Shadow Generation for Composite Image Using Diffusion Model

Supplementary Material

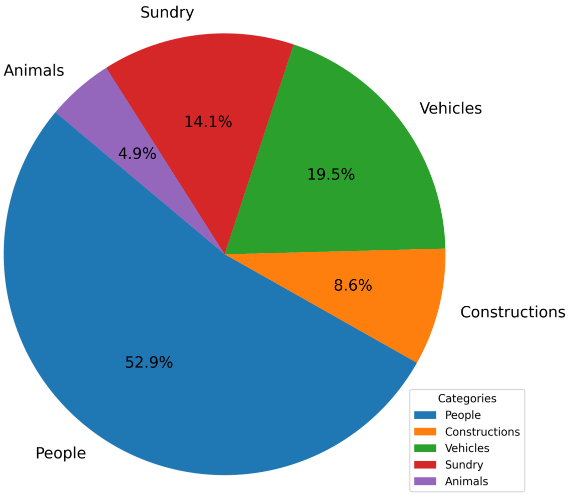

Figure 1. The super-category distribution of foreground object in DESOBAv2 dataset.

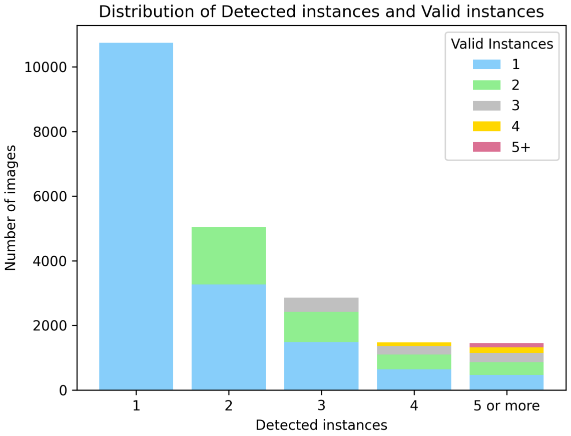

Figure 2. The distribution of images with different numbers of detected instances in DESOBAv2 dataset. Each bar further contains the distribution of images with different numbers of valid instances.

In this document, we provide additional materials to support our main submission. In Section 1, we will provide example images and more statistics of our DESOBAv2 dataset. In Section 2, we will describe the technical details of post-processing. In Section 3, we will show the results of our ablated versions. In Section 4, we will show the qualitative results of our method and baseline methods on DESOBA dataset [2]. In Section 5, we will show more qualitative results of our method and baseline methods on DESOBAv2 dataset. In Section 6, we will show more qualitative results on the real composite images and report the B-T score. In Section 7, we will compare our method with recent generative composition methods [7, 9]. In Section 8, we will show some failure cases of our method.

1. More Statistics of Our DESOBAv2 Dataset

First, we plot the super-category distribution of foreground objects in our DESOBAv2 dataset in Figure 1. We classify all objects into five super-categories (people, constructions, vehicles, sundry, animals). From Figure 1, it can be seen that our DESOBAv2 dataset covers a diversity of categories, in which “people” is the dominant super-category.

We provide some examples from our DESOBAv2 dataset in Figure 3 and Figure 4. For each super-category (people, constructions, vehicles, sundry, animals), we show two tuples in the form of $\{\boldsymbol{I}_c, \boldsymbol{M}_{fo}, \boldsymbol{M}_{fs}, \boldsymbol{M}_{bo}, \boldsymbol{M}_{bs}, \boldsymbol{I}_g\}$, in which $\boldsymbol{I}_c$ is composite image, $\boldsymbol{M}_{fo}$ is foreground object mask, $\boldsymbol{M}_{fs}$ is foreground shadow mask, $\boldsymbol{M}_{bo}$ is background object mask, $\boldsymbol{M}_{bs}$ is background shadow mask, $\boldsymbol{I}_g$ is ground-truth target image.

Then, we summarize the statistics of detected instances and valid instances in our DESOBAv2 dataset. Recall that we use object-shadow detection model [8] to detect object-shadow pairs, and refer to one detected object-shadow pair as one detected instance. After manually filtering the low-quality instances, we refer to the remaining instances as valid instances. Our DESOBAv2 dataset has in total $21,575$ images. In Figure 2, we first plot the distribution of images with different numbers of detected instances, based on which most images have fewer than 5 detected instances. Among the images with specific number of detected instances, we further plot the distribution of images with different numbers of valid instances. Note that all images in our dataset have at least one valid instance, so $10,752$ images with one detected instance all have one valid instance. The images with more than one detected instance have different numbers of valid instances.

2. Technical Details of Post-processing

To address the problem of color shift and background distortion (see Figure 8), we develop a post-processing network which consists of one encoder and two decoders, as illustrated in the left part of Figure 5. The encoder E_p takes the concatenation of generated image $\tilde{\boldsymbol{I}}_g$, composite image $\boldsymbol{I}_c$, and foreground object mask $\boldsymbol{M}_{fo}$ as input. It can be seen that $\tilde{\boldsymbol{I}}_g$ and $\boldsymbol{I}_c$ have notable color discrepancy. One de-

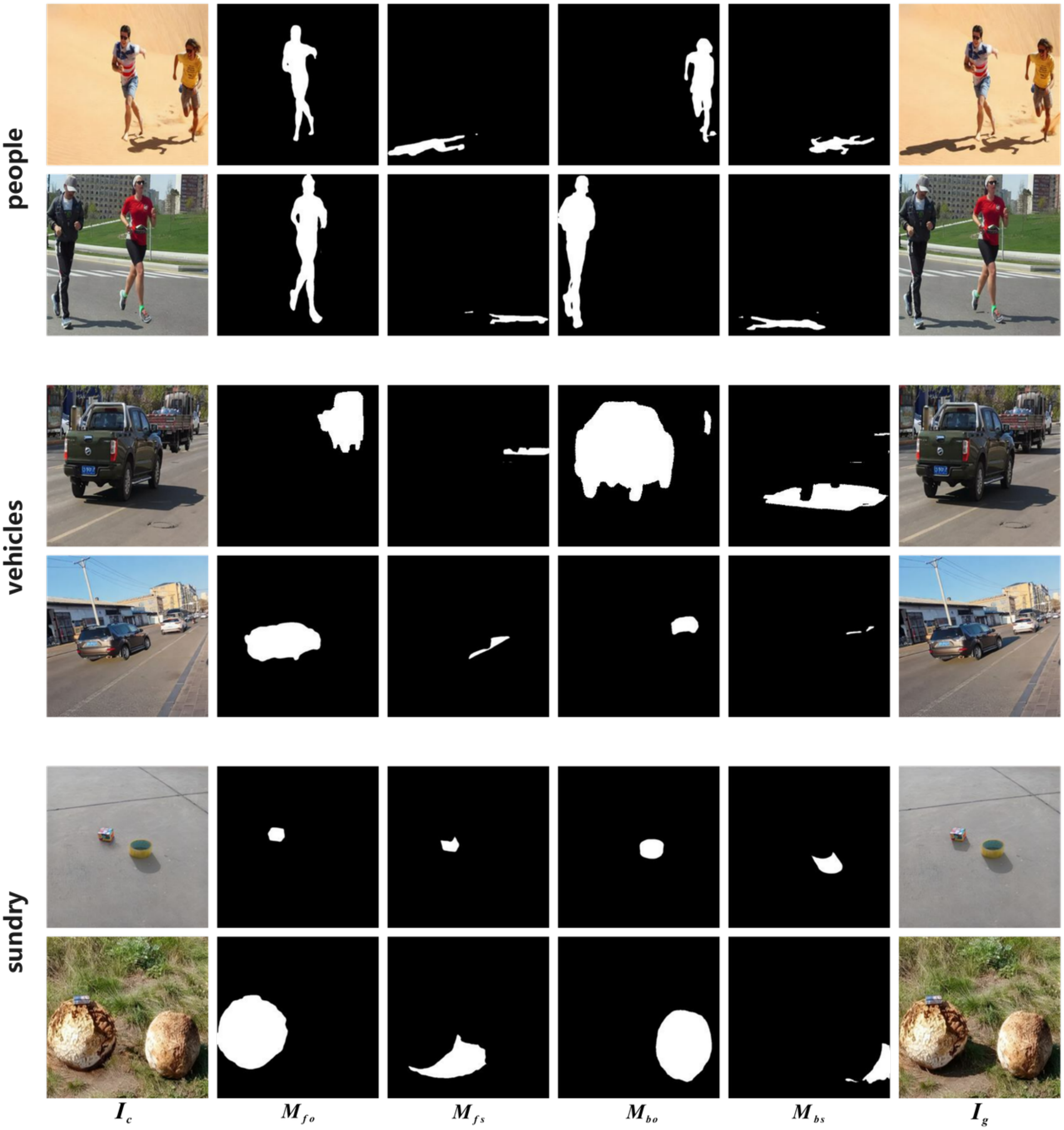

Figure 3. Some examples of “people”, “vehicles”, and “sundry” super-categories in our DESOBAv2 dataset. From left to right in each row, we show the composite image $\boldsymbol{I}_c$, the foreground object mask $\boldsymbol{M}_{fo}$, the foreground shadow mask $\boldsymbol{M}_{fs}$, the background object mask $\boldsymbol{M}_{bo}$, the background shadow mask $\boldsymbol{M}_{bs}$, the ground-truth target image $\boldsymbol{I}_g$.

coder D_i produces the rectified images $\tilde{\boldsymbol{I}}'_g$ to fix the color shift problem. Its main goal is adjusting the color of $\tilde{\boldsymbol{I}}_g$ to match $\boldsymbol{I}_c$, that is, the output $\tilde{\boldsymbol{I}}'_g$ and the input $\boldsymbol{I}_c$ should be the same except the foreground shadow region. When the color of the other regions is rectified, the color of foreground shadow region is rectified synchronously. The rectified foreground shadow region will be included in the final image.

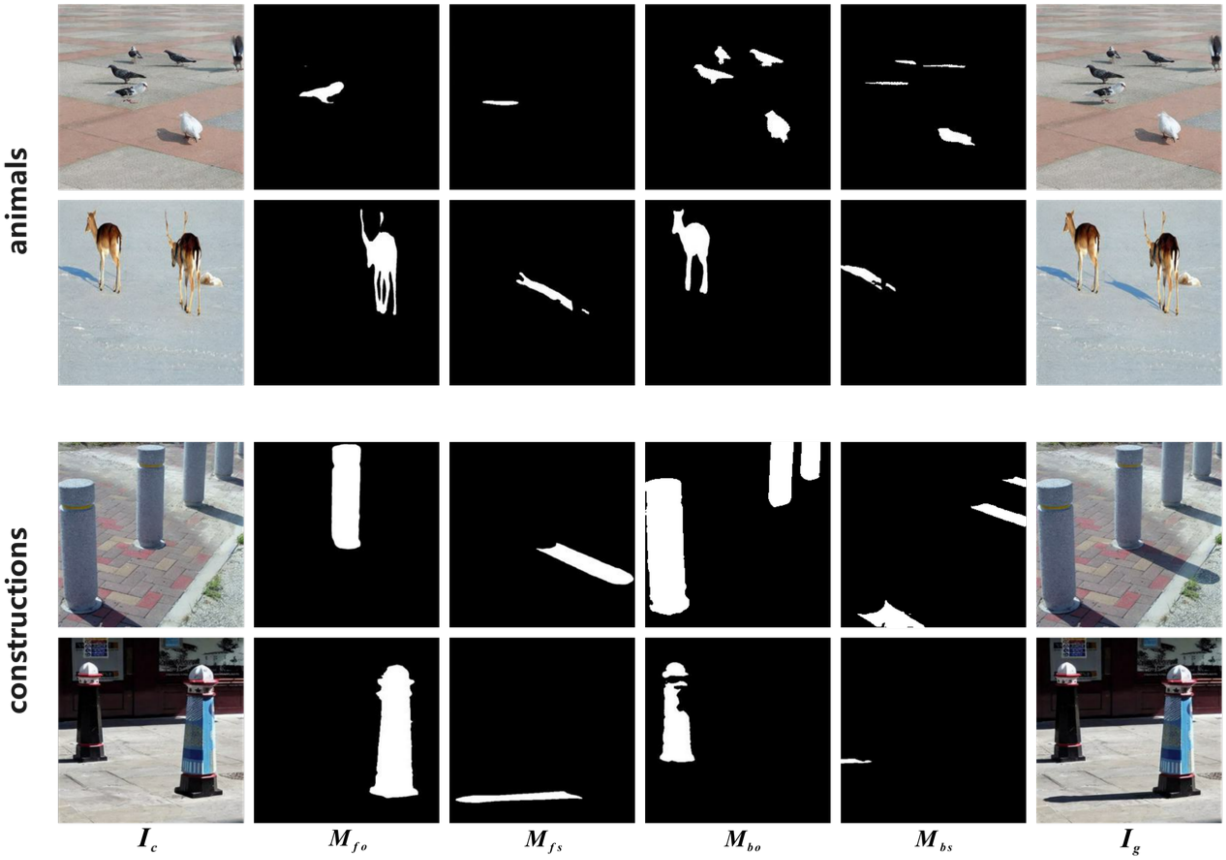

Figure 4. Some examples of "animals" and "constructions" super-categories in our DESOBAv2 dataset. From left to right in each row, we show the composite image $\boldsymbol{I}_c$, the foreground object mask $\boldsymbol{M}_{fo}$, the foreground shadow mask $\boldsymbol{M}_{fs}$, the background object mask $\boldsymbol{M}_{bo}$, the background shadow mask $\boldsymbol{M}_{bs}$, the ground-truth target image $\boldsymbol{I}_g$.

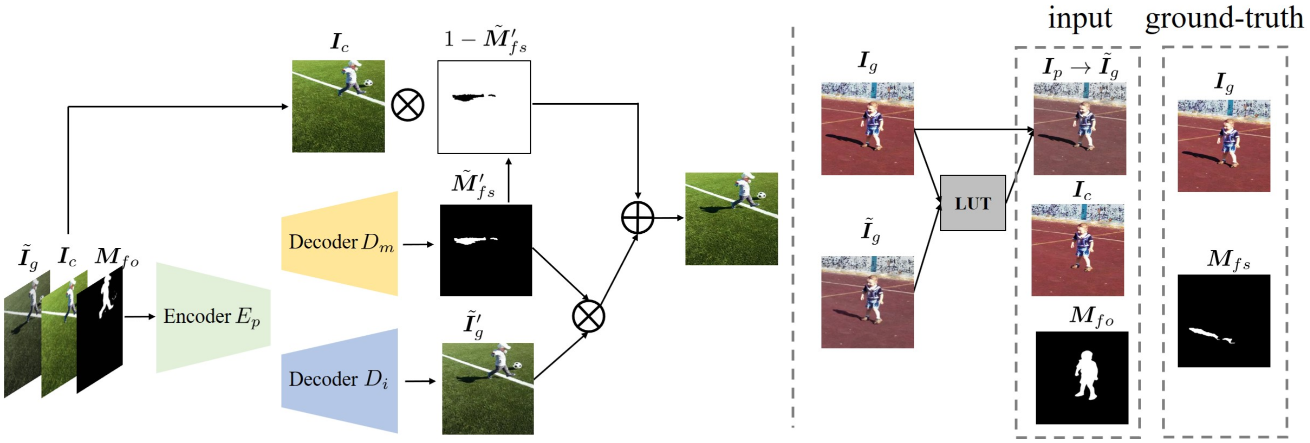

Figure 5. The framework of our post-processing network. In the left part, we show our post-processing network structure which can refine a generated image. In the right part, we show the process of constructing training data to train post-processing network.

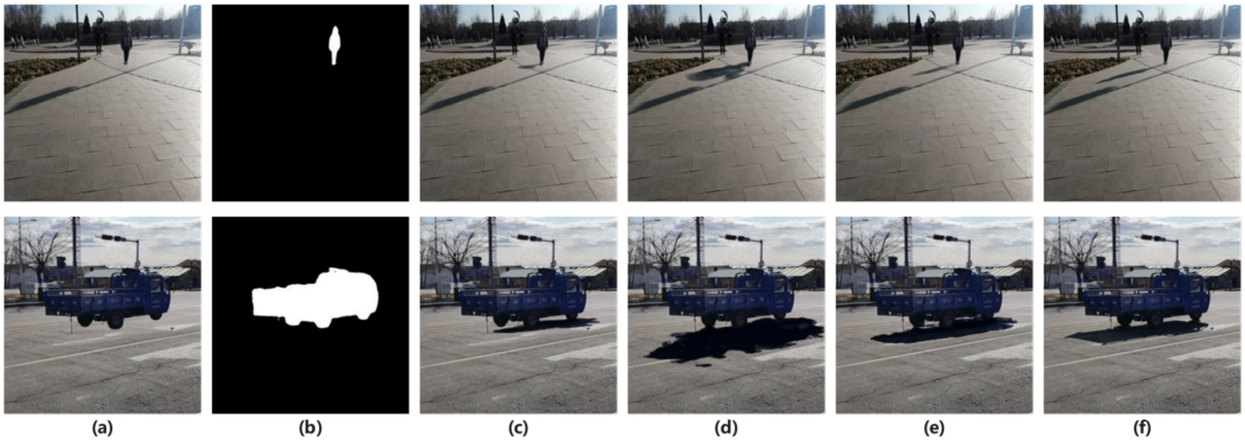

Figure 6. The ablation studies on weighted noise loss. From left to right are input composite image (a), foreground object mask (b), results of row 1 (c), row 2 (d), row 3 (e) in Table 2 in the main paper, and ground-truth (f).

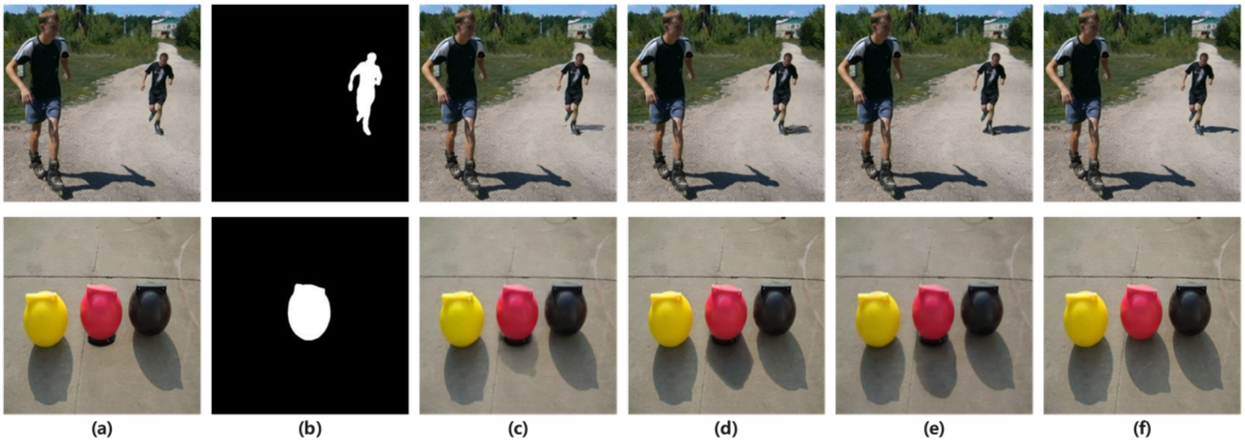

Figure 7. The ablation studies on intensity modulation. From left to right are input composite image (a), foreground object mask (b), results of row 1 (c), row 4 (d), row 5 (e) in Table 2 in the main paper, and ground-truth (f).

The other decoder D_m predicts the foreground shadow mask $\tilde{\boldsymbol{M}}'_{fs}$. Intuitively, the foreground shadow region could be easily localized by spotting the content difference between inputs $\tilde{\boldsymbol{I}}_g$ and $\boldsymbol{I}_c$. The input foreground object mask $\boldsymbol{M}_{fo}$ could also provide useful hints for the location of foreground shadow. Compared with the foreground shadow mask $\tilde{\boldsymbol{M}}_{fs}$ predicted by denoising U-Net, $\tilde{\boldsymbol{M}}'_{fs}$ is more accurate with higher resolution. The final image can be obtained by $\tilde{\boldsymbol{I}}'_g \circ \tilde{\boldsymbol{M}}'_{fs} + \boldsymbol{I}_c \circ (1 - \tilde{\boldsymbol{M}}'_{fs})$, in which $\circ$ is element-wise product. In this way, we rectify the color of foreground shadow region and faithfully preserve the background details.

Both the encoder and the two decoders have four blocks. Each encoder block has three 3×3 conv layers with ReLU followed by a downsampling layer. Each decoder block has three 3×3 conv layers with ReLU followed by an upsampling layer. The whole network structure is U-Net, with skip connections from all encoder blocks to the corresponding decoder blocks.

Next, we discuss the construction process of training data to train the post-processing network. The construction process is illustrated in the right part of Figure 5. We hope that the post-processing network only adjusts the color of $\tilde{\boldsymbol{I}}_g$ without changing the foreground shadow shape. To sim-

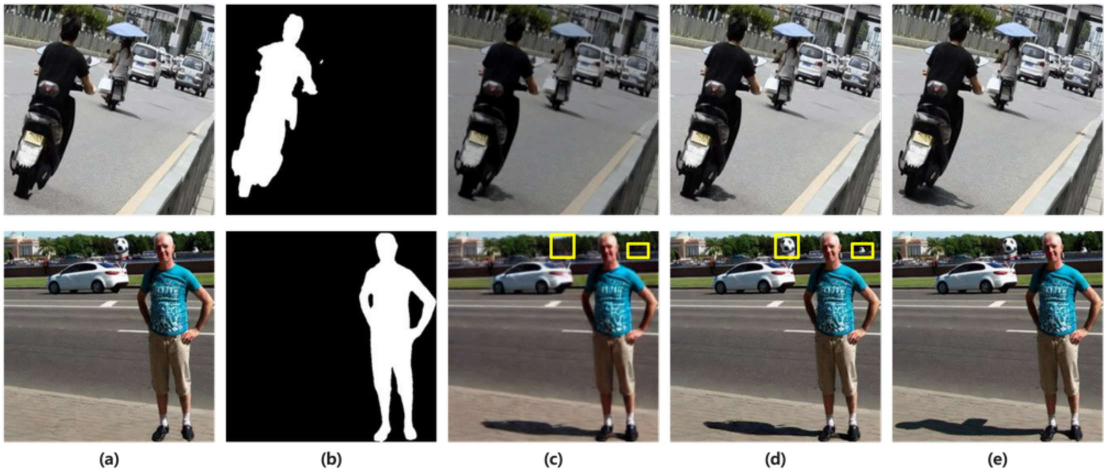

Figure 8. The ablation studies on post-processing. From left to right are input composite image (a), foreground object mask (b), results of row 6 (c), row 7 (d) in Table 2 in the main paper, and ground-truth (e).

Method	B-T score ↑
ShadowGAN	-0.312
Mask-ShadowGAN	-1.284
ARShadowGAN	-0.228
SGRNet	0.062
SGDiffusion	1.763

Table 1. B-T scores of different methods on 100 real composite images.

ulate the color shift issue, we perturb the color of ground-truth image $\boldsymbol{I}_g$ in DESOBAv2 training set. To ensure that the simulated color shift is close to the real color shift, we first obtain our generated result $\tilde{\boldsymbol{I}}_g$, and then optimize an image-specific look-up table (LUT) as the color mapping from ground-truth image $\boldsymbol{I}_g$ to generated result $\tilde{\boldsymbol{I}}_g$. After that, we apply the optimized LUT to $\boldsymbol{I}_g$ to get the perturbed ground-truth image $\boldsymbol{I}_p$. Note that $\boldsymbol{I}_g$ and $\boldsymbol{I}_p$ are only different in color. When training the post-processing network, we treat $\boldsymbol{I}_p$ as the pseudo generated image $\tilde{\boldsymbol{I}}_g$, which is taken along with the composite image $\boldsymbol{I}_c$ and foreground object mask $\boldsymbol{M}_{fo}$ as input. $\boldsymbol{I}_g$ is used to supervise the rectified images $\tilde{\boldsymbol{I}}'_g$, and $\boldsymbol{M}_{fs}$ is used to supervise the predicted foreground shadow mask $\tilde{\boldsymbol{M}}'_{fs}$.

3. Visualization of Ablation Studies

In Table 2 in the main paper, we conduct ablation studies to prove the effectiveness of each design in our method. We show the visual results of our ablated versions on DESOBAv2 test set. We divide all ablated versions into three groups. The first group contains row 1, row 2, row 3, which validates the effectiveness of weighted loss. The second group contains row 1, row 4, row 5, which validates the effectiveness of intensity modulation. The third group contains row 6 and row 7, which validates the effectiveness of post-processing.

The visual results of the first group are shown in Figure 6. By comparing (c) and (e), we can observe that the shapes of generated foreground shadows of (e) are more accurate and closer to the ground-truth (f), which proves that it is useful to pay more attention to the expanded shadow region. However, when not expanding the foreground shadow mask (d), only emphasizing the shadow region makes the model prefer to generate unreasonably large shadows.

The visual results of the second group are shown in Figure 7. By comparing (c) and (e), we can observe that intensity modulation can substantially enhance the intensity of generated shadow, which is more compatible with background shadows. The improvement of (e) over (d) verifies that background shadows could provide useful hints for shadow intensity modulation.

The visual results of the third group are shown in Figure 8. From (c), we can see that the global color tone of generated image severely deviates from the input composite image (a), and some background details (yellow bounding boxes in the second row) are lost. The improvement of (d) over (c) verifies that post-processing is able to address the color shift and preserve the background details.

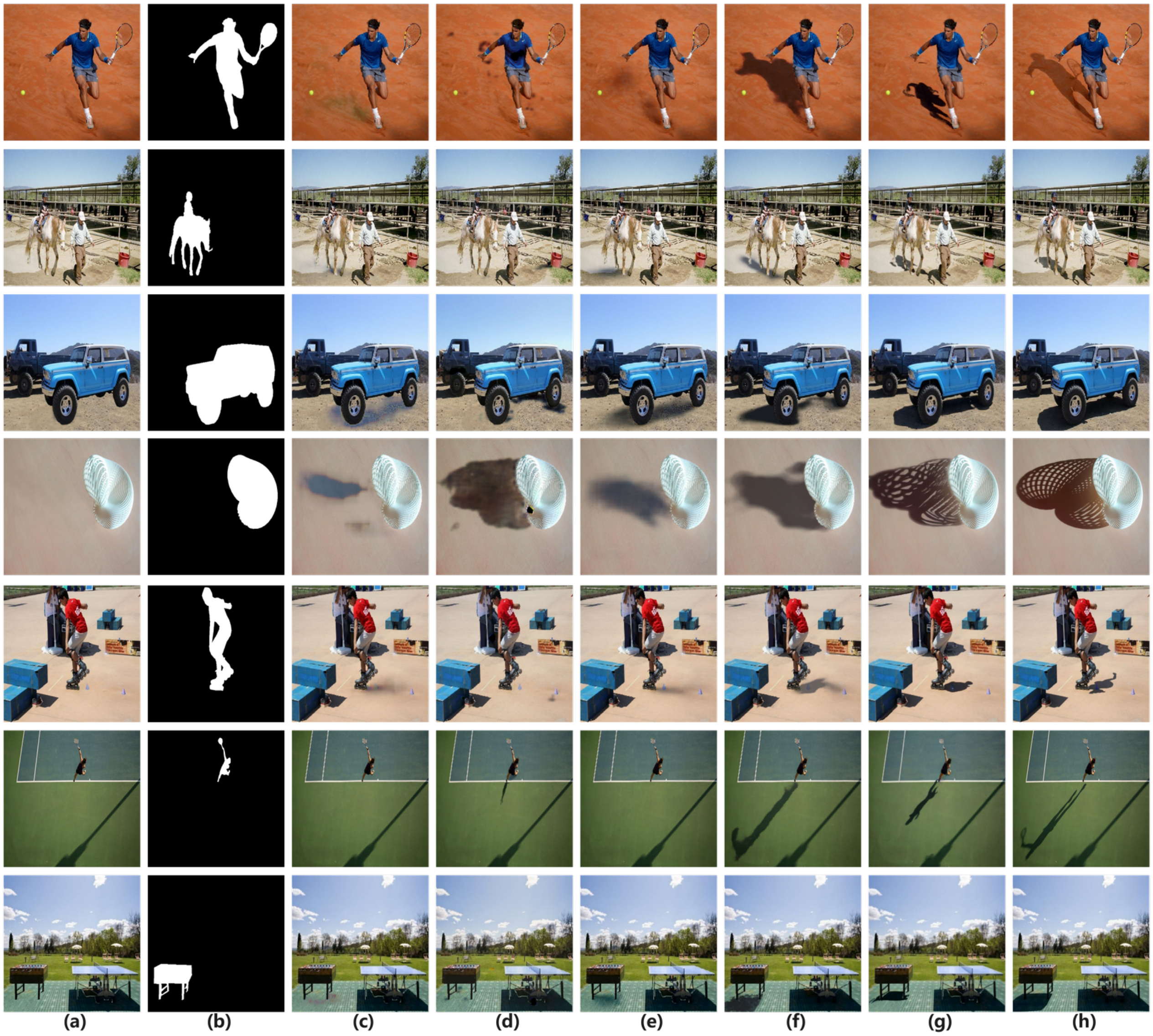

Figure 9. Visual comparison of different methods on DESOBA dataset. From left to right are input composite image (a), foreground object mask (b), results of ShadowGAN [11] (c), MaskshadowGAN [3] (d), ARShadowGAN [6] (e), SGRNet [2] (f), SGDiffusion (g), ground-truth (h).

4. Evaluation on DESOBA Dataset

To further substantiate the robustness of our model, we train all methods on DESOBAv2 training set and then finetune them on DESOBA training set. We test different methods on DESOBA test set, and visualize the results in Figure 9.

We can see that our model excels in generating more accurate and plausible shadow shapes in comparison to the baseline methods. ShadowGAN [11] and Maskshadow-GAN [3] are struggling to produce shadows. ARShadow-GAN [6] tends to produce oval and blurry shadows, regardless of the object shape. In contrast, our model is capable of producing shadows with reasonable shapes and intricate details, even for the person with complicated pose (*e.g.*, row 1) and hollowed-out objects (*e.g.*, row 4).

Besides, we observe that SGRNet is prone to overfit the artifacts in DESOBA dataset. The artifacts are brought by manual shadow removal, based on which the model may find a shortcut for the shadow outline. As shown in row 6 and row 7, the generated shadows of SGRNet are surprisingly close to the ground-truth, while the shadow shapes

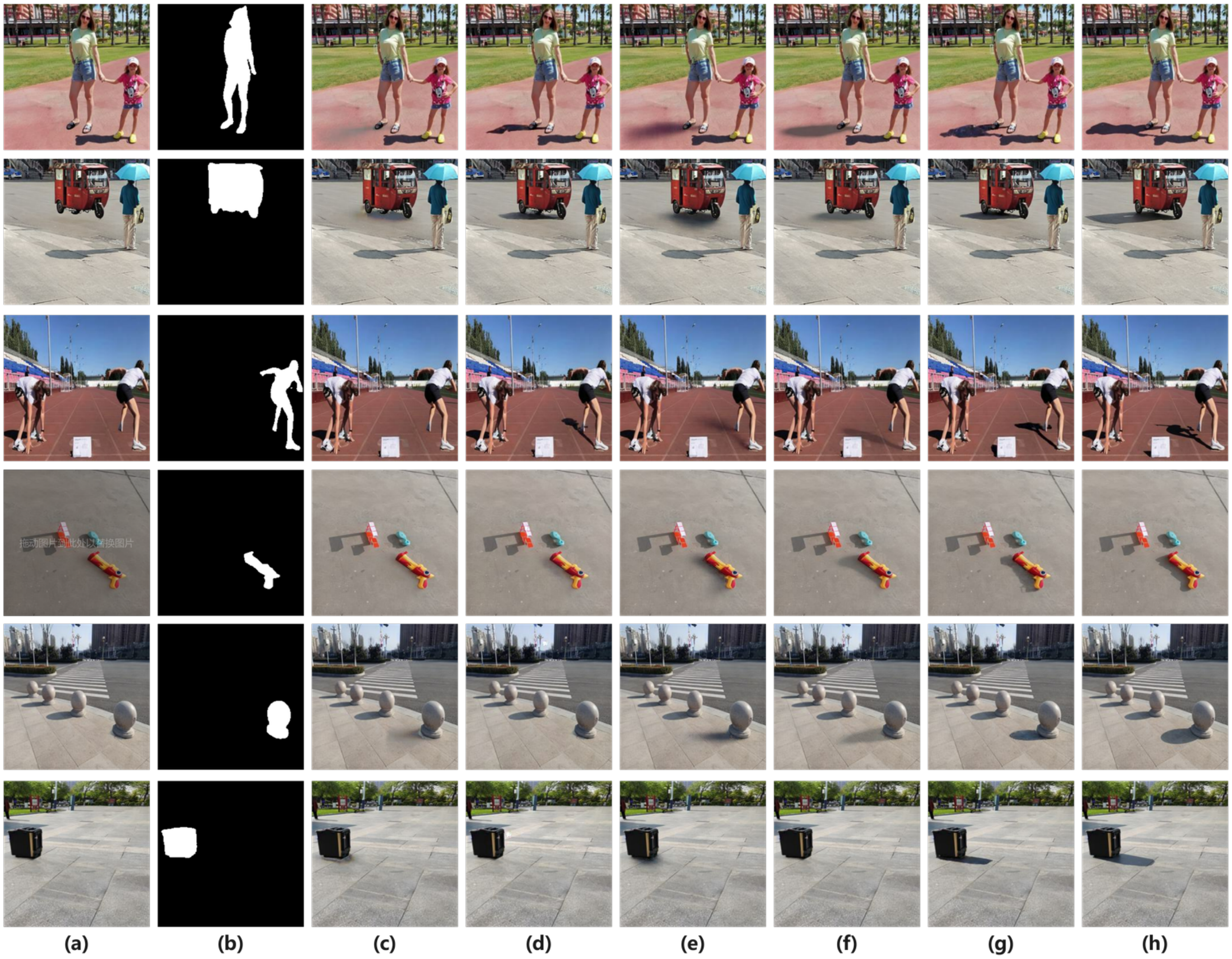

Figure 10. Visual comparison of different methods on DESOBAv2 dataset. From left to right are input composite image (a), foreground object mask (b), results of ShadowGAN [11] (c), MaskshadowGAN [3] (d), ARShadowGAN [6] (e), SGRNet [2] (f), SGDiffusion (g) and ground-truth (h).

could have many possibilities. We conjecture that SGRNet finds a shortcut based on the artifacts in the foreground shadow regions of input composite images.

5. More Visualization Results on DESOBAv2

In the main paper, we have shown the visualization results of different methods on DESOBAv2 test set. Here, we provide more visualization results on DESOBAv2 test set in Figure 10, based on which we have consistent observations as in the main paper.

In particular, our method can generate shadows with more plausible shapes and intensity, while other baseline methods even fail to produce any shadow. For example, in row 1 and row 3, the shadow generated by our method has more delicate shape details associated with human pose (*e.g.*, hand holding).

6. More Results on Real Composite Images

The composite images in DESOBA and DESOBAv2 test sets are synthetic composite images, which may have domain gap with the real composite images. To validate the effectiveness of our method on real composite images, we evaluate different methods on 100 real composite images provided by [2], which are obtained by pasting foreground objects on background images. Since the foregrounds/backgrounds in real composite images are sourced from DESOBA test set, we train all methods on DESOBAv2 training set and finetune them on DESOBA training set. Note that 100 real composite images provided by [2] consist of 74 composite images with one foreground object

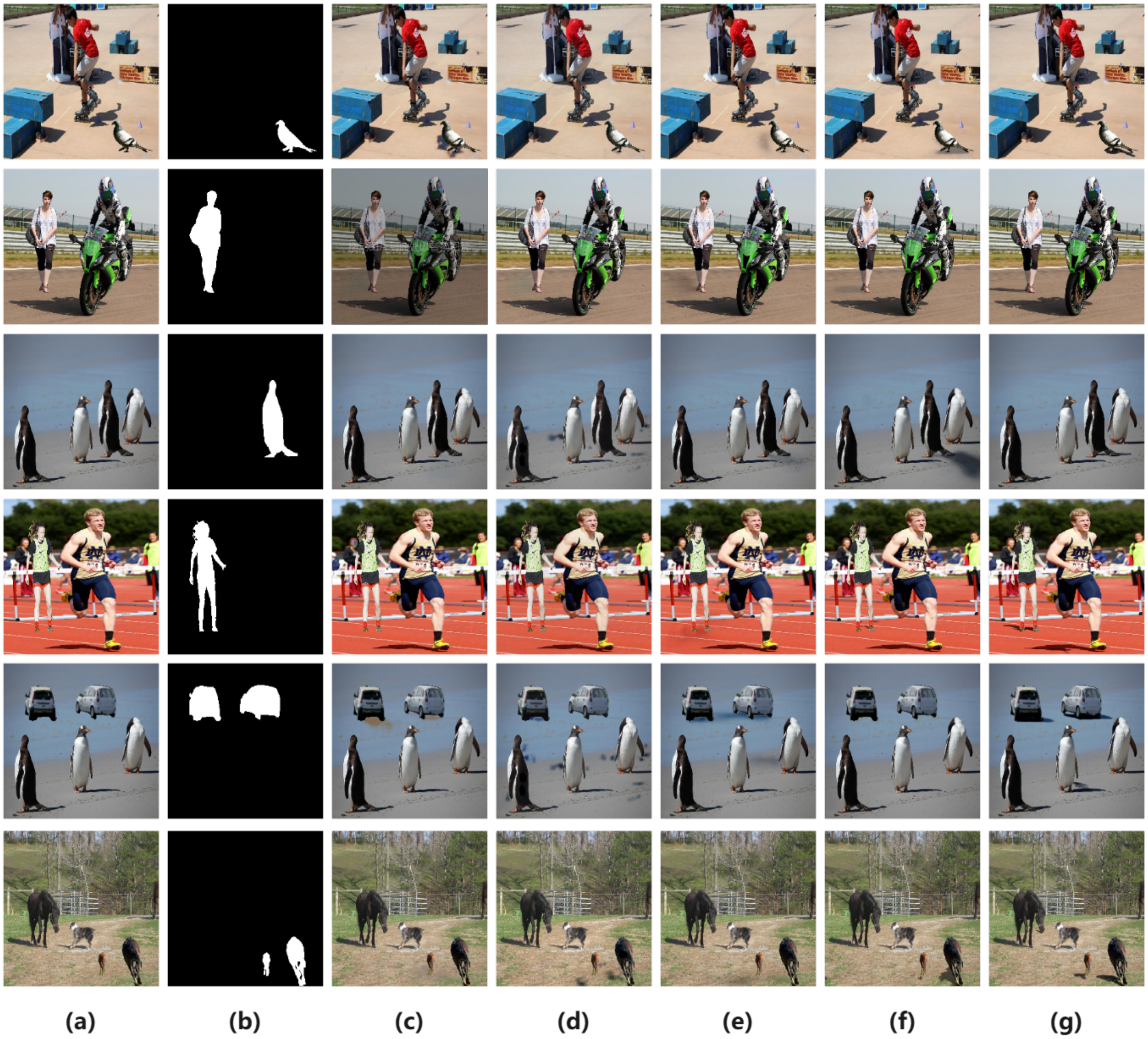

Figure 11. Visual comparison of different methods on real composite images. From left to right are input composite image (a), foreground object mask (b), results of ShadowGAN [11] (c), MaskshadowGAN [3] (d), ARShadowGAN [6] (e), SGRNet [2] (f), SGDiffusion (g).

and 26 composite images with two foreground objects. For the composite images with two foregrounds, we apply our model twice, each time with the shadow generated for one foreground object.

The results of different methods are shown in Figure 11. Our model is capable of generating realistic foreground shadows, far exceeding the existing baselines. We can observe that the results of SGRNet on real composite images are much worse than those on DESOBA, which again verifies that SGRNet is likely to overfit the artifacts in DESOBA dataset. In contrast, our method has better generalization ability and significantly outperforms SGRNet on real composite images.

Since these real composite images do not have ground-truth images, we conduct user study to compare different methods. Following SGRNet [2], given each composite image, we construct 10 image pairs by randomly selecting 2 from 5 results generated by 5 methods. In total, we can construct 1000 image pairs based on 100 real composite images. 50 users are asked to participate in this subjective evaluation. Each user could see an image pair each time and select the one whose foreground shadow is more realistic

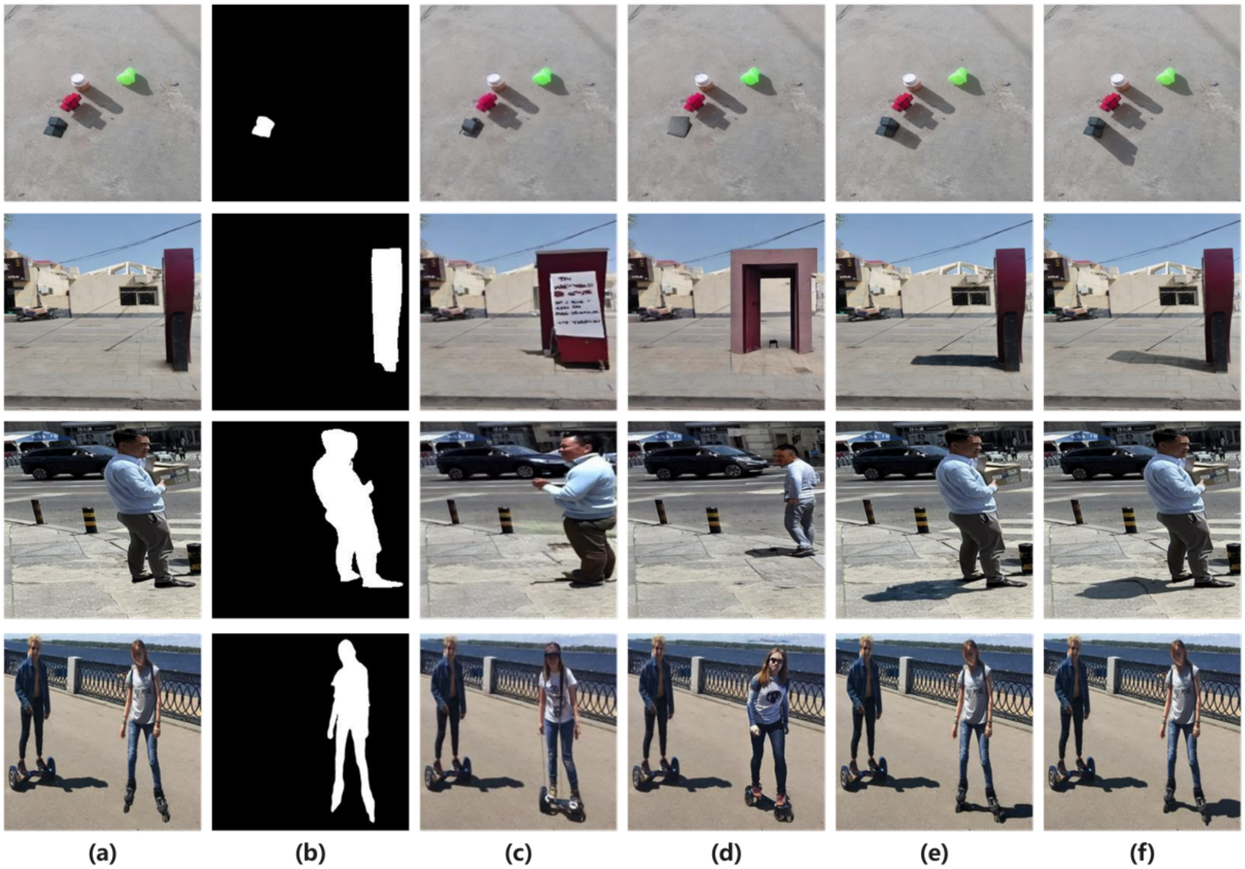

Figure 12. Visual comparison with generative composition methods on DESOBAv2 dataset. From left to right are input composite image (a), foreground object mask (b), results of ObjectStitch [7] (c), PBE [9] (d), SGDiffusion (e) and ground-truth (f).

and compatible with the background. In total, 50 users and 1000 image pairs lead to 50,000 pairwise results, based on which the Bradley-Terry (B-T) model [1, 5] is used to calculate the ranking of all methods. The B-T scores of different methods are reported in Table 1. Our method achieves the highest B-T score, which again proves that our method has outstanding generalization ability and achieves better results on real composite images.

7. Comparison with Generative Composition Methods

With the popularity of generative foundation model, generative image composition has attracted considerable research interest [7, 9, 10]. Specifically, given a pair of background with bounding box and foreground object, they aim to insert the foreground object into the bounding box to produce a realistic composite image, in which the foreground object is seamlessly blended into the background and harmonious with the background. In the generated composite image, the foreground object may have a shadow, even though these methods did not specially consider the shadow problem.

However, these methods have evident drawbacks. Firstly, they could only insert the object into the specified bounding box, but the shadow could fall out of the scope of bounding box, in which case they are unable to generate reasonable shadow. Secondly, the identity of foreground object could be significantly altered, which may be against the user intention. The comparison between our approach and generative image composition methods [7, 9] on DESOBAv2 test set is shown in Figure 12.

For [7, 9], we treat the bounding box enclosing the composite foreground as bounding box and the cropped composite foreground as reference object. We use their released model pretrained on large-scale dataset [4]. Based on the results in Figure 12, we find that these methods [7, 9] are not suitable for our task and their generated shadows have poor quality. Unlike these methods, our approach can generate

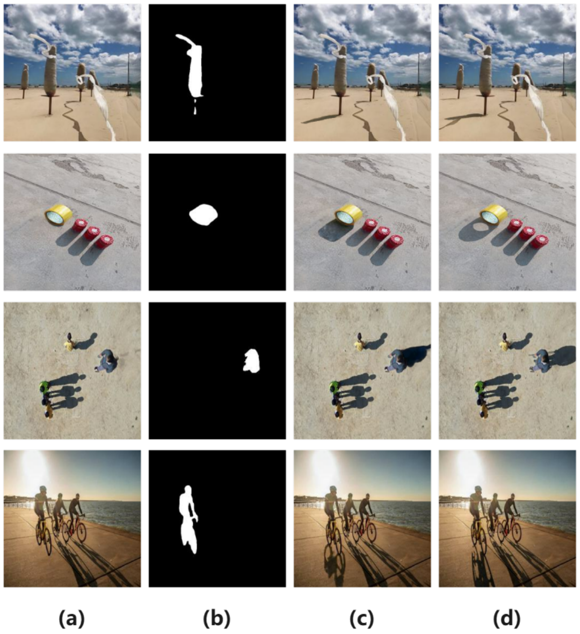

Figure 13. Visualization of failure cases produced by our SGDiffusion. From left to right are input composite image (a), foreground object mask (b), results of SGDiffusion (c), ground-truth (d).

shadows with more precise locations and shapes, while preserving the foreground identity.

8. Failure Cases

Our method can generally achieve satisfactory results. However, for some challenging cases, our method may fail to generate plausible shadows. As shown in Figure 13, for floating objects, our model often fails to generate their shadows (*e.g.*, the white ribbon in the first row). Additionally, our model sometimes cannot capture the internal structure of certain items (*e.g.*, the hollow center of the rolled-up rug is not reflected in the shadow). Moreover, our model struggles to accurately understand the object shapes from the bird view (*e.g.* row 3). Lastly, when the shadows are long and complex, our model may not produce satisfactory results (*e.g.*, row 4).

References

[1] 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. 9

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

[3] 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

[4] Alina Kuznetsova, Hassan Rom, Neil Gordon Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *IJCV*, 128:1956–1981, 2018. 9

[5] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *CVPR*, 2016. 9

[6] 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. 6, 7, 8

[7] 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. 1, 9

[8] Tianyu Wang, Xiaowei Hu, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection with a single-stage detector. *TPAMI*, 2022. 1

[9] 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, 9

[10] 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. 9

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