Shadow Generation for Composite Image Using Diffusion Model

Qingyang Liu[1], Junqi You[1], Jianting Wang[1], Xinhao Tao[1], Bo Zhang[1], Li Niu[1,2*]
[1] Shanghai Jiao Tong University [2] miguo.ai
[1]{narumimaria,yjqsjtu2022,glory1299,taoxinhao,bo-zhang,ustcnewly}@sjtu.edu.cn

Abstract

In the realm of image composition, generating realistic shadow for the inserted foreground remains a formidable challenge. Previous works have developed image-to-image translation models which are trained on paired training data. However, they are struggling to generate shadows with accurate shapes and intensities, hindered by data scarcity and inherent task complexity. In this paper, we resort to foundation model with rich prior knowledge of natural shadow images. Specifically, we first adapt ControlNet to our task and then propose intensity modulation modules to improve the shadow intensity. Moreover, we extend the small-scale DESOBA dataset to DESOBAv2 using a novel data acquisition pipeline. Experimental results on both DESOBA and DESOBAv2 datasets as well as real composite images demonstrate the superior capability of our model for shadow generation task. The dataset, code, and model are released at https://github.com/bcmi/Object-Shadow-Generation-Dataset-DESOBAv2.

1. Introduction

Image composition [28] aims to merge the foreground of one image with another background image to produce a composite image, which has a wide range of applications like virtual reality, artistic creation, and E-commerce. Simply pasting the foreground onto the background often results in visual inconsistencies, including the incompatible illumination between foreground and background [3], lack of foreground shadow/reflection [12, 34], and so on. In this paper, we focus on the shadow issue, *i.e.*, the inserted foreground does not have plausible shadow on the background, which could significantly degrade the realism and quality of composite image.

As illustrated in Figure 1, shadow generation is a challenging task because the foreground shadow is determined by many complicated factors like the lighting information and the geometry of foreground/background. The existing shadow generation methods can be divided into rendering based methods [34–36] and non-rendering based methods [12, 22, 53]. Rendering based methods usually impose restrict assumptions on the geometry and lighting, which could hardly be satisfied in real-world scenarios. Besides, [35, 36] require users to specify the lighting information, which hinders its direct application in our task. Non-rendering based methods usually train an image-to-image translation network, based on pairs of composite images without foreground shadows and real images with foreground shadows. However, due to the training data scarcity and task difficulty, these methods are struggling to generate shadows with reasonable shapes and intensities.

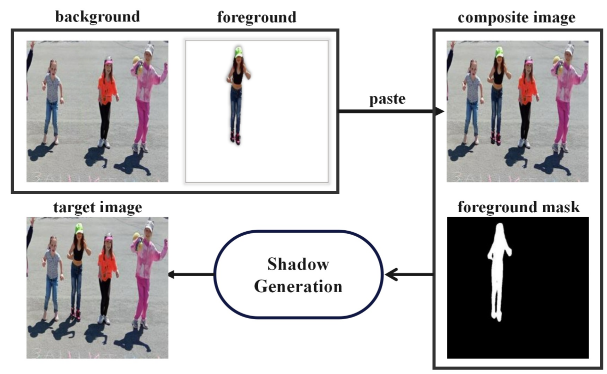

Figure 1. A composite image can be obtained by pasting the foreground on the background. Shadow generation aims to generate plausible shadow for the inserted foreground in the composite image to produce a more realistic image.

Recently, foundation model (*e.g.*, stable diffusion [32]) pretrained on large-scale dataset has demonstrated unprecedented potential for image generation and editing. In previous works [44, 48] on object-guided inpainting or composition, they show that the generated foregrounds are accompanied by shadows even without considering the shadow issue, probably because of the rich prior knowledge of natural shadow images in foundation model. However, they could only generate satisfactory shadows in simple cases and the object appearance could be altered unexpectedly.

*Corresponding author.

We build our method upon conditional foundation model [52] and propose several key innovations. First, we modify the control encoder input and the noise loss to fit our task. Then, we observe that the generated shadow intensity (the level of darkness) is unsatisfactory. Especially when the background objects has shadows, the intensity inconsistency between foreground shadow and background shadows make the whole image unrealistic. Therefore, we introduce another intensity encoder to modulate the foreground shadow intensity. Specifically, the denoising U-Net is modified to output both noise map and foreground shadow mask. The intensity encoder takes in the composite image and background shadow mask, producing the scale/bias to modulate the predicted noise within the foreground shadow region. Finally, we devise a post-processing network to rectify the color shift and background variation.

The model training requires abundant pairs of composite images without foreground shadows and real images with foreground shadows. The existing real-world shadow generation dataset DESOBA [12] is limited by scale (*i.e.*, 1,012 real images and 3,623 pairs) due to the high cost of manual shadow removal, which is insufficient to train our model. To ensure sufficient supervision, we design a novel data construction pipeline, which extends DESOBA to DESOBAv2 (*i.e.*, 21,575 real images and 28,573 pairs) using object-shadow detection and inpainting techniques. Specifically, we first collect a large number of real-world images with one or more object-shadow pairs. Then, we use pretrained object-shadow detection model [41] to predict object and shadow masks for object-shadow pairs. Next, we apply pretrained inpainting model [32] to inpaint the detected shadow regions to get deshadowed images. Finally, based on real images and deshadowed images, we construct pairs of synthetic composite images and ground-truth target images.

We conduct experiments on both DESOBAv2 and DESOBA datasets. The results reveal remarkable improvement in shadow generation task, after leveraging the benefits of large-scale data and foundation model. Our main contributions can be summarized as follows: 1) We contribute DESOBAv2, a large-scale real-world shadow generation dataset, which could greatly facilitate the shadow generation task. 2) We propose a cutting-edge diffusion model specifically designed to produce shadows for the composite foregrounds. 3) Through comprehensive experiments, we validate the efficacy of our dataset construction pipeline and the superiority of our proposed model.

2. Related Work

2.1. Image Composition

Image composition aims to overlay a foreground object on a background image to yield a composite result [20, 22, 42, 46, 47]. Previous research works have tackled different issues that can compromise the quality of composite images. For instance, image blending methods [31, 42, 49, 51] target at combining the foreground and background seamlessly. Image harmonization methods [3–6, 40] aim to rectify the illumination disparity between foreground and background. Nonetheless, the above methods largely overlook the shadow cast by the foreground onto the background. Recently, generative image composition methods [38, 44, 48] can insert a foreground object into a bounding box in the background and the inserted object is likely to have shadow effect. However, they could only generate satisfactory shadows in simple cases and the object appearance could be altered unexpectedly.

2.2. Shadow Generation

In this paper, the goal of shadow generation task is generating plausible shadow for the composite foreground. Existing methods can be broadly categorized into rendering based methods and non-rendering based methods. The rendering based methods necessitate a comprehensive understanding of factors like illumination, reflectance, material properties, and scene geometry to produce shadows for the inserted objects. However, such detailed knowledge relies on user input [15, 16, 21, 35, 36] or model prediction [1, 7, 19, 50], which is either labor-intensive or unreliable [53]. For example, [35, 36] could produce compelling results with user control. However, in the composite image, the lighting information should be inferred automatically from background instead of requested by users.

Non-rendering based methods [12, 22, 25, 53] aim to translate an input composite image without foreground shadow to an output with foreground shadow, bypassing the need for explicit knowledge of the aforementioned factors. For instance, ShadowGAN [53] utilizes both global and local conditional discriminator to enhance the realism of generated shadows. ARShadowGAN [22] emphasizes the importance of background shadow and uses it to guide foreground shadow generation. SGRNet [12] encourages the information exchange between foreground and background, and employs a classic illumination model for better shadow effect. The work [25] produces multiple underexposure images and fuses them to get the final shadow region. DMASNet [39] decomposes shadow mask prediction into box prediction and shape prediction, achieving better cross-domain transferability.

To the best of our knowledge, we are the first diffusion-based method focusing on shadow generation.

2.3. Diffusion Models

In recent years, diffusion models have emerged as a powerful tool in image generation and image editing. These models approach image generation as a series of stochastic transitions, moving from a basic distribution to the desired

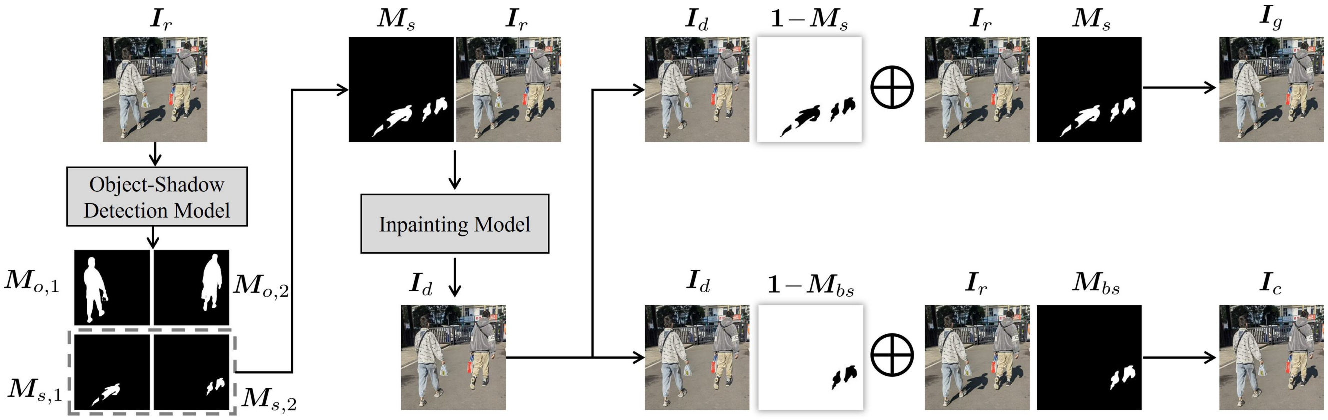

Figure 2. The pipeline of dataset construction. We use object-shadow detection model [41] to predict pairs of object and shadow masks in the real image $\boldsymbol{I}_r$. Then we obtain the union $\boldsymbol{M}_s$ of all shadow masks as the inpainting mask and apply inpainting model [32] to get a deshadowed image $\boldsymbol{I}_d$. After designating a foreground object, we replace the background shadow regions $\boldsymbol{M}_{bs}$ in $\boldsymbol{I}_d$ with the counterparts in $\boldsymbol{I}_r$ to synthesize a composite image $\boldsymbol{I}_c$, and replace all the shadow regions $\boldsymbol{M}_s$ in $\boldsymbol{I}_d$ with the counterparts in $\boldsymbol{I}_r$ to obtain the ground-truth target image $\boldsymbol{I}_g$.

data distribution [11]. Diffusion models can be divided into unconditional diffusion models [11, 37] and conditional diffusion models [27, 32, 52]. Unconditional diffusion models focus on generating realistic images by capturing the distribution of natural images, without the need of any specific input conditions. Conditional diffusion models are designed to produce images under the guidance of specific conditional inputs, such as text descriptions, semantic masks, and so on. ControlNet [52] is a popular conditional diffusion model, which equips large pretrained text-to-image diffusion models with spatial-aware and task-specific conditions. We build our model upon ControlNet and propose several innovations to meet the specific requirements of shadow generation.

3. Dataset Construction

The pipeline of our dataset construction is illustrated in Figure 2, which will be detailed next.

3.1. Shadow Image Collection

We harvest an extensive collection of real-world outdoor images with natural lighting across various scenes from two sources. On one hand, we crawl online images from public websites that have licenses for reuse. On the other hand, we hire photographers to capture photos in the outdoor scenes that satisfy our requirements. We only preserve the images with at least one object-shadow pair, arriving at 44,044 images.

3.2. Shadow Removal

Given a real image $\boldsymbol{I}_r$ with object-shadow pairs, we use the pretrained object-shadow detection model [41] to predict K pairs of object and shadow masks. We use $\boldsymbol{M}_{o,k}$ (*resp.*, $\boldsymbol{M}_{s,k}$) to denote the object (*resp.*, shadow) mask of the k-th object. We refer to one detected object-shadow pair as one detected instance. We eliminate the images without any detected instance.

Subsequently, we attempt to erase all the detected shadows. We have tried some state-of-the-art shadow removal models [8, 9], but the performance in the wild is below our expectation due to poor generalization ability. Considering the recent rapid advance of image inpainting [14, 23, 29, 32, 45, 56] techniques, we resort to image inpainting to remove the shadows. Although image inpainting cannot preserve the background information precisely, we observe that the background textures in the shadow region are usually very simple, and the inpainted result has similar textures with the original background. Thus, we roughly treat the inpainted results as deshadowed results.

We obtain the union of all detected shadow masks $\boldsymbol{M}_s = \boldsymbol{M}_{s,1} \cup \boldsymbol{M}_{s,2} \cup \cdots \cup \boldsymbol{M}_{s,K}$ as the inpainting mask and apply the pretrained inpainting model [32] to get a deshadowed image $\boldsymbol{I}_d$. In practice, we observe that the inpainting model is prone to generate low-quality shadow in the inpainted region in some cases. To prevent the inpainting model from generating undesirable shadows in the inpainted region, we adopt some tricks like dilating the inpainting mask and flipping images vertically, which can effectively obstruct undesirable shadow generation during inpainting. However, there may still exist undesirable shadows or noticeable artifacts in the inpainted region.

After inpainting, we manually filter the object-shadow pairs according to the following rules: 1) We remove the object-shadow pairs with low-quality object masks or shadow masks. 2) We remove those object-shadow pairs with generated shadows or noticeable artifacts in the inpainted region. After manual filtering, we refer to the remaining object-shadow pairs as valid instances. We have 21,575 images with 28,573 valid instances.

3.3. Composite Image Synthesis

Given a pair of a real image $\boldsymbol{I}_r$ and a deshadowed image $\boldsymbol{I}_d$, we randomly select the k-th foreground object from valid instances and synthesize the composite image. $\boldsymbol{M}_{o,k}$ (*resp.*, $\boldsymbol{M}_{s,k}$) is referred to as the foreground object (*resp.*, shadow) mask $\boldsymbol{M}_{fo}$ (*resp.*, $\boldsymbol{M}_{fs}$). One strategy is replacing the shadow region $\boldsymbol{M}_{fs}$ of this foreground object in $\boldsymbol{I}_r$ with the counterpart in $\boldsymbol{I}_d$ to erase the foreground shadow. However, this strategy may leave traces along the shadow boundary, in which case the model may find a shortcut to generate the shadow. Another strategy is replacing the shadow regions $\boldsymbol{M}_{bs} = \boldsymbol{M}_{s,1} \cup \cdots \cup \boldsymbol{M}_{s,k-1} \cup \boldsymbol{M}_{s,k+1} \cup \cdots \cup \boldsymbol{M}_{s,K}$ of the other objects in $\boldsymbol{I}_d$ with the counterparts in $\boldsymbol{I}_r$ to synthesize a composite image $\boldsymbol{I}_c$, in which only the selected foreground object does not have shadow while all the other objects have shadows. We adopt the second strategy.

After inpainting, the background may undergo slight changes, so the background of $\boldsymbol{I}_c$ may be slightly different from that of $\boldsymbol{I}_r$. To ensure consistent background, we obtain the ground-truth target image $\boldsymbol{I}_g$ by replacing the shadow regions $\boldsymbol{M}_s$ of all objects in $\boldsymbol{I}_d$ with the counterparts in $\boldsymbol{I}_r$. Then, $\boldsymbol{I}_c$ and $\boldsymbol{I}_g$ form a pair of input composite image and ground-truth target image. So far, we obtain tuples in the form of $\{\boldsymbol{I}_c, \boldsymbol{M}_{fo}, \boldsymbol{M}_{fs}, \boldsymbol{M}_{bs}, \boldsymbol{I}_g\}$, which will be used for model training. Example images and more statistics of our dataset can be found in the supplementary.

4. Background

Stable Diffusion [32] is latent diffusion model operating in a latent space. First, 512×512 images are converted to 64×64 latent images using VAE [18] with encoder E_r and decoder D_r. The image space is projected to the latent space using E_r, and back to the image space using D_r. Then, the forward diffusion process and backward denoising process are performed in the latent space. The denoising U-Net [33] consists of an encoder with 12 blocks, a middle block, and a skip-connected decoder with 12 blocks.

During training, random Gaussian noise $\boldsymbol{\epsilon}$ is added to the latent image $\boldsymbol{z}_0$ in the denoising step t, producing a noisy latent image $\boldsymbol{z}_t$. Given time step t and text prompt $\boldsymbol{c}_{txt}$, the denoising U-Net with model parameters $\epsilon_{\boldsymbol{\theta}}$ is trained to predict the added noise $\boldsymbol{\epsilon}$.

To support spatial conditional information (*e.g.*, edge, pose, depth), ControlNet [52] integrates a control encoder E_c with pre-trained Stable Diffusion. Specifically, the control encoder contains trainable replicas of its 12 encoding blocks and middle block across four resolutions $(64\times64, 32\times32, 16\times16, 8\times8)$. It takes a 512×512 conditional image as input.

The conditional feature maps $\boldsymbol{c}_{img}$ output from control encoder are used to enhance the 12 skip-connections and middle block in denoising U-Net via zero convolution layers. While the original Stable Diffusion is fixed to retain prior knowledge, control encoder could incorporate additional conditions to guide image generation. The objective could be rewritten as

$$\mathcal{L}_{ctrl} = \mathbb{E}_{t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)}\Big[\|\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{txt}, \boldsymbol{c}_{img})\|_2^2\Big]. \quad (1)$$

5. Method

Given a composite image $\boldsymbol{I}_c$ without foreground shadow as well as the foreground object mask $\boldsymbol{M}_{fo}$, our Shadow Generation Diffusion (SGDiffusion) model aims to produce $\tilde{\boldsymbol{I}}_g$ with plausible foreground shadow. We will adapt ControlNet [52] to shadow generation task in Section 5.1, and propose novel modules to improve the shadow intensity in Section 5.2. Finally, we will briefly introduce post-processing techniques to enhance the image quality in Section 5.3.

5.1. Adapting ControlNet to Shadow Generation

For shadow generation task, the useful conditional information is input composite image $\boldsymbol{I}_c$ and foreground object mask $\boldsymbol{M}_{fo}$, in which the foreground object mask indicates the target object we need to generate shadow for. We concatenate $\boldsymbol{I}_c$ with $\boldsymbol{M}_{fo}$ as the input of control encoder E_c. The control encoder outputs the conditional feature maps $\boldsymbol{c}_{sg}$, which are injected into the denoising decoder to provide guidance. For the text prompt, we have tried several variants like "the [object category] with shadow", but they have no significant impact on the generated shadows. Therefore, we use null text prompt by default.

Given a set of conditions including time step t and conditional feature maps $\boldsymbol{c}_{sg}$, the denoising U-Net with model parameters $\epsilon_{\boldsymbol{\theta}}$ predicts the noise $\boldsymbol{\epsilon}$ added to the noisy latent image $\boldsymbol{z}_t$:

$$\mathcal{L}_{sg} = \mathbb{E}_{t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)}\Big[\|\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{sg}))\|_2^2\Big]. \quad (2)$$

To enforce the model to place more emphasis on the foreground shadow region, we introduce weighted noise loss, which assigns higher weights to the foreground shadow region. We expand the foreground shadow mask by a dilated kernel to get the expanded mask $\hat{\boldsymbol{M}}_{fs}$. The weights in the expanded foreground shadow region are w while the other weights are 1, leading to the weight map $\boldsymbol{W}_{fs}$. If we do not expand the foreground shadow region, the model will be misled to generate large shadows, overlooking the details of shadow shapes and boundaries. By applying weight map $\boldsymbol{W}_{fs}$ to the noise loss, we can arrive at

$$\mathcal{L}_{wsg} = \mathbb{E}_{t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)}\Big[\|\boldsymbol{W}_{fs}\circ(\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{sg}))\,\|_2^2\Big], \quad (3)$$

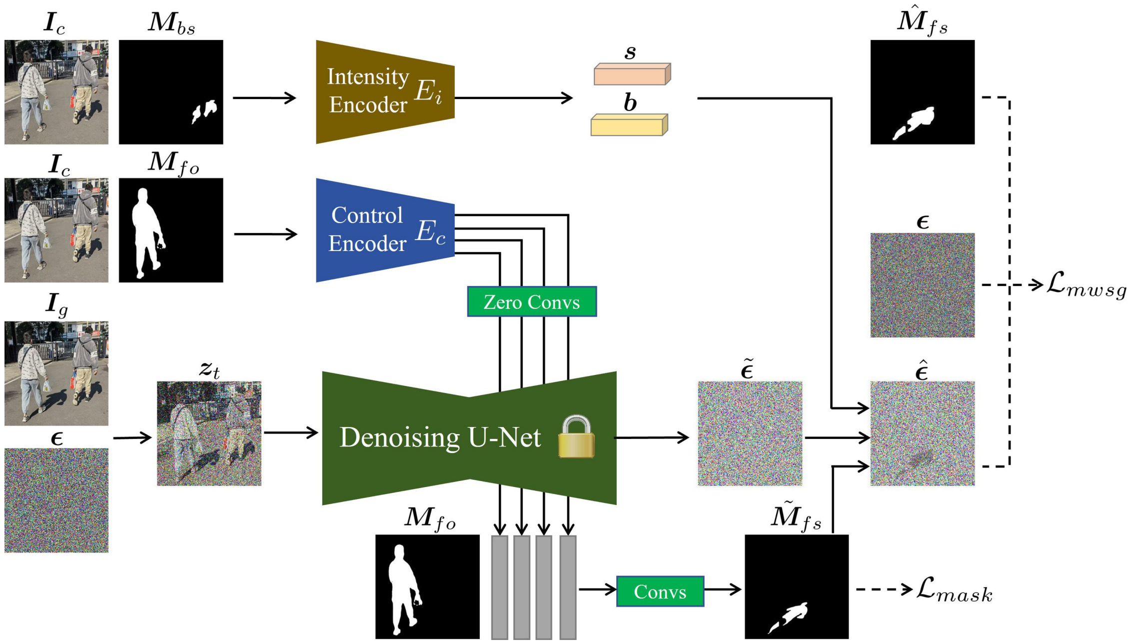

Figure 3. The framework of our SGDiffusion. We adapt ControlNet (Control Encoder and Denoising U-Net) to shadow generation task. We also introduce an intensity encoder to modulate the foreground shadow region in the noise map $\tilde{\boldsymbol{\epsilon}}$, leading to $\hat{\boldsymbol{\epsilon}}$. The output noise $\hat{\boldsymbol{\epsilon}}$ is supervised by weighted noise loss $\mathcal{L}_{mwsg}$ based on the expanded foreground shadow mask $\hat{\boldsymbol{M}}_{fs}$

where $\circ$ denotes element-wise multiplication.

During inference, to retain more information of input composite image $\boldsymbol{I}_c$ in the initial noise, we obtain $\boldsymbol{z}_T$ by adding noise to the latent image of $\boldsymbol{I}_c$, rather than directly sampling from the Gaussian distribution $\mathcal{N}(0, 1)$.

5.2. Shadow Intensity Modulation

By using the adapted ControlNet in Section 5.1, we observe that the intensity of generated foreground shadow is unsatisfactory. Especially when the background has object-shadow pairs, the generated foreground shadow is often notably darker or brighter than background shadows. Such inconsistency between foreground shadow intensity and background shadow intensity makes the whole image unrealistic.

Therefore, we introduce another intensity encoder to modulate the foreground shadow intensity. Specifically, we use encoder E_i to extract intensity-relevant information. Intuitively, by observing background shadows and its surrounding unshadowed areas, we can estimate the intensity of foreground shadows. Thus, the input of intensity encoder E_i should include the composite image $\boldsymbol{I}_c$ and background shadow mask $\boldsymbol{M}_{bs}$. When there is no background shadow, the mask is all black. We concatenate $\boldsymbol{I}_c$ with background shadow mask $\boldsymbol{M}_{bs}$ as the input of intensity encoder.

The intensity encoder outputs scales and biases to adjust the intensity of noise map within the foreground shadow region. The modulated noise map results in the modulated latent image, and further results in the modulated foreground shadow. Therefore, the intensity adjustment of noise map is finally embodied in the intensity variation of generated foreground shadow. Specifically, when the noise map has c channels, E_i outputs the c-dim scale vector $\boldsymbol{s}$ and c-dim bias vector $\boldsymbol{b}$, containing channel-wise scales and biases. $\boldsymbol{s}$ and $\boldsymbol{b}$ are used to modulate the predicted noise map within the foreground shadow region.

One problem is that the foreground shadow region is unknown in the testing stage, so we need to predict the foreground shadow mask. To avoid much extra computational cost, we take advantage of the feature maps in the denoising U-Net to predict the foreground shadow mask. Previous works usually combine different layers of feature maps in denoising U-Net for mask prediction [24, 43]. We try different layers of feature maps and find that decoder feature maps are more effective in shadow mask prediction. We also use foreground object mask, which could provide useful hints for the location of foreground shadow. We resize all decoder feature maps and foreground object mask to the same size, and concatenate them channel-wisely. The concatenation passes through several convolutional layers to predict the foreground shadow mask $\tilde{\boldsymbol{M}}_{fs}$. $\tilde{\boldsymbol{M}}_{fs}$ is supervised with ground-truth foreground shadow mask $\boldsymbol{M}_{fs}$ by Binary Cross-Entropy (BCE) loss and Dice loss [26]:

$$\mathcal{L}_{mask} = \mathcal{L}_{bce}(\tilde{\boldsymbol{M}}_{fs}, \boldsymbol{M}_{fs}) + \mathcal{L}_{dice}(\tilde{\boldsymbol{M}}_{fs}, \boldsymbol{M}_{fs}). \quad (4)$$

When t is large, $\boldsymbol{z}_t$ is close to random noise and thus the decoder feature maps are not informative to predict shadow mask. Hence, we only employ the loss $\mathcal{L}_{mask}$ when the time step t is small. We set the threshold of t as σT, in which T is the total number of steps. Accordingly, shadow intensity modulation is only applied when t is smaller than the threshold σT.

Provided with the predicted foreground shadow mask $\tilde{\boldsymbol{M}}_{fs}$, we can modulate the noise map within the foreground shadow region. Given the predicted noise map $\tilde{\boldsymbol{\epsilon}} = \epsilon_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}_{sg})$, we multiply $\tilde{\boldsymbol{\epsilon}}$ by channel-wise scales $\boldsymbol{s}$ and add channel-wise biases $\boldsymbol{b}$ to get $\tilde{\boldsymbol{\epsilon}}'$. Then, based on $\tilde{\boldsymbol{M}}_{fs}$, we combine the modulated noise map and original noise map to get the final noise map: $\hat{\boldsymbol{\epsilon}} = \tilde{\boldsymbol{\epsilon}}' \circ \tilde{\boldsymbol{M}}_{fs} + \tilde{\boldsymbol{\epsilon}} \circ (1 - \tilde{\boldsymbol{M}}_{fs})$.

We replace the predicted noise map in Eqn. (3) with the final noise map $\hat{\boldsymbol{\epsilon}}$ and get

$$\mathcal{L}_{mwsg} = \mathbb{E}_{t, \boldsymbol{\epsilon} \sim \mathcal{N}(0,1)} \left[\| \boldsymbol{W}_{fs} \circ (\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}) \|_2^2 \right]. \quad (5)$$

We summarize the mask prediction loss in Eqn. (4) and weighted noise loss in Eqn. (5) as

$$\mathcal{L}_{all} = \mathcal{L}_{mask} + \lambda \mathcal{L}_{mwsg}, \quad (6)$$

where λ is a trade-off parameter.

5.3. Post-processing

We observe that the generated images could have color shift and background variation issues. Color shift means that the overall color tone deviates from the input composite image. Background variation means that some background details are changed. To solve these issues, we create a multi-task post-processing network which yields the rectified image together with the foreground shadow mask. Then, we combine input composite image and rectified image based on the predicted foreground shadow mask to produce the final image. The technical details are left to supplementary.

6. Experiments

6.1. Datasets and Evaluation Metrics

We conduct experiments on both DESOBA [12] and our contributed DESOBAv2 dataset. We split DESOBAv2 into 21,088 training images with 27,718 tuples and 487 test images with 855 tuples. Following [12], the test set contains BOS images (with background object-shadow pairs) and BOS-free images. Most of our experiments are based on DESOBAv2 dataset due to the following two concerns: 1) DESOBAv2 has larger test set which supports more comprehensive evaluation. 2) DESOBA has the artifacts caused by manual shadow removal and the existing methods (*e.g.*, SGRNet) tend to overfit such artifacts.

For the generated results, we evaluate both image quality and mask quality. For image evaluation, following [12], we adopt RMSE and SSIM, which are calculated based on the ground-truth target image and the generated image. Global RMSE (GR) and Global SSIM (GS) are calculated over the whole image, while Local RMSE (LR) and Local SSIM (LS) are calculated over the ground-truth foreground shadow region. For the mask evaluation, following [12], we adopt Balanced Error Rate (BER), which is calculated based on the ground-truth binary foreground shadow mask and the predicted foreground shadow mask obtained by threshold 0.5. Global BER (GB) is calculated over the whole image, while Local BER (LB) is calculated over the ground-truth foreground shadow region. Note that diffusion model has stochastic property and shadow generation is a multi-modal task, that is, one input has multiple plausible outputs. Similar to multi-modal inpainting evaluation [54, 55], we generate 5 results for one test image with different random seeds and select the one closest to the ground-truth (the highest Local SSIM) to calculate evaluation metrics.

6.2. Implementation Details

We develop our method with PyTorch 1.12.1 [30]. Our model is trained using the Adam optimizer [17] with a constant learning rate of $1e^{-5}$ over 50 epochs, on four NVIDIA RTX A6000 GPUs. Our method is built upon ControlNet [52]. We employ ResNet18 [10] as the intensity encoder. The mask predictor passes the concatenation of decoder feature maps and foreground object mask through four convolutional layers, with ReLU activation following the first three layers and Sigmoid activation following the last layer. We set the hyper-parameters w, σ, and λ as 10, 0.7, and 1, respectively.

6.3. Comparison with Baselines

Following [12], we compare with ShadowGAN [53], Mask-ShadowGAN [13], ARShadowGAN [22], and SGRNet [12]. We train and test all methods on DESOBAv2 dataset. The quantitative results are summarized in Table 1. We observe that our SGDiffusion achieves the lowest GRMSE, LRMSE and the highest GSSIM, LSSIM, which demonstrates that our method could generate shadow images that are closer to the ground-truth shadow images. The best GB and LB results demonstrate that the shapes and locations of our generated shadows are more accurate.

For qualitative comparison, we show several example results in Figure 4. Compared with the baseline methods, the shadows produced by our model have more reasonable shapes and intensities. Moreover, as shown in row 1, our method can take into account the self-occlusion of objects to generate discontinuous shadows. As shown in row 4, our method can also consider the material of the objects, producing shadows with translucency effects. We provide more examples in the supplementary.

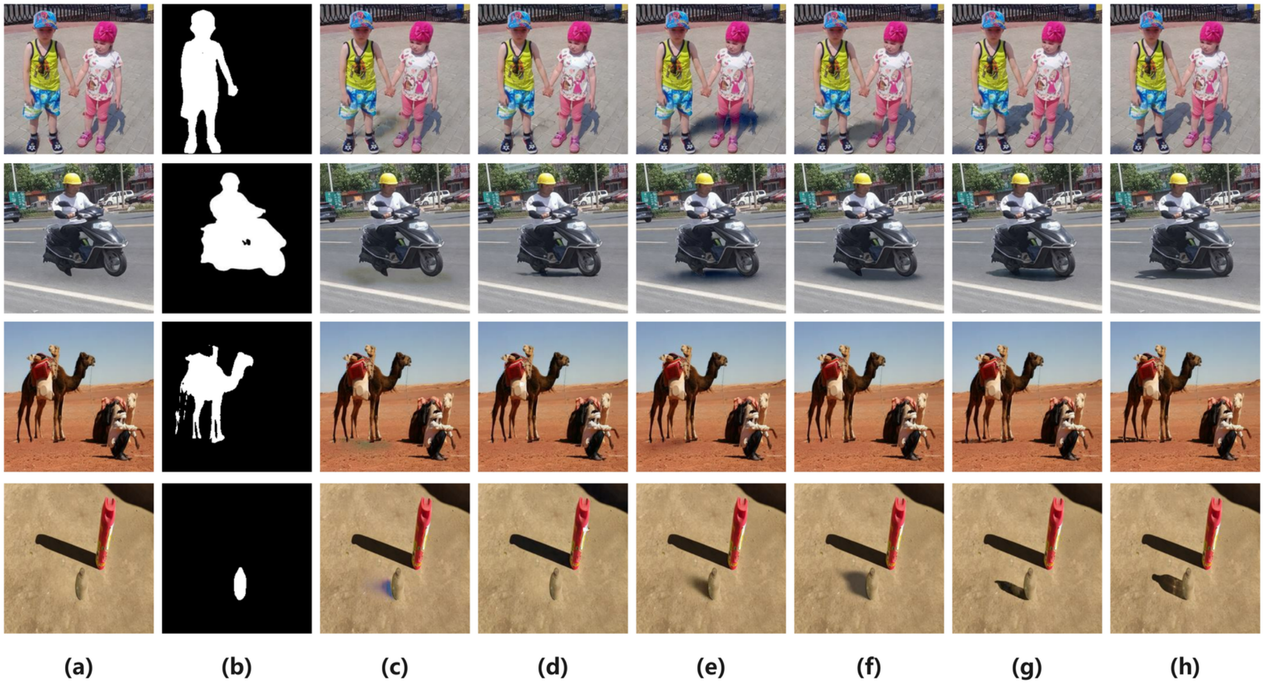

Figure 4. Visual comparison of different methods on DESOBAv2 dataset. From left to right are input composite image (a), foreground object mask (b), results of ShadowGAN [53] (c), MaskshadowGAN [13] (d), ARShadowGAN [22] (e), SGRNet [12] (f), our SGDiffusion (g), ground-truth (h).

Method	BOS Test Images						BOS-free Test Images					
	GR ↓	LR ↓	GS ↑	LS ↑	GB↓	LB↓	GR ↓	LR ↓	GS ↑	LS ↑	GB↓	LB↓
ShadowGAN [53]	7.511	67.464	0.961	0.197	0.446	0.890	17.325	76.508	0.901	0.060	0.425	0.842
MaskshadowGAN [13]	8.997	79.418	0.951	0.180	0.500	1.000	19.338	94.327	0.906	0.044	0.500	1.000
ARShadowGAN [22]	7.335	58.037	0.961	0.241	0.383	0.761	16.067	63.713	0.908	0.104	0.349	0.682
SGRNet [12]	7.184	68.255	0.964	0.206	0.301	0.596	15.596	60.350	0.909	0.100	0.271	0.534
SGDiffusion	**6.098**	**53.611**	**0.971**	**0.370**	**0.245**	**0.487**	**15.110**	**55.874**	**0.913**	**0.117**	**0.233**	**0.452**

Table 1. The results of different methods on DESOBAv2 dataset. The best results are highlighted in boldface.

6.4. Ablation Studies

We study the impact of weighted noise loss (WL), intensity modulation (IM), and post-processing (PP) of our SGDiffusion on BOS test images from DESOBAv2. The quantitative results are summarized in Table 2.

In row 1, we report the results of basic ControlNet without weighted noise loss. For WL, the comparison between row 3 and row 1 emphasizes the importance of paying more attention to the foreground shadow region. We also report a special case † in row 2, where the foreground shadow mask is not expanded when constructing the weight map. The results in row 2 are comparable or even worse than those in row 1, as the model tends to generate larger shadow size while ignoring shape and edge details. For IM, the comparison between row 1 and row 5 shows that the intensity modulation can significantly improve the shadow quality by adjusting the shadow intensity. We also report a special case ○ in row 4, where the intensity encoder input does not contain background shadow mask. The comparison between row 4 and row 5 shows that background shadow mask is helpful, because the background shadow regions and their surrounding regions could provide useful clues to infer shadow intensity. For PP, the comparison between row 6 and row 7 demonstrates that post-processing effectively corrects color shift and background variations, substantially reducing the

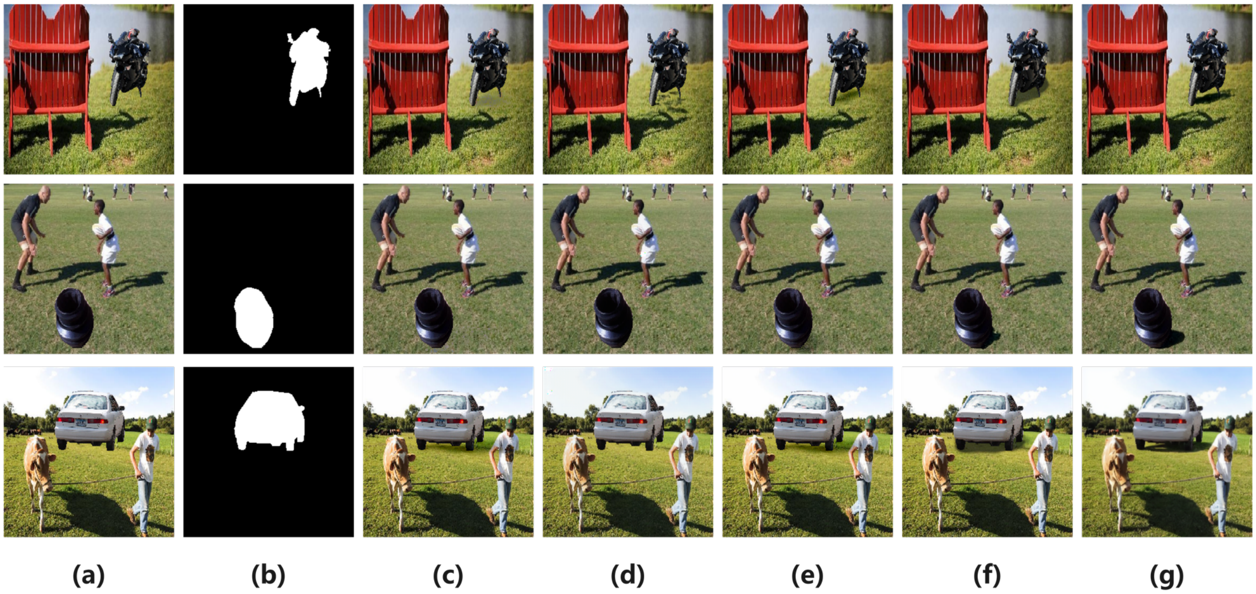

Figure 5. Visual comparison of different methods on real composite images. From left to right are input composite image (a), foreground object mask (b), results of ShadowGAN [53] (c), MaskshadowGAN [13] (d), ARShadowGAN [22] (e), SGRNet [12] (f), SGDiffusion (g).

Row	WL	IM	PP	GR ↓	LR ↓	GB↓	LB↓
1	-	-	+	8.285	59.753	0.271	0.534
2	†	-	+	8.319	59.491	0.282	0.563
3	+	-	+	7.041	53.829	0.249	0.492
4	-	○	+	7.410	56.121	0.269	0.536
5	-	+	+	7.357	54.159	0.262	0.526
6	+	+	-	13.447	55.231	0.245	0.487
7	+	+	+	6.098	53.611	0.245	0.487

Table 2. Ablation studies of our method on BOS test images from DESOBAv2 dataset. WL is short for weighted loss and † means without expanding shadow mask. IM is short for intensity modulation and ○ means without using background shadow mask. PP is short for post-processing.

global RMSE. We also provide the visual results of ablated versions in the supplementary.

6.5. Real Composite Images

We compare different methods on real composite images provided by [12], where background images and foreground objects are from the DESOBA [12] test set. We train all methods on DESOBAv2 and finetune them on DESOBA. The visual results of different methods are showcased in Figure 5. These results confirm that SGDiffusion adeptly synthesizes lifelike shadows with precise contours, locations, and directions, which are compatible with the background object-shadow pairs and foreground object information. In contrast, previous methods often produce vague and misdirected shadows. We provide more examples in the supplementary.

Given the absence of ground-truth images for real composite images, following [12], we opt for subjective evaluation, engaging 50 human raters in the user study. Each participant is presented with image pairs from the results generated by 5 methods, and asked to choose the image with more realistic foreground shadow. Using the Bradley-Terry model [2], we report the B-T scores in the supplementary, which again proves the advantage of our method.

7. Conclusion

In this paper, we have contributed a large-scale shadow generation dataset DESOBAv2. We have also designed a novel diffusion-based shadow generation method. Extensive experimental results show that our method is able to generate plausible shadows for composite foregrounds, significantly surpassing previous methods.

8. Acknowledgement

The work was supported by the National Natural Science Foundation of China (Grant No. 62076162), the Shanghai Municipal Science and Technology Major/Key Project, China (Grant No. 2021SHZDZX0102).

References

[1] Ibrahim Arief, Simon McCallum, and Jon Yngve Hardeberg. Realtime estimation of illumination direction for augmented reality on mobile devices. In *CIC*, 2012. 2

[2] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 8

[3] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, 2020. 1, 2

[4] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *ICME*, 2021.

[5] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *CVPR*, 2022.

[6] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *TIP*, 2020. 2

[7] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-Francois Lalonde. Deep parametric indoor lighting estimation. In *ICCV*, 2019. 2

[8] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. In *AAAI*, 2023. 3

[9] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *CVPR*, 2023. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[12] Yan Hong, Li Niu, and Jianfu Zhang. Shadow generation for composite image in real-world scenes. *AAAI*, 2022. 1, 2, 6, 7, 8

[13] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *ICCV*, 2019. 6, 7, 8

[14] Shaozong Huang and Lan Hong. Diffusion model for mural image inpainting. In *ITOEC*, 2023. 3

[15] Kevin Karsch, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, Hailin Jin, Rafael Fonte, Michael Sittig, and David Forsyth. Automatic scene inference for 3d object compositing. *ACM TOG*, 2014. 2

[16] Eric Kee, James F. O'Brien, and Hany Samir Farid. Exposing photo manipulation from shading and shadows. *ACM TOG*, 2014. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 4

[19] Bin Liao, Yao Zhu, Chao Liang, Fei Luo, and Chunxia Xiao. Illumination animating and editing in a single picture using scene structure estimation. *Computers & Graphics*, 82:53–64, 2019. 2

[20] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *CVPR*, 2018. 2

[21] Bin Liu, Kun Xu, and Ralph R Martin. Static scene illumination estimation from videos with applications. *JCST*, 32 (3):430–442, 2017. 2

[22] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *CVPR*, 2020. 1, 2, 6, 7, 8

[23] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 3

[24] Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. *CoRR*, abs/2303.17870, 2023. 5

[25] Quanling Meng, Shengping Zhang, Zonglin Li, Chenyang Wang, Weigang Zhang, and Qingming Huang. Automatic shadow generation via exposure fusion. *IEEE Transactions on Multimedia*, 2023. 2

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 5

[27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *CoRR*, abs/2302.08453, 2023. 3

[28] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *CoRR*, abs/2106.14490, 2021. 1

[29] Sibam Parida, Vignesh Srinivas, Bhavishya Jain, Rajesh Naik, and Neeraj Rao. Survey on diverse image inpainting using diffusion models. In *PCEMS*, 2023. 3

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32, 2019. 6

[31] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *SIGGRAPH*. 2003. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[34] Yichen Sheng, Jianming Zhang, and Bedrich Benes. Ssn: Soft shadow network for image compositing. In *CVPR*, 2021. 1

[35] Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A Cengiz Oztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. Controllable shadow generation using pixel height maps. In *ECCV*, 2022. 1, 2

[36] Yichen Sheng, Jianming Zhang, Julien Philip, Yannick Hold-Geoffroy, Xin Sun, He Zhang, Lu Ling, and Bedrich Benes. Pixht-lab: Pixel height based light effect generation for image compositing. In *CVPR*, 2023. 1, 2

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *CoRR*, abs/2010.02502, 2020. 3

[38] Yi-Zhe Song, Zhifei Zhang, Zhe L. Lin, Scott D. Cohen, Brian L. Price, Jianming Zhang, Soo Ye Kim, and Daniel G. Aliaga. Objectstitch: Generative object compositing. In *CVPR*, 2023. 2

[39] Xinhao Tao, Junyan Cao, Yan Hong, and Li Niu. Shadow generation with decomposed mask prediction and attentive shadow filling. In *AAAI*, 2024. 2

[40] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017. 2

[41] Tianyu Wang, Xiaowei Hu, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection with a single-stage detector. *TPAMI*, 2022. 2, 3

[42] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *ACM MM*, 2019. 2

[43] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 5

[44] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, 2023. 1, 2

[45] Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *ACM MM*, 2023. 3

[46] Fangneng Zhan, Jiaxing Huang, and Shijian Lu. Adaptive composition gan towards realistic image synthesis. *CoRR*, abs/1905.04693, 2019. 2

[47] Fangneng Zhan, Shijian Lu, Changgong Zhang, Feiying Ma, and Xuansong Xie. Towards realistic 3d embedding via view alignment. *CoRR*, abs/2007.07066, 2020. 2

[48] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023. 1, 2

[49] He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, and Vishal M Patel. Deep image compositing. In *WACV*, 2021. 2

[50] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *CVPR*, 2019. 2

[51] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *WACV*, 2020. 2

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 4, 6

[53] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5:105–115, 2019. 1, 2, 6, 7, 8

[54] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *CVPR*, 2020. 6

[55] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, 2019. 6

[56] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Image inpainting with cascaded modulation gan and object-aware training. In *ECCV*, 2022. 3