# **CFDiffusion: Controllable Foreground Relighting in Image Compositing via Diffusion Model**

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

Gang Fu xyzgfu@gmail.com Department of Computing The Hong Kong Polytechnic University, China Jing Zhou zhoujing@whu.edu.cn School of Computer Science Wuhan University, China

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

Chunxia Xiao\* cxxiao@whu.edu.cn School of Computer Science Wuhan University, China Zhongyun Bao tantouxy@163.com

School of Computer Science Wuhan University, China

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



Figure 1: We unify image harmonization and shadow generation into one model, adjusting foreground object to match the background and generating appropriate shadows. The red rectangles highlight the foreground object.

# ABSTRACT

Inserting foreground objects into specific background scenes and eliminating the illumination inconsistency (eg., color, brightness) between them is an important and challenging task. It typically involves multiple processing tasks, such as image harmonization and shadow generation. In these two domains, there are already many mature solutions, but they often only focus on one of the tasks. Recently, some image composition methods have utilized diffusion models to address both of these issues simultaneously, but they cannot guarantee complete reconstruction of the foreground content. In this work, we propose CFDiffusion, which can simultaneously handle image harmonization and shadow generation. We first employ a shadow mask predictor to estimate the shadow mask

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

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0686-8/24/10 https://doi.org/10.1145/3664647.3681283 of the foreground object. Next, we design a harmonization-shadow generator based on a diffusion model to harmonize the foreground and generate shadows concurrently. Additionally, we propose a foreground content enhancement module to ensure the complete preservation of foreground content at the insertion location, and we also develop an adaptive encoder to guide the harmonization process in the foreground area. The experimental results on the iHarmony4 dataset and the IH-SG dataset demonstrate the superiority of our CFDiffusion approach.

# **CCS CONCEPTS**

• Computing methodologies → Image manipulation.

# **KEYWORDS**

Image composition, Image harmonization, Shadow generation

#### **ACM Reference Format:**

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 (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681283

<sup>\*</sup>Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

Ziqi Yu, Jing Zhou, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao

# **1** INTRODUCTION

Image composition stands as a fundamental task in computer vision and augmented reality, aimed at seamlessly integrating objects from one image into another to produce a convincingly realistic composite image. Merely inserting the foreground object into a background image without careful consideration results in noticeable discrepancies between the foreground and background, such as differences in color, brightness, and shadows. Based on this, we can decompose the image compositing process into dual subtasks, each addressing specific issues: image harmonization [2, 3, 8, 9, 12, 24, 28, 53, 57] and shadow generation [21, 33, 44, 49, 54, 61].

Image harmonization aims to adjust the compatibility between the foreground and background in terms of color and brightness, while shadow generation ensures that the inserted foreground objects cast realistic and reasonable shadows. Various practical methods exist for both of these subtasks. However, employing multiple models to address each subproblem individually is both cumbersome and impractical. What we need is a unified network that can address both of these issues simultaneously, achieving excellent results for each. Figure 1 illustrates the problems we need to deal with and the results we should achieve.

In recent years, generative models such as GANs [4, 23, 25, 48] and diffusion models [1, 20, 35, 39, 41, 44, 45] have demonstrated significant potential in image composition. Particularly, diffusion models have surpassed various preceding methods in image editing [1, 27, 39] and other applications [19, 38, 42]. Conditional diffusion model aims to generate images under the guidance of conditional information, such as text or semantic masks. Among them, Stable Diffusion (SD) [44] stands out as one of the most popular models, successfully integrating text from CLIP [43] into latent diffusion.

Some existing works have introduced the diffusion model into related research domains, particularly in image editing and image composition. For example, SDEdit [39] composites images by adding noise to the input image and then iteratively denoises it through stochastic differential equations. However, these methods lacks proper and sufficient guidance during the denoising process, resulting in the final image lacking sufficient content fedelity. Besides, most diffusion models for image editing focus on manipulating images using text input, which is inappropriate for image composition.

Recently, some image compositing methods [51, 58] have attempted to address these issues within a unified model, which can significantly simplify both model size and complexity. For example, ObjectStitch [51] utilizes a bounding box that encompasses the foreground object to specify the region for foreground object insertion and shadow generation, then processes the foreground within this designated area. Given the recent successful applications of diffusion models in image processing, these methods typically rely on pretrained diffusion models. However, in practice, these methods result in uncontrollable adjustments to the foreground in terms of both position and content texture, raising concerns about preserving the fidelity and credibility of the foreground object.

In this paper, our aim is to address the issue of inadequate foreground fidelity observed in previous discussions on image composition. We introduce a method called CFDiffusion that concurrently handles image harmonization and shadow generation tasks. Building upon stable diffusion, we introduce a foreground content enhancement module (FCEM), which utilizes a foreground content encoder to extract foreground content information, thus guiding the reconstruction of foreground content. Furthermore, we equip SD with a lightweight adaptive encoder designed to extract crucial conditional information from the composite image, such as background style and color, to guide the denoising process of SD.

To validate the effectiveness of our approach, we compare it with state-of-the-art methods and conduct experiments on benchmark datasets such as iHarmony4 [10] and the IH-SG dataset. The experimental results demonstrate that our method achieves more realistic image harmonization and produces shadows that are both genuine and believable.

Our contributions can be summarized as follows:

- We introduce a novel image composition method called CFDiffusion. This method simultaneously handles image harmonization and shadow generation tasks for foreground objects with masked insertion points.
- We design a foreground content enhancement module to fully reconstruct the foreground content and texture details.

Extensive experiments conducted on both public datasets and the IH-SG dataset [62], validate the effectiveness of our proposed method.

# 2 RELATED WORK

### 2.1 Image Harmonization

As a subtask of image compositing, the objective of image harmonization is to integrate objects from a given foreground image into a background image to create a cohesive composite image. This process involves adjusting the color and lighting information of the foreground object to ensure its compatibility with the background of the composite image.

Traditional methods [31, 52] rely on adjusting the appearance of the foreground to match the color statistics of the background, typically focusing on obtaining color statistics and then transferring this information between the foreground and background. These methods are fast and straightforward but often struggle with complex scenes and produce artifacts because the realism of the image is often not well captured by these statistics. Particularly, with the release of the first large-scale image harmonization dataset, iHarmony4 [10], supervised image harmonization methods [7, 11, 13, 14, 16, 32] have garnered increasing attention. For instance, Hao et al. [14] employed attention blocks to compute non-local information for foreground adjustment. SSAM [11] integrates them using a dual-path attention model, focusing on the relationship between spliced and unspliced regions. DoveNet [10] treats image harmonization as a domain translation task. CDT-Net [9] combines pixel-to-pixel and RGB-to-RGB transformations for high-resolution image harmonization. Ling et al. [32] introduced the concept of style from the background image, treating the harmonization task as a style transfer problem. They first introduced the concept of background image style and proposed a new regionaware adaptive instance normalization (RAIN) method, which can extract style information such as hue hidden in background features and adaptively transfer it to the foreground area. Hang et al. [17] proposed a background attention adaptive instance normalization (BAIN) based on RAIN to achieve attention-weighted

CFDiffusion: Controllable Foreground Relighting in Image Compositing via Diffusion Model

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



Figure 2: Overview of our CFDiffusion. It consists of two stages: foreground object shadow mask prediction stage and shadowharmonization generation stage.  $E_f$  and  $E_b$  respectively represent the foreground encoder and background encoder of the foreground object shadow mask generator.

background feature distribution according to the feature similarity of foreground and background. However, when the task extends to shadow generation, these methods do not scale well to handle both tasks simultaneously.

#### 2.2 Shadow Generation

Previous work on shadow generation can be categorized into two main approaches: rendering-based methods and image-to-image translation methods.

Rendering-based methods [26, 29] relies on a clear understanding of lighting, reflectance, material properties, and scene geometry to generate shadows for inserted virtual objects using rendering techniques. However, such knowledge is often unavailable or impractical for applications in real-world scenarios. Image-to-image translation methods predominantly employ deep learning techniques, characterized by encoder-decoder architectures. By training on paired images, including images with shadows and those without, these methods directly learn the mapping from shadow-free images to shadowed images from the input data. Importantly, this approach typically eliminates the need for explicit knowledge about lighting, reflectance, material properties, and scene geometry. AR-ShadowGAN [33] introduces an attention-guided network capable of directly modeling the mapping relationship between the shadows of foreground objects and their corresponding real environments, accompanied by the release of the Shadow-AR dataset. SGRNet [21] promotes comprehensive information interaction between the foreground and background. It initially predicts a mask for shadow regions and subsequently forecasts shadow parameters to fill these regions. Additionally, a new shadow training dataset, DESOBA, was introduced. ShadowGAN [61] combines both global conditional discriminators and local conditional discriminators to generate shadows for inserted 3D foreground objects without relying on background lighting information. In addition, some shadow removal methods and shadow detection methods [6] can also give us reverse

inspiration on shadow generation, such as Mask-ShadowGAN [22], and CANet [5]. Unlike the approach in this paper, which involves filling shadow regions, Zhou *et al.* [62] proposed a coarse shadow prediction module to generate coarse shadows and use them as prior knowledge to guide the generation of the diffusion model. Shadow generation is associated with the foreground objects, but it targets different aspects than image harmonization. We aim to combine both aspects using a single network framework.

#### **3 PROPOSED METHOD**

Given a composite image  $I_c$ , a binary mask  $M_{bs}$  representing the background object-shadow pair, and a binary mask  $M_f$  indicating the foreground object, our goal is to obtain an image  $\tilde{I}$  that harmonizes the foreground object and produces reliable shadows under background illumination conditions.

As illustrated in Figure 2, Our method consists of two stages: the foreground object shadow mask prediction stage, foreground harmonization and shadow generation stage. It mainly comprises five components: foreground object shadow generator  $G_{fs}$ , foreground content encoder  $E(\cdot)$ , adaptive encoder  $E_a$ , foreground content enhancement module *FCEM*, and shadow-harmonization generator  $G(\cdot)$  based on a stable diffusion model.

The network workflow is as follows: Firstly, we use the background object-shadow data pair as reference to predict the shadow mask of the foreground object, identifying the approximate location for shadow generation. Then, we input the synthesized image into the shadow-harmonization generator to produce the final result. Simultaneously, we use the foreground content encoder to extract the foreground content embedding, inputting it into the foreground content enhancement module to constrain and complete foreground texture details. The adaptive encoder transfers the background style to the foreground region, providing additional generation guidance for the harmonization-shadow generator.



Figure 3: The display of two six-tuples of the IH-SG dataset [62], from left to right: ground truth, composite image, foreground object mask, shadow mask of foreground object, background object mask, shadow mask of the background object. The foreground and background are captured separately under different lighting environments, paying attention to collecting various scenes, ground surfaces, and shadow casting situations.



Figure 4: Overview of Cross-Attention Integration (Cross-Attention Integration) layer [21]. The g, f, h, v shown in the figure represent  $1 \times 1$  convolution,  $\otimes$  represents matrix multiplication,  $\bigcirc$  represents concatenate.

#### 3.1 Harmonization-Shadow Generator

Recently, diffusion models have shown remarkable performance in many fields: image generation [20, 50], text-to-image generation [44], image translation [30], image inpainting [37, 46], and image editing [15, 39]. The backbone of our Harmony-Shadow Generator is built upon Stable Diffusion (SD) model [44].

SD is a latent diffusion model that undergoes a two-stage pretraining process, involving an autoencoder and a denoising U-Net [20]. In the first stage, the SD model trains an autoencoder: the encoder  $\mathcal{E}$  converts the images *I* into a latent representation  $z'_0 = \mathcal{E}(I)$ , and then the decoder  $\mathcal{D}$  reconstructs the images, resulting in  $\hat{I} = \mathcal{D}(z'_0)$ . In the second stage, the autoencoder's parameters are fixed, and SD introduces noise to the latent space representation  $z'_0$ over *T* steps to generate  $z'_t$ . This process involves the creation of a denoising U-Net  $\epsilon_{\theta}$ , which is trained using a latent denoising loss

$$\mathcal{L}_{LDM}: = \mathbb{E}_{z'_0, y, \epsilon} \mathcal{N}_{(0,1), t}[\| \epsilon - \epsilon_{\theta_1}(z'_t, t, \tau_{\theta_2}(y)) \|_2^2], \quad (1)$$

here  $\epsilon$  is the noise added to the latent space feature  $z'_0$  at each noise step,  $\epsilon_{\theta_1}$  is the denoising U-Net that predicts the noise  $\epsilon$  at

the current step t, and y represents additional conditions (e.g. text, mask, etc.),  $\tau_{\theta_2}$  is instead a domain-specific encoder that projects y to an intermediate representation. In this work, we add conditional information using an adaptive encoder similar to that of composite images with foreground masks. During the inference process, noise is first added to  $z'_0$  to generate  $z'_T$ , and then  $z'_T$  is used as  $z_T$ , which is the initial input of  $\epsilon_{\theta_1}$ . Then we iteratively use  $\epsilon_{\theta_1}$  to estimate the noise at each denoising step t, thereby gradually refining the latent map  $z_T$ , and ultimately become a clean latent feature  $z_0$ . Finally, the clean latent features  $z_0$  are fed to the decoder  $\mathcal{D}$  to generate images.

#### 3.2 Foreground shadow mask generator

Inspired by Hong *et al.* [21], we apply a shadow mask predictor to generate foreground object shadow mask. First, we predict the foreground shadow mask  $\widetilde{M}_{fs}$  through the foreground shadow mask generator  $G_s$ .  $G_s$  consists of an encoder E and a decoder D, and the encoder is divided into a foreground encoder  $E_f$  and a background encoder  $E_b$ . We believe that the background objectshadow pair contains clues that are beneficial for inferring the foreground shadow area. In order to generate the shadow mask of the foreground object, we take the concatenation of the composite image  $I_c$  and the background object-shadow mask  $M_{bs}$  as the input of the background encoder  $E_b$ , and generate the background feature map  $X_b$ . At the same time, the concatenation of the composite image  $I_c$  and the foreground object mask  $M_f$  is used as the input of the foreground encoder  $E_f$  to obtain the foreground feature map  $X_f$ . The process is summarized as follows:

$$X_b = E_b \left( M_{bs}, \ I_c \right), \tag{2}$$

$$X_f = E_f \left( M_f, \ I_c \right). \tag{3}$$

Following Hong *et al.* [21], we use a Cross-Attention Integration (CAI) [21, 55, 56, 60] layer to help the foreground feature map notice the relevant lighting information of the background feature map. As the Figure 4 shows, the input of the CAI layer consists of  $X_f$  and  $X_b$ , which are outputs of foreground encoder  $E_f$  and a background encoder  $E_b$ , and the output feature map is denoted as X. Then X is fed into the decoder D to obtain the mask of the foreground object shadow. Subsequently, we add it to the foreground object mask  $M_f$  to obtain the foreground object-shadow mask  $\tilde{M}_{fs}$ , which serves



A : Attention map  $\bigotimes \bigoplus$  : Element-wise multipliaction/addition

Figure 5: Overview of FCEM module. A in the figure is the attention map output by the cross attention part.  $F_i \in \mathbb{R}^{h_i \times w_i \times c_i}$ is feature map of i-th transformer block

as one of the inputs for the subsequent shadow-harmonization generator. The process is summarized as follows:

$$M_{fs} = D(X) + M_f. \tag{4}$$

### 3.3 Foreground Encoder

Following Zhang *et al.* [59], in order to further enhance the detailed texture of the foreground generation, we employ the pre-trained model ViT - L/14 from CLIP [43] as the foreground image encoder *E*. Initially, we extract the foreground object region from the synthesized image  $I_c$  using the foreground object mask  $M_f$ , which is then inputted into the foreground image encoder *E* to extract the local content embedding of the foreground  $E_l$ . This process can be represented as follows:

$$E_l = E(I_f). (5)$$

The intermediate layer of the CLIP encoder outputs 256 patch tokens containing local details. We extract the information of these patch tokens and integrate these foreground content embeddings into the Foreground Content Enhancement Module of the denoising U-Net model to help us control the generation of foreground content details. The specific details of the FCEM module is located in Section 3.4.

# 3.4 Foreground Content Enhancement Module

Following Zhang *et al.* [59], we utilize a foreground content enhancement module to embed foreground content into the intermediate features of the diffusion model, thereby constraining the stable diffusion model for foreground appearance generation and promoting the composite generation of foreground appearance with high fidelity.

Our foreground content enhancement module is built upon the publicly released v1-4 SD model. To identify foreground regions that need to be constrained, we append the binary foreground object-shadow mask  $\widetilde{M}_{fs}$  to the model input. To achieve this, in the first convolutional layer of U-Net, we attach two additional input channels to respectively contain the foreground object mask  $M_{fs}$  and the predicted foreground object-shadow mask  $\widetilde{M}_{fs}$ . Eventually, the input images are uniformly resized to the resolution of 256×256.

Denoising U-Net of SD consists of a series of basic blocks, each block includes a residual block and a transformer block. The transformer block consists of a self-attention module, a cross-attention module, and a feedforward network.

As illustrated in Figure 5. We record the features from the i - th transformer block as  $F_i \in \mathbb{R}^{h_l \times w_i \times c_i}$ , where  $h_i, w_i, c_i$  represent its height, width, and channel dimensions respectively. We first use the foreground local feature  $F_i^l$  intercepted by the foreground object mask  $M_f$  resized to  $h_i, w_i, c_i$ . The feature map is flattened to  $\overline{F}_i^l \in \mathbb{R}^{N \times c_i}$ , and passed through cross attention together with the foreground local embedding  $E_l$ , then we get an attention map A and refined foreground local feature map  $\widetilde{F}_i^l$ . Then we align the foreground local embedding  $E_l$  with  $\widetilde{F}_i^l$  to obtain the aligned foreground embedding map  $\widetilde{E}_l$ . Further we use  $\widetilde{E}_l$  to modulate  $\widetilde{F}_i^l, \widetilde{E}_l$  is passed through a convolutional layer of  $3 \times 3$  to obtain the spatial awareness modulation weight, and the modulation is normalized. The transformed  $\widetilde{F}_i^l$  is as follows:

$$\hat{F}_{i}^{l} = norm(\tilde{F}_{i}^{l}) \bullet conv(\tilde{E}_{l}).$$
(6)

Finally, after resizing  $\hat{F}_i^l$ , it is added to the foreground object region of  $F_i$ . The output of the Foreground Content Enhancement Module (FCEM) is then delivered as the enhanced foreground content features  $\tilde{F}_i$  to the next residual block.

## 3.5 Adaptive Encoder

Following Lu *et al.* [36] and Mou *et al.* [40], we adopt an adaptive encoder, which is a lightweight model that can align the internal knowledge in the SD model with external control signals. Through this adaptive encoder, we can achieve rich control effects on the color and structure of the SD generation results.

The adaptive encoder takes into account encoding additional conditions and provides multi-step guidance for denoising U-Net in the denoising step. Previous adaptive encoder implementations focus more on coarse structures (e.g., sketches, poses, semantic masks) and exploit textual conditions to indicate additional requirements (e.g., style or context). Different from previous work, we abandon the text CLIP model, splice the composite image  $I_c$  and the foreground mask  $M_f$ , and use a lightweight adaptive encoder to encode while retaining content details and extracting background styles. The structure of the adaptive encoder includes four feature extraction blocks and three DownSample (DS) blocks.

First, the input image is resized to  $64 \times 64$ , and we name it  $F_c^0$ . Each feature extraction module (EM) includes a convolutional layer and two residual blocks. The generation process of  $F_c^i$ ,  $i \in 1, 2, 3, 4$ can be expressed as follows:

$$F_c^1 = EM_1\left(F_c^0\right),\tag{7}$$

$$F_c^i = EM_i \left( DS(F_c^{i-1}) \right). \tag{8}$$



Figure 6: Experimental results on the IH-SG dataset. (a) is the input composite image, (h) is the GT. From (b) to (g), they are the generated results of SGDiffusion [34], ARShadowGAN [33], ObjectStitch [51], DucoNet [53], DIH-GAN [3], and our proposed method, respectively.

The resolutions of  $F_c^1, F_c^2, F_c^3$ , and  $F_c^4$  are 64×64, 32×32, 16×16, 8×8 respectively. Then we use foreground mask  $M_f$  to separate the foreground features  $F_{c,f}^i$  and background features  $F_{c,b}^i$ :

$$F_{c,f}^{i} = Flatten(F_{c}^{i} \circ M_{f}), \tag{9}$$

$$F_{c,b}^{i} = Flatten(F_{c}^{i}(1 - M_{f})).$$

$$(10)$$

Among them,  $M_f$  represents the foreground object mask scaled to the corresponding size of  $F_c^i$ ,  $\circ$  represents element-wise product, and  $Flatten(\cdot)$  represents expanding a 2D feature map into a 1D feature feature sequence.

We use a transformer layer to extract and transfer the style of the background to the foreground area to achieve harmonious processing of the foreground. In addition, the parts of the background area that are related to the foreground can provide more references when harmonizing the foreground, so they are very important. We will also pay more attention to the areas that are more related to the background and foreground.  $F_{c,f}^i$  is used as query,  $F_{c,b}^i$  is used as keys/values, and the final background stylized foreground feature  $\hat{F}_{c,f}^i$  can be expressed as:

$$\hat{F}_{c,f}^{i} = Transformer(F_{c,f}^{i}, F_{c,b}^{i}, F_{c,b}^{i}).$$

$$(11)$$

# 3.6 Traning Losses and Details

Our total loss function  $L_{total}$  consists of the standard noise loss  $\mathcal{L}_{LDM}$  of the diffusion model and a reconstruction loss  $L_{rec}$ . Therefore, the final loss function of our CFDiffusion is:

$$L_{total} = \lambda_1 L_{rec} + \lambda_2 \mathcal{L}_{LDM} + \lambda_3 L_{fs}, \tag{12}$$

where  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  are hyper-parameters which control the influence of each term.

**Noise Loss.** First, we adopt the standard noise loss of the diffusion model, aiming to reconstruct the image features in the latent

space:

$$\mathcal{L}_{LDM}: = \mathbb{E}_{z'_0, y, \epsilon} \underset{\mathcal{N}(0,1), t}{\underset{\mathcal{N}(0,1), t}{[\parallel \epsilon - \epsilon_{\theta_1}(z'_t, t, \tau_{\theta_2}(y)) \parallel_2^2]}.$$
(13)

**Reconstruction loss.** It is a classical  $L_1$  loss between the generator output image  $\hat{I}$  and real ground-truth image I, to further constrain the generated image towards the ground truth, which is expressed as:

$$L_{rec} = \| \widetilde{I} - I \|_1 . \tag{14}$$

**MSE Loss.** Additionally, we compute the loss for the foreground mask prediction module using the following method:

$$L_{fs} = \| M_{fs} - M_{fs} \|_2^2 .$$
 (15)

#### **4 EXPERIMENTS**

To verify the superiority of our proposed CFDiffusion, we compare CFDiffusion with state-of-the-arts on the real-world iHarmony4 [10] dataset and IH-SG dataset [62], and provide assessments both quantitatively and qualitatively.

# 4.1 Experimental Settings

The proposed method is implemented using PyTorch, which is trained using one NVIDIA RTX 3090 GPU. All images are resized to  $256 \times 256$  for training and testing. We adopt adam optimizer with the momentum as (0.9,0.999), and the learning rate initialized as 0.00003. Following [47], we use the Kaiming initialization technique [18] to initialize the weights of the proposed model and use a 0.9999 Exponetial Moving Average (EMA) for all our experiments. We use 1000 diffusion steps T and noise schedule  $\beta_t$  linearly increasing from 0.0001 to 0.002 for training, and 25 steps for inference. After a few trials, we set  $\lambda_2 = \lambda_3 = 10$ ,  $\lambda_1 = 1$  by observing the grenerated images. The training epoch is set as 1500.

CFDiffusion: Controllable Foreground Relighting in Image Compositing via Diffusion Model

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



Figure 7: The visual results of harmonization experiments on iHarmony4 [10]. (a) represents the input image, (f) is the ground truth (GT), and (b) to (e) show the generated results from CDT-Net [9], Harmonizer [28], DucoNet [53], and our method, respectively. It can be seen that our results are closest to the Ground Truth.

Table 1: Results of quantitative comparison on IH-SG dataset. " $\uparrow$ " indicates the higher the better, and " $\downarrow$ " indicates the lower the better. The best results are marked in bold.

| Method            | RMSE↓ | SSIM ↑ | fMSE ↓   | fSSIM ↑ |
|-------------------|-------|--------|----------|---------|
| SGDiffusion [34]  | 8.591 | 0.825  | 875.882  | 0.809   |
| ARShadowGAN [33]  | 9.164 | 0.817  | 942.154  | 0.816   |
| ObjectStitch [51] | 9.357 | 0.773  | 1145.116 | 0.773   |
| DucoNet [53]      | 7.346 | 0.861  | 454.213  | 0.915   |
| DIH-GAN [3]       | 6.145 | 0.847  | 567.311  | 0.894   |
| Ours              | 5.582 | 0.917  | 367.919  | 0.937   |

Table 2: Results of quantitative comparison on iHarmony4.

| Method          | RMSE ↓ | SSIM ↑ | fMSE↓   | fSSIM ↑ |
|-----------------|--------|--------|---------|---------|
| CDT-Net [9]     | 6.847  | 0.804  | 379.187 | 0.858   |
| Harmonizer [28] | 6.308  | 0.854  | 410.847 | 0.821   |
| DucoNet [53]    | 6.152  | 0.876  | 365.236 | 0.915   |
| Ours            | 5.582  | 0.917  | 367.919 | 0.937   |

Table 3: Ablation study of FCEM and adaptive encoder.

| Method               | RMSE ↓ | SSIM ↑ | fMSE ↓  | fSSIM ↑ |
|----------------------|--------|--------|---------|---------|
| baseline             | 11.192 | 0.687  | 543.417 | 0.675   |
| w/o FCEM             | 7.264  | 0.746  | 398.542 | 0.785   |
| w/o adaptive encoder | 8.437  | 0.727  | 485.534 | 0.841   |
| Ours                 | 5.582  | 0.917  | 367.919 | 0.937   |

**Compared methods.** We compared our model with five deep learning-based methods from the related fields: two image harmonization methods including DucoNet [53], DIH-GAN [3], three shadow generation methods include ARShadowGAN [33], SGDiffusion [34], and ObjectStitch [51]. Among them, SGDiffusion is the latest method using a diffusion model for shadow generation tasks, while ObjectStitch is the latest image composition method that addresses both image harmonization and shadow generation tasks. DucoNet works on image harmonization based on dual color spaces. ARShadowGAN makes full use of the background information to guide the shadow generation of foreground objects. We train and test all these methods based on the IH-SG dataset [62].

In addition, we also tested our image harmonization capabilities on the iHarmony4 dataset, and compared the three image harmonization methods of CDT-Net [9], Harmonizer [28] and DucoNet [53] to further prove the superiority of our CFDiffusion. CDT-Net coherently combines pixel-to-pixel conversion and RGB-to-RGB conversion in an end-to-end network.

**Evaluation metrics.** We use four metrics to evaluate the image illumination harmonization results, which are Relative Mean Square Error (RMSE), Structural Similarity Index Measure (SSIM), foreground Mean Square Error (fMSE), foreground Structural Similarity Index Measure (fSSIM). Generally, the smaller RMSE and fMSE, and the larger fSSIM and SSIM indicate the better image illumination harmonization results.

**IH-SG Dataset.** To better train our model, we used the IH-SG dataset [62] that can address both of image harmonization and shadow generation concurrently. In total, the dataset comprises over 1000 outdoor scenes and more than 10000 six-tuples. Zhou *et al.* [62] also captured numerous shadow scenes under complex conditions, such as shadows casting on walls and steps, to enrich the shadow data samples, making the dataset more realistic and diverse. Two six-tuples are illustrated in the Figure 3.

#### 4.2 Comparison with State-of-the-Arts

**Experiments on IH-SG dataset.** The visualization results of different methods on the testing set are shown in Figure 6, and the quantitative comparison results are summarized in Table 1. Apparently, our method CFDiffusion achieves the better quantitative

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

Ziqi Yu, Jing Zhou, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao

results than other state-of-the-art methods on all four metrics. This is primarily due to the powerful image generation capability of the diffusion model and the effective utilization of foreground and background information by our FCEM and adaptive encoder.

**Experiments on iHarmony4.** The iHarmony4 dataset is one of the most popular large-scale datasets in the field of image harmonization. Therefore, we compared our method with several image harmonization methods on the iHarmony4 dataset, and the quantitative comparison results are shown in Table 2, visualization results are shown in Figure 7. Our proposed approach also achieves the best results.

**Experiments on DESOBAv2.** Recently, Liu *et al.* [34] extended the DESOBA dataset [21] to DESOBAv2, which has become the latest shadow generation dataset. To validate the generalization ability of our model, we conducted experiments on the DESOBAv2 dataset by applying slight perturbations to foreground objects, and the results are shown in Figure 8.

# 4.3 Ablation Study

In order to verify the effectiveness of each component of our method, we conduct ablation studies by modifying the CFDiffusion architecture. Specifically, we set the following variants:

To verify the crucial roles played by our FCEM module and adaptive encoder module in the overall model, we set up several variants. Firstly, we chose the original diffusion model as the baseline, which is referred to as "*baseline*" in Table 3. To demonstrate the pivotal roles of the FCEM module and adaptive encoder module in the entire model, we remove these two modules separately from the complete model, which are referred to as "*w/o FCEM*" and "*w/o adaptive encoder*" in Table 3. Finally, we compare these variants with the full model "*Ours (full model*)" for comprehensive analysis, and some results are shown in Figure 9.

We train these variants using the same training data and quantitatively evaluated their impact on the test results. The evaluation results are presented in Table 3. From the table, we can observe that: after introducing guided supervision with FCEM, the model's quantitative performance has made significant strides, sufficiently demonstrating the strong guiding role of the FCEM module in capturing texture details of foreground objects. Moreover, with the inclusion of the adaptive encoder, the model's performance has also noticeably improved compared to the original diffusion model, confirming its guiding role in the harmonization generation of foreground objects.

With the simultaneous introduction of both the FCEM and adaptive encoder modules, our full model achieves the best performance, demonstrating the effectiveness of our approach. Additionally, incorporating the FCEM and adaptive encoder modules into the original diffusion model significantly improves performance.

#### 4.4 Limitations

As shown in Figure 10, our model fails to effectively generate shadows in scenarios lacking background object references. During the foreground shadow mask prediction stage, we only use the foreground encoder module without referring to background shadow information. As a result, the inferred shadow shapes significantly deviate from the ground truth.



Figure 8: The experimental results on DESOBAv2 dataset.



Figure 9: Ablation study results. From (a) to (f), they are respectively the composite images, baseline method, w/o FCEM, w/o adaptive encoder, our complete method, and GT.



Figure 10: Failed example. From (a) to (c) are the input, generated results, and GT, respectively.

# **5 CONCLUSION AND FUTURE WORK**

We have presented an image compositing method using diffusion models to resolve illumination inconsistencies between foreground and background while generating realistic shadows. Our approach employs an adaptive encoder for background style extraction and the FCEM module to better preserve foreground details. Extensive comparison and ablation experiments validate the effectiveness of CFDiffusion. Future work will focus on extending our CFDiffusion to real-time video lighting harmonization and shadow generation. CFDiffusion: Controllable Foreground Relighting in Image Compositing via Diffusion Model

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

## ACKNOWLEDGEMENT

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

#### REFERENCES

- Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended Diffusion for Text-driven Editing of Natural Images. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022. 01767
- [2] 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.
- [3] Zhongyun Bao, Chengjiang Long, Gang Fu, Daquan Liu, Yuanzhen Li, Jiaming Wu, and Chunxia Xiao. [n. d.]. Deep Image-based Illumination Harmonization. ([n. d.]).
- [4] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. International Conference on Learning Representations, International Conference on Learning Representations (Sep 2018).
- [5] Zipei Chen, Chengjiang Long, Ling Zhang, and Chunxia Xiao. 2021. Canet: A context-aware network for shadow removal. In Proceedings of the IEEE/CVF international conference on computer vision. 4743–4752.
- [6] 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.
- [7] Wenyan Cong, Junyan Cao, Li Niu, Chenglong Zhang, Xuesong Gao, Zhiwei Tang, and Liqing Zhang. 2021. Deep Image Harmonization by Bridging the Reality Gap. Cornell University - arXiv, Cornell University - arXiv (Mar 2021).
- [8] 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.
- [9] 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 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01792
- [10] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. 2020. DoveNet: Deep Image Harmonization via Domain Verification. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr42600.2020.00842
- [11] Xiaodong Cun and Chi-Man Pun. 2020. Improving the Harmony of the Composite Image by Spatial-Separated Attention Module. *IEEE Transactions on Image Processing* (Jan 2020), 4759–4771. https://doi.org/10.1109/tip.2020.2975979
- [12] JulianJorgeAndrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. [n.d.]. PCT-Net: Full Resolution Image Harmonization Using Pixel-Wise Color Transformations. ([n.d.]).
- [13] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. 2021. Image Harmonization With Transformer. International Conference on Computer Vision, International Conference on Computer Vision (Jan 2021).
- [14] Hao Guoqing, Satoshi Iizuka, and Kiichi Fukui. 2020. Image Harmonization with Attention-based Deep Feature Modulation. British Machine Vision Conference, British Machine Vision Conference (Jan 2020).
- [15] Roy Hachnochi, Mingrui Zhao, Nadav Orzech, Rinon Gal, Ali Mahdavi-Amiri, Daniel Cohen-Or, and AmitHaim Bermano. 2023. Cross-domain Compositing with Pretrained Diffusion Models. (Feb 2023).
- [16] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. 2022. SCS-Co: Self-Consistent Style Contrastive Learning for Image Harmonization. (Apr 2022).
- [17] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. 2022. Scs-co: Selfconsistent style contrastive learning for image harmonization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 19710– 19719.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In 2015 IEEE International Conference on Computer Vision (ICCV). https://doi.org/ 10.1109/iccv.2015.123
- [19] Liu He, Yijuan Lu, John Corring, Dinei Florencio, and Cha Zhang. 2023. Diffusionbased Document Layout Generation. (Mar 2023).
- [20] Jonathan Ho, Ajay Jain, Pieter Abbeel, and UC Berkeley. [n.d.]. Denoising Diffusion Probabilistic Models. ([n.d.]).
- [21] Yan Hong, Li Niu, and Jianfu Zhang. 2022. Shadow Generation for Composite Image in Real-World Scenes. Proceedings of the AAAI Conference on Artificial Intelligence (Jul 2022), 914–922. https://doi.org/10.1609/aaai.v36i1.19974
- [22] 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 International Conference on Computer Vision.

- [23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/ cvpr.2017.632
- [24] 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.
- [25] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr. 2019.00453
- [26] Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. 2014. Automatic Scene Inference for 3D Object Compositing. ACM Transactions on Graphics (May 2014), 1–15. https://doi.org/10.1145/2602146
- [27] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2022. Imagic: Text-Based Real Image Editing with Diffusion Models. (Oct 2022).
- [28] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and RynsonW.H. Lau. 2022. Harmonizer: Learning to Perform White-Box Image and Video Harmonization. (Jul 2022).
- [29] Eric Kee, James F. O'brien, and Hany Farid. 2014. Exposing Photo Manipulation from Shading and Shadows. ACM Transactions on Graphics (Sep 2014), 1–21. https://doi.org/10.1145/2629646
- [30] Gihyun Kwon and JongChul Ye. 2022. Diffusion-based Image Translation using Disentangled Style and Content Representation. (Sep 2022).
- [31] Jean-Francois Lalonde and Alexei A. Efros. 2007. Using Color Compatibility for Assessing Image Realism. In 2007 IEEE 11th International Conference on Computer Vision. https://doi.org/10.1109/iccv.2007.4409107
- [32] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. 2021. Region-aware Adaptive Instance Normalization for Image Harmonization. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/ cvpr46437.2021.00924
- [33] 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/ cvpr42600.2020.00816
- [34] 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).
- [35] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. 2023. More Control for Free! Image Synthesis with Semantic Diffusion Guidance. In 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). https: //doi.org/10.1109/wacv56688.2023.00037
- [36] Lingxiao Lu, Jiangtong Li, Junyan Cao, Li Niu, and Liqing2023 Zhang. [n.d.]. Painterly Image Harmonization using Diffusion Model. ([n.d.]).
- [37] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01117
   [38] Shitong Luo and Wei Hu. 2021. Diffusion Probabilistic Models for 3D Point
- [38] Shitong Luo and Wei Hu. 2021. Diffusion Probabilistic Models for 3D Point Cloud Generation. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr46437.2021.00286
- [39] Chenlin Meng, Ya-Ling He, Song Yang, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. *Cornell University - arXiv,Cornell University - arXiv* (Aug 2021).
- [40] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. (Feb 2023).
- [41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. [n. d.]. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. ([n. d.]).
- [42] Ben Poole, Ajay Jain, JonathanT Barron, Ben Mildenhall, Google Research, and UC Berkeley. [n.d.]. DREAMFUSION: TEXT-TO-3D USING 2D DIFFUSION. ([n.d.]).
- [43] Alec Radford, JongWook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Askell Amanda, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *Cornell University - arXiv,Cornell University arXiv* (Feb 2021).
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models.

Ziqi Yu, Jing Zhou, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao

In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01042

- [45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. (Aug 2022).
- [46] 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 Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings. https://doi.org/10.1145/3528233.3530757
- [47] 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* (Jan 2022), 1–14. https://doi.org/10.1109/tpami.2022.3204461
- [48] Mathew Salvaris, Danielle Dean, and Wee Hyong Tok. 2018. Generative Adversarial Networks. 187–208. https://doi.org/10.1007/978-1-4842-3679-6\_8
- [49] Yichen Sheng, Jianming Zhang, and Bedrich Benes. 2021. SSN: Soft Shadow Network for Image Compositing. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr46437.2021.00436
- [50] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. arXiv: Learning, arXiv: Learning (Oct 2020).
- [51] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, SooYe Kim, and Daniel Aliaga. 2022. ObjectStitch: Generative Object Compositing. (Dec 2022).
- [52] Kalyan Sunkavalli, Micah K. Johnson, Wojciech Matusik, and Hanspeter Pfister. 2010. Multi-scale image harmonization. ACM Transactions on Graphics (Jul 2010), 1–10. https://doi.org/10.1145/1778765.1778862
- [53] Linfeng Tan, Jiangtong Li, Li Niu, and Liqing Zhang. 2023. Deep Image Harmonization in Dual Color Spaces. (Aug 2023).

- [54] 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.
- [55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. Neural Information Processing Systems, Neural Information Processing Systems (Jun 2017).
- [56] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. [n. d.]. Non-local Neural Networks. ([n. d.]).
- [57] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. 2022. DCCF: Deep Comprehensible Color Filter Learning Framework for High-Resolution Image Harmonization. (Jul 2022).
- [58] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2022. Paint by Example: Exemplar-based Image Editing with Diffusion Models. (Nov 2022).
- [59] 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).
- [60] Han Zhang, Ian Goodfellow, DimitrisN. Metaxas, and Augustus Odena. 2019. Self-Attention Generative Adversarial Networks. International Conference on Machine Learning, International Conference on Machine Learning (May 2019).
- [61] Shuyang Zhang, Runze Liang, and Miao Wang. 2019. ShadowGAN: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media* (Mar 2019), 105–115. https://doi.org/10.1007/s41095-019-0136-1
- [62] Jing Zhou, Ziqi Yu, Zhongyun Bao, Gang Fu, Weilei He, Chao Liang, and Chunxia Xiao. 2024. Foreground Harmonization and Shadow Generation for Com-posite Image. In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24).