

Supplementary Materials: Foreground Harmonization and Shadow Generation for Composite Image

Anonymous Author(s)

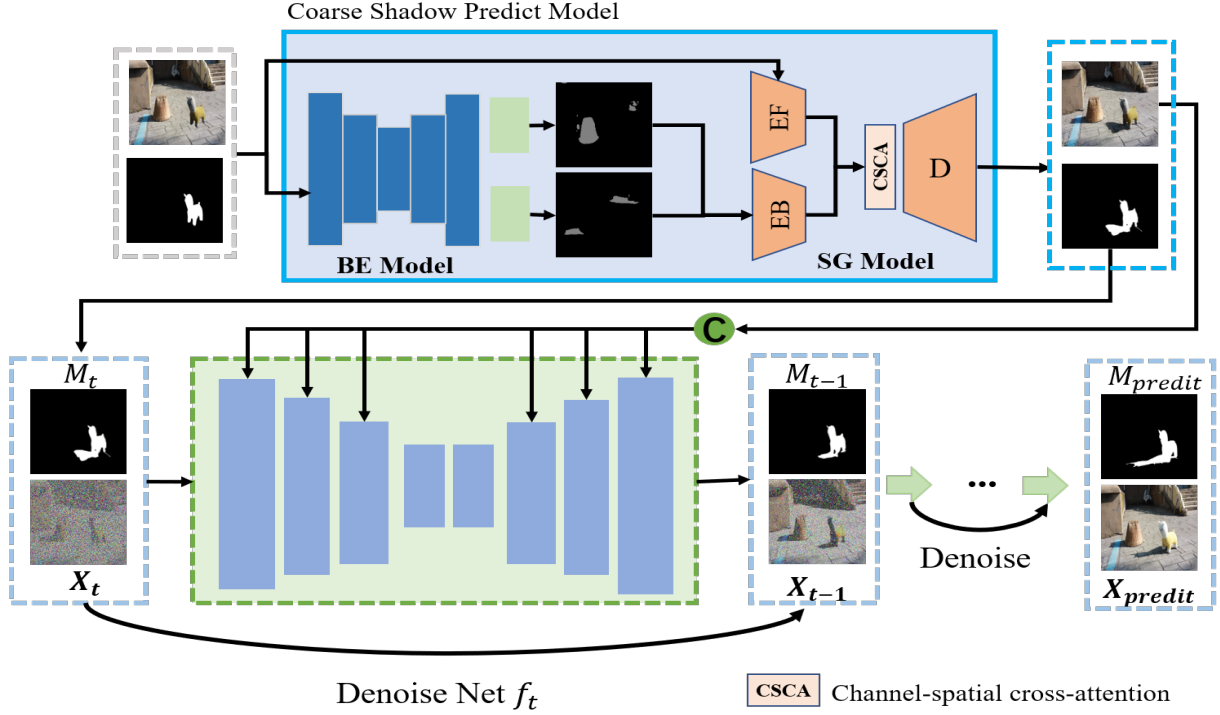


Figure 1: Pipeline of the proposed method. IH-SG Diffusion model includes coarse shadow prediction network (SP) and denoise network f_t . Given a disharmonious image, our model can generate a harmonious image with controllable foreground objects and reasonable cast shadows.

In this supplementary document, we first provide a detailed structure of our IH-SGDiffusion network to ensure better understanding and reproducibility. And in Figure 1, we have made some corrections to the rough shadow prediction module, as the original foreground extraction module EF was mistakenly marked as EB . In addition, we will provide more visual comparison results of different methods on IH-SG dataset. Our code, trained models, and dataset will be released upon the acceptance of our manuscript.

1 NETWORK STRUCTURE

The detailed structure of our IH-SG diffusion network is shown in Table 1. Symbols of the operators are listed as follows:

- $\text{Conv}(\text{cin}, \text{cout}, k, s, p)$: a convolution operation with cin input channels, cout output channels, kernel size of k , stride size of s , and padding p .
- $\text{ConvT}(\text{cin}, \text{cout}, k, s, p)$: a transposed convolution operation with cin input channels, cout output channels, kernel size of k , stride size of s , and padding p .
- BN: Batch Normalization.

- ReLU: Rectified Linear Unit, a widely used activation function in neural networks.
- Resblock: a Residual Block (Resblock), proposed by He et al. [3], extracts feature maps. Each Residual Block consists of consecutive operations including convolution (Conv), batch normalization (BN), and Leaky ReLU activation function.
- Concat: a concatenation operation in the channel dimension.

Overall, the Background Extraction module can be seen as a U-shaped network with two shared encoders, so it is not listed repeatedly. For the shadow generation module, it consists of two identical encoders and a decoder, as well as a channel space cross attention mechanism.

2 QUALITATIVE ANALYSIS

We have demonstrated more methods and the light and shadow editing results of IH-SG Diffusion. As shown in Figure 2. Compared with other methods, our results are more reasonable, harmonious, and visually closer to real images.

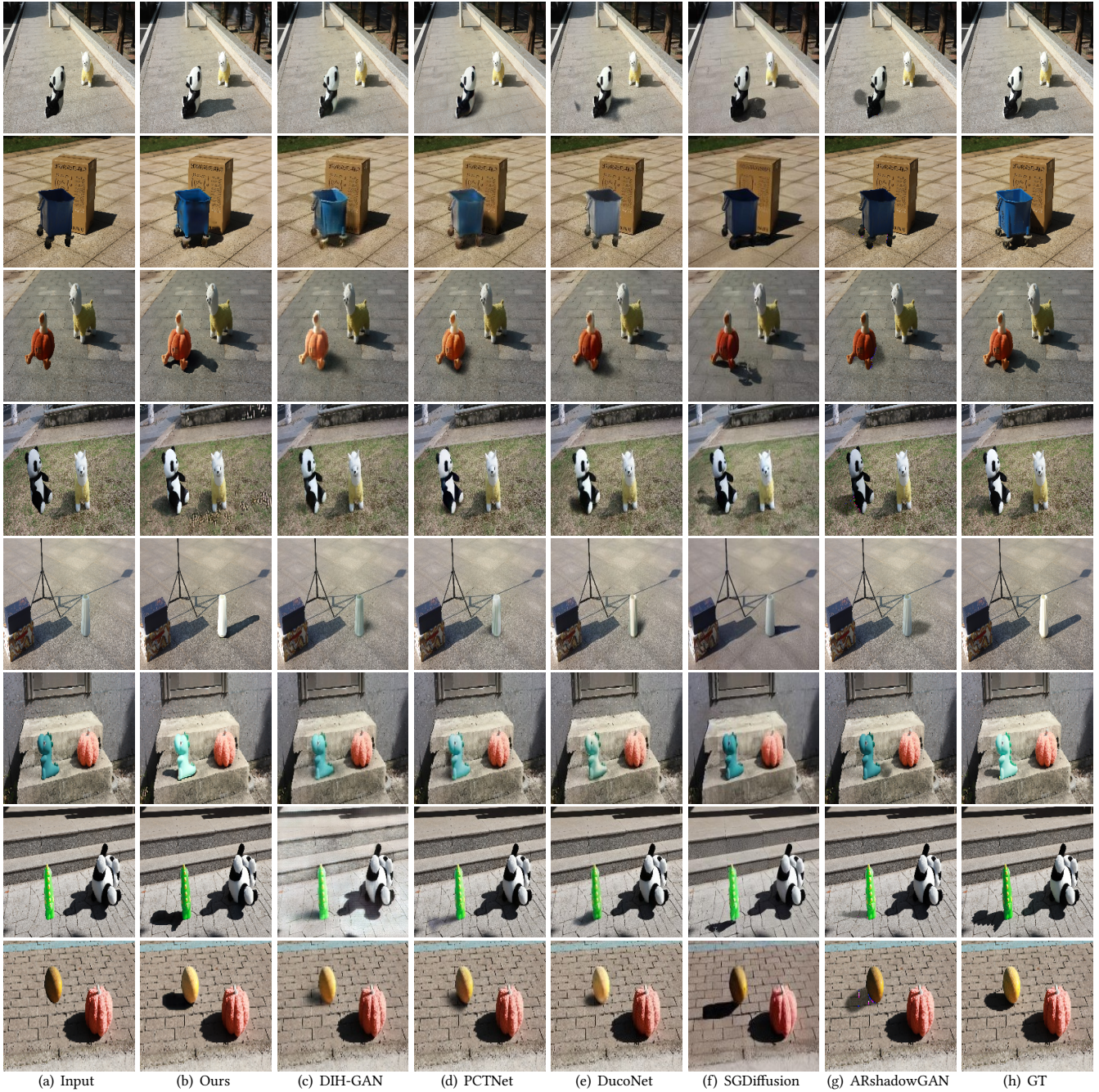


Figure 2: Three testing cases of different methods on IH-SG dataset. From left to right are composite images, the results of our results, DIH-GAN [1], PCTNet [2], DucoNet [6] and the SGDiffusion [5], ARshadowGAN [4] and ground truth, respectively.

REFERENCES

- [1] Zhongyun Bao, Chengjiang Long, Gang Fu, Daquan Liu, Yanzhen Li, Jiaming Wu, and Chunxia Xiao. 2022. Deep Image-based Illumination Harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18542–18551.
- [2] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. 2023. PCT-Net: Full Resolution Image Harmonization Using Pixel-Wise Color Transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5917–5926.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [4] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. 2020. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8139–8148.

Table 2: The structure of shadow generation module.

Layer name(s)	
1	Conv(3, 64, 3, 1, 1) + ReLU
D1-1	Conv(64, 64, 3, 1, 1) + ReLU
D1-2	Conv(64, 64, 3, 1, 1) + ReLU + maxpool
D2-1	Conv(64, 128, 3, 1, 1) + ReLU
D2-2	Conv(128, 128, 3, 1, 1) + ReLU + maxpool
D3-1	Conv(128, 256, 3, 1, 1) + ReLU
D3-2	Conv(256, 256, 3, 1, 1) + ReLU + maxpool
D4-1	Conv(256, 512, 3, 1, 1) + ReLU
D4-2	Conv(512, 512, 3, 1, 1) + ReLU + maxpool
M1	Conv(512, 1024, 3, 1, 1)
M2	Conv(1024, 1024, 3, 1, 1)
U4-1	ConvT(1024, 512, 3, 2, 1)
C1	Concat(U4-1, D4-2)
U3-1	Conv(1024, 512, 3, 1, 1) + ReLU
U3-2	Conv(512, 512, 3, 1, 1) + ReLU
U3-3	ConvT(512, 256, 3, 2, 1)
C2	Concat(U3-3, D3-2)
U2-1	Conv(512, 256, 3, 2, 1) + ReLU
U2-2	Conv(256, 256, 3, 1, 1) + ReLU
U2-3	ConvT(256, 128, 3, 2, 1)
C3	Concat(U2-3, D2-2)
U1-1	Conv(256, 128, 3, 2, 1) + ReLU
U1-2	Conv(128, 128, 3, 1, 1) + ReLU
U1-3	ConvT(128, 64, 3, 2, 1)
C4	Concat(U1-3, D1-2)
U0-1	Conv(128, 64, 3, 1, 1) + ReLU
U0-2	Conv(64, 64, 3, 1, 1) + ReLU
out	Conv(64, 1, 3, 1, 1)

Table 1: The structure of background extraction module.

Layer name(s)	
1	Conv(4, 64, 7, 2, 3) + BN + ReLU + Avgpool
D1	Resblock + Avgpool
D2	Resblock + Avgpool
D3	Resblock + Avgpool
D4	Resblock + Avgpool
M1	Conv(512, 1024, 3, 1, 1)
C1	ConvT(1024, 512, 3, 2, 1)
U1	Concat(D4, C1) + Conv + BN + ReLU
C2	ConvT(512, 256, 3, 2, 1)
U2	Concat(D3, C2) + Conv + BN + ReLU
C3	ConvT(256, 128, 3, 2, 1)
U3	Concat(D2, C3) + Conv + BN + ReLU
C4	ConvT(128, 64, 3, 2, 1)
U4	Concat(D1, C4) + Conv + BN + ReLU
out	Conv(64, 2, 3, 1, 1)

[5] 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).

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