HomoFormer: Homogenized Transformer for Image Shadow Removal

Supplementary Material

1. Random Shuffle is Efficient

Random shuffle operation pair is efficient since it does not incur any extra parameters or FLOPs. It only involves swapping elements of a tensor. As shown in Table Tab. 2, the only extra cost of random shuffle is 0.01s running time on a single GTX 1080Ti GPU.

Table 1. The efficiency of random shuffle. The input size is 256×256.

Method	#Params. (M)	FLOPs (G)	Time(s)
w/o random shuffle	17.8	38.4	0.07
w random shuffle	17.8	38.4	0.08

2. Model Efficiency.

To get a full picture of our model, we include comparisons of the number of parameters and FLOPs with other methods in Tab. 2. FLOPs is calculated with the input size 256×256. Although the main concern is not efficiency, HomoFormer, as shown in Tab. 2, is still highly competitive compared with previous methods.

Table 2. The comparison of the number of parameters and FLOPs with other methods. The input resolution is 256×256 for FLOPs.

Method	#Params. (M)	FLOPs (G)
ST-CGAN [14] (CVPR'18)	29.2	**17.9**
SP+M-Net [10] (ICCV'19)	141.2	39.8
DSC [6] (TPAMI'19)	22.3	123.5
DHAN [1] (AAAI'20)	21.8	262.9
G2R [12] (CVPR'21)	22.8	113.9
Fu *et al.* [2] (CVPR'21)	142.2	104.8
Jin *et al.* [8] (ICCV'21)	21.2	105.0
SG-ShadowNet [13] (ECCV'22)	**6.2**	39.7
HomoFormer(Ours)	<u>17.8</u>	<u>38.4</u>

3. Implementation details.

Our HomoFormer employs a four-level encoder-decoder structure. The numbers of blocks are $\{2, 2, 2, 2\}$ for level-1 to level-4 of Encoder and the corresponding blocks for Decoder are mirrored. The number of channel is set to 32 and the window size for local self-attention is 8. Our proposed method is implemented using PyTorch. Following the prior [15], we adopt the Adam optimizer [9] with the momentum ($\beta_1 = 0.9, \beta_2 = 0.999$). The initial learning rate is $2e^{-4}$, then gradually reduces to $1e^{-6}$ with the cosine annealing strategy. The training samples are augmented by the horizontal flipping and rotation of 90°, 180°, or 270°. The patch size is 384×384 and the batch size is 8. We train our HomoFormer with two NVIDIA GeForce GTX 3090 GPUs. For all the datasets, the total number of training epoch is set to 600.

4. Additional Result on ISTD

Tab. 3 provides quantitative comparison on ISTD dataset. The performance of HomoFormer is competitive but not optimal. The reason we speculate is that there is illumination inconsistency in the training pair of ISTD. However, shuffling makes HomoFormer difficult to adapt to the illumination shift.

Table 3. Quantitative comparisons with the SOTA methods on the ISTD datasets. The best and the second results are **boldfaced** and <u>underlined</u>, respectively.

Region / Method	Shadow MAE↓	Non-Shadow MAE↓	All MAE↓
Input images	40.2	2.6	8.5
Guo *et al.* [5] (TPAMI'12)	18.65	7.76	9.26
ST-CGAN [14] (CVPR'18)	9.99	6.05	6.65
ShadowGAN [7] (ICCV'19)	12.67	6.68	7.41
G2R [12] (CVPR'21)	10.72	7.55	7.85
Fu *et al.* [2] (CVPR'21)	7.91	5.51	5.88
Jin *et al.* [8] (ICCV'21)	11.43	3.81	6.57
BMNet [15] (CVPR'22)	7.60	4.59	5.02
ShadowFormer [3] (AAAI'23)	<u>6.16</u>	**3.90**	**4.27**
ShadowDiffusion [4] (CVPR'23)	**4.13**	<u>4.14</u>	<u>4.12</u>
HomoFormer(ours)	6.85	4.24	4.49

5. More Visual Results

5.1. De-shadowing Results on ITSD+ Dataset

Figs. 2 to 5 present visual comparisons with other state-of-the-art methods on the ISTD+ dataset. We can observe that our HomoFormer can restore clean images more faithfully (see boundary artifacts in Figs. 3 and 5).

5.2. De-shadowing Results on SRD Dataset

Figs. 6 to 9 show visual comparisons with other state-of-the-art methods on the SRD dataset. We can observe that the restored images output by our HomoFormer are more close to the ground truth images.

5.3. De-shadowing Results on SBU Dataset

To validate the generalization, we evaluate the model pre-trained on ISTD+ on the SBU dataset [11]. Fig. 10 shows the visual comparison.

5.4. Failure Case

HomoFormer may struggle to remove shadow clearly when facing with real-world complex scenarios. Fig. 1 present a case that HomoFormer pretrained on ISTD+ fails to lighten the large shadowed region of a building, which is absent in training data. The main reason we guess derives from the huge gap between training and testing data and HomoFormer cannot effectively bridge this gap relying only on its shuffling behaviour.

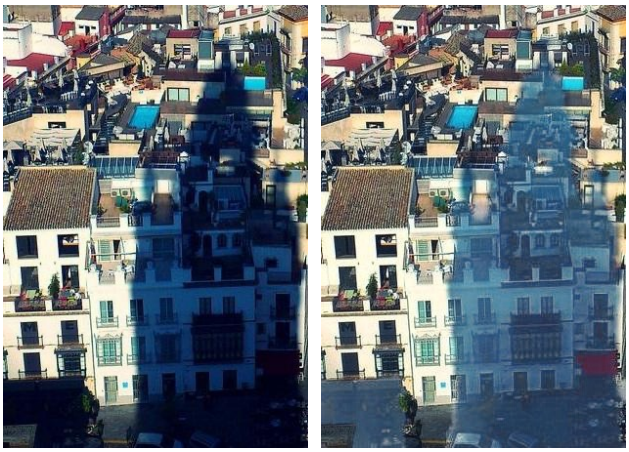


Figure 1. Failure Case of HomoFormer.

5.5. Uncertainty Distribution

Uncertainty can play a significant role for computer vision. For shadow removal, we can leverage to uncertainty can estimate the degree of *confidence* about the prediction. The presented HomoFormer provides a natural approach to estimate its uncertainty due to its inherent random shuffle behaviour. For example, we can evaluate an image M times and compute the standard deviation as the uncertainty. Specifically, let X denote the input and F denote the function of HomoFormer , the uncertainty $\mathcal{U}$ can be estimated by:

$$\mathcal{U} \approx \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(F(X;\mathcal{S}_i)-\bar{F}(X)\right)^2}, \tag{1}$$

where $\bar{F}(X)$ is the averaged output:

$$\bar{F}(X)=\frac{1}{M}\sum_{i=1}^{M}F(X;\mathcal{S}_i). \tag{2}$$

Fig. 11 suggests that *without* resorting to groundtruth image, the estimated uncertainty can predict where errors are prone to take place. We can observe that the model often struggles [1] to restore pixels in high-frequent (e.g., alphabets in Fig. 11) and shaded regions (e.g., shaded regions in Fig. 11). Generally, we can put less confidence to those regressed pixels with more uncertainty. For real-world scenarios where the groundtruth images are lacked, this property is of great practical significance.

6. Broader Impacts

Generally, image acquisition system tends to suffer from shadow degradations. Therefore, image shadow removal is of importance in research and application. Our proposed HomoFormer can attain outstanding performance for de-shadowing images. Nevertheless, some negative consequences may come along. For instance, the deviation from the actual image textures may lead to unfair judgments in criminal situations. In these scenarios, it is required to consult with human experts to avoid misjudgments.

[1] The corresponding uncertainty is high.

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, 2020.

[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, 2021.

[3] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. In AAAI, 2023.

[4] 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 CVPR, 2023.

[5] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. TPAMI, 2013.

[6] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. TPAMI, 2019.

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

[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 ICCV, 2021.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[10] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In ICCV, 2019.

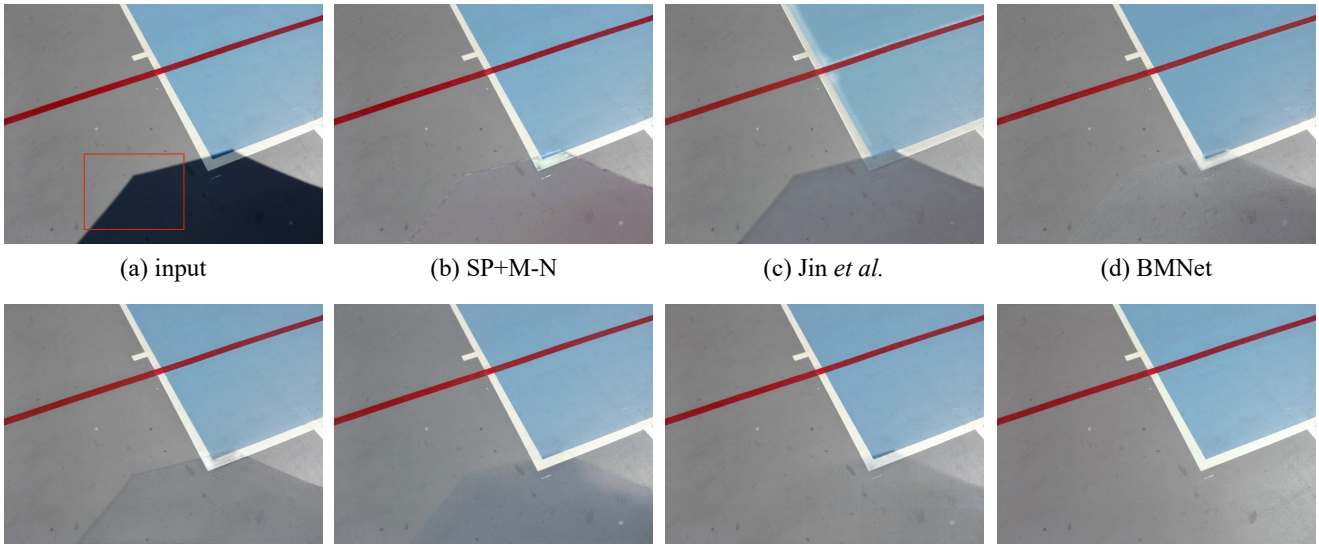
[11] Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. TPAMI, 2022.

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

[13] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In ECCV, 2022.

[14] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In CVPR, 2019.

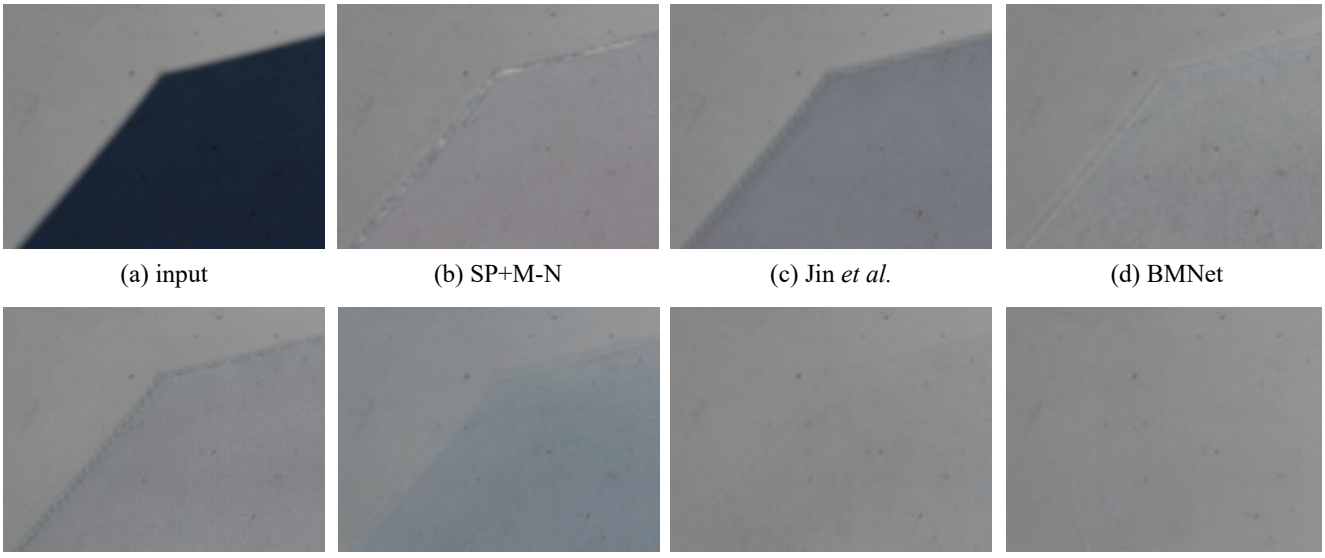
[15] Yurui Zhu, Jie Huang, Xueyang Fu, Feng Zhao, Qibin Sun, and Zheng–Jun Zha. Bijective mapping network for shadow removal. In CVPR, 2022.



(a) input (b) SP+M-N (c) Jin *et al.* (d) BMNet

(e) SG-ShadowNet (f) ShadowFormer (g) HomoFormer (h) GT

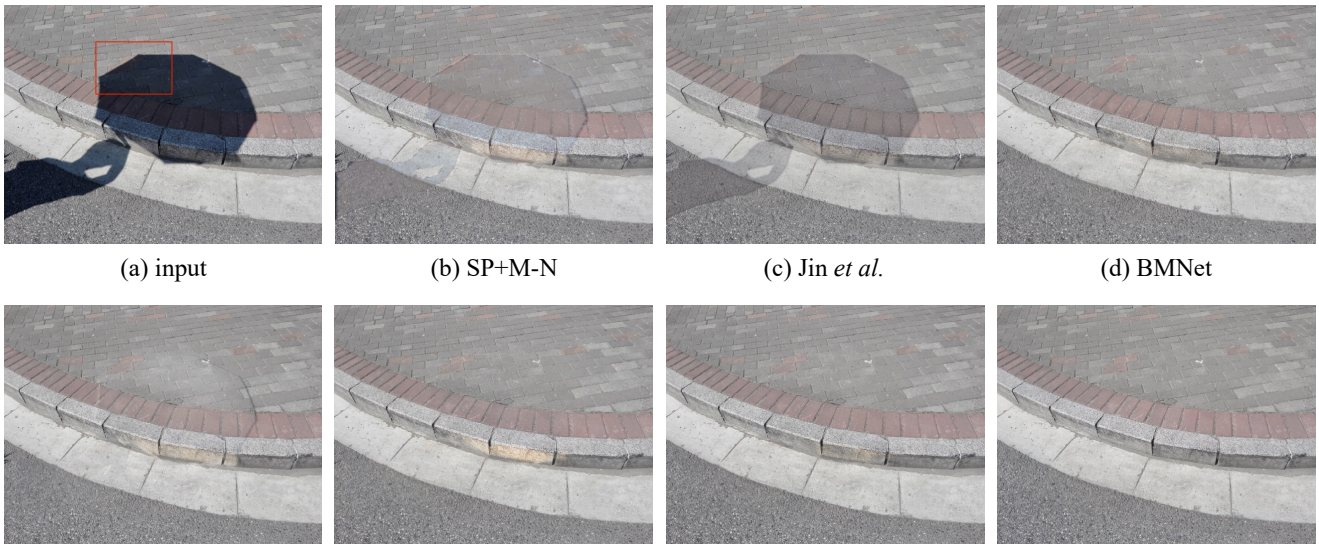
Figure 2. Visual comparisons with state-of-the-art methods on the ISTD+ dataset.



(a) input (b) SP+M-N (c) Jin *et al.* (d) BMNet

(e) SG-ShadowNet (f) ShadowFormer (g) HomoFormer (h) GT

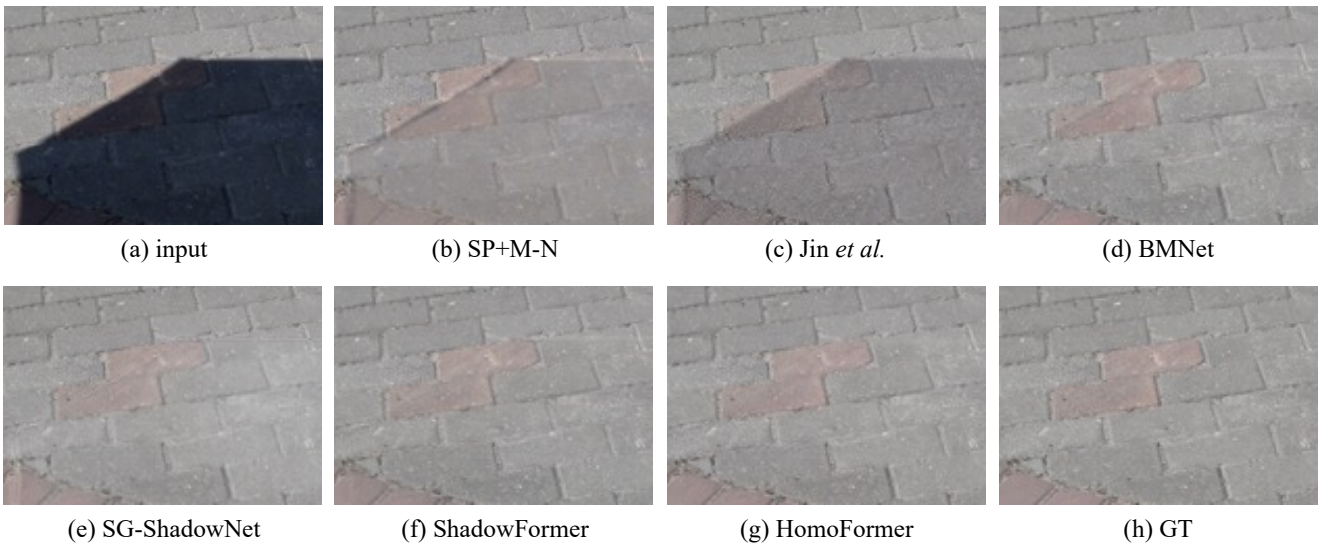
Figure 3. Enlarged region of Fig. 2.



(a) input (b) SP+M-N (c) Jin *et al.* (d) BMNet

(e) SG-ShadowNet (f) ShadowFormer (g) HomoFormer (h) GT

Figure 4. Visual comparisons with state-of-the-art methods on the ISTD+ dataset.



(a) input (b) SP+M-N (c) Jin *et al.* (d) BMNet

(e) SG-ShadowNet (f) ShadowFormer (g) HomoFormer (h) GT

Figure 5. Enlarged region of Fig. 4.

(a) input (b) DSC (c) DHAN (d) BMNet

(e) SG-ShadowNet (f) UniShadow (g) GT

Figure 6. Visual comparisons with state-of-the-art methods on the SRD dataset.

(a) input (b) DSC (c) DHAN (d) BMNet

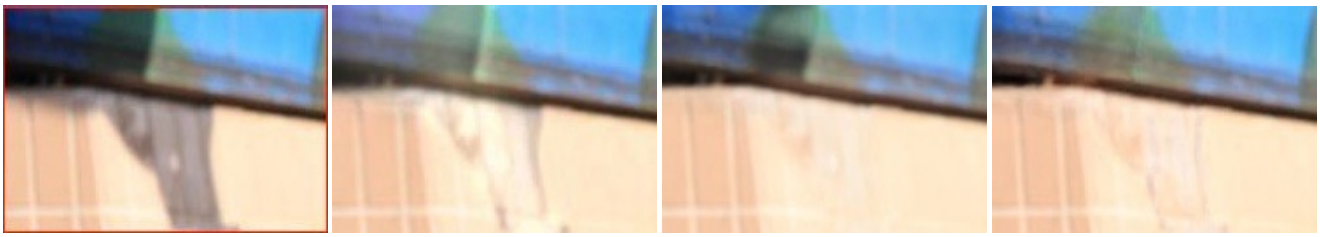
(e) SG-ShadowNet (f) HomoFormer (g) GT

Figure 7. Enlarged region of Fig. 6.

(a) input (b) DSC (c) DHAN (d) BMNet

(e) SG-ShadowNet (f) HomoFormer (g) GT

Figure 8. Visual comparisons with state-of-the-art methods on the SRD dataset.



(a) input (b) DSC (c) DHAN (d) BMNet

(e) SG-ShadowNet (f) HomoFormer (g) GT

Figure 9. Enlarged region of Fig. 8.

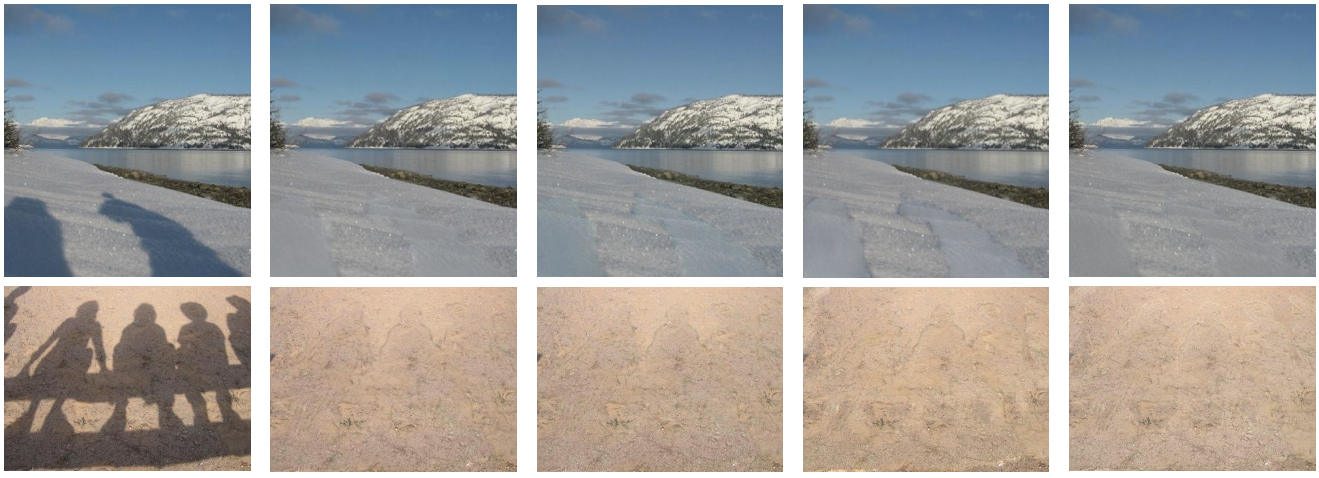
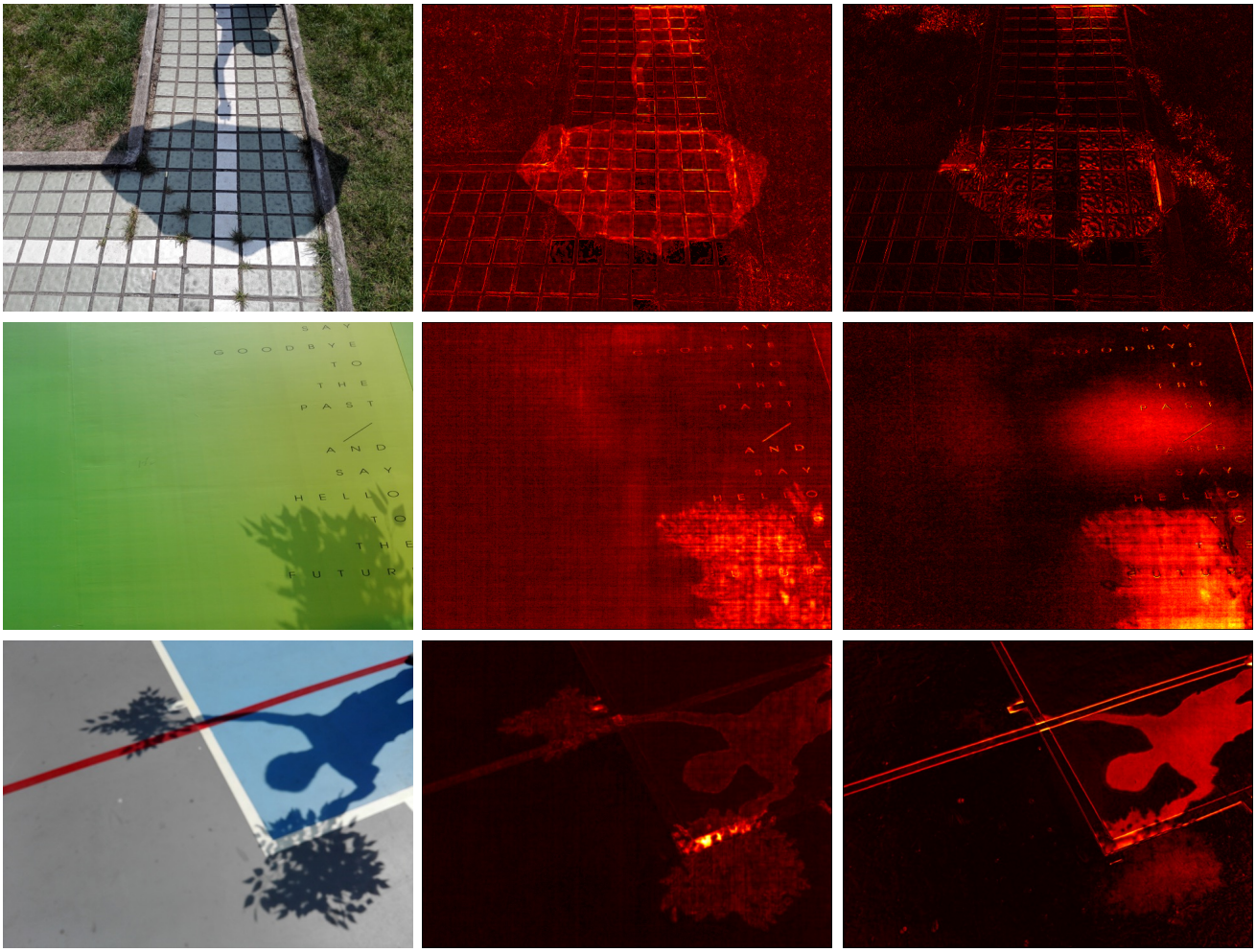


Figure 10. Visual comparisons with state-of-the-art methods on the SBU dataset [11].



(a) input (b) uncertainty distribution (c) error distribution

Figure 11. Uncertainty originating from random shuffle can be used to predict the error distribution (*i.e.*, $|I - I'|$) between the evaluated clean image I' and shadow-free image I. The HomoFormer often struggles to restore pixels in high-frequent (e.g., alphabets in the second row) and shaded regions (e.g., shaded regions).