Panoramic Image Reflection Removal (Supplementary Material)

Yuchen Hong[1#] Qian Zheng[2#] Lingran Zhao[1] Xudong Jiang[2] Alex C. Kot[2] Boxin Shi[1,3,4*]

[1] NELVT, Department of Computer Science and Technology, Peking University, Beijing, China

[2] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

[3] Institute for Artificial Intelligence, Peking University, Beijing, China [4] Peng Cheng Laboratory, Shenzhen, China

yuchenhong.cn@gmail.com, {zhengqian, exdjiang, eackot}@ntu.edu.sg, {calvinzhao, shiboxin}@pku.edu.cn

In the supplementary material, we show a comparison on reflection alignment with an existing alignment approach, introduce more details about our method, provide an ablation study on transmission recovery, and conduct more comprehensive comparisons with state-of-the-art reflection removal methods. We further provide a demonstration video to show two user scenarios of using a panoramic camera and a mobile phone.

6. Comparison on Reflection Alignment

In this section, we conduct experiments on reflection alignment comparing to an two-step alignment approach RANSAC-Flow [4] (denoted as 'RF20' for brevity), corresponding to footnote 1 in the paper. In panoramic image reflection removal, we need to align the panoramic reflection scene $\mathbf{R}_{\mathrm{P}}$ and the mixture image $\mathbf{M}$, which have different content as $\mathbf{M}$ is the combination of the transmission scene $\mathbf{T}$ and the glass-reflected image $\mathbf{R}_{\mathrm{G}}$. As shown in Figure 9, since RF20 [4] is developed to work on consistent image content without the impact of glass, it mismatches image features of $\mathbf{R}_{\mathrm{P}}$ and $\mathbf{T}$, resulting in its failure in our case. After taking the photometric and geometric discrepancy into consideration, our coarse-to-fine alignment algorithm correctly matches image features of $\mathbf{R}_{\mathrm{P}}$ and $\mathbf{R}_{\mathrm{G}}$ to achieve more precise alignment, which provides reliable guidance for the following transmission recovery.

7. Analysis about Scale Discrepancy

In this section, we provide more details about the scale discrepancy between the reflection image $\mathbf{R}_{\mathrm{G}}$ and the reflection scene $\mathbf{R}_{\mathrm{P}}$, which plays a significant role in their geometric misalignment, corresponding to footnote 2 in the paper. As shown in Figure 10, patches with the same size are extracted in both $\mathbf{R}_{\mathrm{G}}$ and $\mathbf{R}_{\mathrm{P}}$, however, an object (*e.g.*, plant) can appear with quite different spatial scales in these

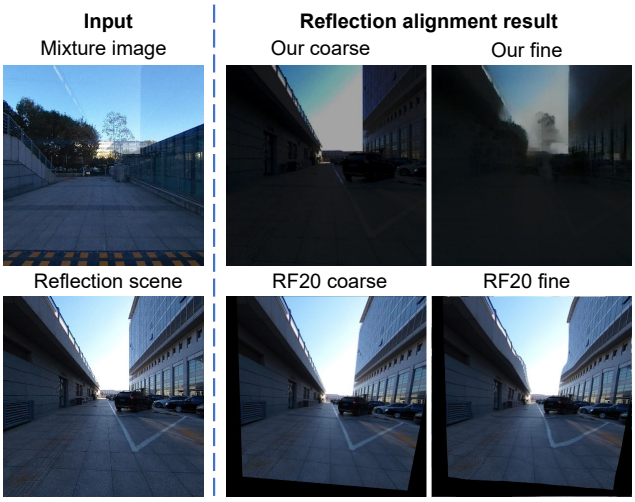

Figure 9. Qualitative comparison of our coarse-to-fine reflection alignment algorithm with the two-step alignment approach RF20 [4] on PORTABLE dataset.

two images. We define the above scale discrepancy through scalar s, implying that objects of $\mathbf{R}_{\mathrm{P}}$ can match the scale of their corresponding ones in $\mathbf{R}_{\mathrm{G}}$ if the image size of $\mathbf{R}_{\mathrm{G}}$ is reduced by such scalar. According to the camera imaging model, different distances from the camera to the glass plate and the reflection scene with different FoV (field of view) between $\mathbf{R}_{\mathrm{G}}$ and $\mathbf{R}_{\mathrm{P}}$ influence the scale discrepancy, which can be summarized as followed:

$$s = \frac{d_{cr} + 2d_{cg}}{d_{cr}} \cdot \frac{\tan(\phi_{\mathbf{R}_{\mathrm{G}}}/2)}{\tan(\phi_{\mathbf{R}_{\mathrm{P}}}/2)}, \tag{10}$$

where d_{cg} and d_{cr} represents distances from the camera to the glass plate and the reflection scene, with $\phi_{\mathbf{R}_{\mathrm{G}}}$ and $\phi_{\mathbf{R}_{\mathrm{P}}}$ to be the FoV of $\mathbf{R}_{\mathrm{G}}$ and $\mathbf{R}_{\mathrm{P}}$, respectively, as illustrated in Figure 10(a).

In the geometric alignment procedure of our method, we employ an ergodic searching and matching process, where the sizes of slide windows are varied with s (to be $sh \times sw$

#Equal contribution. *Corresponding author.

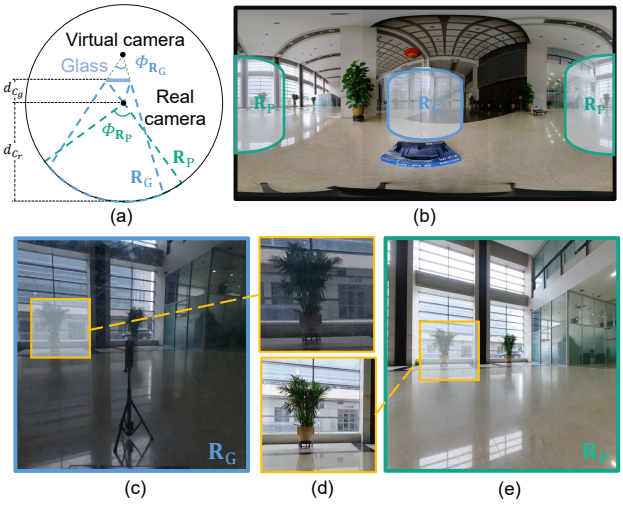

Figure 10. (a) Camera model of capturing a scene containing a glass plate by a panoramic camera. (b) Captured panoramic image. (c) Glass-reflected reflection image $\mathbf{R}_{\mathrm{G}}$, which is 'captured' by the virtual camera. (d) Illustration of the geometric misalignment. (e) Panoramic reflection scene $\mathbf{R}_{\mathrm{P}}$, which is captured by the real camera.

as in Section 3.2.1 of the main paper). In practice, d_{cr} is generally much larger than d_{cg}, then we assume that d_{cr} is at least four times larger than d_{cg}. Meanwhile, as the initial image extraction step generates $\mathbf{M}$ with user interaction and the glass plate is usually in a limited size, the FoV of the cropped $\mathbf{M}$ is generally between $45°$ and $60°$. Therefore, in our setup, s ranges from 0.4 to 0.85, and we sample s with the step of 0.05 for the ergodic searching and matching procedure to balance the potential scale discrepancy range and the computational cost.

8. Details about Training Data Synthesizing

In this section, we introduce more details about training data synthesizing, corresponding to footnote 3 in the paper. Transmission scenes and reflection scenes are selected from SUN RGB-D [5] and Cityscapes [1], respectively, to cover various real scenarios. Reflection images are generated from reflection scenes, considering geometric and photometric misalignment in real data. For geometric misalignment, we randomly sample s in the range of $[0.4, 0.85]$ according to the analysis in Section 7. Then we randomly crop a patch with size of $sh \times sw$ (h and w are the height and width of the reflection scene) from the reflection scene to simulate the spatial translation. Afterwards, a random perspective transformation is conducted to imitate visual parallax. For photometric misalignment, we utilize the transformation function estimated from [6] as mentioned in Section 3.2.2 of the paper, to simulate the photometric discrepancy between $\mathbf{R}_{\mathrm{P}}$ and $\mathbf{R}_{\mathrm{G}}$. Finally, mixture images are generated with transmission scenes and reflection images by the same blending formulation in [7].

Table 3. Quantitative comparison of estimated reflection images between our method and several state-of-the-art reflection removal methods including IBCLN [3], KH20 [2], and CoRRN [7] on our PORTABLE dataset. ↑ (↓) indicates larger (smaller) values are better. Bold numbers indicate the best performing results.

Method	Error Metric			
	PSNR↑	SSIM↑	NCC↑	LMSE↓
Ours	**20.790**	**0.649**	**0.809**	**0.071**
ICBLN [3]	16.675	0.497	0.473	0.122
KH20 [2]	17.295	0.564	0.453	0.080
CoRRN [7]	15.127	0.457	0.467	0.107

Table 4. Quantitative comparison between our transmission network and its variant *w/o*-Saliency on our PORTABLE dataset.

Method	Error Metric			
	PSNR↑	SSIM↑	NCC↑	LMSE↓
Ours	**23.986**	**0.749**	**0.926**	**0.021**
w/o-Saliency	23.348	0.741	0.917	**0.021**

9. Evaluation on Reflection Recovery

In this section, we compare the estimated reflection images by our reflection refinement network with several state-of-the-art reflection removal methods including IBCLN [3], KH20 [2], and CoRRN [7] in both quantitative results and visual quality on our PORTABLE dataset, corresponding to footnote 4 in the paper. The comparison does not include another state-of-the-art method ERRNet [8] as it only estimates the transmission scene. As shown in Table 3, comparing to other single-image methods, quantitative results demonstrate the prominent advantage of our method on reflection image estimation, thanks to the content information about reflection scenes from the panoramic image. As can be observed from visual quality results in Figure 11, our method is able to estimate distinct and precise reflection images, while other methods generate results with low image quality and incorrect artifacts from transmission scenes due to the content ambiguity, which further demonstrates the effectiveness of our reflection refinement network.

10. Ablation Study on Transmission Recovery

In this section, we conduct an ablation study on our transmission recovery network, corresponding to footnote 6 in the paper. As introduced in Section 3.4, the content information of the reflection scene is used as the saliency information to indicate reflection regions in $\mathbf{M}$ and guide reflection removal (through the attention map) in our transmission network. To validate the effectiveness of such guidance, we compare the transmission network with its variant denoted

as '*w/o*-Saliency' which removes attention maps but concatenates features of $\mathbf{R}_{\mathrm{G}}$ and $\mathbf{M}$ directly. Quantitative results in Table 4 show advantages of the attention map guidance strategy, and Figure 12 suggests that the saliency information from reflection images helps generate more pleasant results in visual quality, which is mainly due to the nonlinear relation of $\mathbf{T}$, $\mathbf{M}$, and $\mathbf{R}_{\mathrm{G}}$.

11. More Visual Quality Results

In this section, we provide more qualitative reflection removal results on our PORTABLE, NATURAL, and PHONE dataset in Figure 13, Figure 14, and Figure 15, compared with several state-of-the-art reflection removal methods including IBCLN [3], KH20 [2], CoRRN [7], and ERR-Net [8]. We also exhibit some challenging examples that our method cannot handle well in Figure 16 and Figure 17, where certain regions of strong reflections still exist. Even so, the results are still better than single-image methods. The reason can be that the corresponding regions in mixture images are almost saturate, which renders transmission recovery to be more similar to inpainting. Our future work will try to solve the problem to generate more visual pleasant results.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[2] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single image reflection removal with physically-based training images. In *CVPR*, June 2020. 2, 3, 4, 5, 6

[3] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *CVPR*, pages 3565–3574, 2020. 2, 3, 4, 5, 6

[4] X Shen and F Darmon. Ransac-flow: Generic two-stage image alignment. In *ECCV*, volume 12349, 2020. 1

[5] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 2015. 2

[6] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, and Alex C. Kot. Reflection scene separation from a single image. In *CVPR*, 2020. 2

[7] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, Ah-Hwee Tan, and Alex Kot Chichung. CoRRN: Cooperative reflection removal network. *IEEE TPAMI*, 2019. 2, 3, 4, 5, 6

[8] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *CVPR*, pages 8178–8187, 2019. 2, 3, 5, 6

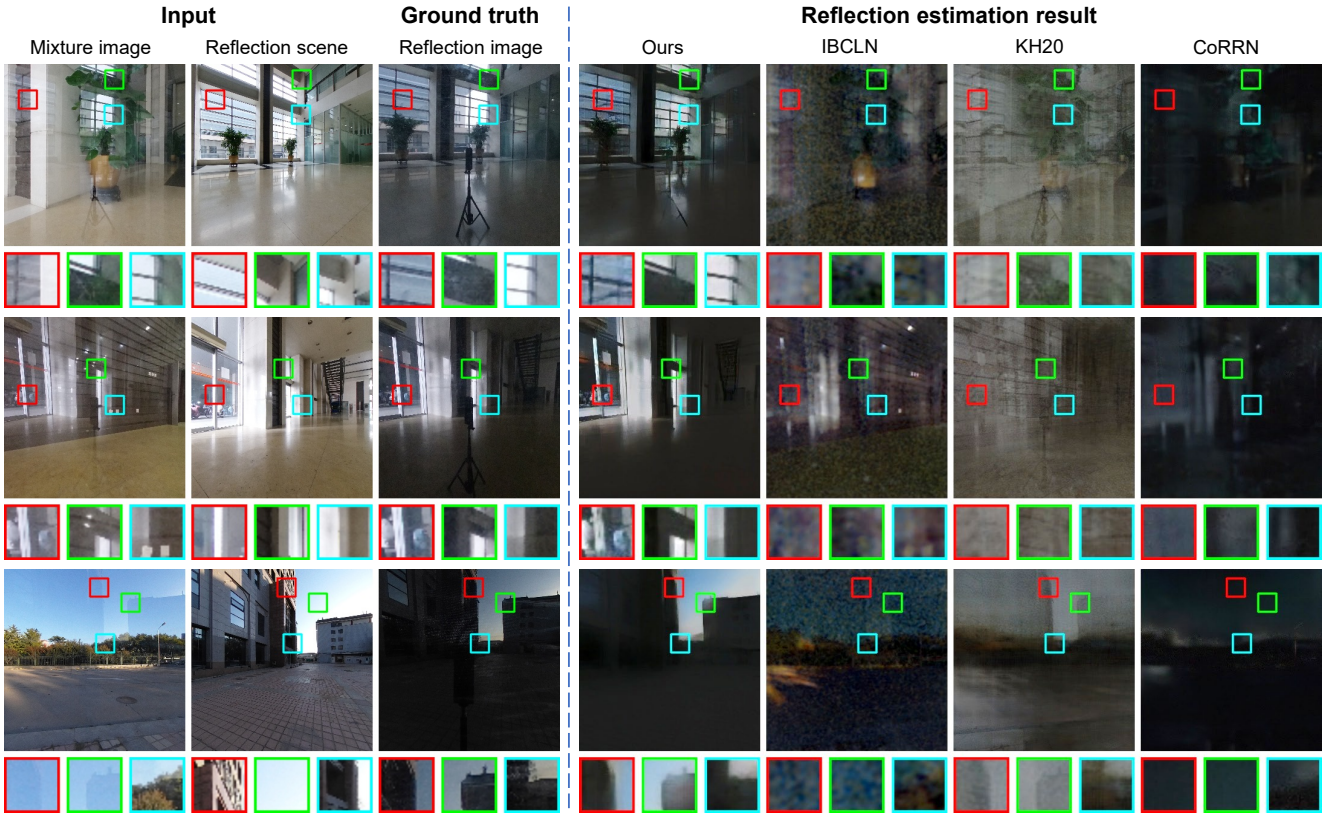

Figure 11. Examples of reflection estimation results on PORTABLE dataset, compared with IBCLN [3], KH20 [2], and CoRRN [7]. Close-up views are displayed at the bottom of each image. Zoom in for better details.

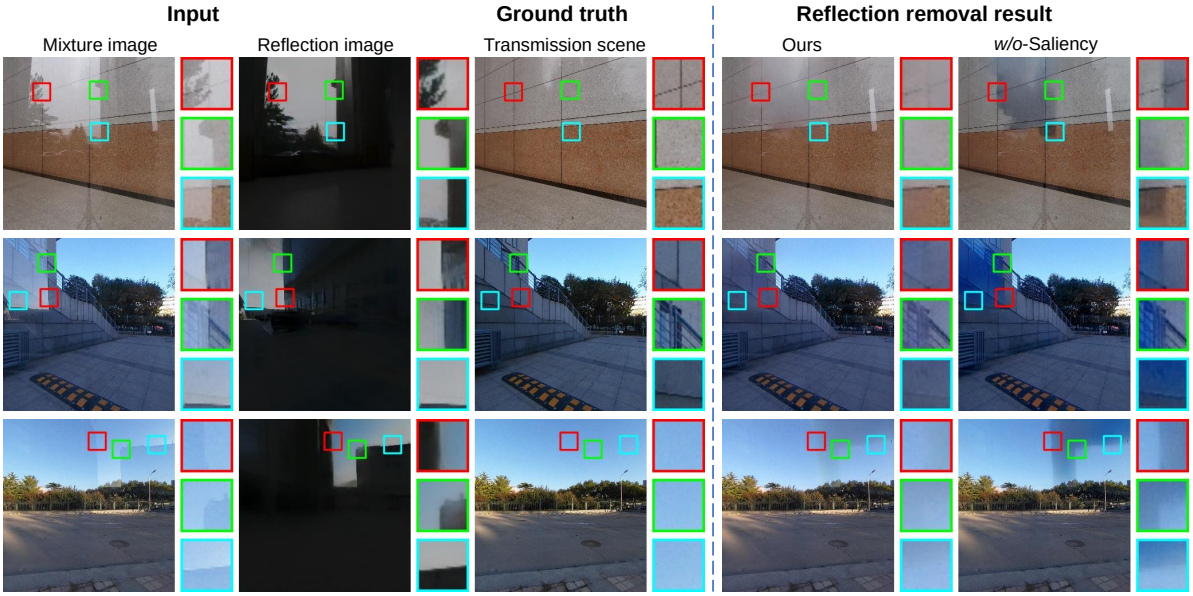

Figure 12. Visual quality comparison of our transmission network with its variant *w/o*-Saliency. Close-up views are displayed at the right side of each image. Zoom in for better details.

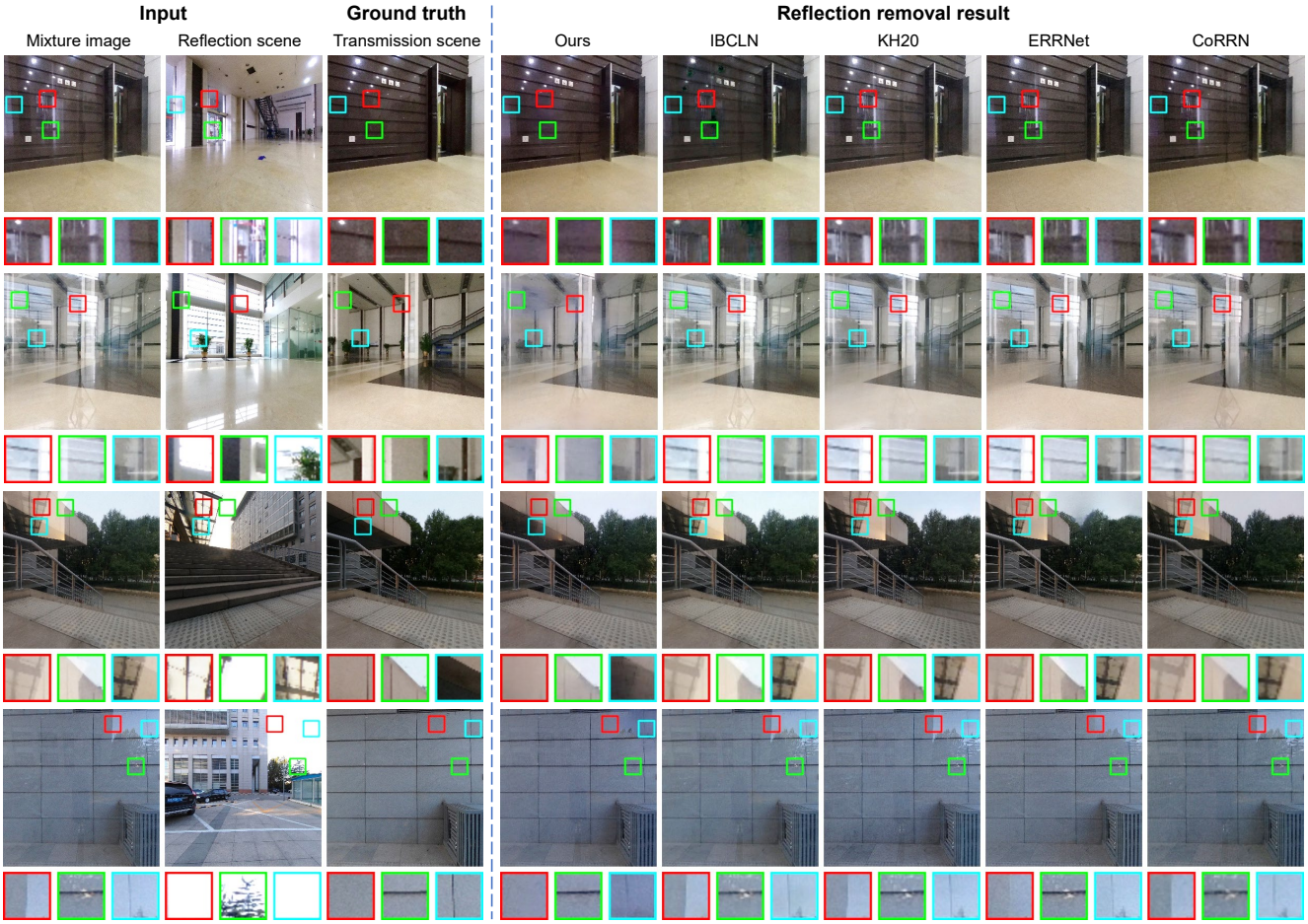

Figure 13. More examples of reflection removal results on PORTABLE dataset, compared with IBCLN [3], KH20 [2], ERRNet [8], and CoRRN [7]. Close-up views are displayed at the bottom of each image. Zoom in for better details.

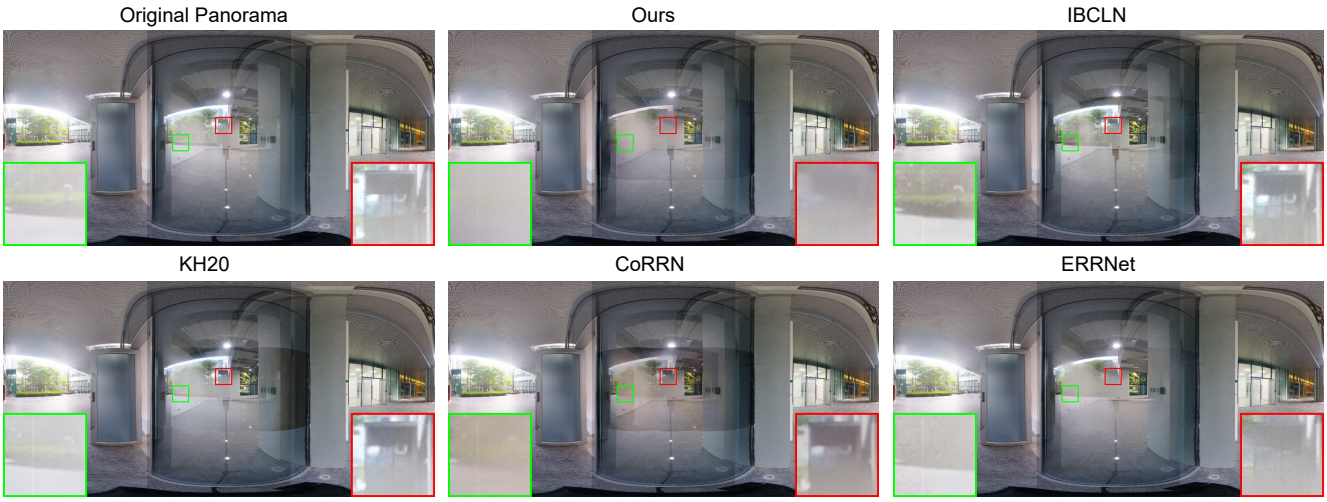

Figure 14. Another example on the NATURAL dataset, compared with several state-of-the-art single-image methods. Close-up views are displayed at the bottom of each image. Zoom in for better details.

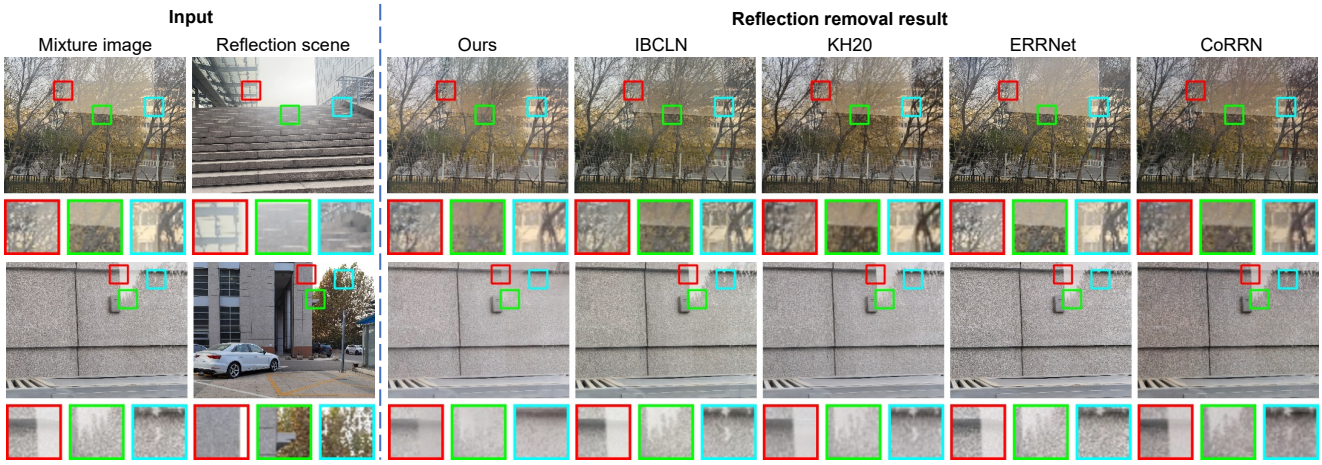

Figure 15. More examples of reflection removal results on PHONE dataset, compared with IBCLN [3], KH20 [2], ERRNet [8], and CoRRN [7]. Close-up views are displayed at the bottom of each image. Zoom in for better details.

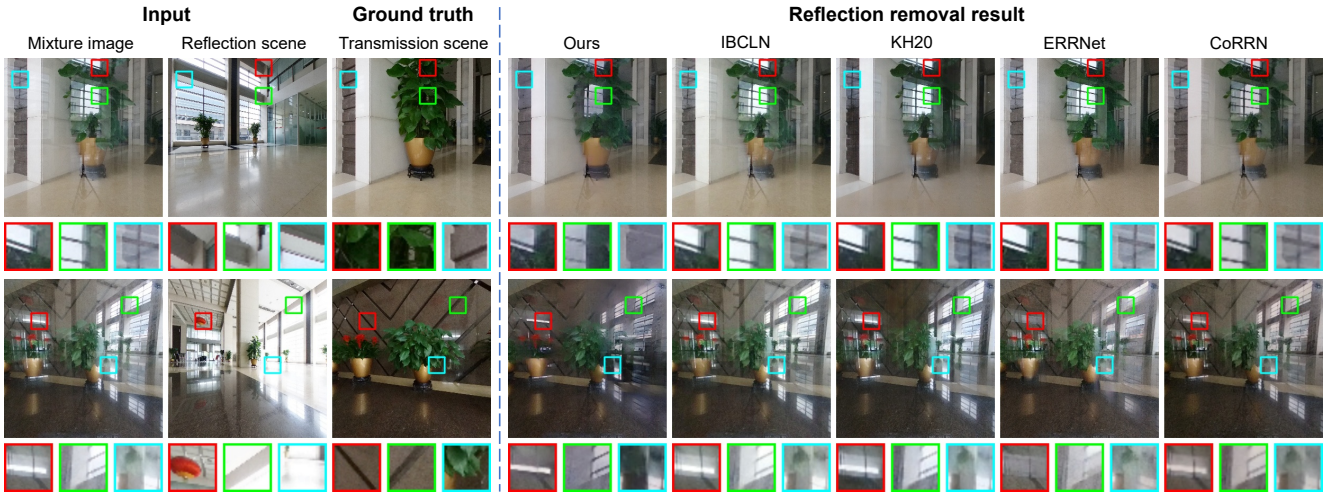

Figure 16. Challenge cases PORTABLE dataset, compared with IBCLN [3], KH20 [2], ERRNet [8], and CoRRN [7]. Close-up views are displayed at the bottom of each image. Zoom in for better details.

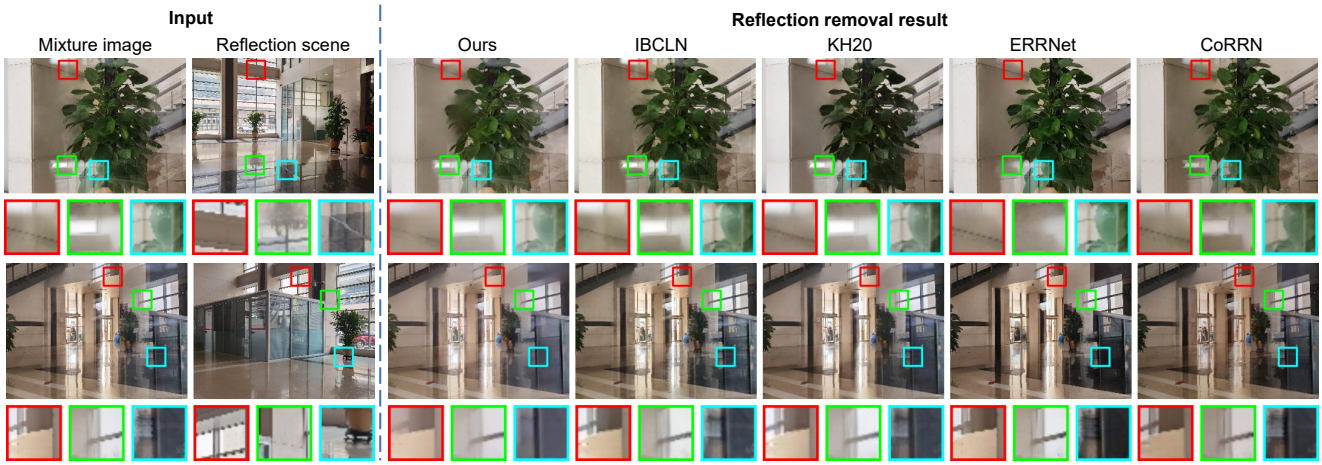

Figure 17. Challenge cases on PHONE dataset, compared with IBCLN [3], KH20 [2], ERRNet [8], and CoRRN [7]. Close-up views are displayed at the bottom of each image. Zoom in for better details.