Shadow Removal Refinement via Material-Consistent Shadow Edges Supplementary Material

Shilin Hu[1], Hieu Le[2], ShahRukh Athar[1,3], Sagnik Das[1], Dimitris Samaras[1]
[1]Stony Brook University [2]EPFL [3]Captions

In this supplementary material, we provide the following:

1. Details of material-consistent shadow edge extraction.
2. Correctness of the proposed CDD metric.
3. Using shadow edges from shadow detection models.
4. Further details of implementation.
5. Quantitative results on ISTD+ test set.
6. More results on the cross-dataset testing.
7. More qualitative results.

1. Details of Material-Consistent Shadow Edge Extraction

We select only material-consistent shadow edges (MC-Edges) and enforce color and texture consistency on both sides of these edges in the shadow-removed outputs. These constraints should not be enforced on shadow edges aligning with object boundaries, as both sides of those edges should exhibit different shadow-free textures and colors. To extract MC edges, we first use the provided $SamAutomaticMaskGenerator$ function from the Segment Anything Model (SAM) [5] to predict material masks, and then sample edge pixels and patches where the material masks intersect with the shadow mask. Details of the process are described in Algorithm 1. More visual examples of improved material mask segmentation by our fine-tuned SAM are shown in Fig. 1.

To demonstrate the effectiveness of the fine-tuned SAM in extracting material-consistent shadow edges, we compare its performance to vanilla SAM on different materials. Tab. 1 presents the percentage of edge pixels extracted by each SAM model across various materials in the ISTD+ dataset [6]. The results indicate that our proposed method significantly outperforms vanilla SAM, successfully extracting shadow edge pixels across different material types.

Algorithm 1 Material-Consistent Shadow Edge Extraction

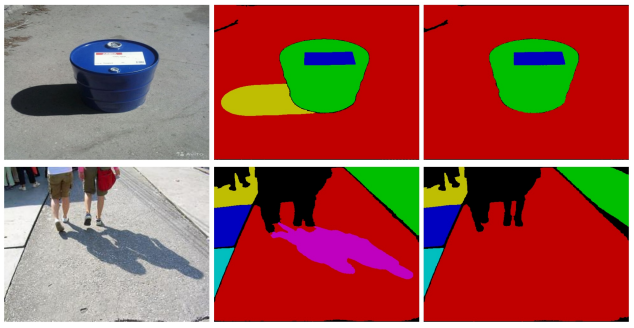
Data: Input shadow image I, shadow mask M; **Model:** Fine-tuned SAM f_{SAM}

Result: Sampled shadow/shadow-free pairs, $Pixel_{in/out}$ and $Patch_{in/out}$

EdgePixels $= \{M - erode(M)\} + \{dilate(M) - M\}$

SegMasks $= MaskGenerator(f_{SAM}, I, M)$

for $i = 1 \rightarrow n$ **do**
  if *SegMasks*$[i]$ *overlaps with* M **then**
    Get $Pixel_{in/out}$ in EdgePixels $\in$ SegMasks$[i]$
    Get $Patch_{in/out}$ in SegMasks$[i]$
  else
    continue



Input Vanilla SAM Fine-tuned SAM

Figure 1. Visual examples of improved material-consistent mask segmentation by our fine-tuned SAM.

2. Correctness of the Proposed CDD Metric

To justify the correctness of our proposed Color Distribution Difference (CDD) metric, we first show that the CDD value corresponds to the shadow intensity. As shown in the top row of Fig. 2, we choose a shadow image and manually adjust the shadow intensity. We find that the weaker the shadow effect, the lower the CDD value, indicating that the CDD metric can effectively represent the

Table 1. Percentage of material-consistent shadow edge pixels detected (recall%) via vanilla SAM and our fine-tuned SAM.

Material Type	Vanilla SAM	Fine-tuned SAM
grass	85.7	98.6
cement	63.6	96.0
ceramic	66.1	71.0
playground	75.6	82.8

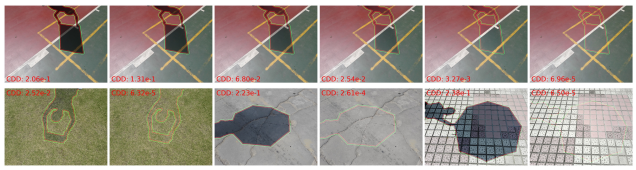


Figure 2. Correctness of the Color Distribution Difference metric. (*top*) We manually adjust the shadow intensity from strong to weak, and the CDD values are lower when the shadow effect is weaker. (*bottom*) We compare the CDD values of shadow images and their shadow-free counterparts, the CDD value of the shadow-free version is at least two orders of magnitude lower than that of the shadow version. CDD is computed using the pixels marked in the images and the values are reported in the images.

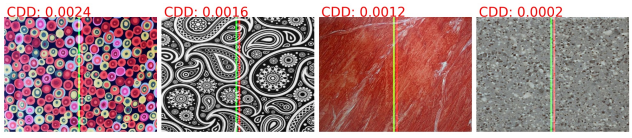


Figure 3. Validation of our proposed CDD metric on DTD [1]. We show examples from the subset and apply a simple method to compute the CDD values on the DTD data using parallel lines in the middle (see the red and $green$ lines in each image). We can see that this simple annotation still yields low CDD scores.

quality of shadow removal.

Additionally, we show the CDD results on ground truth images from ISTD+ [6]. We compare the CDD values of the shadow-free images against their shadow counterparts, as shown in the bottom row of Fig. 2. The CDD values for shadow-free images are at least two orders of magnitude lower than those for shadow images. Therefore, we believe CDD can serve as a valuable metric for evaluating shadow removal performance when ground truth images are unavailable.

Finally, to validate the adequacy of our proposed CDD metric in measuring consistency alongside shadow edges, we select a subset from the DTD dataset [1], which we believe showcases more complex textures than our proposed shadow image dataset (Fig. 3). We evaluate the CDD score on pixels from parallel lines in the middle of the image. This result can be considered the upper bound of the ground truth CDD score for our proposed dataset. The CDD measurement on this selected subset is 0.0022, significantly lower than the shadow removal results in the main paper (0.0157). This evaluation further demonstrates that the proposed CDD metric can effectively serve as a valuable shadow removal evaluation metric.

Table 2. Shadow detection results on the ISTD+ [6] and our proposed dataset using SILT [11]. Following [11], the performance is evaluated by the Balanced Error Rate (BER).

Dataset	BER	S	NS
ISTD+	1.12	0.80	1.44
Proposed	4.05	4.02	4.09

3. Shadow Removal Refinement using Shadow Edges from Shadow Detection Models

In our main paper, we conduct experiments using ground truth shadow masks sourced from established datasets [4, 6, 10]. These masks might not be available for ideal automated shadow removal systems used in real-world scenarios. In this section, we show that our method can be used with shadow masks detected from a shadow detection method. We note that shadow detection is a relatively easier task compared to shadow removal and a robust, scalable shadow detection system is more feasible since shadow detection training data is much easier to obtain. At some point, one can expect to obtain accurate shadow masks automatically, which could be directly incorporated into our system to improve shadow removal.

We use the state-of-the-art shadow detection method, SILT [11], to generate shadow masks for each testing image. Tab. 2 presents the detection performance on the two testing datasets using SILT. Then, we compare the performance of ShadowFormer [2] using ground truth shadow masks and these detected shadow masks in Tab. 3. We find that on both the ISTD+ dataset [6] and our proposed dataset, using detected masks results in worse shadow removal performance compared to using ground truth masks. A typical failure case is shown in the top row of Fig. 4. Nevertheless, applying our method atop ShadowFormer [2] can improve performance in both cases, as shown in Tab. 3. The bottom row of Fig. 4 visualizes an example of how our method improves the shadow-removal result.

4. Further Details of Implementation

4.1. Computation Overhead

Our approach iteratively refines the pre-trained model using extracted self-supervision. During testing on our proposed dataset, we update the entire model for 20 iterations per image, resulting in an average overhead of 24 seconds on an NVIDIA TITAN RTX GPU.

Table 3. Quantitative comparison of results using ground truth shadow masks and SILT-detected shadow masks on [2] and [2]+Ours. MAE and CDD values are reported. Note that CDD is reported as $1000\times$ the original value.

Methods	Proposed		ISTD+				
	CDD		MAE			CDD	
	Mean	Var	S	NS	A	Mean	Var
[2] w. GT mask	25.0	40.8	5.3	2.2	2.7	1.5	2.6
Ours w. GT mask	15.9	30.2	5.0	2.2	2.7	1.0	1.6
[2] w. detected mask	25.8	42.4	6.2	2.6	3.1	2.7	9.8
Ours w. detected mask	19.3	38.5	5.7	2.5	3.0	2.4	8.4

Input Detected mask [2] Ours

Figure 4. Examples of using SILT [11] detected masks for shadow removal refinement. The top row shows a failure case from ISTD+ [6], where the dark region is predicted as the shadow region, leading to an inaccurate shadow removal result that our refinement method cannot rectify. The bottom row shows a successful case from the proposed dataset, where an accurate shadow mask is predicted, and our refinement method improves the shadow removal performance.

We further investigate the refining performance with different numbers of iterations and different model update policies (*e.g.* updating only the last decoder layer of ShadowFormer [2]). The results are shown in Tab. 4. Breaking down the overhead, our shadow edge extraction process takes 2 seconds, and each model update iteration requires approximately 1 second. It is evident that the number of iterations is the primary cause of our computational overhead. Additionally, we found that partially updating the model does not improve the efficiency of our refinement and leads to decreased performance.

4.2. Number and Size of sampled patches

In the main paper, we sample 8 patches of size 16×16 in our final configuration. Here, we experiment with different numbers and sizes of patches and compare the shadow removal performance on our proposed dataset. As shown in Tab. 5, these hyperparameters do not significantly affect overall performance.

Table 4. Comparison of performance and computation overhead using different numbers of iterations and update policies. We report the CDD mean values on the full proposed test set.

# of Iter	Update	CDD Mean	Overhead (s)
10	whole	17.6	13
20	whole	15.7	24
	partial	19.1	24
30	whole	18.1	36

Table 5. Quantitative comparison of using different numbers and sizes of sampled patches in our final configuration. Note that CDD values are reported as $1000\times$ the original value.

CDD Mean/Std. Number \ Size	8×8	16×16	32×32
4	15.7/29.1	15.7/29.1	15.7/29.1
8	15.7/29.2	15.7/29.1	15.7/29.0
16	15.8/29.2	15.7/29.1	15.7/29.1

Table 6. Detailed hyperparameter settings used in SAM [5], SP+M-Net [6], and ShadowFormer [2].

Hyperparameter	Value	Hyperparameter	Value
SAM		**ShadowFormer**	
point_per_side	16	input_size	(640,480)
predict_iou_thres	0.90	train_ps	320
stability_score_thres	0.90	embed_dim	32
min_mask_region_area	500	win_size	10
SP+M-Net		token_projection	*linear*
input_size	(512,512)	token_mlp	*leff*

4.3. Test Settings

We provide detailed hyperparameter settings for SAM [5], SP+M-Net [6], and ShadowFormer [2] in Tab. 6. For SAM, we use the *vit_b* model with the learning rate set to $1e^{-4}$. During fine-tuning, we extract 32 points per image as prompts. By enabling $multimask\_output = True$ for the mask decoder, we calculate the mean Dice loss [9] over three output masks per prompt.

4.4. MC-Edge Annotations for the ISTD+

To provide CDD evaluation on the ISTD+ [6] dataset, we annotate the MC-edges in each image in a semi-automatic manner. Among the 46 unique scenes in the ISTD+ test set, we observe that 42 do not exhibit partial shadow edges coinciding with object boundaries. For these scenes, we simply erode and dilate the original shadow mask, then perform subtraction to get the pixels near the shadow edge. In the re-

(a) Automatic annotation (b) Manual annotation

Figure 5. Examples of shadow edge pixel annotation in the ISTD+ [6]. (a) shows automatic annotation on 42 scenes using eroded and dilated shadow masks, (b) shows manual annotation of material-consistent pixels on 4 scenes where shadow edges coincide with object boundaries.

Table 7. Comparison with SOTA models. We compare the performance of pre-trained models and models using our adaptation method on the ISTD+ [6] dataset. MAE and CDD are reported, note that CDD is reported in $1000\times$ the original value.

Methods	ISTD+				
	MAE			CDD	
	Shadow	NonShadow	All	Mean	Var
Input	40.2	2.6	8.5	148.5	90.0
Inpaint4Shadow [7]	5.9	2.9	3.4	2.1	4.3
ShadowDiffusion [3]	**4.9**	2.3	**2.7**	/	/
SP+M-Net [6]	7.3	2.5	3.3	3.2	3.8
SP+M-Net+Ours	6.1	2.5	3.1	1.6	2.6
ShadowFormer [2]	5.3	**2.2**	**2.7**	1.5	2.6
ShadowFormer+Ours	5.0	**2.2**	**2.7**	**1.0**	**1.6**

maining four scenes, we manually annotate the MC-edges (as depicted in Fig. 5).

5. Quantitative Results on ISTD+ Test Set

Recent improvements in shadow removal on ISTD+ [6] have reached saturation. The test set comprises multiple images from scenes similar to those in the training set, with shadows cast by objects outside the captured scene. State-of-the-art (SOTA) methods effectively learn the mapping between shadow and shadow-free pairs, already yielding satisfactory results. In Tab. 7, we compare our method with SOTA methods. Our method achieves performance comparable to SOTA methods. Although our method improves shadow removal on several challenging cases in ISTD+, the overall performance does not significantly surpass the well-trained existing methods on simple shadow images. This further supports the need to extend shadow removal techniques to general shadow images in real-world scenarios.

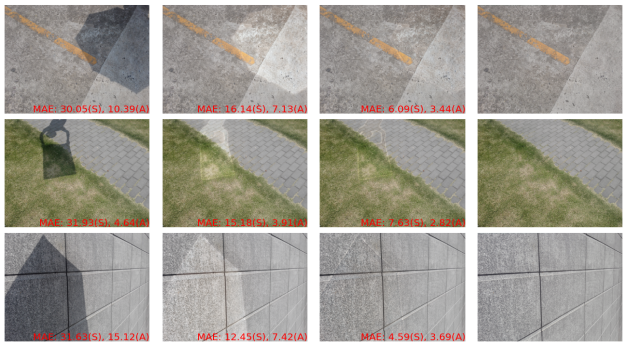
6. More Results on the Cross-Dataset Testing

6.1. Effect of Non-Shadow Loss

$\mathcal{L}_{nonshadow}$ is specifically designed to mitigate the errors caused by prior models pre-trained on data pairs with light intensity inconsistencies. In Tab. 8, we can see that our proposed method outperforms the pre-trained models, and incorporating $\mathcal{L}_{nonshadow}$ further improves performance by correcting the non-shadow region.

Table 8. Quantitative results on cross-dataset testing comparing the effect of $\mathcal{L}_{nonshadow}$. ISTD pre-trained ShadowFormer and SRD pre-trained ShadowFormer are tested on the ISTD+ test set.

Trained On	Tested On	Methods	MAE			CDD	
			S	NS	A	Mean	Var
SRD	ISTD+	prior model	13.7	3.4	5.1	55.0	43.3
		w.o. $\mathcal{L}_{nonshadow}$	6.4	2.9	3.5	15.8	13.6
		w. $\mathcal{L}_{nonshadow}$	6.2	2.4	3.0	8.0	9.4
ISTD	ISTD+	prior model	10.6	6.3	7.0	11.8	17.7
		w.o. $\mathcal{L}_{nonshadow}$	6.2	6.5	6.3	9.8	15.2
		w. $\mathcal{L}_{nonshadow}$	6.3	2.7	3.4	1.0	3.1



(a) Input (b) [2] (c) [2]+Ours (d) GT

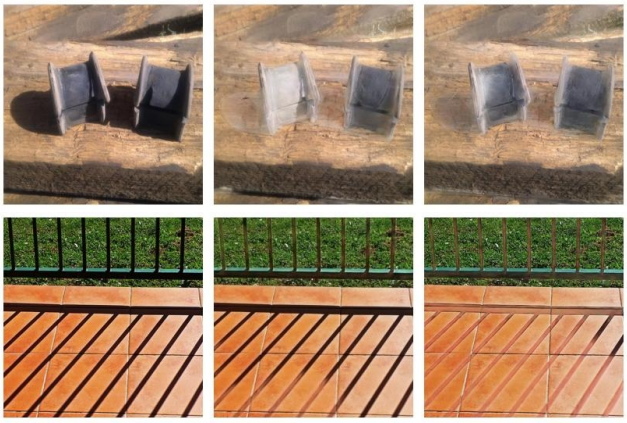
Figure 6. Qualitative comparison of cross-dataset testing. We use the SRD [8] pre-trained ShadowFormer [2] and test it on the ISTD+ test set. (a) shows the input image, (b) shows the ShadowFormer result, (c) presents the results with refinement, and (d) presents the ground truth. We also report shadow region (S) and overall (A) MAE results.

6.2. Qualitative results on SRD pre-trained model

Here, we show additional qualitative results from cross-dataset testing, where we use a ShadowFormer [2] pre-trained on the SRD [8] to test on ISTD+ [6] images, both with and without our refinement method. As depicted in Fig. 6, the pre-trained model does not perform well on out-of-distribution shadow images, while our refinement method significantly improves the performance.

7. More Qualitative Results

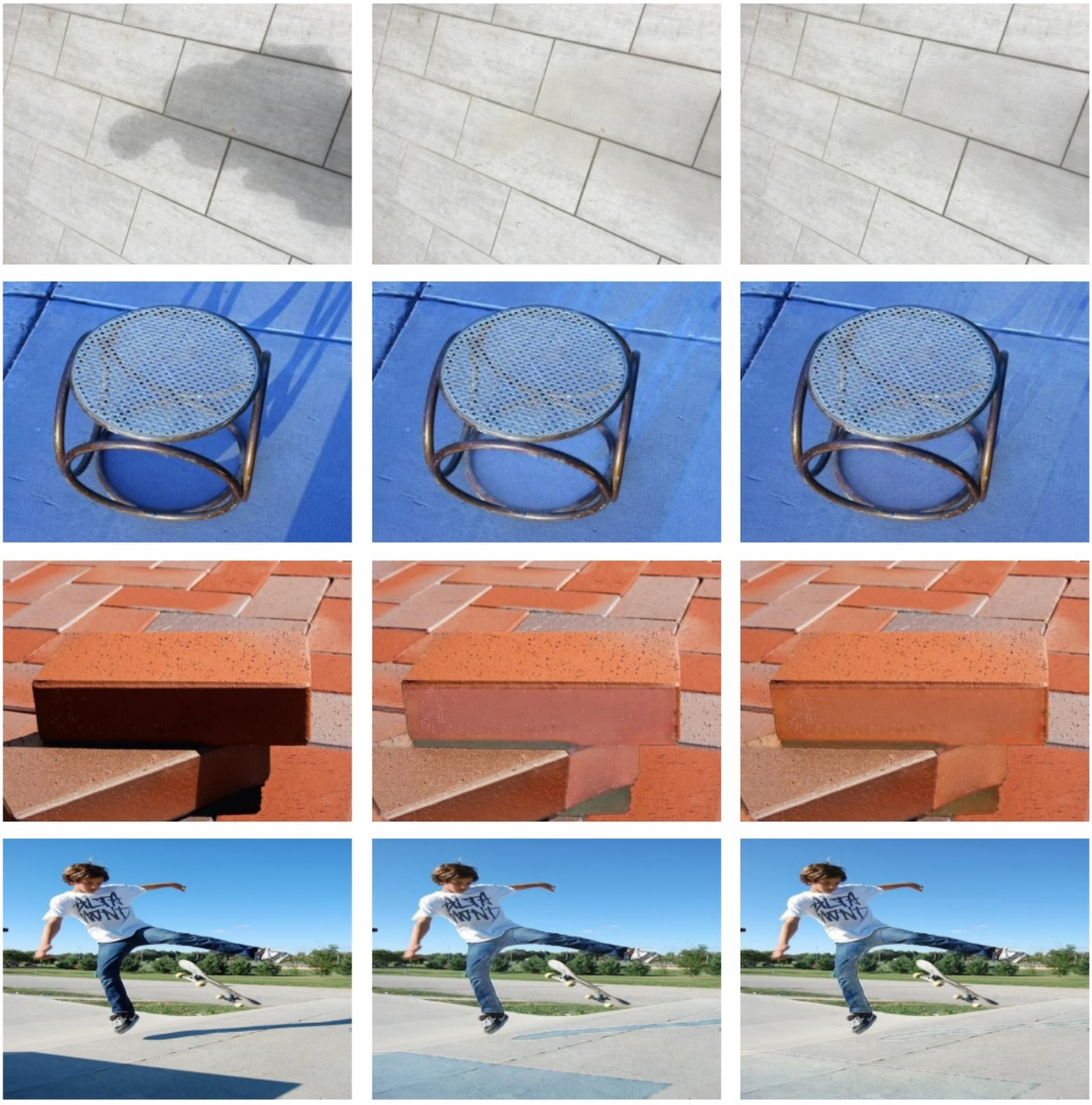
Qualitative results on refining the SRD pre-trained ShadowDiffusion [3] are provided in Fig. 7. We then present qualitative results on test cases containing soft shadows and attached shadows in Fig. 8. Additional results on our proposed test set, using our method applied to SP+M-Net [6]



Input [3] [3]+Ours

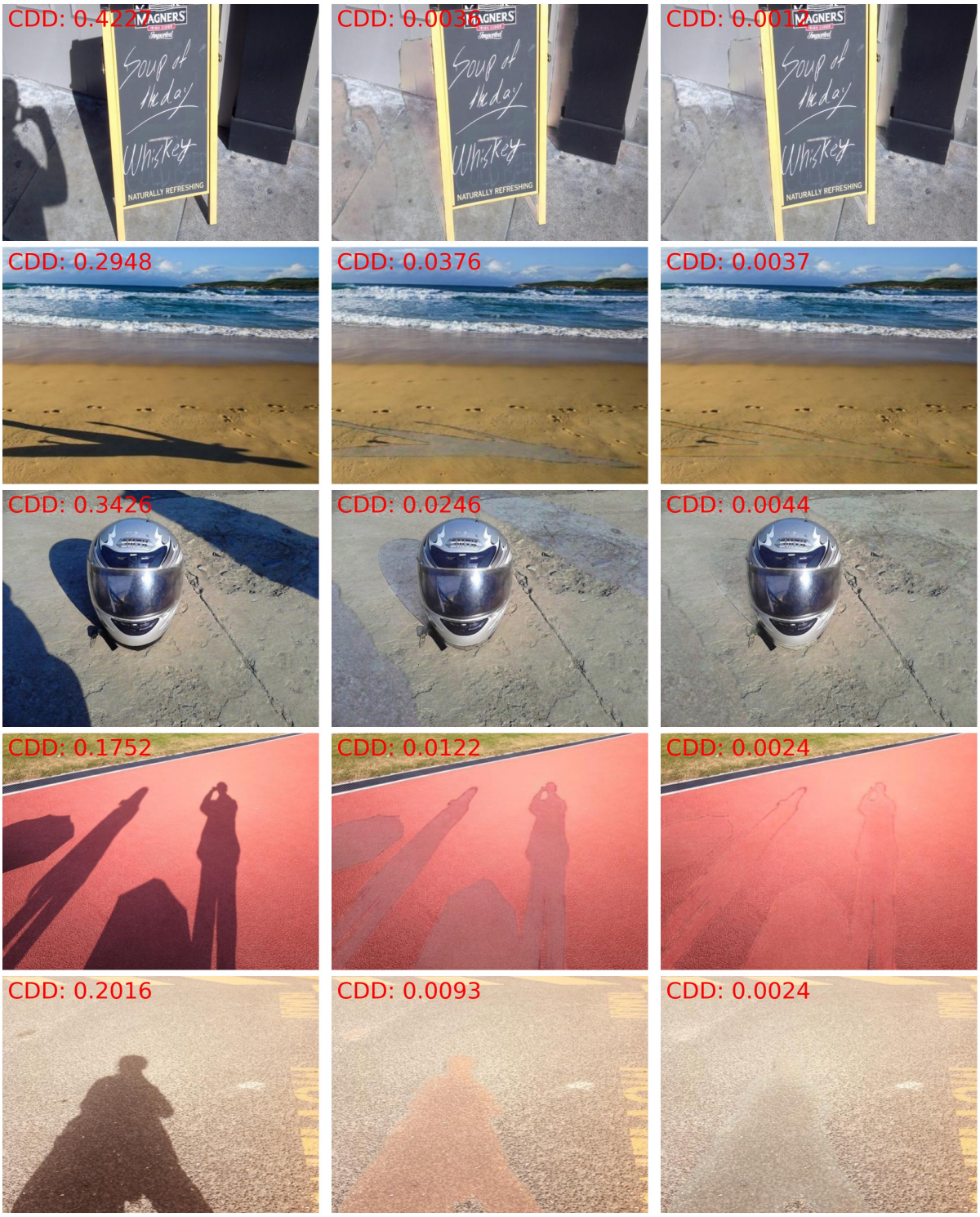
Figure 7. Qualitative results on our proposed dataset. (a) shows input image, (b) shows ShadowDiffusion [3] result, and (c) presents the results after refinement.

and ShadowFormer [2], are shown in Fig. 9 and Fig. 10. Finally, Fig. 11 presents qualitative results on the ISTD+ test set using our method applied to ShadowFormer.



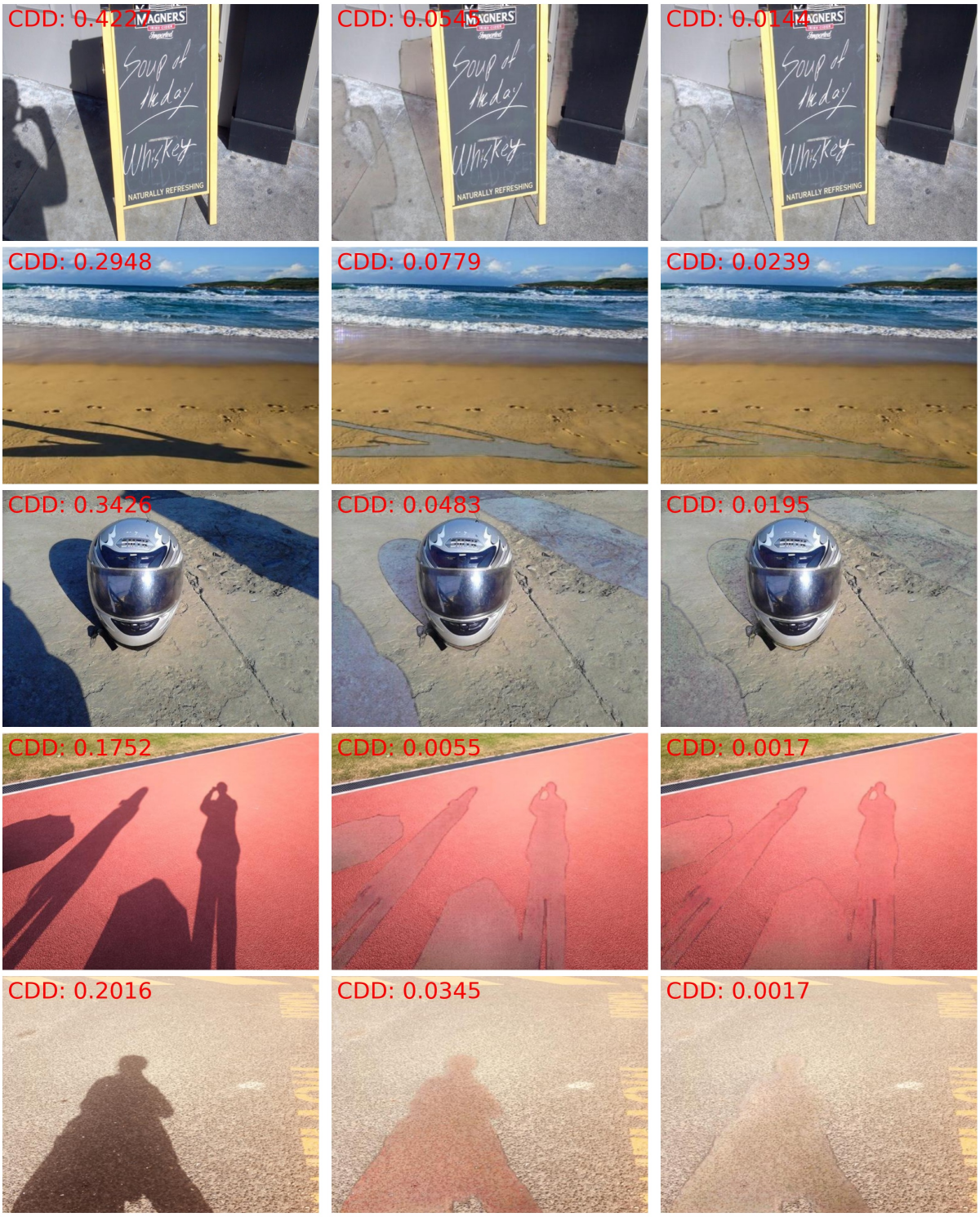
(a) Input (b) ShadowFormer (c) ShadowFormer+Ours

Figure 8. Qualitative results on soft shadows and attached shadows. The top two rows show results on soft shadows which are easier cases for the prior model [2], and the bottom two rows show results on self-cast shadows which confuses the prior model. However, our method also cannot fully address the attached shadows.



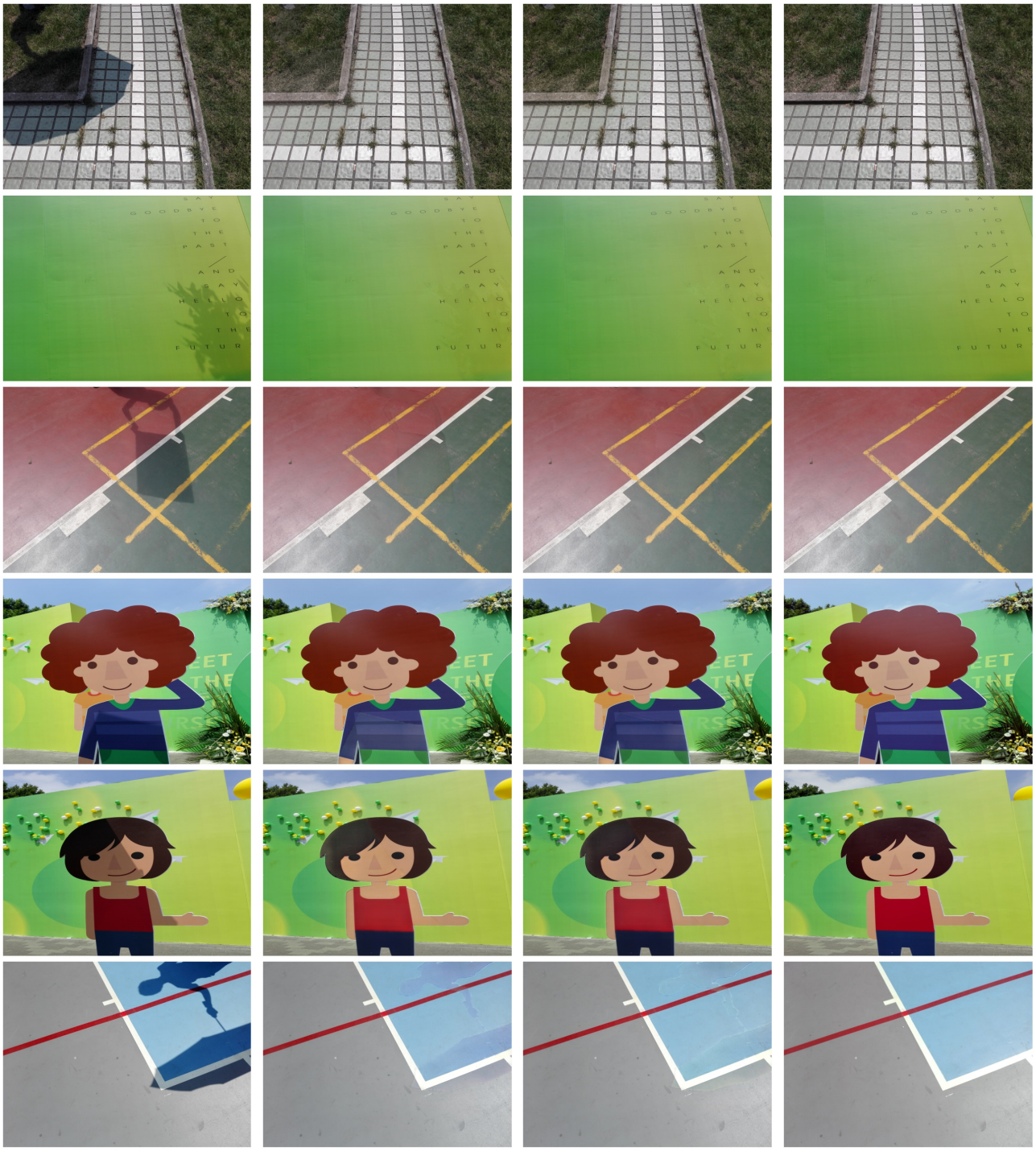
(a) Input (b) ShadowFormer (c) ShadowFormer+Ours

Figure 9. Qualitative results on our proposed dataset. (a) shows input image, (b) shows ShadowFormer [2] result, and (c) presents the results after refinement.



(a) Input (b) SP+M-Net (c) SP+M-Net+Ours

Figure 10. Qualitative results on our proposed dataset. (a) shows input image, (b) shows SP+M-Net [6] result, and (c) presents the results after refinement.



(a) Input (b) [2] (c) [2]+Ours (d) GT

Figure 11. More qualitative results on ISTD+ [6]. (a) shows input image, (b) shows ShadowFormer [2] result, (c) presents the results using our refinement method, and (d) shows the ground truth.

References

[1] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 2

[2] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023. 2, 3, 4, 5, 6, 7, 9

[3] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14049–14058, 2023. 4, 5

[4] Xiaowei Hu, Tianyu Wang, Chi-Wing Fu, Yitong Jiang, Qiong Wang, and Pheng-Ann Heng. Revisiting shadow detection: A new benchmark dataset for complex world. *IEEE Transactions on Image Processing*, 30:1925–1934, 2021. 2

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 3

[6] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8578–8587, 2019. 1, 2, 3, 4, 8, 9

[7] Xiaoguang Li, Qing Guo, Rabab Abdelfattah, Di Lin, Wei Feng, Ivor Tsang, and Song Wang. Leveraging inpainting for single-image shadow removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13055–13064, 2023. 4

[8] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4067–4075, 2017. 4

[9] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 3

[10] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 816–832. Springer, 2016. 2

[11] Han Yang, Tianyu Wang, Xiaowei Hu, and Chi-Wing Fu. Silt: Shadow-aware iterative label tuning for learning to detect shadows from noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12687–12698, 2023. 2, 3