Supplementary for Shadow Generation Using Diffusion Model with Geometry Prior

Haonan Zhao[1], Qingyang Liu[1], Xinhao Tao[1], Li Niu[1,2]*, Guangtao Zhai[1]
[1] Shanghai Jiao Tong University [2] miguo.ai
[1]{2zz-n-24,narumimaria,taoxinhao,ustcnewly,zhaiguangtao}@sjtu.edu.cn

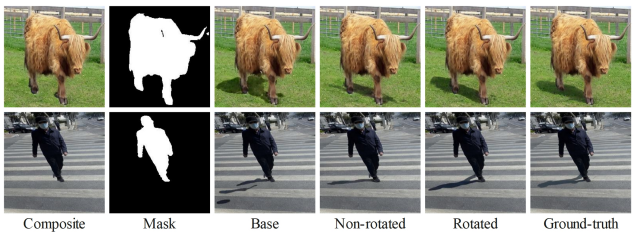

Figure 1. Visual results of the first group of ablation studies. From left to right, we show composite image, foreground object mask, the results of basic ControlNet, ControlNet with non-rotated bounding box prior, ControlNet with rotated bounding prior, and ground-truth.

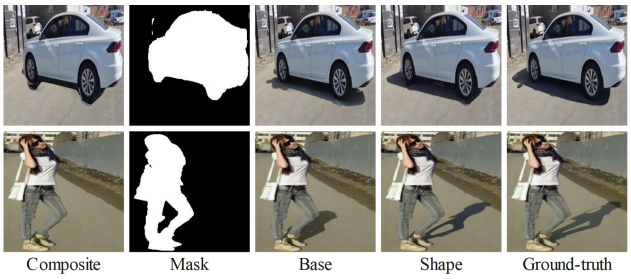

Figure 2. Visual results of the second group of ablation studies. From left to right, we show composite image, foreground object mask, the results of basic ControlNet, ControlNet with shadow shape prior, and ground-truth.

In this supplementary section, we first present visual results from the ablation studies discussed in the main paper in Section 1, which help to evaluate the contributions of different components of our model. Section 2 shows additional visual comparisons with baselines on DESOBAv2 dataset [8]. Furthermore, Section 3 presents the results on real composite images together with the B-T scores [1] from user study. Section 4 discusses shadow generation under different Lighting Directions and Viewpoints. Then, Section 5 shows the comparison with generative composition methods. Section 6 delves into the robustness of our model through multiple random generations. In Section 7, we introduce the details of our shape auto-encoder G_r and geometry encoder E_g. Section 8 presents results for different values of the hyper-parameter K and N. Section 9 supplements details on rotated bounding box regression. Finally, we discuss some limitations of our method in Section 10.

*Corresponding author.

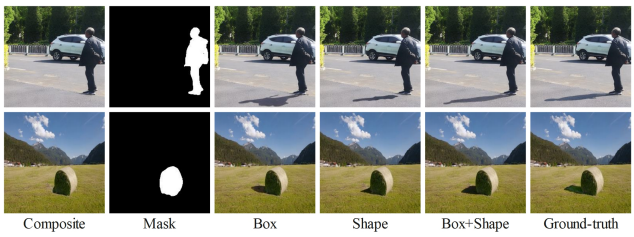

Figure 3. Visual results of the third group of ablation studies. From left to right, we show composite image, foreground object mask, the results of ControlNet with rotated bounding box prior, ControlNet with shadow shape prior, ControlNet with both priors, and ground-truth.

1. Visualization of Ablation Studies

In this section, we provide visual results of the ablation studies from Table 2 of the main paper. Specifically, the visualizations are divided into three groups: The first group corresponds to rows 1-3 of Table 2, exploring the impact of different types of bounding box regression on shadow quality. The second group, corresponding to rows 1 and 4 of Table 2, investigates the effect of incorporating shadow shape embeddings classification on the results. The third group, which corresponds to rows 3-5 of Table 2, discusses how the combination of these techniques can complement each other to improve the quality of generated shadows.

The results of the first group are shown in Figure 1. It can be observed that using rotated bounding box prior pro-

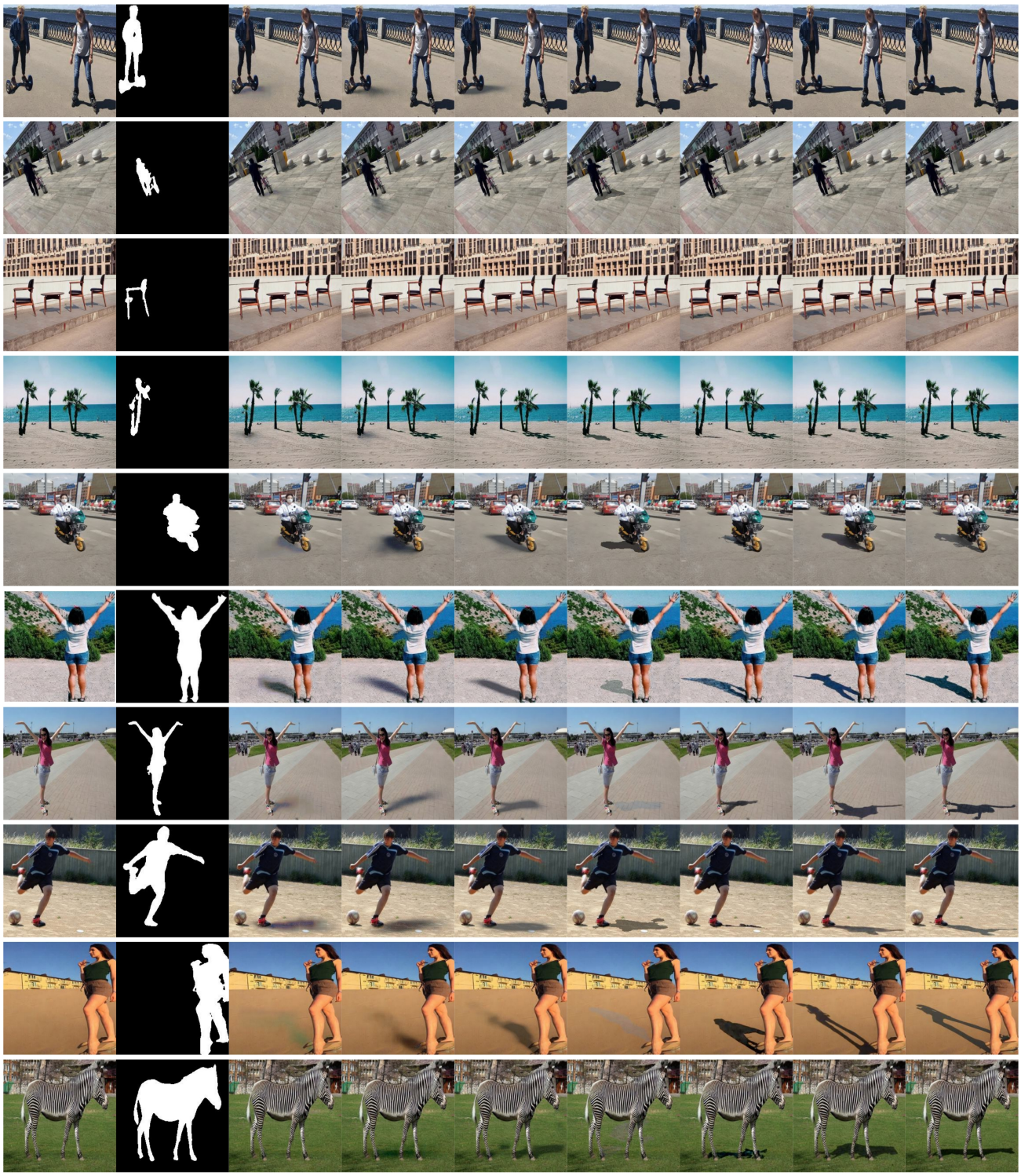

Figure 4. Visual comparison with state-of-the-art methods on DESOBAv2 dataset. From left to right, we show composite image, foreground object mask, results of ShadowGAN [18], AR-SG [7], SGRNet [4], DMASNet [13], SGDiffusion [8], our GPSDiffusion, and Ground-truth.

Figure 5. Visual comparison with state-of-the-art methods on real composite images. From left to right, we show composite image, foreground object mask, results of ShadowGAN [18], AR-SG [7], SGRNet [4], DMASNet [13], SGDiffusion [8], our GPSDiffusion, and Ground-truth.

vides the most accurate shadow positioning. Additionally, as shown in the second row, the model equipped with rotated bounding box prior generates shadows with more pronounced angles. In contrast, shadows generated with standard non-rotated bounding box prior tend to be parallel to the stripes, demonstrating the effectiveness of the rotated bounding box prior.

The results of the second group are illustrated in Figure 2. When using only the basic ControlNet, the generated results are generally reasonable and capture the general appearance of the shadows. However, these results often lack detailed shadow shapes, leading to less precise representations of complex shadow structures. In contrast, incorporating shadow shape prior significantly enhances the model's ability to predict more accurate and detailed shadow shapes. This improvement highlights the effectiveness of integrating matched shape embeddings into the ControlNet framework, as it provides more precise shadow representations, aligning more closely with the ground-truth. This demonstrates that the matched shape embeddings can help capture finer shadow details and improve overall shadow quality.

The results of the third group are presented in Figure 3. While bounding box prior helps the model predict shadow regions with reasonable accuracy, the generated shadows can be somewhat rough. Conversely, the shadow shape prior improves the model's ability to capture shadow shapes, although it can cause some positional shifts. The combination of two priors enhances both the precision of shadow placement and the quality of shape details, demonstrating that jointly using two priors provide more comprehensive geometry priors for the model.

2. More Visualization Results on DESOBAv2

Figure 4 shows more qualitative comparison results with other methods on DESOBAv2 [8]. For shadow generation in various scenes, our method demonstrates significant improvements in geometry details (*e.g.*, location, scale, shape), highlighting the effectiveness of using geometry priors in our method.

3. More Results on Real Composite Images

Figure 5 shows more qualitative comparison results with other methods on real composite images from [4]. It can be seen that our method generates the most reasonable and realistic shadows. Due to the lack of ground-truth, we conduct user study to quantitatively compare all methods on real composite images: ShadowGAN [18], Mask-SG [5], AR-SG [7], SGRNet [4], DMASNet [13] and SGDiffusion [8].

Specifically, for each real image, we generate results

Figure 6. Visual comparison with state-of-the-art methods with different lighting directions and viewpoints. From left to right, we show composite image, foreground object mask, results of ShadowGAN [18], AR-SG [7], SGRNet [4], DMASNet [13], SGDiffusion [8], our GPSDiffusion, and Ground-truth.

Method	B-T score ↑
ShadowGAN [18]	-1.061
Mask-SG [5]	-1.712
AR-SG [7]	-0.653
SGRNet [4]	-0.109
DMASNet [13]	0.353
SGDiffusion [8]	0.652
Ours	2.529

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

from all 7 methods and create 21 pairs by randomly selecting two results. We evaluated 100 real images, resulting in a total of 2100 pairs. 50 users are invited to assess the pairs, with each user selecting the more realistic shadow for the foreground object in each pair. This leads to 105,000 pairwise results in total. We then calculate the scores for each method based on the Bradley-Terry (B-T) model [1] using the obtained pairwise results. The results reported in Table 1 show that our method achieves the highest score, indicating that our method generates the most reasonable shadows and aligns best with human visual perception.

4. Experiments with Different Lighting and Viewing Directions

Figure 6 shows the visual comparisons between our GPSDiffusion and baselines under different lighting and viewing directions. Our method can generate realistic shadows when the viewing direction is perpendicular to the lighting direction (row 1) or parallel with the lighting direction (row 3). Even for the unusual bird-view (rows 2, 4, and 5), our method can still generate visually realistic shadows, while other methods struggle to produce satisfactory shadows, or may not generate any shadows at all.

5. Comparison with Generative Image Composition Methods

Recently, due to the increasingly popularity of foundation diffusion model, generative image composition [2, 6, 9–12, 16] has attracted more and more attention. This task aims to naturally insert given foreground objects into background images, which has certain overlap with shadow generation task. While these methods may incidentally gener-

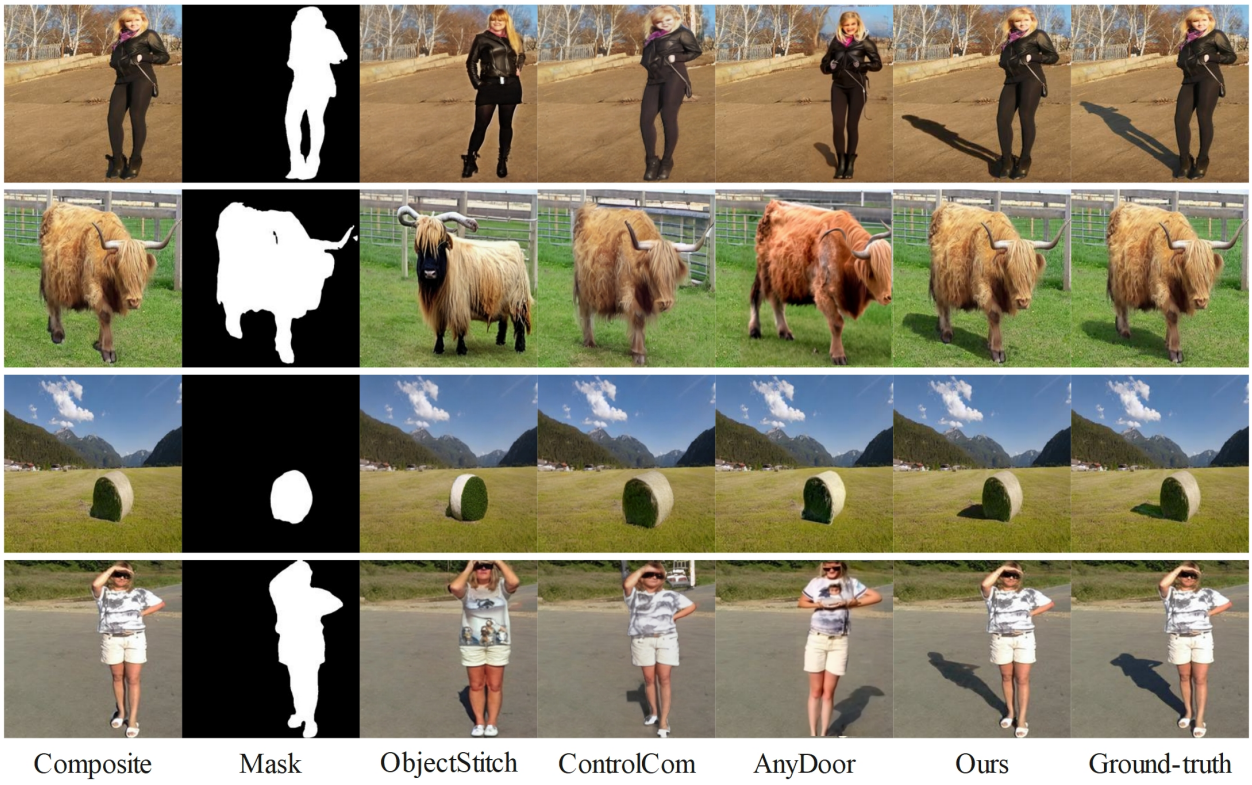

Figure 7. Visual comparison with generative composition methods on DESOBAv2 dataset. From left to right, we show composite image, foreground object mask, results of ObjectStitch [11], ControlCom [17], AnyDoor [2], our GPSDiffusion, and Ground-truth.

ate shadows with simple shapes for the inserted foreground objects without specific training. However, the quality of generated shadow is very low and the model has no ability to generate complicated shadows. Moreover, the details of foreground objects may be changed unexpectedly.

We choose open-sourced ObjectStitch [11], ControlCom [17], and AnyDoor [2] as baselines for comparison. Although ObjectDrop [15] and TOB [14] discuss the shadow issues, but they did not release code or model. For these baselines, we take the bounding box enclosing the composite foreground as the reference bounding box and the cropped composite foreground as the reference object, utilizing their released models.

The results of different methods are shown in Figure 7, revealing that the shadows generated by these methods are incomplete due to bounding box constraints or incompatible with the geometry of foreground object. For those complicated shadows, the baselines perform significantly worse than our method. Besides, for the baseline methods, the detailed information cannot be well preserved. For example, the human faces are distorted and the clothes patterns are changed in the results of [2, 11, 17]. Therefore, our method has clear advantage over these methods for shadow generation.

6. Multiple Results of Our Method

Considering the stochastic property of diffusion model, we display five randomly generated results using our method in Figure 8.

Particularly, in the cases where the input image exhibits prominent lighting direction (row 1, 2), which can be inferred based on background object-shadow pairs, the model adeptly captures the lighting information and generate shadows in the proper directions. The consistency in shadow directions across multiple results demonstrates the model's ability to handle lighting variations effectively. In the cases without explicit lighting information or background object-shadow pairs, the model estimates a reasonable range of possible lighting directions and generates multiple plausible shadows (rows 3, 4). It is worth noting that this range is learnt from the prior knowledge in the training set, thus the generated shadows are very likely to be located near the ground-truth shadows.

7. Details of Network Architecture

Shape auto-encoder Our shape auto-encoder consists of 4 encoder layers and 4 decoder layers, making a total of 8 layers. The encoder includes the follow-

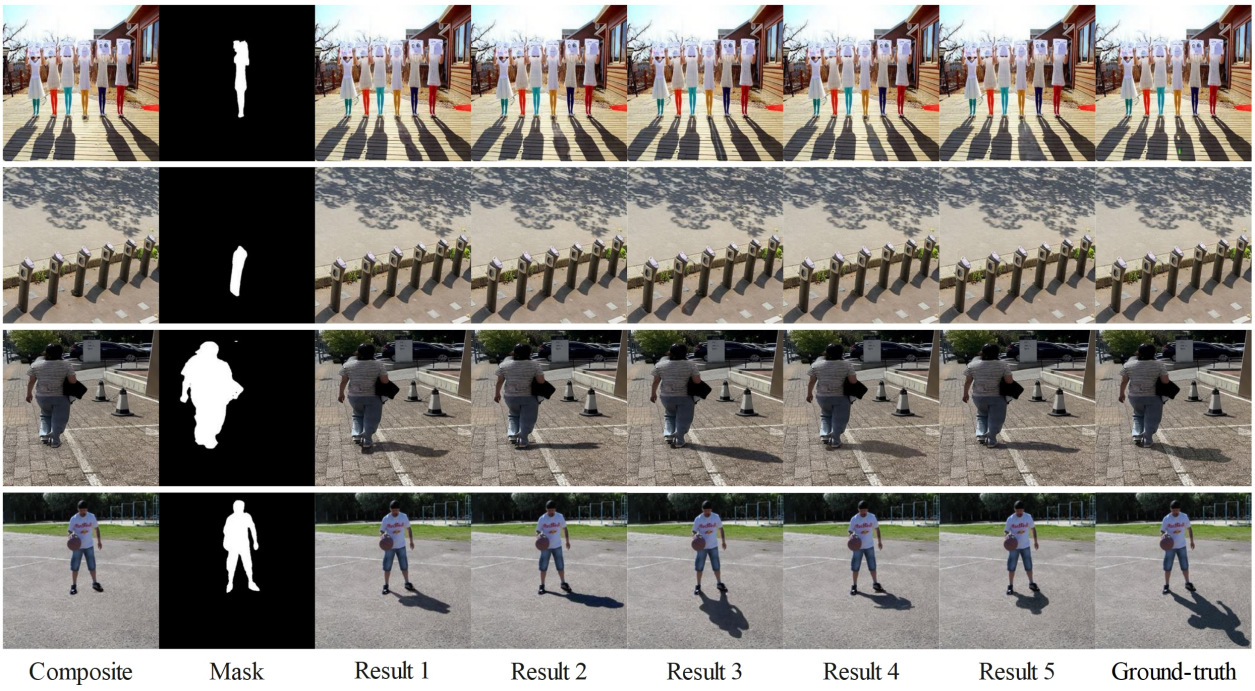

Figure 8. Visual results of our method using different random seeds. From left to right, we show composite image, foreground object mask, five results randomly generated by our method, and ground-truth.

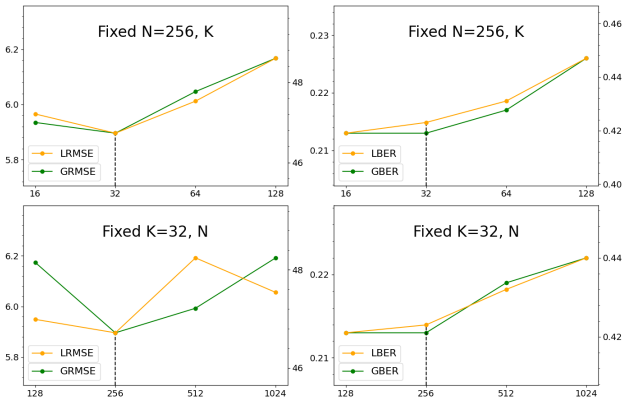

Figure 9. The hyper-parameter analyses of matched shape embedding number K and cluster number N on BOS test images from DESOBAv2. The default values are indicated by dashed vertical lines.

ing layers: $conv(1, 16)$, $conv(16, 32)$, $conv(32, 64)$ and $conv(64, 128)$, where each $conv(c1, c2)$ represents a 3×3 Convolution-ReLU layer with input and output channels $c1$ and $c2$, a stride of 2, and padding of 1. The decoder includes: $deconv(128, 64)$, $deconv(64, 32)$, $deconv(32, 16)$ and $deconv(16, 1)$, where each $deconv(c1, c2)$ represents a 3×3 TransposeConvolution-ReLU layer with input and output channels $c1$ and $c2$, a stride of 2, padding of 1, and output padding of 1. The final layer of the decoder uses a Sigmoid activation function instead of ReLU.

Geometry encoder Our geometry encoder E_g is based on ResNet-34 [3] and features two specialized heads: the box head and the shape head. The box head is designed for rotated bounding box regression and consists of three 3×3 convolution layers, each followed by Instance Normalization and ReLU activation functions. After these operations, the output is subjected to adaptive average pooling, which is then processed by a linear layer that generates five parameters corresponding to the bounding box. The shape head is used for shadow shape embeddings classification, and modifies the final fully connected layer of the ResNet-34 to output class scores based on the number of classes.

8. Hyper-parameter Analyses

Recall that we group all shadow shape embeddings into N clusters and retrieve top-K matched shape embeddings to be used for diffusion model. By default, we set $N = 256$ and $K = 32$. We explore the effect of using different N and K based on four metrics: Global RMSE (GR), Local RMSE (LR), Global BER (GB), and Local BER (LB). We vary K in the range of [16, 32, 64, 128] and N in the range of [128, 256, 512, 1024] respectively, while keeping the other one fixed as the default value. Figure 9 presents the

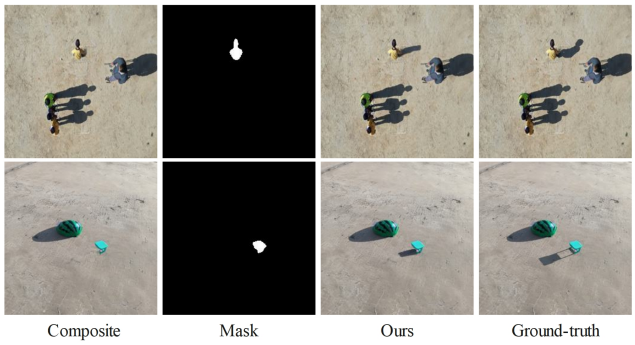

Figure 10. Visual results of failure cases produced by our GPSDiffusion. From left to right, we show composite image, foreground object mask, our GPSDiffusion, and Ground-truth.

results of these experiments, which shows that our method is relatively robust and generally outperforms the baselines when setting the hyper-parameters in a reasonable range.

9. Details of Rotated Bounding Box Regression

We obtain the ground-truth rotated bounding boxes of foreground objects and shadows using the built-in functions in $OpenCV$. Specifically, we merge the effective regions of the foreground object mask and shadow mask, respectively. Then, we calculate the minimum bounding rotated rectangle using $cv2.minAreaRect$ to obtain the five parameters of (x, y, w, h, θ) of the bounding box. This process yields B_o and B_s as described in Section 3.1 in the main paper, allowing us to supervise the rotated bounding box regression as in Eqn. (1) in the main paper.

10. Failure Cases

Our method generally produces reasonable and realistic shadows, but the results can be less satisfactory in certain complex scenes. Figure 10 shows some examples where the results fall short. For instance, in row 1, our model struggles to generate accurate shadows for non-upright human poses captured from top-down viewpoint. Additionally, as shown in row 2, for the objects that are suspended, if the camera viewpoint does not sufficiently convey the suspension information, our model fails to learn the internal structure and generates shadows like those of solid objects.

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. 1, 4

[2] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2024. 4, 5

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[4] Yan Hong, Li Niu, and Jianfu Zhang. Shadow generation for composite image in real-world scenes. In *Proceedings of the AAAI conference on artificial intelligence*, pages 914–922, 2022. 2, 3, 4

[5] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2472–2481, 2019. 3, 4

[6] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17089–17099, 2023. 4

[7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8139–8148, 2020. 2, 3, 4

[8] Qingyang Liu, Junqi You, Jianting Wang, Xinhao Tao, Bo Zhang, and Li Niu. Shadow generation for composite image using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2024. 1, 2, 3, 4

[9] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 4

[10] Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. Collage diffusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4208–4217, 2024.

[11] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 5

[12] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8048–8058, 2024. 4

[13] Xinhao Tao, Junyan Cao, Yan Hong, and Li Niu. Shadow generation with decomposed mask prediction and attentive shadow filling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5198–5206, 2024. 2, 3, 4

[14] Gemma Canet Tarrés, Zhe Lin, Zhifei Zhang, Jianming Zhang, Yizhi Song, Dan Ruta, Andrew Gilbert, John Col-

lomosse, and Soo Ye Kim. Thinking outside the bbox: Unconstrained generative object compositing. *arXiv preprint arXiv:2409.04559*, 2024. 5

[15] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. *arXiv preprint arXiv:2403.18818*, 2024. 5

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 4

[17] 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. 5

[18] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5:105–115, 2019. 2, 3, 4