PAR^{2}Net: End-to-end Panoramic Image Reflection Removal (Supplementary Material)

Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, *Fellow, IEEE*, Alex C. Kot, *Fellow, IEEE*, and Boxin Shi*, *Senior Member, IEEE*

Appendix A
Additional qualitative comparisons on real data

To evaluate the performance of the proposed method (PAR^{2}Net), we conduct more qualitative comparisons on the PORTABLE and NATURAL datasets in Fig. 13 and Fig. 14. We compare PAR^{2}Net with our preliminary work HZ21 [3] and a single-image method DX21 [1] which is selected to represent state-of-the-art single-image methods (since it performs best among the five single-image methods in the quantitative comparison, *i.e.*, in Table 1 of the main paper). In addition, we display more results on the PHONE dataset in Fig. 15 by comparing PAR^{2}Net with DX21 [1] to show our generalization capacity to limited-FoV images.

Appendix B
Details of the unsupervised version for ablation study

In the ablation study (Sec. 6.2 in the main paper), we implement an unsupervised version of the proposed method inspired by Han *et al.* [2], which compares the different learning strategies on the panoramic image reflection removal task. We retain the network architecture of the proposed method and employ the loss functions in [2] to adapt the unsupervised learning strategy. We update the network parameters with 1000 iterations for each test image.

Training recovery modules. Following Han *et al.* [2], we first train recovery modules for reflection refinement and transmission recovery by recovering input images, *i.e.*, recovering mixture images $\mathbf{M}$ and reflection scenes $\mathbf{R}_{\mathrm{S}}$ by using features extracted from the feature extraction stage (*i.e.*, $\mathbf{F}_{\mathbf{M}}$ and $\mathbf{F}_{\mathbf{R}_{\mathrm{S}}}$ in Sec. 4.2.1 of the main paper). In detail, we utilize the auto-encoder loss $\mathcal{L}_{\mathrm{A}}$ [2] defined as follows:

$$\mathcal{L}_{\mathrm{A}} = \mathcal{L}_{\mathrm{rec}}(\mathbf{M}, \mathbf{M}^{\mathrm{est}}) + \mathcal{L}_{\mathrm{rec}}(\mathbf{R}_{\mathrm{S}}, \mathbf{R}_{\mathrm{S}}^{\mathrm{est}}), \tag{18}$$

where $\mathbf{M}^{\mathrm{est}}$ and $\mathbf{R}_{\mathrm{S}}^{\mathrm{est}}$ denote mixture images and reflection scenes obtained by the recovery module, and $\mathcal{L}_{\mathrm{rec}}$ measures the differences in the color and gradient domains between two images [2].

Training the complete network. After training the recovery modules, we train the complete network module as a whole. The reconstruction loss $\mathcal{L}_{\mathrm{recon}}$ proposed in Sec 4.3 of the main paper is retained to constrain the search space for estimating reflection layers and transmission scenes. Besides, we adopt the gradient prior loss $\mathcal{L}_{\mathrm{grad}}$ in [2] to leverage the independence of two estimated components (*i.e.*, $\mathbf{R}_{\mathrm{L}}^{\mathrm{est}}$ and $\mathbf{T}_{\mathrm{S}}^{\mathrm{est}}$) in the gradient domain. For exploiting the correlations of reflection scenes and layers, we use the reflection loss $\mathcal{L}_{\mathrm{ref}}$ in [2] which is defined as:

$$\mathcal{L}_{\mathrm{ref}} = \mathcal{L}_{\mathrm{mse}}(\mathbf{C}^{\mathrm{ref}}, \mathbf{R}_{\mathrm{L}}^{\mathrm{est}}) + \alpha\mathcal{L}_{\mathrm{mse}}(\mathbf{G}^{\mathrm{ref}}, \nabla\mathbf{R}_{\mathrm{L}}^{\mathrm{est}}), \tag{19}$$

where $\mathbf{C}^{\mathrm{ref}}$ and $\mathbf{G}^{\mathrm{ref}}$ denote reference images in the color domain and gradient domain (obtained by the reference image generation method of [2]), respectively, and we set α as 10 following [2]. In general, the total loss for training the complete network is defined as:

$$\mathcal{L}_{\mathrm{total}} = \omega_1\mathcal{L}_{\mathrm{recon}} + \omega_2\mathcal{L}_{\mathrm{grad}} + \omega_3\mathcal{L}_{\mathrm{ref}}. \tag{20}$$

Following previous methods [1], [2], the weights are empirically set as $\omega_1 = 1$, $\omega_2 = 3$, and $\omega_3 = 5$.

*Corresponding author.

- *Yuchen Hong, Lingran Zhao, and Boxin Shi are with the National Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China. Email: yuchenhong.cn@gmail.com, {calvinzhao, shiboxin}@pku.edu.cn.*
- *Qian Zheng is with the State Key Lab of Brain-Machine Intelligence, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China. Email: qianzheng@zju.edu.cn.*
- *Xudong Jiang and Alex C. Kot are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore. Email: {exdjiang, eackot}@ntu.edu.sg.*

References

[1] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *Proc. International Conference on Computer Vision (ICCV)*, 2021.
[2] Byeong-Ju Han and Jae-Young Sim. Zero-shot learning for reflection removal of single 360-degree image. In *Proc. European Conference on Computer Vision (ECCV)*, 2022.
[3] Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, Alex C Kot, and Boxin Shi. Panoramic image reflection removal. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2021.

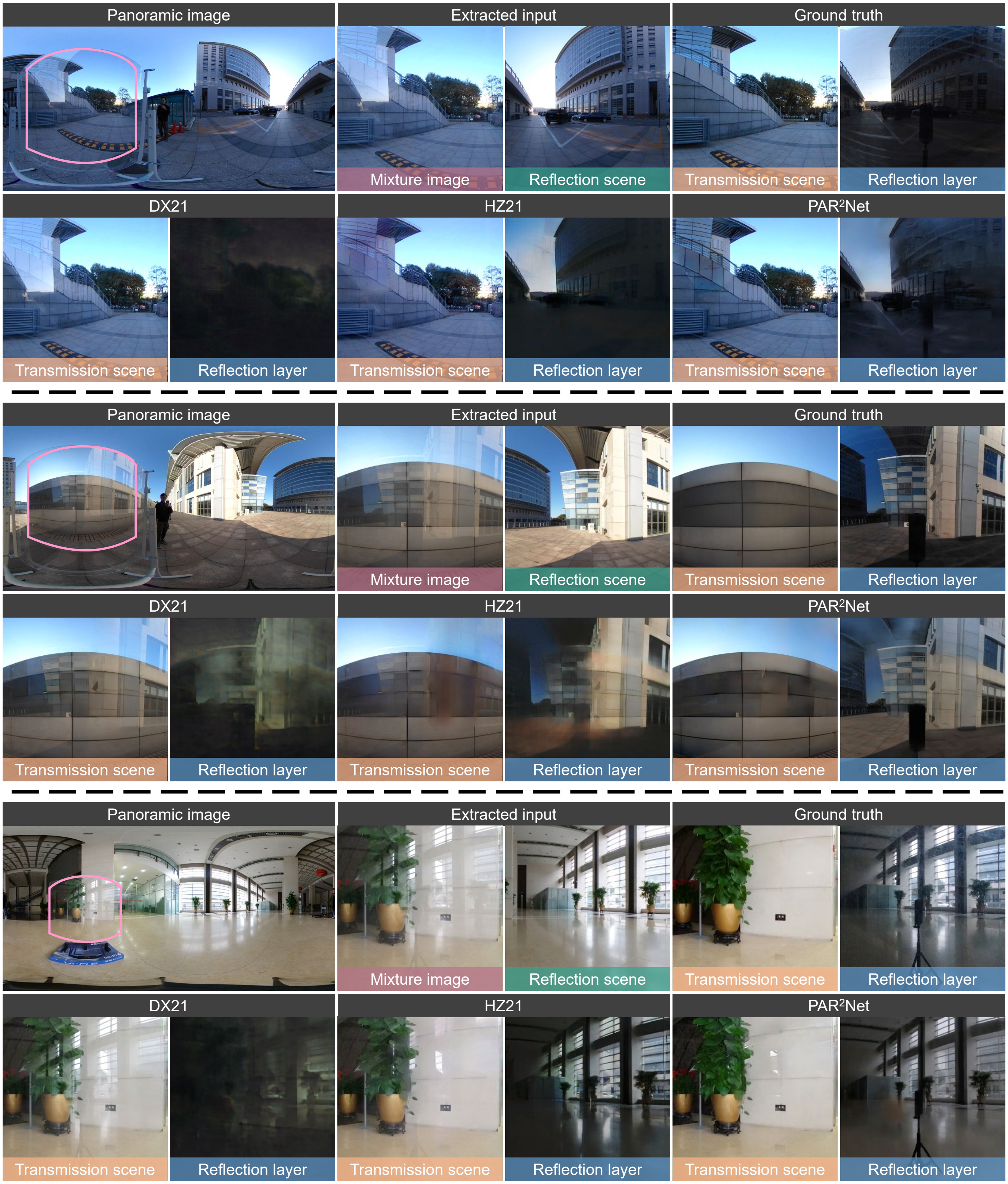

Fig. 13: More qualitative results on the PORTABLE dataset. Inputs and results are shown in the same manner as Fig. 10 of the main paper. Please zoom in for details.

Fig. 14: More qualitative results on the NATURAL dataset. Inputs and results are shown in the same manner as Fig. 11 of the main paper. Please zoom in for details.

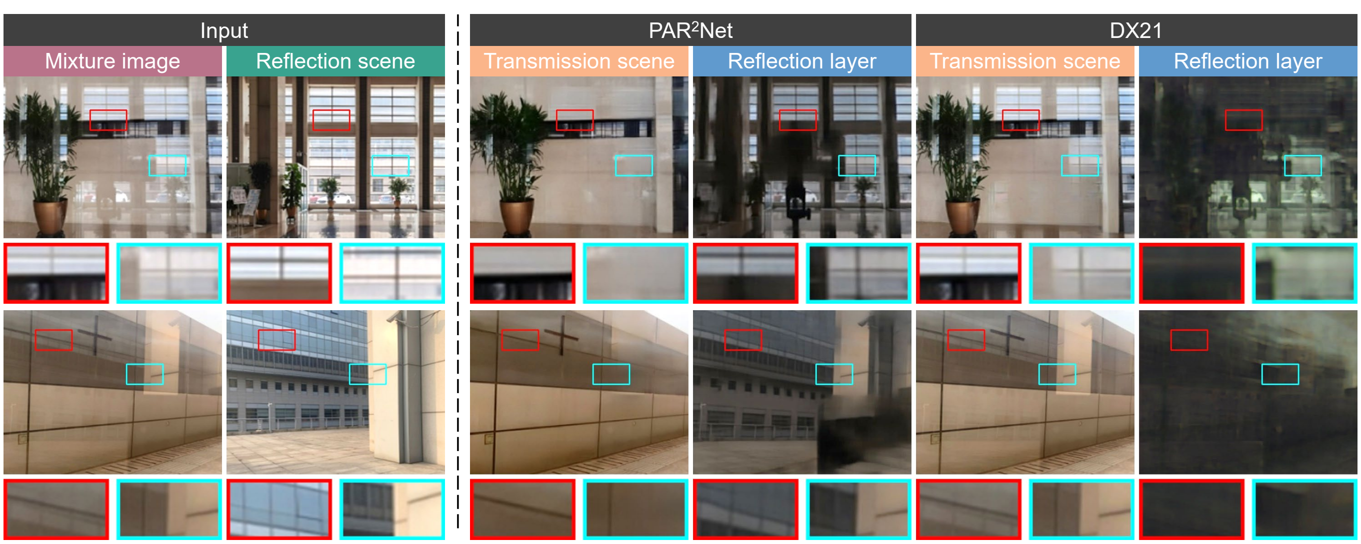

Fig. 15: More qualitative results on the PHONE dataset, compared with the state-of-the-art single-image method DX21 [1]. Close-up views are displayed at the bottom of images. Please zoom in for details.