SoftShadow: Leveraging Soft Masks for Penumbra-Aware Shadow Removal

Supplementary Material

In this supplementary material, we include more implementation details on the proposed SoftShadow model (Section 8), mask sensitivity test for SOTA methods (Section 9), soft and hard shadow image examples on well-used datasets (Section 10), and more visual results of our SoftShadow on the SRD dataset (Section 11).

8. Implementation details

Due to the computational cost, there are three training stages. We first finetune SAM [19] for 100 epochs with a fixed learning rate of 1e-4, and we set $\lambda_1 = 0.1$. Next, we use the soft masks predicted by the tuned SAM to finetune ShadowDiffusion [9] for another 500 epochs because the diffusion model [12] requires more epochs to reach convergence. The learning rate is set to 1e-5. Finally, we train SAM and ShadowDiffusion jointly for 500 epochs, with $\lambda_1 = 0.1$, $\lambda_2 = 1$, and the learning rate is set to 1e-5. We use Adam [18] optimizer for all training stages. We experiment with the hyperparameters λ_1 and λ_2 within the range of 0.01 to 1. To find the optimal learning rate, we tried setting the learning rate to values between 1e-4 and 1e-6. Empirically, the models work best when the learning rate is 1e-5 while a smaller or larger learning rate could lead to underfitting or model collapse.

For the input size, SAM requires a 1024×1024 RGB image as the input, so we resize our image from 640×840 to 1024×1024. The output mask size from SAM is 256×256. In ShadowDiffusion, we train with image and mask size set to 256×256. We train the model with 2 GPUs and the total batchsize is set to 16 for all three stages. When finetuning SAM, we accumulate the gradient for 4 steps and the total batchsize for each step is 4. Compared to SAM Adapter [1], our implementation requires much less video memory to train the entire model, making our method more efficient and hardware-friendly. SAM Adapter requires around 60GB of video memory when batchsize is 4, which can only be met by GPUs such as A100 80GB, and H100. In contrast, our method requires 37GB of video memory and can be run on GPUs such as A100 40GB. In practice, we train the model on two A100 40GB and it is totally feasible to train the model on RTX series GPUs.

9. Mask Sensitivity Evaluation

In Section 5, we explore the mask sensitivity task within the ISTD+ dataset. The mask comparison is shown in Figure 10. Since the ISTD+ dataset includes ground truth masks, it is not feasible for our methods to outperform the latest state-of-the-art (SOTA) approaches that rely on these ground truth masks during inference. However, we achieve comparable results without utilizing external masks, demonstrating the robustness of our approach.

Input Masks	ShadowDiffusion	HomoFormer	Ours
Otsu mask	35.02	26.26	**35.57**
DHAN mask	34.73	<u>35.37</u>	
FDRNet mask	33.39	21.01	
Pretrained SAM	31.30	18.22	
Mean	33.61	25.48	N/A
Std Dev	1.469	6.519	N/A

Table 6. The stability test on SRD dataset. The values in the table are PSNR results. The best result is **boldfaced**. The second best result is <u>underlined</u>, respectively.

We also evaluate the sensitivity on the SRD dataset, which does not have include ground truth masks provided. Most methods rely on the DHAN mask during inference. To provide a comprehensive evaluation, we test masks detected by DHAN [3], Otsu [14], FDRNet [38] and the pretrained SAM [19]. Note that the detector FDRNet is trained on the ISTD [31] datasets. We do not distinguish between the ISTD and ISTD+ datasets here because the difference in illumination does not affect the position of the shadows. We use the same pretrain model trained on the ISTD dataset to generate the mask on the SRD dataset. So the mask quality decreases because of the domain gap. The results are shown in Table 6, the values shown are PSNR. Since our method does not require mask inputs, our results remain consistent across all tests. Our method outperforms all others. Although HomoFormer achieves comparable results with DHAN masks [3], it struggles to remove shadows with other mask inputs. The PSNR results for HomoFormer drop considerably when using alternative masks, despite showing strong performance with DHAN masks. This discrepancy might be due to HomoFormer being slightly overfitted to the style of DHAN masks, which include many small regions in the shadow masks, as illustrated in Figure 9.

The mask we used is shown in the Figure 10. From the mask, because Otsu use shadow-free image information, its mask is better than other methods. Compared with other mask detecters, we extract the more accurate mask.

In sum, the sensitivity test suggests that previous methods, e.g., HomoFormer, can be sensitive to the quality of the input mask. This can be a problem for end users as variations in the input mask could lead to drastically different results while our method does not require external masks, making the usage more straightforward and robust against user error.

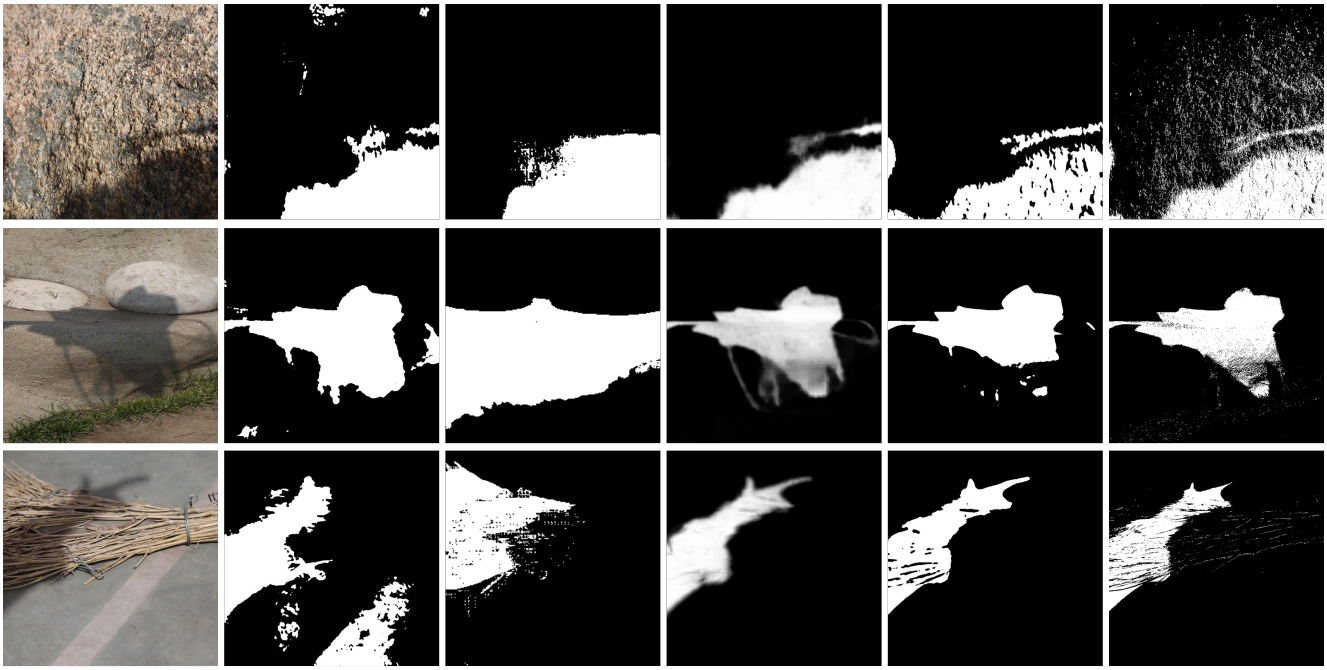

Figure 9. Illustrate the soft shadow image examples in the SRD dataset. The red boxes highlight the boundaries and regions of soft shadows, showcasing their blurry boundaries and gradual transitions in the penumbra area.

10. Shadow Image Examples

In this section, we present examples of soft shadows from the SRD [27] and LRSS [7] datasets, as well as hard shadow examples from the ISTD+ 11, the SRD dataset contains numerous soft shadow scenarios, the image varies in color, texture, and illumination. These soft shadows can be categorized into two types: blurry boundaries of large object shadows and small soft shadow regions. Shown in the first row and second row of Figure 11, respectively. In Figure 12, the boundary of shadows is more blurry than it is in the SRD dataset. The LRSS dataset has limited diversity in background colors and textures. Figure 13 shows that the shadows in the ISTD+ dataset are more hard shadows, which have sharp boundaries compared with shadow images on SRD or LRSS datasets. The ISTD+ dataset also offers less diversity in image scenarios compared to the SRD dataset.

11. More Visual Results

In this section, we present more soft shadow examples on SRD datasets. Figure 14 shows visual comparisons with other SOTA methods that require external masks on the SRD dataset, the masks they used are DHAN masks [3]. We can observe that our SoftShadow has better removal results on the soft boundaries of large object shadows and in small shadow regions. We show more visual comparison in Figure 15 with other end-to-end methods that do not need shadow mask inputs. We can observe that our Softshadow removes shadow more precisely in shadow areas.

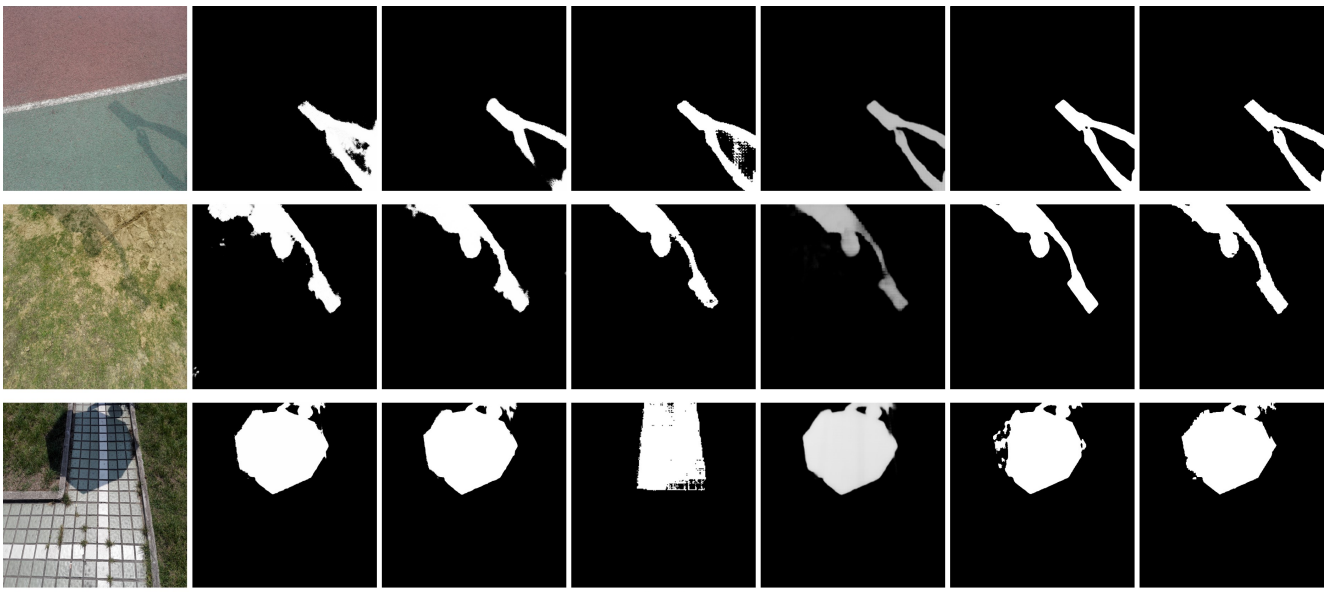

Figure 10. Examples of shadow masks in the ISTD+ dataset. Our mask is an intermediate soft shadow mask represented as a grayscale image, while all other masks are binary.

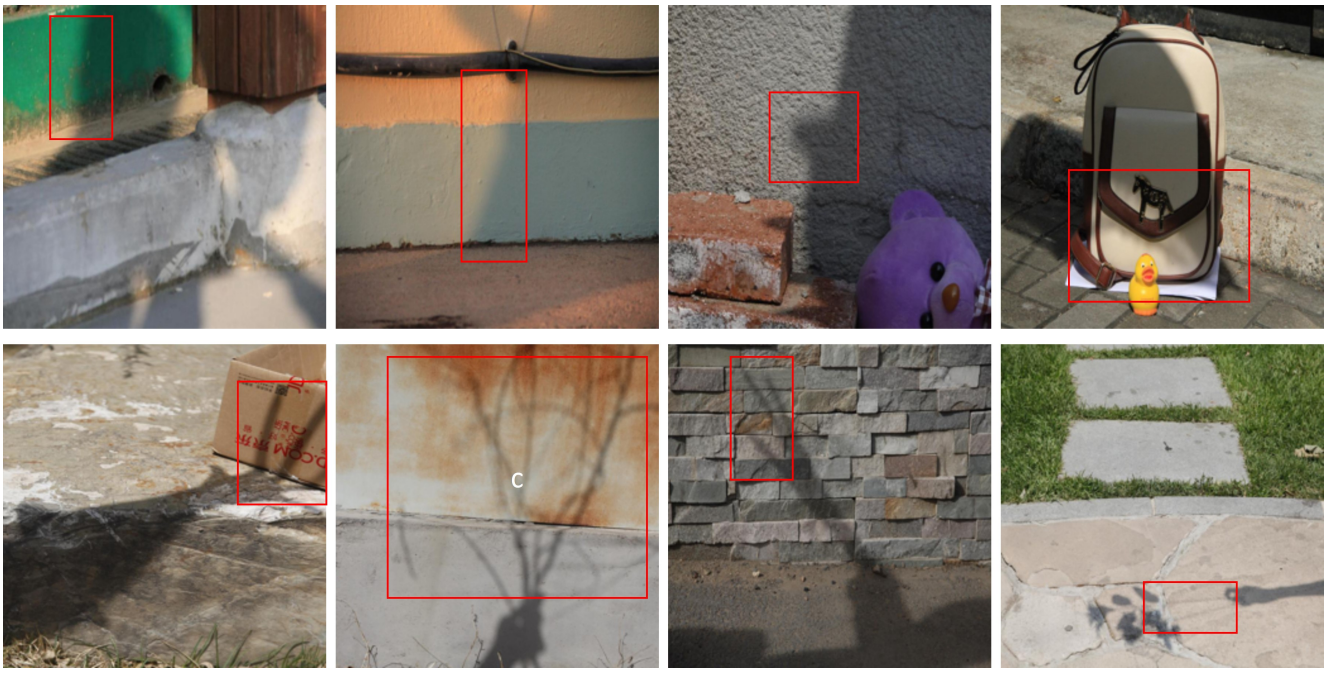

Figure 11. Illustrate the soft shadow image examples in the SRD dataset. The red boxes highlight the boundaries and regions of soft shadows, showcasing their blurry boundaries and gradual transitions in the penumbra area.

Figure 12. Illustrate the soft shadow image examples in LRSS dataset.

Figure 13. Illustrate the hard shadow image examples in ISTD+ dataset.

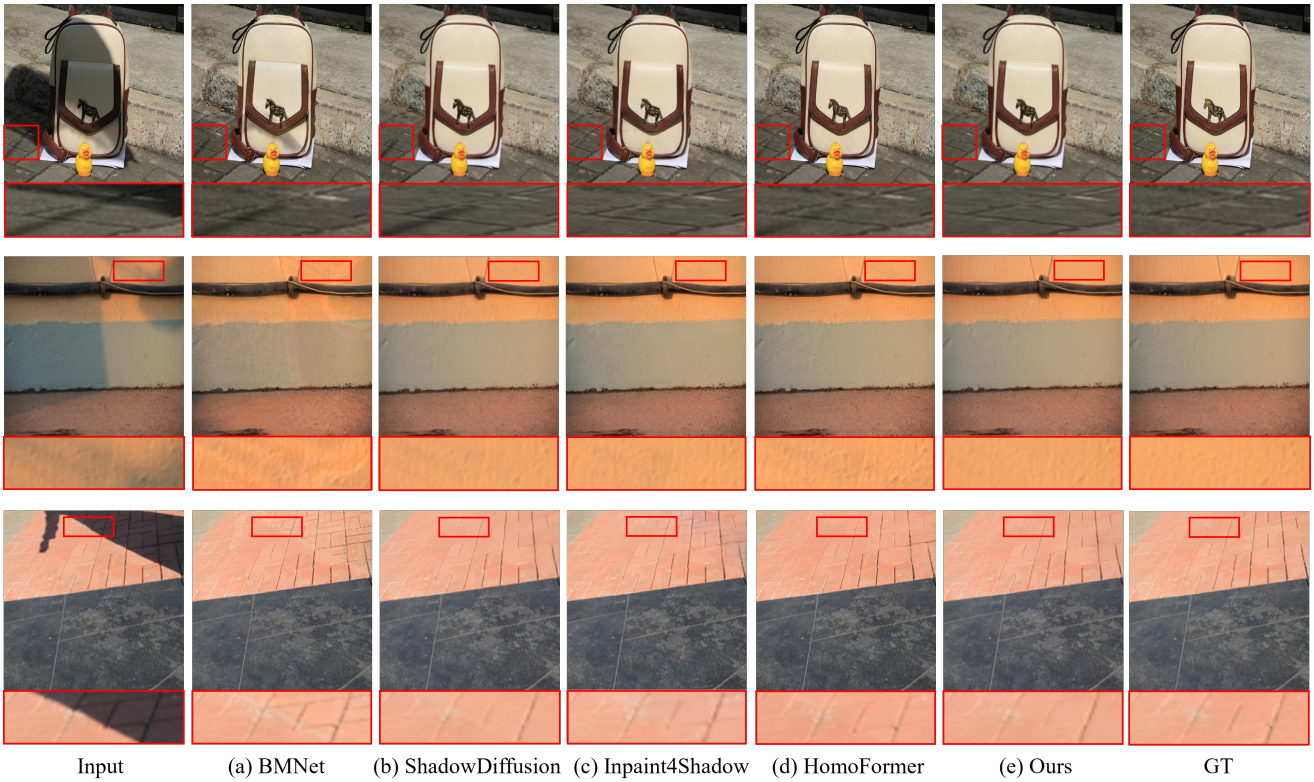

Figure 14. Examples of soft shadow image removal results on the SRD dataset [27]. The input shadow image, the estimated results of (a) BMNet [39], (b) ShadowDiffusion [9], Inpaint4Shadow [22], (d) Homoformer [34], and (e) Ours, as well as the ground truth image, respectively.

Figure 15. Examples of shadow removal results on SRD datasets [27]. The input shadow image, the estimated results of (a) DC-ShadowNet [16], (b) DeS3 [17], and (c) Ours, as well as the ground truth image, respectively.