Supplementary Material
Boundary-Aware Divide and Conquer: A Diffusion-based Solution for Unsupervised Shadow Removal

Lanqing Guo[1], Chong Wang[1], Wenhan Yang[2], Yufei Wang[1], Bihan Wen[1*]
[1]Nanyang Technological University, Singapore [2]Peng Cheng Laboratory, China
{lanqing001, wang1711, yufei001, bihan.wen}@ntu.edu.sg, yangwh@pcl.ac.cn

In this supplementary material, we include more implementation details of the architecture of the proposed Shadow-Invariant Intrinsic Decomposition (SIID) model and the shadow simulation process (Section A), more ablation studies to verify the effects of each loss function employed in SIID (Section B), more cross-dataset evaluations to verify the generalizability of our proposed method (Section C), as well as more visual comparisons on ISTD [15, 9], and the video shadow removal [10] datasets (Section D).

A. Implementation Details

A.1. Architecture of SIID

Our proposed Shadow-Invariant Intrinsic Decomposition (SIID) model is built based on one recent transformer-based image restoration model [16] to exploit the multi-scale non-local similarity. SIID consists of one encoder $\mathcal{D}$ and two decoders $\mathcal{D}_r$ and $\mathcal{D}_l$ for reconstructing reflectance and illumination, respectively. Figure A illustrates the detailed architecture and Figure B shows the network components of *LeWin Block* and its corresponding *LeFF*.

A.2. Details of Shadow Simulation

We select a non-learning based shadow synthetic method [7] without any external dataset, which is built on a physically-grounded shadow illumination model that can synthesize a shadow image given an arbitrary combination of a shadow-free image, a shadow mask, and shadow attenuation parameters. Given a shadow-free image $y \in \mathbb{R}^{H\times W\times 3}$, we first obtain an image $x^{dark} \in \mathbb{R}^{H\times W\times 3}$, where all the pixels are darkened with the same attenuation property, according to the illumination affine model as follow:

$$y_{ijk} = w_k + b x^{dark}_{ijk} \Longleftrightarrow x^{dark}_{ijk} = \frac{1}{b} x^{ns}_{ijk} - \frac{w_k}{b} , \tag{A}$$

where i and j indicate indices for row and column location indexes, respectively. $k = 0, 1$ and 2 indicates the k-th color channel (red, green, and blue).

The final synthesized shadow image x is corrupted in some regions, can be obtained by composing shadow-free image y and darkened image x^{dark}, by alpha composition using the shadow mask $m \in \mathbb{R}^{H\times W}$ as the alpha factor:

$$x_{ijk} = (1 - m_{ij}) y_{ijk} + m_{ij} x^{dark}_{ijk} . \tag{B}$$

Then the attenuation parameters w_k and b can be sampled according to [7] to simulate plausible and diverse shadows.

B. More Ablation Studies

The effects of losses for training SIID. We thoroughly investigate the impact of each loss function applied in the training of the proposed shadow-invariant intrinsic decomposition (SIID). Table A shows the evaluation results on different combinations of loss functions. Our complete model achieves the best performance among those invariants.

*Corresponding author: Bihan Wen.

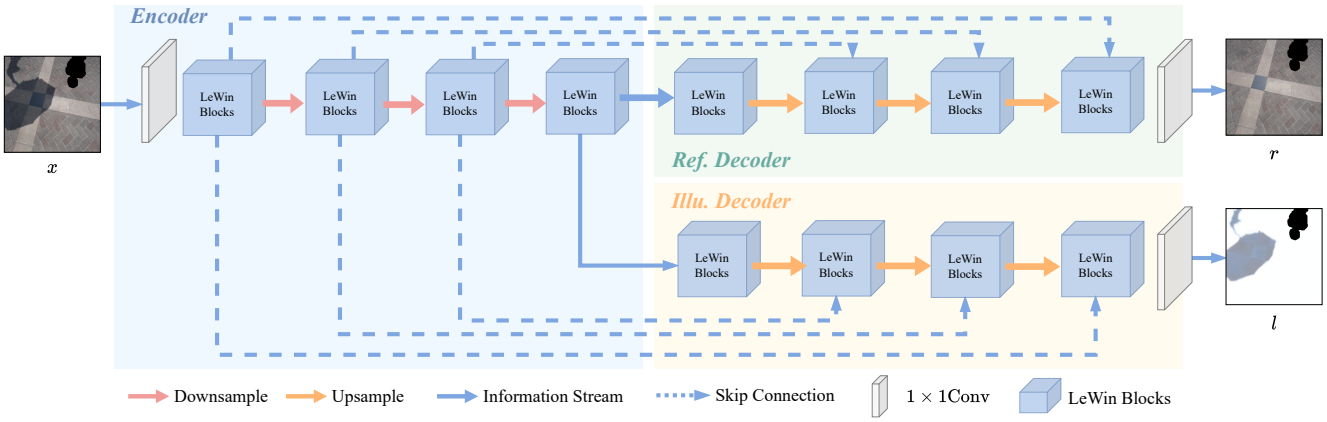
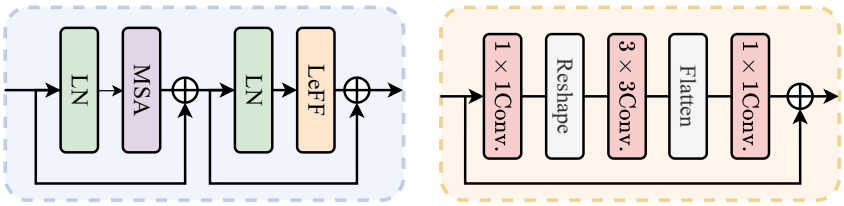


Figure A: Detailed architecture of the proposed Shadow-Invariant Intrinsic Decomposition (SIID) model, which consists of one encoder $\mathcal{D}$ and two decoders $\mathcal{D}_r$ and $\mathcal{D}_l$ for reconstructing reflectance r and illumination l, respectively.



(a) LeWin Block (b) LeFF

Figure B: Detailed architecture of the proposed Shadow-Invariant Intrinsic Decomposition (SIID) model, which consists of one encoder $\mathcal{D}$ and two decoders $\mathcal{D}_r$ and $\mathcal{D}_l$ for reconstructing reflectance and illumination, respectively.

Table A: Ablation study to verify the effectiveness of each loss for training shadow-invariant intrinsic decomposition (SIID) model over the ISTD dataset.

	$\mathcal{L}_{recon}$	$\mathcal{L}_{cons}$	$\mathcal{L}_{smo}$	$\mathcal{L}_{boun}$	Shadow RMSE↓	Shadow PSNR↑	Shadow SSIM↑	All RMSE↓	All PSNR↑	All SSIM↑
①		✓			8.1	35.45	0.983	3.6	32.22	0.955
②	✓	✓			8.0	35.89	0.984	3.6	32.33	0.956
③	✓	✓	✓		7.6	**36.22**	0.979	3.5	32.60	0.958
④	✓	✓	✓	✓	**7.6**	35.91	**0.986**	**3.3**	**32.73**	**0.962**

C. More Evaluations

Cross-dataset evaluation. Supervised learning methods [11, 8] learn the pixel-wise mapping between the shadow images and ground-truth shadow-free ones in a fully-supervised manner, which easily results in overfitting the training dataset with poor generalization. While certain supervised methods may exhibit strong performance in cases where the training and testing scenes are highly similar, if there exist substantial disparities between the distributions of the training and testing sets, performance may suffer significantly. To further explore the generalizability of our method, we conduct a cross-dataset evaluation on ISTD dataset [15, 9].

In detail, we select the recent reproducible supervised learning method, *i.e.*, Fu *et al.* [2], and we directly apply the pre-trained model of supervised methods to the ISTD dataset, where that pre-trained model is trained on the SRD dataset [13] with different scenes and styles from the ISTD dataset. On the contrary, the unsupervised and classic methods do not require paired training data, which can be adapted to any testing set. Table B summarizes the quantitative results on the ISTD dataset. We find an obvious drop for the supervised learning method when the testing distribution is different from the training one.

Method	Setting	Shadow Region (S)			Non-Shadow Region (NS)			All Image (ALL)		
		PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
Yang *et al.* [17]	Classic	21.57	0.878	23.2	22.25	0.782	14.2	20.26	0.706	15.9
Gong *et al.* [3]		30.53	0.972	13.0	36.63	0.982	2.6	28.96	0.943	4.3
Guo *et al.* [4]		26.89	0.960	20.1	35.48	0.975	3.1	25.51	0.924	6.1
DHAN [1]	Supervised	32.92	0.988	9.6	27.15	0.971	7.4	25.66	0.956	7.8
SP+M-Net [9]		**37.60**	**0.990**	**6.3**	**36.02**	0.976	**2.9**	**32.94**	**0.962**	**3.5**
Fu *et al.* [2]		36.04	0.978	6.7	31.16	0.892	3.8	29.45	0.861	4.2
Fu *et al.* [2]*		32.67	0.969	12.0	29.99	0.876	4.5	27.14	0.832	5.8
SG-ShadowNet [14]		36.80	0.990	6.5	35.57	**0.978**	**2.9**	32.46	0.962	**3.5**
MaskShadow-GAN [6]	Unsupervised	32.19	0.984	10.8	33.44	0.974	3.8	28.81	0.946	4.8
LG-ShadowNet [11]		32.44	0.982	9.9	33.68	0.971	3.4	29.20	0.945	4.4
DC-ShadowNet [8]		31.06	0.976	12.2	27.03	0.961	6.8	25.03	0.926	7.8
Le *et al.* [10]		33.09	0.983	10.4	35.26	0.977	2.9	30.12	0.950	4.0
G2R [12]		33.58	0.979	8.9	35.52	0.976	2.9	30.52	0.944	3.9
Ours (w/ detected mask)		35.71	**0.986**	**7.6**	36.39	0.981	2.7	32.11	0.959	3.5
Ours (w/ GT mask)		**35.91**	**0.986**	**7.6**	**37.27**	**0.984**	**2.4**	**32.73**	**0.962**	**3.3**

Table B: Quantitative comparison results of the proposed method with the state-of-the-art methods on ISTD [15, 9] dataset. The best performances for supervised learning and unsupervised learning methods are highlighted in **Bold**. * indicates the pre-trained model is trained using other datasets instead of ISTD, *e.g.*, SRD dataset.

D. More Visual Examples

Figure C, Figure D, and Figure E illustrate some visual results on ISTD [15, 9]. Our method can effectively suppress the boundary artifacts and achieve better visual results, especially compared to those unsupervised methods. Besides, to verify the effectiveness of our method on video shadow removal, we also provide some visual examples on the video shadow dataset [10] as shown in Figure F. For some extremely dark cases, our method can restore clearer structures compared to the competing methods as shown in Figure F.

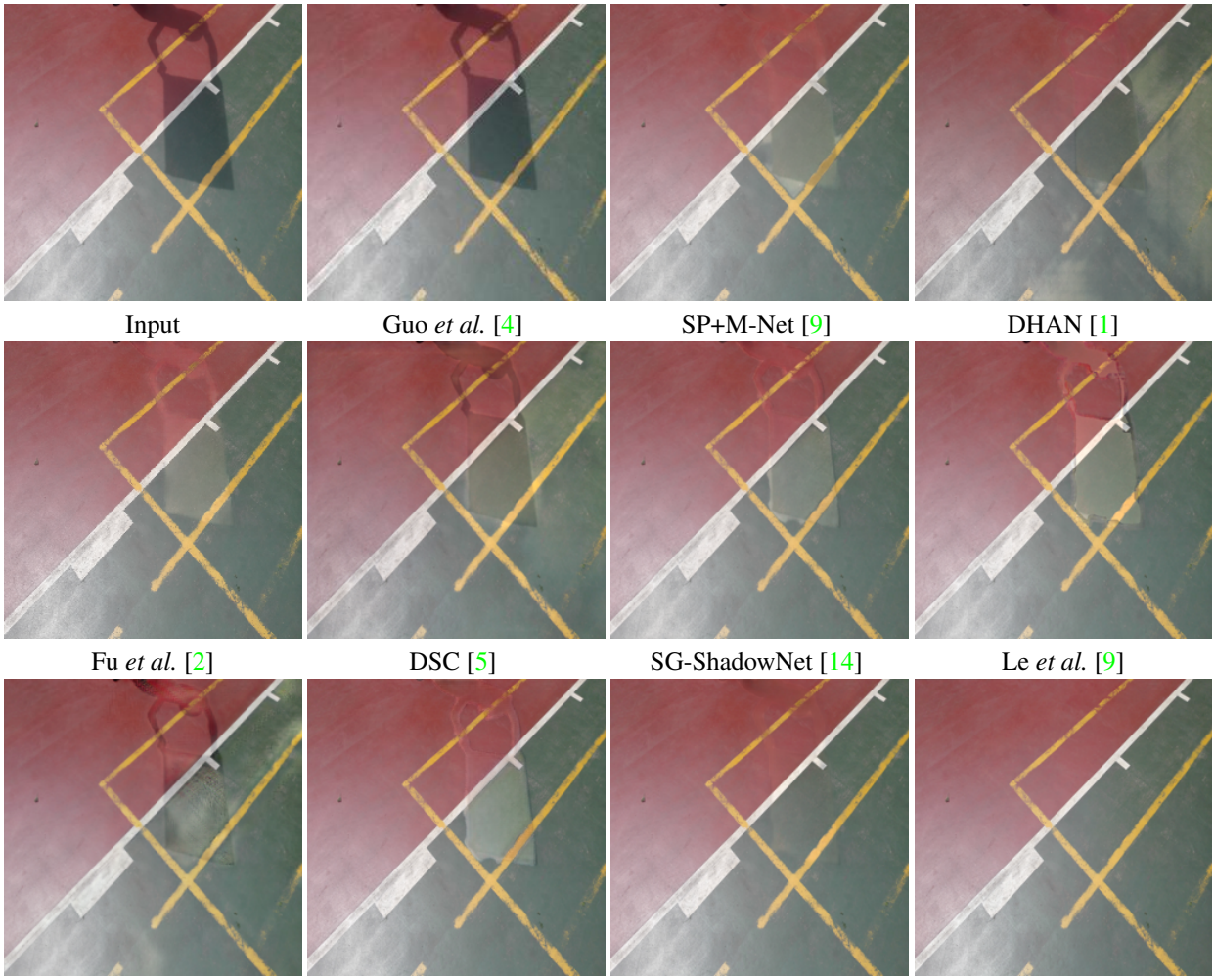


Figure C: One example of shadow removal results on the ISTD [15, 9] dataset. The input shadow image, the estimated results of classic methods: Guo *et al.*[4]; supervised learning methods: SP+M-Net [9], DHAN [1], Fu *et al.* [2], DSC [5], SG-ShadowNet [14]; unsupervised learning methods: Le *et al.*[10], DC-ShadowNet [8], G2R [12], Our proposed method; and the ground truth, respectively. Please zoom in to see the details.

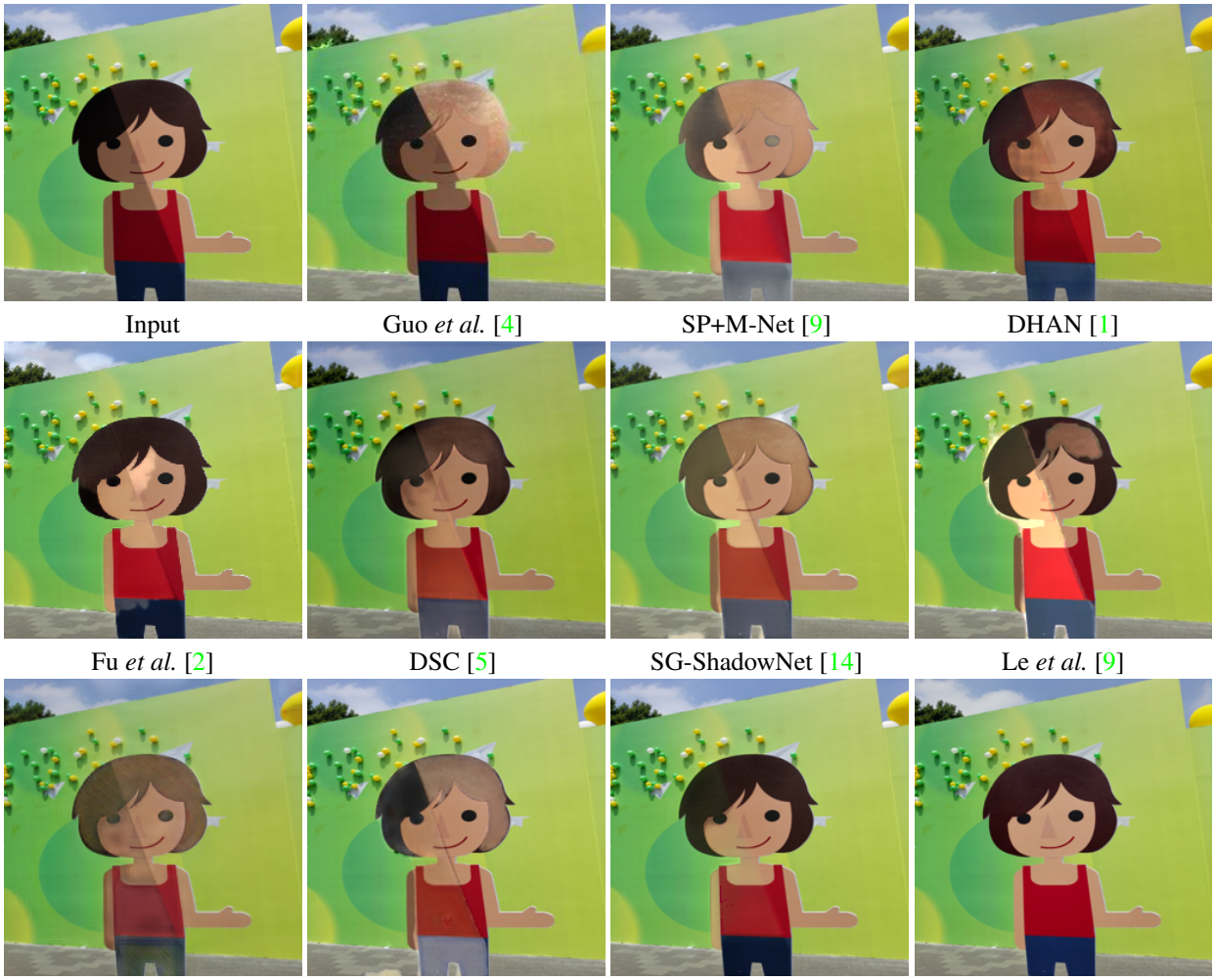


Figure D: One example of shadow removal results on the ISTD [15, 9] dataset. The input shadow image, the estimated results of classic methods: Guo *et al.*[4]; supervised learning methods: SP+M-Net [9], DHAN [1], Fu *et al.* [2], DSC [5], SG-ShadowNet [14]; unsupervised learning methods: Le *et al.*[10], DC-ShadowNet [8], G2R [12], Our proposed method; and the ground truth, respectively. Please zoom in to see the details.

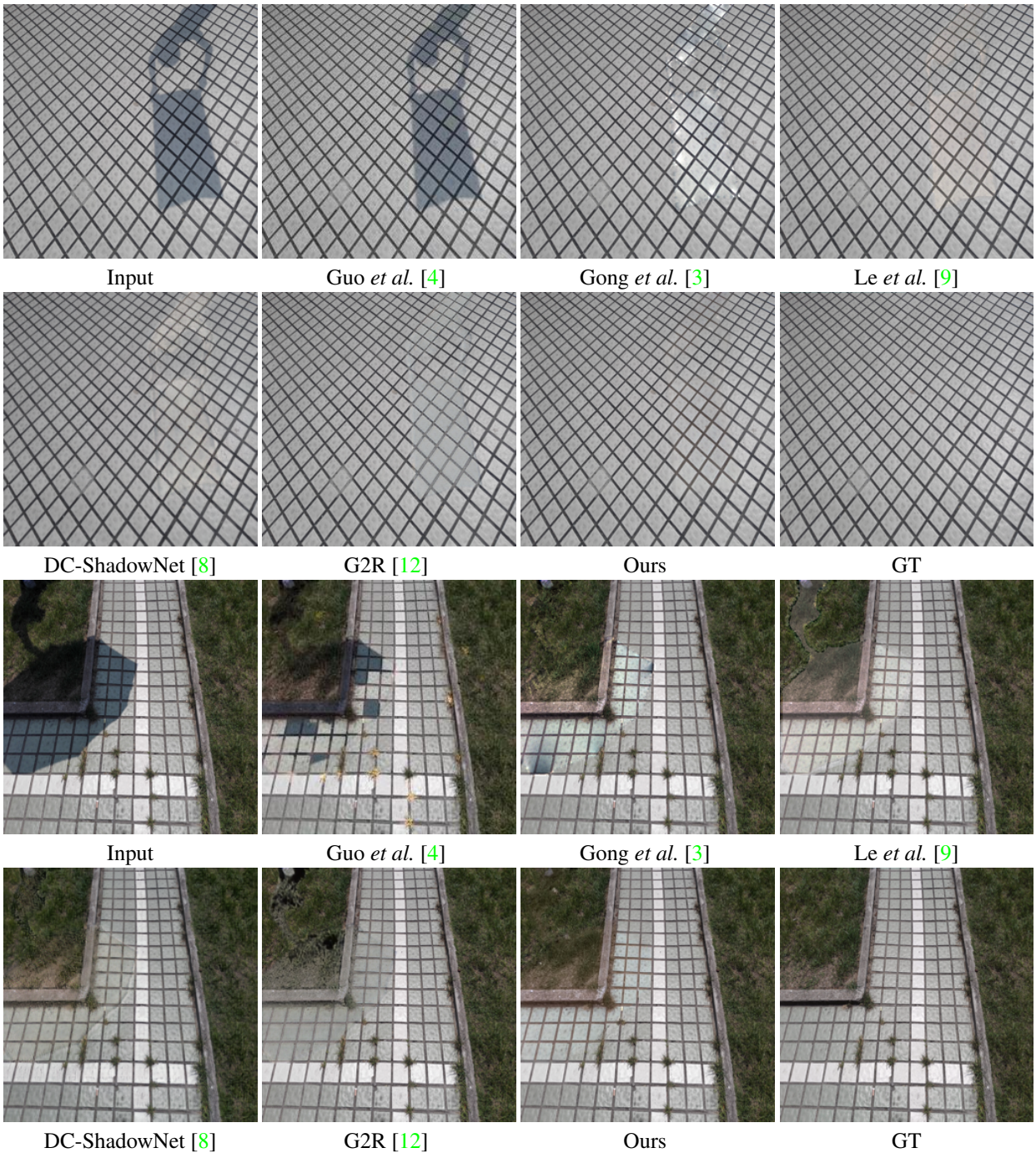


Figure E: Two examples of shadow removal results on the ISTD [15, 9] dataset. The input shadow image, the estimated results of classic methods: Guo *et al.* [4], Gong *et al.* [3]; unsupervised learning methods: Le *et al.* [9], DC-ShadowNet [8], G2R [12], our proposed method; and the ground truth, respectively. Please zoom in to see the details.

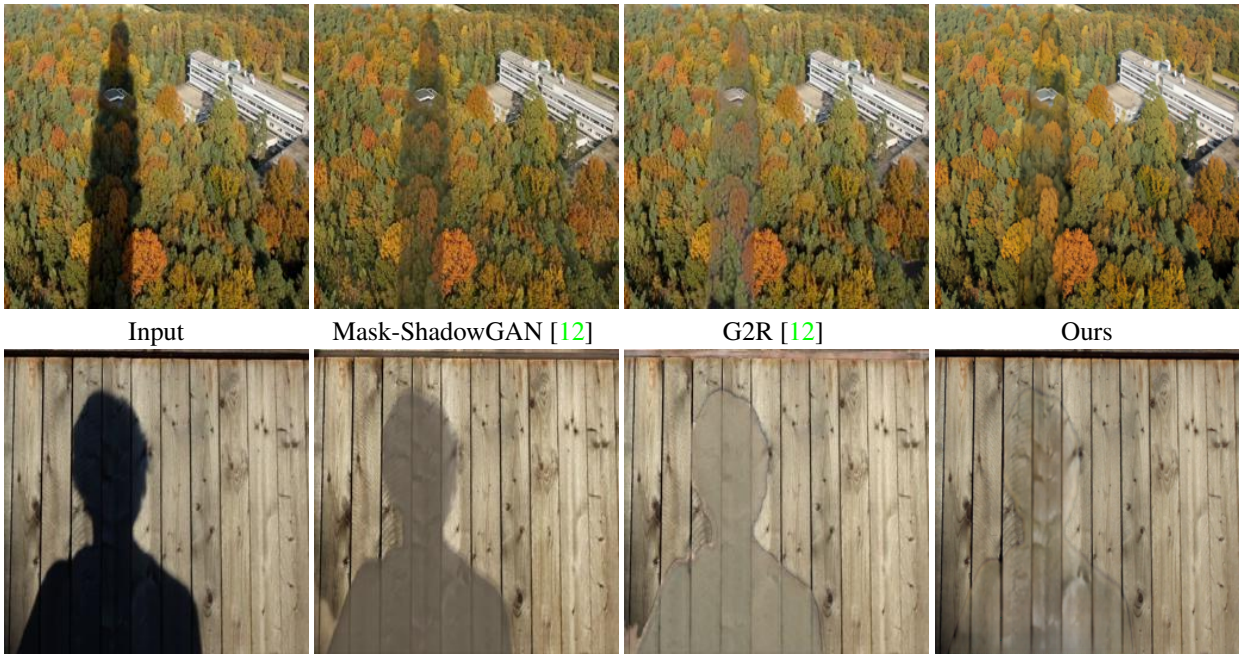


Figure F: Two examples of shadow removal results on the video shadow removal dataset [10]. The input shadow image, the estimated results of unsupervised learning methods: Mask-ShadowGAN [6], G2R [12], and our proposed method, respectively. Please zoom in to see the details.

References

[1] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *AAAI*, pages 10680–10687, 2020. 3, 4, 5

[2] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *CVPR*, pages 10571–10580, 2021. 2, 3, 4, 5

[3] Han Gong and Darren Cosker. Interactive removal and ground truth for difficult shadow scenes. *JOSA A*, 33(9):1798–1811, 2016. 3, 6

[4] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. 35(12):2956–2967, 2012. 3, 4, 5, 6

[5] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. 42(11):2795–2808, 2020. 4, 5

[6] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *ICCV*, pages 2472–2481, 2019. 3, 7

[7] Naoto Inoue and Toshihiko Yamasaki. Learning from synthetic shadows for shadow detection and removal. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4187–4197, 2020. 1

[8] Yeying Jin, Aashish Sharma, and Robby T Tan. Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5027–5036, 2021. 2, 3, 4, 5, 6

[9] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *ICCV*, pages 8578–8587, 2019. 1, 2, 3, 4, 5, 6

[10] Hieu Le and Dimitris Samaras. From shadow segmentation to shadow removal. In *ECCV*, 2020. 1, 3, 4, 5, 7

[11] Zhihao Liu, Hui Yin, Yang Mi, Mengyang Pu, and Song Wang. Shadow removal by a lightness-guided network with training on unpaired data. *IEEE Transactions on Image Processing*, 30:1853–1865, 2021. 2, 3

[12] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *CVPR*, 2021. 3, 4, 5, 6, 7

[13] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*, pages 4067–4075, 2017. 2

[14] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 361–378. Springer, 2022. 3, 4, 5

[15] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, pages 1788–1797, 2018. 1, 2, 3, 4, 5, 6

[16] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 1

[17] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. 21(10):4361–4368, 2012. 3