, respectively.

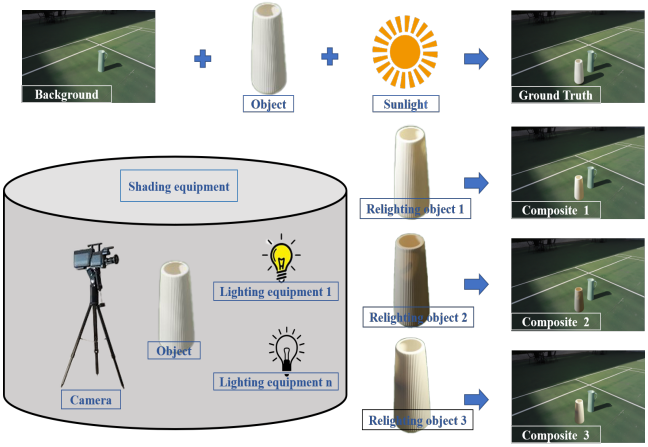


Figure 3: The pipeline of dataset construction. This figure shows the process of obtaining real images and composite images.

.

the illumination of images through diffusion model to generate illumination for the foreground object consistent with the background, as well as corresponding cast shadow.

2.2 Shadow Generation

The existing work on shadow generation can be divided into two categories: rendering-based methods and image-to-image translation methods. Rendering-based methods require explicit knowledge of lighting, reflectance, and scene geometry to generate shadows for inserted virtual objects using rendering techniques. However, such detailed knowledge relies on user input [24, 27] or model prediction [11, 25]. Sheng et al. [43] explored the generation of controllable soft shadows, introduced the concept of pixel height [42, 44] and explored the correlation between objects, ground, and camera poses. In the absence of user interaction, Gardner [11] attempted to recover explicit lighting conditions and scene geometry based on a single image, but inaccurate estimates may lead to unsatisfactory results.

Image-to-image translation methods learn the mapping from input images without foreground shadows to output images with foreground shadows, without requiring explicit knowledge of lighting, reflectance, etc. Hu et al. [20] proposed a method that can adapt to different scenarios, but failed to generate shadows in complex scenes. ShadowGAN [58] utilizes both global and local conditional discriminator to enhance the realism of generated shadows. Liu et al. [28] released the ShadowAR dataset and proposed an attention-guided network for shadow generation. Yan et al. [19] addressed real-world scenes and generated plausible shadows. SGDiffusion [29] focuses on the shadow generation problem based on a diffusion model. DMASNet [49] decomposes shadow mask prediction into box prediction and shape prediction. However, the shadows generated by these methods are still not accurate enough.

2.3 Diffusion Model

Diffusion-based generative models recently produced amazing results with improvements adopted in denoising diffusion probabilistic models [18], which becomes increasingly influential in the field of low-level vision tasks, such as superresolution [40], inpainting [31], and colorization [39]. The methods [35, 37, 38] explored different modal conditions into the diffusion process, achieving controllability of the generated content of the generative model. Pallette [39] was proposed as a general image-to-image framework to solve the image restoration with conditional denoising diffusion probability models. The methods [56, 57] were proposed to generate results towards expectations. However, most of these methods focus on synthetic degradation, such as image coloring, image restoration, and super-resolution. In this paper, we explore the problem of foreground harmony and shadow generation in the real world with limited training pairs. We build our model upon shadowdiffusion [13] to address the above issues.

3 IH-SG DATASET

Lighting not only results in different color brightnesses of foreground objects but also leads to the generation of corresponding cast shadows. In this paper, we simultaneously focus on the harmonization of foreground objects and the generation of corresponding realistic shadows. Therefore, we have constructed a high-quality real outdoor dataset IH-SG, including composite images I_c, real images I_{real}, foreground obeject masks M_{fo}, foreground shadow masks M_{fs}, background object masks M_{bo} and background shadow masks M_{bs}.

3.1 Image Collection

We take photos outdoors that meet our requirements, including background images, real images, and relighting images. The pipeline of data cconstruction is illustrated in Figure 3.

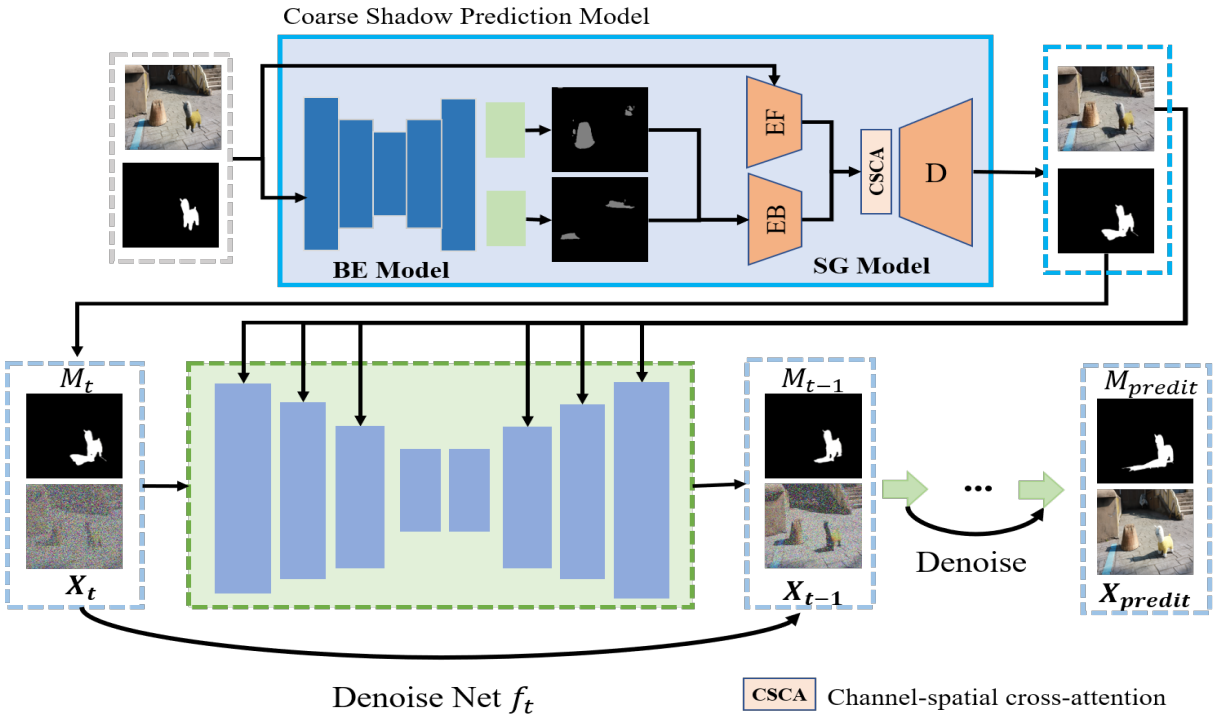


Figure 4: Pipeline of the proposed method. IH-SG Diffusion model includes coarse shadow prediction network (SP) and denoise network f_t. Given a disharmonious image, our model can generate a harmonious image with controllable foreground objects and reasonable cast shadows.

Background images I_{back}: During the shooting process, it is necessary to choose appropriate weather conditions and time period to avoid excessively dim or harsh lighting conditions. Rainy days, sunrise or sunset periods are not suitable for data shooting because the lighting during these periods undergoes significant changes. Moreover, the appropriate shooting angle and position are crucial. To ensure the stability of the camera, we stabilize it on a tripod and control it through a mobile device.

Real images I_{real}: Without altering camera parameters, positions, etc., placing foreground objects and capturing images to be used as real images in the training set. The camera remains stable throughout the process, ensuring consistency between background and real images, thus reducing alignment and correction efforts during subsequent image synthesis. Rapid placement of foreground objects generally assumes minimal changes in lighting between background and real images captured in a short time.

Relighting foreground images $I^i_{relight}$: Without altering the shooting scene or camera position, we employ appropriately sized and shaped shading equipment to shade the scene, ensuring that neither the camera nor the objects are affected. Shading equipment is utilized to effectively block external light interference. Placing lights and adjusting their brightness and direction to illuminate foreground objects. Camera parameters such as exposure time, aperture size, and ISO sensitivity are adjusted based on the actual shooting environment and lighting conditions to achieve the desired exposure effects. Subsequently, adjust the lighting conditions to capture different relighted images $I^i_{relight}$, $i \in N$, where N represents different lighting conditions.

3.2 Image Synthesis

Based on background image I_{back} and relighting images $I^i_{relight}$, composite image I_c can be obtained. To obtain refined data, we used Photoshop to obtain corresponding masks, including foreground object mask M_{fo}, foreground shadow mask M_{fs}, background object mask M_{bo}, and background shadow mask M_{bs}. Then, we obtain composite images:

$$I_c = I_{relight} \times M_{fo} + I_{back} \times \left(1 - M_{fo}\right). \tag{1}$$

Then, I_c and I_{real} form a pair of input composite image and ground-truth target image. Due to shooting conditions, there may be significant differences between the background images and the real images in the background area. If there are differences in color or brightness, some image processing can be use, such as color transfer [53]. Additionally, some unsuitable images could be filtered out. After that, we obtained 15k tuples in the form of $\{I_c, M_{fo}, M_{fs}, M_{bo}, M_{bs}, I_{real}\}$, which will be used for model training.

4 METHOD

4.1 Problem Definition

The input is a tuple (I_c, M_{fo}), where $I_c \in R^{H\times W\times C}$, with H and W representing the height and width of image, and $M_{fo} \in R^{H\times W\times 1}$.

This model aims to generate foreground object that is consistent with the background, and to generate reasonable cast shadow.

4.2 Coarse Shadow Predict Module

As shown in Figure 4, the coarse shadow prediction module aims to predict cast shadows for foreground objects, including a background feature extraction network (BE) and a shadow generation network (SG). The composite image and foreground object mask are the inputs.

4.2.1 ***Background Extraction Module.*** Inspired by [4, 10], we know the key areas in the image are crucial. For the shadow generation task, although complete background information may provide more details, it does not directly yield reasonable shadows for image-to-image transformation networks. This is because it may not adequately focus on objects and their shadow information. Therefore, we propose a BE module to learn relevant information from the background image to generate attention maps for reference objects and their shadows.

The module adopts an encoder-decoder network with an attention mechanism as basic architecture, comprising an encoder and two decoders. The composite image without foreground object shadow and foreground object mask are concatenated along the channel dimension and serve as input to the encoder E. The extracted high-level features are fed into two separate branches of decoders. One decoder D_1 predicts the reference object mask M_{bo}, while the other decoder D_2 predicts the corresponding shadow mask M_{bs}:

$$M_{bo} = D_1(E(I_c)), \tag{2}$$

$$M_{bs} = D_2(E(I_c)). \tag{3}$$

4.2.2 ***Shadow Generation Module.*** Given a composite image I_c without foreground shadow and a foreground object mask M_{fo}, this module aims to generate coarse foreground shadow I_{shadow}. The specific network structure is as follows: Two same encoders, one decoder, and one special channel-spatial cross-attention mechanism (CSCA).

Through the BE module, we can identify key areas in the background image that are beneficial for shadow generation. Inspired by [19], in order to better utilize the information in the background, we adopt foreground encoder E_F and background encoder E_B, respectively. The foreground encoder E_F takes the concatenation of the composite image I_c and the foreground object mask M_{fo} as input, generating a foreground feature map X_f. The background encoder E_B takes the concatenation of I_c and M_{bos} as input to generate a background feature map X_b:

$$X_f = E_F(I_c, M_{fo}), \tag{4}$$

$$X_b = E_b(I_c, (M_{bo} + M_{bs})). \tag{5}$$

Inspired from the existing attention methods [51], we introduce a channel-spatial cross-attention (CSCA) to assist the foreground feature map X_f in obtaining relevant reference information from the background feature map X_b. Then, the decoder D is used to predict coarse shadow images I_{shadow} for foreground objects and refine the corresponding mask M_{shadow}:

$$I_{shadow}, M_{shadow} = D(CSCA(X_f, X_b)). \tag{6}$$

4.2.3 ***Channel-Spatial Cross-Attention Module.*** Obtaining relevant illumination information is crucial for generating accurate foreground shadows. Inspired by previous attention-based methods Equation (13), we used a Channel-Spatial Cross-Attention Module (CSCA), as shown in Figure 5, to help the foreground feature map X_f extract relevant illumination information from the background feature map X_b. By constructing the relative positional relationship between reference information and foreground through this module, it effectively guides the generation of foreground shadows in a reasonable direction.

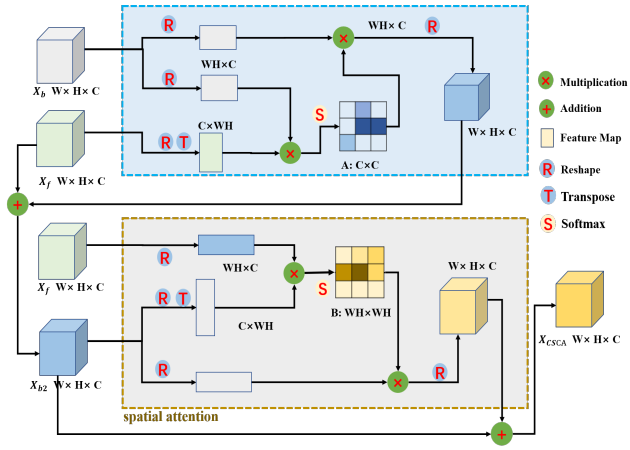


Figure 5: Channel-Spatial Cross-Attention Module. It includes channel cross-attention and spatial cross-attention sub-modules.

Channel cross-attention: To project foreground and background features into a common space, we reshape $X_f \in R^{W\times H\times C}$ to $X_f^r \in R^{WH\times C}$ and $X_b \in R^{W\times H\times C}$ to $X_b^r \in R^{WH\times C}$. Then, we compute the dependencies between any two elements of X_f and X_b in the global context:

$$A = softmax\left((X_f^r)^T X_b^r\right). \tag{7}$$

Using the obtained similarity map A, we incorporate information from X_f^r, then reshape it, and obtain the weighted feature map X_{b2}:

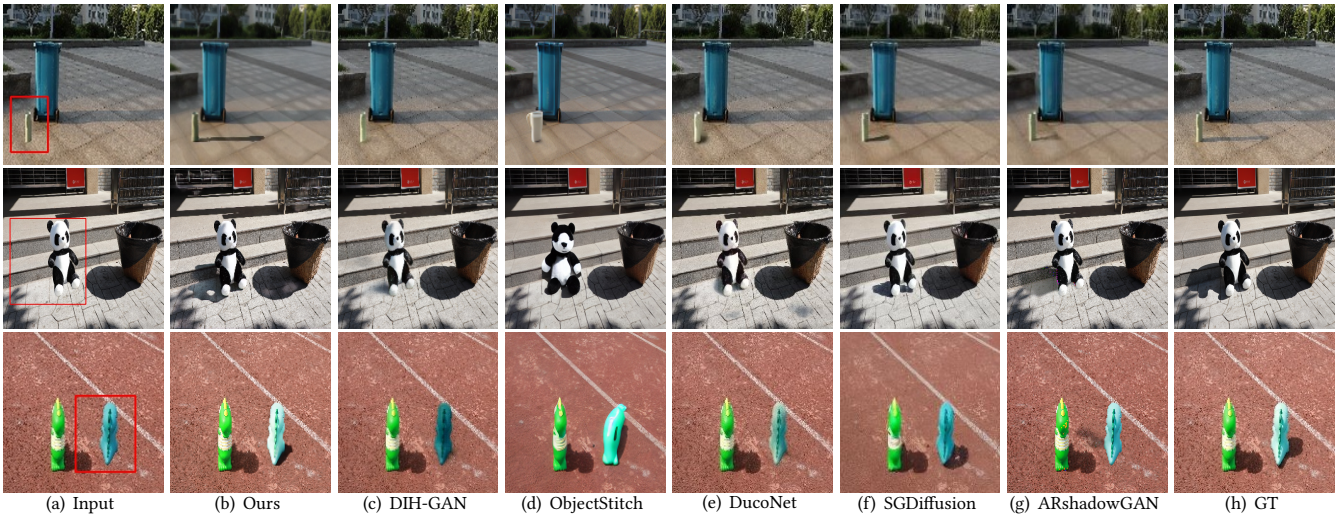
$$X_{b2} = X_f + reshape\left(X_b^r A\right). \tag{8}$$

Spatial cross-attention: Similar to the channel cross-attention, we reshape $X_f \in R^{W\times H\times C}$ to $X_f^r \in R^{WH\times C}$ and $X_{b2} \in R^{W\times H\times C}$ to $X_{b2}^r \in R^{WH\times C}$. Then we compute the similarity between feature maps:

$$B = softmax\left(X_f^r (X_{b2}^r)^T\right). \tag{9}$$

Using the obtained similarity image B and weight X_{b2}^r, we then reshape it to obtain the weighted feature map X_{CSCA}:

$$X_{CSCA} = X_{b2} + reshape\left(BX_{b2}^r\right). \tag{10}$$



(a) Input (b) Ours (c) DIH-GAN (d) ObjectStitch (e) DucoNet (f) SGDiffusion (g) ARshadowGAN (h) GT

Figure 6: Three testing cases of diferent methods on IH-SG dataset. From left to right are composite images, our results, DIH-GAN [2], ObjectStitch [45], DucoNet [47] and the SGDiffusion [29], ARshadowGAN [28] and ground truth, respectively.

(a) Input (b) DucoNet (c) CDTNet (d) Harmonizer (e) Ours (f) GT

Figure 7: Comparisons on iharmony4 dataset. From left to right are composite images, the results of DucoNet [47], CDTNet [7] , our results, and ground truth, respectively.

4.3 Harmony Diffusion Module

Controlling the generation of desired images in a controllable manner poses a challenging task for diffusion models. Especially when the objective is to obtain harmonious foreground images, it is crucial to ensure that the foreground and background share the same lighting distribution while preserving the content and structural information of the foreground objects. With the introduction of CLIP technology, text-guided diffusion models offer some controllable guidance. However, we recognize that images often provide more information than long texts. Therefore, in this module, we use compsite images with coarse shadows as conditions to guide the controllable generation of the diffusion model.

Diffusion model generates an image x_0 by denoising a random image following a Gaussian distribution $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This model mainly uses multiple denoising steps$x_{T-1}, ..., x_0$ to gradually bring the image x_0 closer to the data distribution. Diffusion model is divided into forward diffusion and inverse denoising phases.

Forward process. To construct training data, the forward process involves adding noise perturbations to the training image x_0 to generate noisy data $x_1, ..., x_T$:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + (1 - \bar{\alpha}_t)\epsilon, \tag{11}$$

with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s = \prod_{s=0}^{t}(1 - \beta_s)$.

Reverse process. The reverse process aims to derive the posterior distribution for the less noisy image x_{t-1} given the more noisy image x_t using the denoising network f_θ:

$$p(x_{t-1}|x_t, x_0) \sim \mathcal{N}(x_{t-1}; \mu_t(x_0, x_t), \sigma_t^2 \mathbf{I}). \tag{12}$$

In addition to adjusting the lighting effect of foreground objects, there are also issues with predicting rough shadows in the previous stage, which require further refinement through the network. We have observed the following issues:

- The shape of the shadows generated in the previous stage is unrealistic.
- The lighting of the foreground is inconsistent with the background image.

We use composite images with coarse foreground object shadow (y) as conditional guidance to generate harmonized foreground object with realistic cast shadow. We train the denoising network f_θ to predict x_0 instead of the noise ϵ:

$$x_0, m_{t-1} = f_\theta(x_t, y, m, t). \tag{13}$$

where m is the predicted foreground object-shadow mask. Following [18], our harmony diffusion model objective function is:

$$\mathcal{L}_{pix} = \left\| x_{gt} - x_0 \right\|_2^2, \tag{14}$$

where x_{gt} is the real image. Considering that we need to iteratively optimize the shadow area, we also need to calculate the loss between the foreground object-shadow mask m_f and the generated foregroung object-shadow mask m_t for our method :

$$\mathcal{L}_{mask} = \left\| m_f - m_t \right\|_2^2. \tag{15}$$

Therefore, the total loss can be formulated as:

$$\mathcal{L}_{Total} = \mathcal{L}_{pix} + 0.2 \times \mathcal{L}_{mask}. \tag{16}$$

5 EXPERIMENTS

5.1 Experimental Setups

The proposed method is implemented using PyTorch, and training is performed using two GeForce RTX 3090. The training epoch is set to 1000. We utilize the Adam optimizer with a momentum of $(0.9, 0.999)$. The initial learning rate is set to 0.9. We employ the Kaiming initialization technique to initialize the weights of the proposed model, and use a 0.9999 exponential moving average (EMA) throughout all experiments. The diffusion model adopts DDIM. We adopt a U-Net architecture similar to the denoiser ϵ_θ in [13]. Training is carried out with 200 diffusion steps T and a noise schedule β_t that linearly increases from 0.0001 to 0.02, and inference is performed with 200 steps.

5.2 Dataset and Evaluation Metrics

We evaluated the performance of our method on IH-SG for image harmonization and shadow generation tasks. We resized the images to a size of 256 × 256 pixels. We calculated the Root Mean Square Error (RMSE), the Structural Similarity Index (SSIM) , fMSE, fSSIM for the generated images. And fMSE (resp., fSSIM) means MSE (resp., SSIM) within the foreground regions. In general, smaller values of RMSE and fMSE and larger values of SSIM and fSSIM indicate better quality of the generated images.

5.3 Comparison with Baselines

We compare with following methods: DIH-GAN [2], ObjectStitch [45], DucoNet [47], SGDiffusion [29] and ARshadowGAN [28].

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

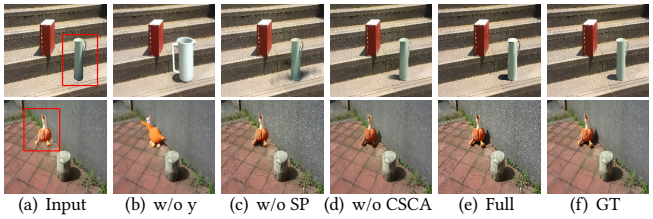
Method	RMSE ↓	SSIM ↑	fMSE ↓	fSSIM ↑
DucoNet [47]	7.249	0.858	452.65	0.917
DIH-GAN [2]	6.108	0.849	579.12	0.886
ObjectStitch [45]	9.487	0.762	1249.48	0.794
ARshadowGAN [28]	9.146	0.812	977.81	0.807
SGDiffusion [29]	8.727	0.833	868.92	0.811
Ours	**5.248**	**0.923**	**374.89**	**0.935**

(a) Input (b) M_{bo} (c) M_{bs} (d) SP result (e) GT

Figure 8: Shadow prediction module. The second and third columns reflect the module's attention to background objects and their shadows.

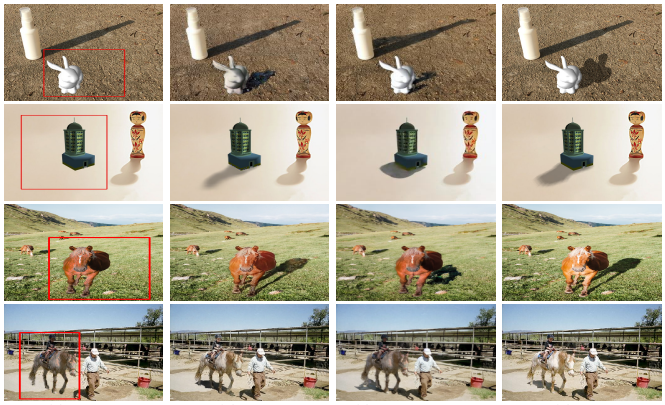
Quantitative comparison. Table 1 reports the comparison results on IH-SG test set. It can be observed that our method achieves the best quantitative results across all four evaluation metrics. This is mainly because the existing image harmonization methods struggle to generalize well to outdoor real-world datasets, while the existing shadow generation methods either rely on simple estimations of foreground shadow masks or directly generate shadows using learned data distributions. Such inaccurate estimations often lead to inferior results. In contrast, our method leverages the coarse shadow prediction module (SP) to effectively utilize background information, and the harmonization diffusion model can better guide the lighting editing of inserted objects, bridge the lighting gap between inserted objects and background environments. Additionally, by iteratively refining shadow regions, our method achieves more realistic shadow effects closer to real images.

Visual comparison. We provide some visual comparison results in Figure 6. It can be observed that our method not only achieves lighting variations across different scenes but also achieves the best visual effects of realistic shadows. Among these competing methods, for ARShadowGAN, it is difficult to edit object lighting, and the generated shadows are not accurate in shape and direction. On the other hand, SGDiffusion can generate relatively accurate shadows but still lacks in shape and shadow color accuracy. As for DucoNet, they fail to generalize well to outdoor real-world datasets. It aims to achieve visual harmony in images, which does not effectively address the problem of object shadows. The semantics of the image generated by the ObjectStitch have changed. In contrast, DIH-GAN, with its multi-scale attention mechanism and lighting feature exchange mechanism, can automatically infer object shadows and lighting generation. However, the shadows generated by this method lack completeness in details. In comparison, our model can harmonize foreground objects and generate realistic and reasonable cast shadows.



(a) Input (b) w/o y (c) w/o SP (d) w/o CSCA (e) Full (f) GT

Figure 9: Ablation study results. The pictures fully demonstrate the effectiveness of image condition (y), coarse shadow prediction module (SP) and channel-spatial cross-attention module (CSCA).



(a) Input (b) Our (c) SGDiffusion (d) GT

Figure 10: The results on the shadow generation dataset. The first row is image from the shadowAR dataset, and the last row is image from the DESOBAv2 dataset.

5.4 Ablation Study

We study the impact of image condition y, shadow prediction module (SP), and channel-spatial cross-attention (CSCA) mechanism of our method on test images from IH-SH. The results are shown in Table 2, Figure 9.

Figure 8 visualizes the coarse shadow prediction module. It can be observed that model can effectively focus on the relevant areas in the background image, such as background objects and their shadows, and predict approximately correct shadows.

To demonstrate the effectiveness of image conditions, we removed the guidance from images, denoted as "w/o y". The performance of "w/o y" is inferior compared to other models, indicating that utilizing image-condition guidance better preserves content structural information.

To investigate the necessity of the coarse shadow prediction module SP, we removed this module, referred to as "w/o SP". It can be observed that without the SP module, there is a slight deficiency in shadow generation, and even the direction may be inaccurate. The performance of "w/o SP" is inferior to that of the full model, demonstrating the advantage of extracting background information and estimating coarse shadow regions.

To demonstrate the effectiveness of the CSCA mechanism, it was removed and replaced with a CAI layer [19], denoted as "w/o CSCA". The results are not as good as the entire model, indicating that CSCA can help generate more realistic images.

Table 2: Ablation study results. "↑" indicates the higher the better, and "↓" indicates the lower the better. The best results are marked in bold.

Method	RMSE ↓	SSIM ↑	fMSE ↓	fSSIM ↑
w/o y	9.372	0.669	1027.164	0.811
w/o SP	8.994	0.726	783.271	0.893
w/o CSCA	6.532	0.873	390.661	0.923
Full	**5.524**	**0.915**	**362.713**	**0.935**

(a) Input (b) Our (c) GT

Figure 11: Failed cases. There are difficulties in generating non-planar cast shadows.

5.5 Discussion

Comparison on iHarmony4 [8] dataset. Figure 7 demonstrates the applicability of our method in image harmonization task. It can be observed that DucoNet [47] and CDTNet [7] do not effectively transfer low-level illumination to the foreground, while our method achieves the best results. Our method can bridge the lighting gap between foreground objects and background environment, achieving lighting effects closer to ground truth (GT) images. It can also preserve the structural information of foreground objects without changing their structure and details.

Comparison on DESOBAv2 [29] and shadowAR [28] dataset. We perturbed the foreground objects in the DESOBAv2. Figure 10 demonstrates that our method can learn illumination information in the background to generate harmonious foreground objects and shadows for foreground objects. However, it is also observed that the generated shadows are somewhat unrealistic in few cases.

Limitations: As depicted in Figure 11, the proposed method has been successfully applied to image harmonization and shadow generation tasks in various environments. However, our method faces challenges in generating non-planar projection shadows. This is because generating non-planar shadows requires more information, such as object geometry and environmental depth information.

6 CONCLUSION

In this work, we have introduced a diffusion model-based method to edit the lighting of foreground objects and generate visually reasonable cast shadows as well as preserving the structure of the image. In addition, we have proposed a large-scale high-quality outdoor real-world dataset IH-SG for image harmonization and shadow generation tasks. Our future work is to solve the generation of non-planar cast shadows of foreground objects.

7 ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (**No. 62372336, No. 62372339 and No. 61972298**) and Wuhan University Huawei GeoInformatices Innovation Lab.

REFERENCES

[1] Zhongyun Bao, Gang Fu, Zipei Chen, and Chunxia Xiao. 2024. Illuminator: Image-based illumination editing for indoor scene harmonization. *Computational Visual Media* (2024), 1–19.

[2] Zhongyun Bao, Chengjiang Long, Gang Fu, Daquan Liu, Yuanzhen Li, Jiaming Wu, and Chunxia Xiao. 2022. Deep Image-based Illumination Harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18542–18551.

[3] Junyan Cao, Wenyan Cong, Li Niu, Jianfu Zhang, and Liqing Zhang. 2021. Deep image harmonization by bridging the reality gap. *arXiv preprint arXiv:2103.17104* (2021).

[4] Zipei Chen, Xiao Lu, Ling Zhang, and Chunxia Xiao. 2022. Semi-supervised video shadow detection via image-assisted pseudo-label generation. In *Proceedings of the 30th acm international conference on multimedia*. 2700–2708.

[5] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. 2006. Color harmonization. In *ACM SIGGRAPH 2006 Papers*. 624–630.

[6] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. 2021. Bargainnet: Background-guided domain translation for image harmonization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[7] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. 2022. High-resolution image harmonization via collaborative dual transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18470–18479.

[8] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. 2020. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8394–8403.

[9] Xiaodong Cun and Chi-Man Pun. 2020. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing* 29 (2020), 4759–4771.

[10] Gang Fu, Qing Zhang, Lei Zhu, Qifeng Lin, Yihao Wang, Siyuan Fan, and Chunxia Xiao. 2024. Towards high-resolution specular highlight detection. *International Journal of Computer Vision* 132, 1 (2024), 95–117.

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

[12] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. 2023. PCT-Net: Full Resolution Image Harmonization Using Pixel-Wise Color Transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5917–5926.

[13] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. 2023. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14049–14058.

[14] Zonghui Guo, Zhaorui Gu, Bing Zheng, Junyu Dong, and Haiyong Zheng. 2022. Transformer for image harmonization and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[15] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. 2021. Image harmonization with transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 14870–14879.

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

[17] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. 2022. Scs-co: Self-consistent style contrastive learning for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19710–19719.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[19] Yan Hong, Li Niu, and Jianfu Zhang. 2022. Shadow generation for composite image in real-world scenes. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 914–922.

[20] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. 2019. Mask-shadowgan: Learning to remove shadows from unpaired data. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2472–2481.

[21] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. 2006. Drag-and-drop pasting. *ACM Transactions on graphics (TOG)* 25, 3 (2006), 631–637.

[22] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. 2021. Ssh: A self-supervised framework for image harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4832–4841.

[23] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. 2022. Harmonizer: Learning to perform white-box image and video harmonization. In *European Conference on Computer Vision*. Springer, 690–706.

[24] Eric Kee, James F O'brien, and Hany Farid. 2014. Exposing Photo Manipulation from Shading and Shadows. *ACM Trans. Graph.* 33, 5 (2014), 165–1.

[25] Bin Liao, Yao Zhu, Chao Liang, Fei Luo, and Chunxia Xiao. 2019. Illumination animating and editing in a single picture using scene structure estimation. *Computers & Graphics* 82 (2019), 53–64.

[26] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. 2021. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9361–9370.

[27] Bin Liu, Kun Xu, and Ralph R Martin. 2017. Static scene illumination estimation from videos with applications. *Journal of Computer Science and Technology* 32 (2017), 430–442.

[28] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. 2020. 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*. 8139–8148.

[29] Qingyang Liu, Junqi You, Jianting Wang, Xinhao Tao, Bo Zhang, and Li Niu. 2024. Shadow Generation for Composite Image Using Diffusion model. *arXiv preprint arXiv:2403.15234* (2024).

[30] Sheng Liu, Cong Phuoc Huynh, Cong Chen, Maxim Arap, and Raffay Hamid. 2023. LEMaRT: Label-efficient masked region transform for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18290–18299.

[31] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11461–11471.

[32] Li Niu, Junyan Cao, Wenyan Cong, and Liqing Zhang. 2023. Deep Image Harmonization with Learnable Augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7482–7491.

[33] Patrick Pérez, Michel Gangnet, and Andrew Blake. 2023. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 577–582.

[34] Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. 2005. N-dimensional probability density function transfer and its application to color transfer. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, Vol. 2. IEEE, 1434–1439.

[35] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

[36] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. 2001. Color transfer between images. *IEEE Computer graphics and applications* 21, 5 (2001), 34–41.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

[39] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–10.

[40] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4713–4726.

[41] Xintian Shen, Jiangning Zhang, Jun Chen, Shipeng Bai, Yue Han, Yabiao Wang, Chengjie Wang, and Yong Liu. 2023. Learning Global-aware Kernel for Image Harmonization. *arXiv preprint arXiv:2305.11676* (2023).

[42] Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A Cengiz Oztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. 2022. Controllable shadow generation using pixel height maps. In *European Conference on Computer Vision*. Springer, 240–256.

[43] Yichen Sheng, Jianming Zhang, and Bedrich Benes. 2021. SSN: Soft shadow network for image compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4380–4390.

[44] Yichen Sheng, Jianming Zhang, Julien Philip, Yannick Hold-Geoffroy, Xin Sun, He Zhang, Lu Ling, and Bedrich Benes. 2023. PixHt-Lab: Pixel Height Based Light Effect Generation for Image Compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16643–16653.

[45] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. 2023. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition. 18310–18319.

[46] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. 2010. Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 1–10.

[47] Linfeng Tan, Jiangtong Li, Li Niu, and Liqing Zhang. 2023. Deep Image Harmonization in Dual Color Spaces. In *Proceedings of the 31st ACM International Conference on Multimedia*. 2159–2167.

[48] Michael W Tao, Micah K Johnson, and Sylvain Paris. 2013. Error-tolerant image compositing. *International journal of computer vision* 103 (2013), 178–189.

[49] Xinhao Tao, Junyan Cao, Yan Hong, and Li Niu. 2024. Shadow generation with decomposed mask prediction and attentive shadow filling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 5198–5206.

[50] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. 2017. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3789–3797.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[52] Ke Wang, Michaël Gharbi, He Zhang, Zhihao Xia, and Eli Shechtman. 2023. Semi-supervised Parametric Real-world Image Harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5927–5936.

[53] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. 2022. CT 2: Colorization transformer via color tokens. In *European Conference on Computer Vision*. Springer, 1–16.

[54] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. 2012. Understanding and improving the realism of image composites. *ACM Transactions on graphics (TOG)* 31, 4 (2012), 1–10.

[55] Ziqi Yu, Jing Zhou, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao. 2024. CFDiffusion: Controllable Foreground Relighting in Image Compositing via Diffusion Model. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

[56] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. 2023. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040* (2023).

[57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

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