# PAR$^2$Net: End-to-end Panoramic Image Reflection Removal

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

**Abstract**—In this paper, we investigate the problem of panoramic image reflection removal to relieve the content ambiguity between the reflection layer and the transmission scene. Although a partial view of the reflection scene is attainable in the panoramic image and provides additional information for reflection removal, it is not trivial to directly apply this for getting rid of undesired reflections due to its misalignment with the reflection-contaminated image. We propose an end-to-end framework to tackle this problem. By resolving misalignment issues with adaptive modules, high-fidelity recovery of the reflection layer and the transmission scenes are accomplished. We further propose a new data generation approach that considers the physics-based formation model of mixture images and the in-camera dynamic range clipping to diminish the domain gap between synthetic and real data. Experimental results demonstrate the effectiveness of the proposed method and its applicability for mobile devices and industrial applications.

**Index Terms**—Reflection removal, panoramic image, deep learning

## 1 Introduction

When taking photos in front of semi-reflectors like glass windows, photographers prevailingly attempt to capture transmission scenes behind the glass, while reflection contamination often degrades the image quality. Consequently, single-image reflection removal has become an attractive topic in computational photography [4], [5], [15], [22], [47], [64], which aims at removing undesirable glass reflections and recovering the clear transmission scene from a contaminated mixture image. The mixture image (denoted as $\mathbf{M}$) can be considered as the combination of two components: the transmission scene (denoted as $\mathbf{T}_\mathrm{S}$) and the reflection layer (denoted as $\mathbf{R}_\mathrm{L}$) [4], [22]. The major challenge of such an ill-posed layer separation problem is that both transmission scenes and reflection layers are from natural scenes, whose image content can be arbitrary, arousing the difficulty of differentiating the dominant content for mixture images. We call it *content ambiguity* in this paper.

To address this challenging task, handcrafted content-free priors derived from natural image statistics and image formation models, *e.g.*, the gradient sparsity [21], relative smoothness [24], and ghosting effects [37], are adopted as useful constraints by non-learning methods, while performances of these methods decrease significantly when the priors they depend on are not observed, *e.g.*, scenar-

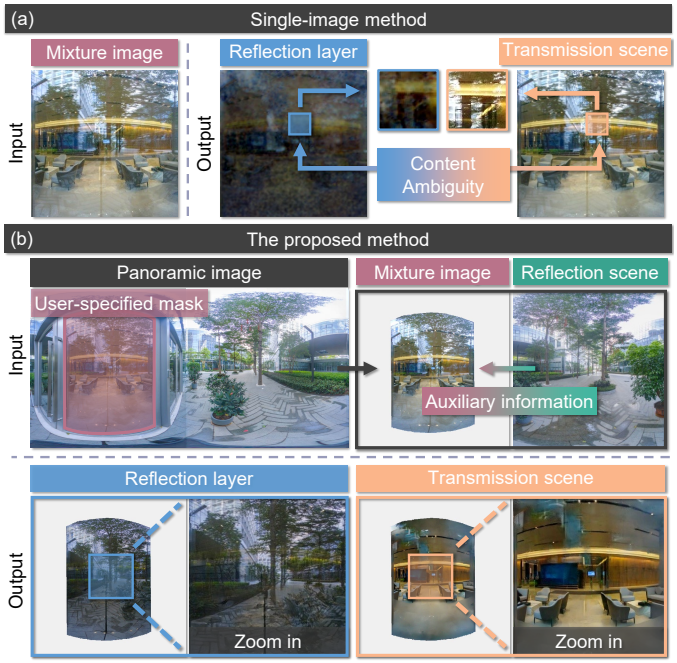


Fig. 1: Illustration of different setups of the single-image reflection removal method (we show the result from [4] as an example) and the proposed method. (a) The single-image method only inputs with a mixture image, and fails in distinguishing the reflection layer and transmission scene due to the content ambiguity. (b) The proposed method inputs with a panoramic image and a user-specified mask, which utilizes auxiliary reflection content information to relieve the content ambiguity, and recovers the much clearer reflection layer and transmission scene than (a).

*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.*

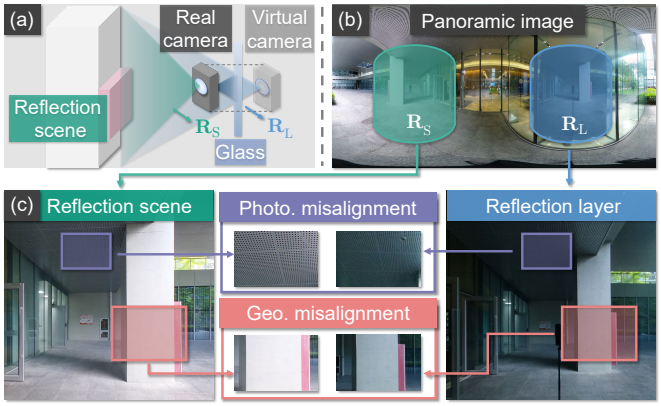


Fig. 2: (a) Illustration of photographing by a panoramic camera in a scene containing a piece of glass. The reflection scene $\mathbf{R}_S$ is captured by the real camera. The reflection layer $\mathbf{R}_L$ can be regarded as captured by the virtual camera. (b) The panoramic image which captures both $\mathbf{R}_S$ and $\mathbf{R}_L$ in a single shot. (c) Photometric and geometric misalignments between the reflection scene $\mathbf{R}_S$ and reflection layer $\mathbf{R}_L$ (images are extracted from the panoramic image for better visualization).

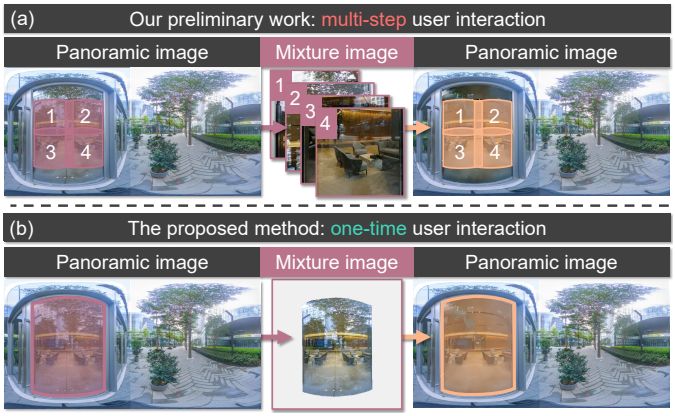


Fig. 3: Illustration of required user interactions of (a) our preliminary work [12] and (b) the proposed method in this paper. Our preliminary work [12] requires iteratively multi-step labeling for post-processing, while the proposed method relaxes requirements on user interactions by using a one-time user-specified mask to handle larger glass regions in a more complete manner.

ios where reflections are with sharp edges. Thanks to the rapid development of deep learning, the latest single-image methods [4], [22] leverage the strong modeling capacity of neural networks to implicitly learn priors from a large scale of training data. However, as the image content of transmission scenes and reflection layers can be rather irrelevant, existing single-image methods are still likely to fail in distinguishing the two components. As shown in Fig. 1(a), when reflections in the mixture image $\mathbf{M}$ are with high intensities and complex textures, the single-image solution may not correctly tell apart the reflection layer from the transmission scene due to high content ambiguity.

To relieve the content ambiguity in a mixture image with reflection contamination, effective constraints should be considered to apply to reflections. A direct solution is to introduce an additional view of the reflection scene, since it provides partial content information about the reflection layer. Fortunately, a panoramic image has a 360° field-of-view (FoV), which naturally contains the reflection-contaminated glass region and an additional view of the reflection scene, providing useful cues for the relief of the content ambiguity.

Though an additional view of the reflection scene is attainable in a panoramic image, it is not trivial to automatically and accurately identify reflections in contaminated regions. As shown in Fig. 2, the major challenge is that the reflection layer $\mathbf{R}_L$ observed from glass and the reflection scene (denoted as $\mathbf{R}_S$) directly captured by the camera are not equivalent, since there exist *geometric and photometric misalignments* between them. According to the law of reflection, a virtual camera can be assumed to locate at the virtual point, which is symmetric with the real camera by the plane of the glass. Though the real and virtual cameras both capture the reflection scene, the captured image contents are geometrically misaligned at the image plane since they are viewed from different positions. Meanwhile, lights attenuate when reflecting at the glass surface [67], causing the reflection layer $\mathbf{R}_L$ (with attenuation) to be obviously darker than the reflection scene $\mathbf{R}_S$ (without attenuation). As a consequence, the geometric and photometric misalignments between the reflection layer $\mathbf{R}_L$ and the reflection scene $\mathbf{R}_S$ hinder the utilization of auxiliary reflection content information for panoramic image reflection removal.

To address above issues, our preliminary work [12] proposes the first *two-step* panoramic image-based reflection removal method focusing on tackling the content ambiguity. Image patches identified from user interactions are rectified from panoramic images and utilized as mixture images $\mathbf{M}$ and reflection scenes $\mathbf{R}_S$. By applying the coarse-to-fine reflection alignment and learning-based transmission recovery in a two-step manner, reflection layers $\mathbf{R}_L$ and transmission scenes $\mathbf{T}_S$ are recovered. However, such a pipeline owns several drawbacks: **1) Multi-step user interaction.** As shown in Fig. 3(a), for panoramic images where glass occupies large regions, more than one patches need to be iteratively rectified to cover the whole glass regions, which increases the computational cost and requests many times of user interaction. **2) Fragile alignment.** The coarse reflection alignment, including the photometric alignment by global polynomial function fitting and the geometric alignment by patch-level matching, is unstable since it is sensitive to hyperparameters, which may degrade the performance of the whole framework if incorrect alignment is calculated. **3) Large domain gap.** The employed data generation procedure [5] lacks the consideration of the physics-based reflection formation model and the in-camera dynamic range clipping, resulting in the domain gap between synthetic and real data, which weakens the generalization capacity of the trained model.

In this paper, we extend our preliminary work [12] to an end-to-end **PA**noramic image **R**eflection **R**emoval **Net**work (PAR$^2$Net) with three improved advantages: **1) One-time user interaction.** As shown in Fig. 3(b), PAR$^2$Net takes in a single panoramic image and a single-step user-specified mask with an arbitrary shape, which avoids repeated user interactions. **2) Robust alignment.** We resolve photometric

and geometric misalignment issues by elaborately improving network modules with an adaptive strategy, which ensures the higher stability of the proposed method. **3) Small domain gap**. From the perspective of the physics-based and high dynamic range (HDR) image formation model, we analyze the formation procedure of mixture images with reflection contaminations and further propose a data generation approach for synthesizing mixture images more realistically to diminish the domain gap in data and facilitate network training. Experiments on real data demonstrate that PAR$^2$Net not only outperforms our preliminary work [12] and state-of-the-art single-image methods (Fig. 1(b) shows an example with [4]), but also generalizes well to limited-FoV images captured without panoramic cameras. Our contributions are summarized as follows:

- We present the first *end-to-end* framework to relieve the content ambiguity for reflection removal by using panoramic images.
- We employ an adaptive strategy to alleviate the impact of photometric and geometric misalignments between reflection scenes and layers, which contributes to high-fidelity transmission recovery.
- We propose a realistic data generation approach by considering the physics-based and high dynamic range image formation model to diminish the domain gap in data and facilitate panoramic image reflection removal.

The remainder of this paper is organized as follows. In Sec. 2, we start with an introduction of existing reflection removal methods and applications of panoramic images. Then Sec. 3 analyzes misalignment issues between reflection scenes and layers. Sec. 4 introduces the network architecture with objective functions. We present details of our data generation method and our dataset in Sec. 5. Quantitative and qualitative experiments and ablation studies are conducted in Sec. 6. Finally, we conclude the paper in Sec. 7.

## 2 Related Work

A panoramic image is generated by stitching multiple images from different viewpoints, but the overlap and correspondence information across different viewpoints have been lost after merging the panoramic image, so it cannot provide motion cues [23], [27], [58], parallax cues [35], polarization information [17], [20], [28], [29], [56], or reflection-free contextual information by using active light sources [3], [11], [19] for multi-image reflection removal. Moreover, since a panoramic image can be handily captured in a single shot, we still focus on the discussion of single-image reflection removal methods because they address similar technical problems as panoramic image reflection removal. We refer readers to [48] for a comprehensive and up-to-date survey on reflection removal.

### 2.1 Reflection removal

Existing methods for single-image reflection removal rely on the assumption of different distributions of transmission scenes and reflection layers, *i.e.*, reflection layers are likely to be more blurry and with lower intensity compared with transmission scenes [45]. Traditional methods formulate this assumption in their optimization pipeline, *e.g.*, gradient sparsity priors [21], smoothness priors [23], [24], ghosting cues [37], image content [44], and penalty on the gradient of recovered transmission scenes [2], [60].

Learning-based methods are developed to generalize the knowledge learned from training data. CEILNet [5] adopts the traditional two-stage framework which predicts edge maps and transmission scenes successively. Zhang *et al.* [64] propose a neural network with perceptual loss to emphasize the independence of transmission scenes and reflection layers in the gradient domain. CRRN [46] and CoRRN [47] combine the gradient inference and the image inference in one unified mechanism to remove reflections concurrently. ERRNet [54] embeds context modules in the network and exploits the unaligned data to enhance the generality of the model. Wen *et al.* [55] synthesize mixture images with learned non-linear blending masks and accomplish reflection removal based on such non-linearity. LBCLN [22] proposes a cascaded refinement approach with the convolutional LSTM network structure to refine the estimation of transmission scenes and reflection layers iteratively. Kim *et al.* [15] generate data with physically-based rendering and restore the transmission scenes considering the various impacts of glass and lens. Dong *et al.* [4] propose to introduce a probabilistic reflection confidence map for telling regions to be reflection-dominated or transmission-dominated. Wang *et al.* [49] consider commonly-confronted colored glass and propose to recover transmission scenes from mixture images taken in front of colored glass. CGDNet [65] proposes a model-driven deep network architecture for reflection removal which unfolds the iterative steps of the optimization algorithm into network layers.

### 2.2 Applications of panoramic images

Thanks to the $360^\circ$ FoV, panoramic images are useful in various computer vision applications. Panoramic images provide complete observation for geometry layouts of scenes, so there are methods studying scene understanding from a single panoramic image, *e.g.*, indoor layout estimation [42], [50], [59], indoor depth reconstruction [1], [34], [43], and vehicle detection [18]. Panoramic images also provide complete observation for environment maps as lighting representation. Some research attempts to recover the environment map only from a partial observation, *e.g.*, a 3D structure and a probability distribution of semantic labels from an RGB-D image [40], lighting represented by an HDR panoramic image for either indoor [25], [39] or outdoor [10], [62] scenarios. Our preliminary work [12] proposes the first two-step solution to investigate how partial views of the reflection scene in a single panoramic image could be utilized to relieve the content ambiguity of reflection removal. Han *et al.* [6] adopt the zero-shot learning scheme to leverage auxiliary contextual information in reflection scenes for panoramic image reflection removal. In this paper, we simplify user interactions compared with our preliminary work [12] and propose a unified framework to achieve high-fidelity reflection removal and transmission recovery in panoramic images.

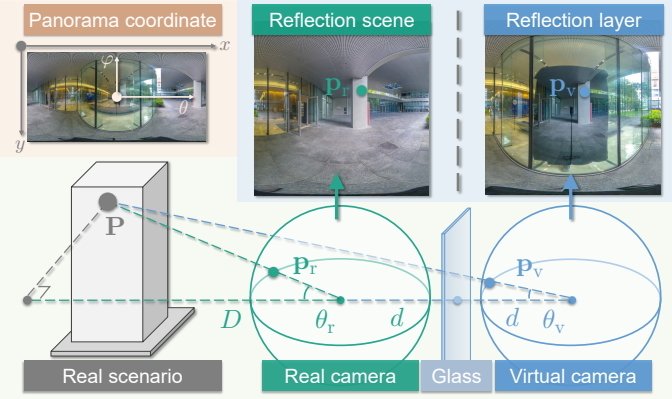


Fig. 4: Illustration about the formation of the geometric misalignment between the reflection scene $\mathbf{R}_{\mathrm{S}}$ and reflection layer $\mathbf{R}_{\mathrm{L}}$.

## 3 Problem Formulation

Given a panoramic image containing a glass region with reflection contaminations, we denote the contaminated region as the mixture image $\mathbf{M}$, which can be formulated as [4]:

$$\mathbf{M} = \mathbf{\Omega} \odot \mathbf{T}_{\mathrm{S}} + \mathbf{R}_{\mathrm{L}}, \tag{1}$$

where $\odot$ is the element-wise multiplication operator, $\mathbf{\Omega}$ is a spatially-varying coefficient map, and $\mathbf{T}_{\mathrm{S}}$ and $\mathbf{R}_{\mathrm{L}}$ are the corresponding transmission scene (if there is no glass) and reflection layer (if there is a piece of black cloth behind the glass), respectively. Besides, thanks to the 360° FoV of the panoramic camera, a view of the reflection scene is also captured, and we denote it as $\mathbf{R}_{\mathrm{S}}$. Intuitively, $\mathbf{R}_{\mathrm{S}}$ can be utilized straightforward to facilitate reflection removal since it provides contextual information about $\mathbf{R}_{\mathrm{L}}$. However, as shown in Fig. 2(b), there exists photometric and geometric misalignment between $\mathbf{R}_{\mathrm{S}}$ and $\mathbf{R}_{\mathrm{L}}$, thus bringing challenges for the exploitation of the auxiliary content. The following sections will introduce the causes of such geometric and photometric misalignment between $\mathbf{R}_{\mathrm{S}}$ and $\mathbf{R}_{\mathrm{L}}$.

### 3.1 Geometric misalignment

As illustrated in Fig. 4, assuming a three-dimensional point $\mathbf{P}$ in a real scenario, which is captured by both the real and virtual camera (corresponding to the reflection scene $\mathbf{R}_{\mathrm{S}}$ and reflection layer $\mathbf{R}_{\mathrm{L}}$), and its projected points in $\mathbf{R}_{\mathrm{S}}$ and $\mathbf{R}_{\mathrm{L}}$ are $\mathbf{p}_{\mathrm{r}}$ and $\mathbf{p}_{\mathrm{v}}$, respectively. The geometric misalignment between $\mathbf{R}_{\mathrm{S}}$ and $\mathbf{R}_{\mathrm{L}}$ can be represented by spatial pixel shifts between corresponding points like $\mathbf{p}_{\mathrm{r}}$ and $\mathbf{p}_{\mathrm{v}}$, which are actually caused by different viewpoints of the real and virtual camera. Since a panoramic image can be considered as a sphere [41], the pixel shift $\triangle x$ (along the horizontal axis for example) between points $\mathbf{p}_{\mathrm{r}}$ and $\mathbf{p}_{\mathrm{v}}$ can be formulated as:

$$\begin{aligned} \triangle x &= W \cdot \frac{\theta_{\mathrm{r}} - \theta_{\mathrm{v}}}{2\pi} \\ &= W \cdot \frac{\arctan[\frac{D+2d}{D} \cdot \tan(\theta_{\mathrm{v}})] - \theta_{\mathrm{v}}}{2\pi}, \end{aligned} \tag{2}$$

where $W$ is the width of the panoramic image, $D$ and $d$ denote the distance from the real camera to the real scenario and the glass plane, and $\theta_{\mathrm{r}}$ and $\theta_{\mathrm{v}}$ denote the azimuthal angle of $\mathbf{p}_{\mathrm{r}}$ and $\mathbf{p}_{\mathrm{v}}$ in the spherical coordinate, respectively.

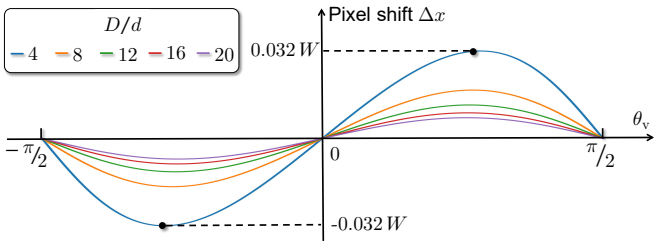


Fig. 5: Curves that plot the relationship between the pixel shift $\triangle x$ (along the horizontal axis for example) with different azimuthal angles $\theta_{\mathrm{v}}$ of points $\mathbf{p}_{\mathrm{v}}$ in $\mathbf{R}_{\mathrm{L}}$ under different ratios of $D/d$.

By assuming that the camera-scene distance $D$ is much larger than the camera-glass distance $d$ (*i.e.*, $D/d \geqslant 4$ as in [12]), we plot the curves of the pixel shift $\triangle x$ in Fig. 5. Since the curves are symmetric about the origin, we pick the first quadrant for analysis. When $D/d$ is fixed, curve tends to increase first and then decrease. When $D/d$ increases, the misalignment between $\mathbf{p}_{\mathrm{r}}$ and $\mathbf{p}_{\mathrm{v}}$ becomes smaller. We select the maximum value of the curve $D/d = 4$, *i.e.*, $0.032W$, as the maximum pixel shift between $\mathbf{p}_{\mathrm{r}}$ and $\mathbf{p}_{\mathrm{v}}$ (*e.g.*, for a panoramic image with the resolution of $512 \times 1024$ pixels, the assumed maximum pixel shift is about 32 pixels). This indicates the potential geometric misalignment between corresponding points of $\mathbf{R}_{\mathrm{S}}$ and $\mathbf{R}_{\mathrm{L}}$ in panoramic images, and is used as the reference bound for the design of our alignment mechanism (introduced in Sec. 4.2.3).

### 3.2 Photometric misalignment

Suppose a situation in which the camera-scene distance is far enough and the reflection scene $\mathbf{R}_{\mathrm{S}}$ and reflection layer $\mathbf{R}_{\mathrm{L}}$ are perfectly aligned in geometry, then we define the photometric misalignment as the pixel-wise discrepancy in intensities between $\mathbf{R}_{\mathrm{S}}$ and $\mathbf{R}_{\mathrm{L}}$. The photometric misalignment is mainly caused by the intensity attenuation when lights reflect from the glass surface [67], turning $\mathbf{R}_{\mathrm{S}}$ much brighter than $\mathbf{R}_{\mathrm{L}}$, and we formulate the photometric misalignment by using a spatially-varying coefficient map $\mathbf{\Phi}$ [28] as follows:

$$\mathbf{R}_{\mathrm{L}} = \mathbf{\Phi} \odot \mathbf{R}_{\mathrm{S}}. \tag{3}$$

## 4 Proposed Method

### 4.1 Input setting

Different from the inputs of single-image reflection removal methods [4], [22], [65] which consider the whole images to be reflection-contaminated, a piece of plate glass in a panoramic image can at most occupy a half of the image region (when the surface area of the glass is infinitely large) due to the 360° FoV of panoramic cameras. Therefore, labeling glass regions in panoramic images is a necessary preprocessing operation before conducting reflection removal[1], and we introduce user interaction to label the reflection-contaminated glass regions by using the `labelme` toolbox[2].

1. Our focus is reflection removal, and detecting the glass region (*e.g.*, GDNet [30]) is beyond the scope.
2. https://github.com/wkentaro/labelme

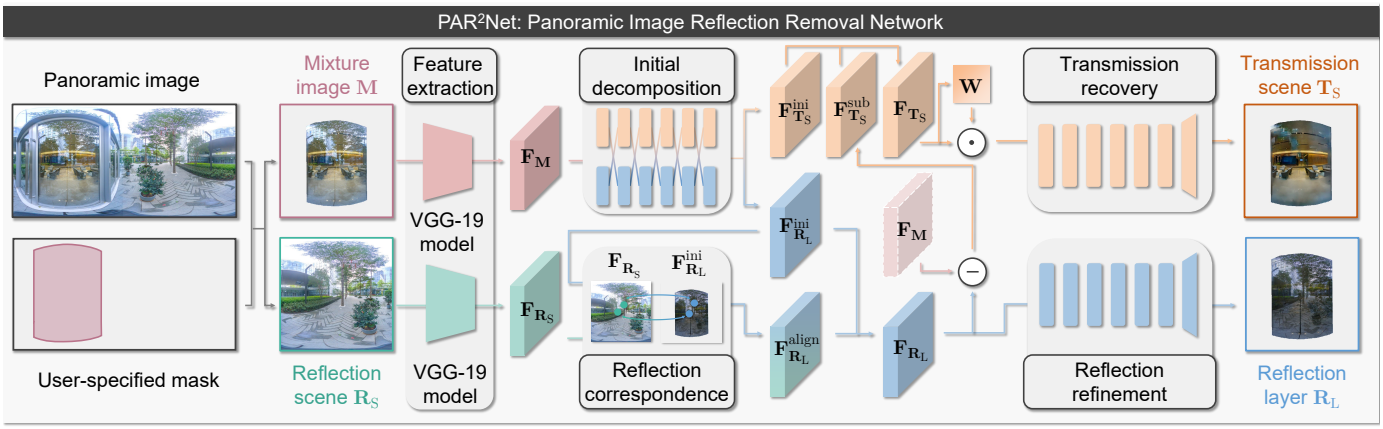


Fig. 6: Framework of the proposed method (PAR$^2$Net), which is designed in an end-to-end manner. Inputting with a panoramic image and a mask indicating the glass region (only requires one-time user interaction), PAR$^2$Net uses the mixture image $\mathbf{M}$ and the reflection scene $\mathbf{R}_\mathrm{S}$ to recover of the reflection layer $\mathbf{R}_\mathrm{L}$ and transmission scene $\mathbf{T}_\mathrm{S}$ via feature extraction (in Sec. 4.2.1), initial decomposition (in Sec. 4.2.2), reflection correspondence (in Sec. 4.2.3), reflection refinement (in Sec. 4.2.4), and transmission recovery (in Sec. 4.2.5).

For a panoramic image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ (where $W = 2H$), a binary mask with arbitrary shapes to indicate the glass region is generated from user interaction, then we multiply the panoramic image with the mask and crop the result as the mixture image $\mathbf{M} \in \mathbb{R}^{H\times H\times 3}$. Naturally, the reflection scene $\mathbf{R}_\mathrm{S} \in \mathbb{R}^{H\times H\times 3}$ locates at the rest half of the panoramic image, so we crop $\mathbf{R}_\mathrm{S}$ with a horizontal flipping operation to diminish the influence of symmetry. Note that unlike our preliminary work [12] which requires multi-step user interactions (in Fig. 3(a)), the proposed method only needs one-time user interaction (in Fig. 3(b)), which relaxes the requirement on users and tackles larger glass regions. The design methodology of the proposed method (in Sec. 4.2) and the optimization with complementary loss functions (in Sec. 4.3) are as follows.

## 4.2 Network architecture

Considering the image formation model of mixture images with the geometric and photometric misalignment between reflection scenes and layers, we propose a unified framework, *i.e.*, PAR$^2$Net, to recover reflection layers and transmission scenes in an end-to-end manner. As the network architecture shown in Fig. 6, we use $\mathbf{M}$ and $\mathbf{R}_\mathrm{S}$ as inputs and recover the reflection layer $\mathbf{R}_\mathrm{L}$ and transmission scene $\mathbf{T}_\mathrm{S}$ via feature extraction (in Sec. 4.2.1), initial decomposition (in Sec. 4.2.2), reflection correspondence (in Sec. 4.2.3), reflection refinement (in Sec. 4.2.4), and transmission recovery (in Sec. 4.2.5). Compared with our preliminary work [12], the initial decomposition stage is added to facilitate its subsequent stages, the reflection correspondence stage adopts an adaptive module to replace the "coarse alignment" stage in [12] and turns the new framework end-to-end, and other stages are adjusted accordingly to fit the new framework.

### 4.2.1 Feature extraction

The multi-level image feature pyramids of inputs (*i.e.*, the mixture image $\mathbf{M}$ and the reflection scene $\mathbf{R}_\mathrm{S}$) are firstly extracted by the widely-used VGG-19 network [38], whose last four layers (*i.e.*, three fully connected layers and a Softmax layer) are removed to adapt to our image-to-image translation problem. The extracted feature pyramids are then transformed into hypercolumn features [7], which have been proved to be effective in aggregating contextual information for reflection removal [54], [64]. To balance the efficiency and effectiveness, hypercolumn features are condensed by a convolutional block with a $1\times 1$ kernel, and then downsampled by another convolutional block with a $3\times 3$ kernel. Each convolutional block is composed of a convolution layer and an activation layer using the ReLU function [31]. We denote the extracted features from mixture images and reflection scenes as $\mathbf{F}_\mathbf{M}$ and $\mathbf{F}_{\mathbf{R}_\mathrm{S}}$, respectively.

### 4.2.2 Initial decomposition

Mixture images are blended with contextual information from both two components, *i.e.*, reflection layers and transmission scenes, thus directly using mixture image features $\mathbf{F}_\mathbf{M}$ and reflection scene features $\mathbf{F}_{\mathbf{R}_\mathrm{S}}$ to seek their mutual correspondence on reflection contents suffers from the interference of transmission contents. Intuitively, it is easier to find correspondence between $\mathbf{R}_\mathrm{S}$ and $\mathbf{R}_\mathrm{L}$ than between $\mathbf{R}_\mathrm{S}$ and $\mathbf{M}$. Therefore, the process for seeking reflection correspondence can be facilitated by using features refined from $\mathbf{F}_\mathbf{M}$ where reflection contents dominate and transmission contents are suppressed. To conduct the initial decomposition of reflection layers and transmissions scenes in the feature domain, we apply a two-stream module composed of YTMT blocks [13], which exploit the additive property of two components and force communications between the two streams.

Each YTMT block [13] is composed of two branches, one corresponding to reflection layers and the other corresponding to transmission scenes. For each branch, a convolutional layer with a $3\times 3$ kernel is first utilized to process features from the corresponding branch of the former YTMT block. The processed features are then passed through an activation layer with a ReLU function and another with a negative ReLU function [13], obtaining activated features (beneficial for the current branch) and deactivated features (will be

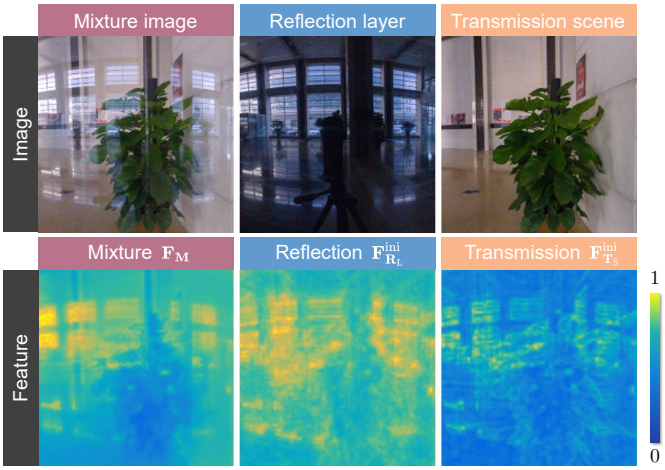


Fig. 7: A visual example of the initial decomposition. First row: the mixture image, reflection layer, and transmission scene. Second row: the mixture image feature $\mathbf{F}_{\mathbf{M}}$ and initially decomposed features of the reflection layer and transmission scene, *i.e.*, $\mathbf{F}^{\text{ini}}_{\mathbf{R}_\text{L}}$ and $\mathbf{F}^{\text{ini}}_{\mathbf{T}_\text{S}}$, respectively.

discarded by the current branch but contain auxiliary information for the other branch), respectively. Next, the two branches exchange deactivated features. Activated features of a branch are merged with deactivated features from the other branch by channel concatenation, and the merged features are then fused and condensed by a convolutional layer with a $1 \times 1$ kernel. Finally, a channel and spatial attention module [57] is employed to re-weight the fused features. For the first YTMT block, we use the mixture image feature $\mathbf{F}_{\mathbf{M}}$ as the same input for both two branches, and the following YTMT blocks are stacked successively. After six YTMT blocks, the initial decomposition in the feature domain is achieved, and we denote the decomposed features of reflection layers and transmission scenes as $\mathbf{F}^{\text{ini}}_{\mathbf{R}_\text{L}}$ and $\mathbf{F}^{\text{ini}}_{\mathbf{T}_\text{S}}$, respectively. An example that visualizes the initial decomposition is shown in Fig. 7. As can be observed, strong reflections are initially decomposed with transmission scenes in the feature domain, which is benefited from the capacity of YTMT blocks [13] to exploit complementary information between two branches.

#### 4.2.3 Reflection correspondence

As discussed in Sec. 3, geometric and photometric misalignments commonly exist between reflection scenes $\mathbf{R}_\text{S}$ and reflection layers $\mathbf{R}_\text{L}$, which can be formulated as:

$$\mathbf{R}_\text{L} = \mathbf{\Phi} \odot \mathcal{F}_{S \to L}(\mathbf{R}_\text{S}), \tag{4}$$

in which $\mathcal{F}_{S \to L}(\cdot)$ presents a pixel-wise geometric transformation operation, and $\mathbf{\Phi}$ is the spatially-varying coefficient map describing the intensity attenuation when lights reflect at the glass surface, which is the same as in Eqn. (3). To tackle misalignment issues and exploit correspondences between reflection scenes and layers in the feature domain, a correspondence module [61] is employed, whose mechanism is similar to the widely-used self-attention module [51] that exploits non-local correlations inside or between images. Besides, the process of utilizing the correspondence module to address misalignment issues exactly matches the formulation in Eqn. (5), which will be discussed as follows.

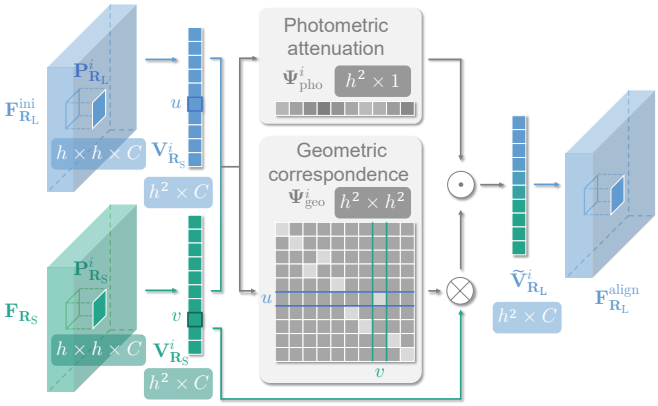


Fig. 8: Overview of the reflection correspondence module, which exploits the initially decomposed features $\mathbf{F}^{\text{ini}}_{\mathbf{R}_\text{L}}$ and reflection scene features $\mathbf{F}_{\mathbf{R}_\text{S}}$ to achieve geometric and photometric alignments in the feature domain.

Inputs of the correspondence module are features of the reflection scene (*i.e.*, $\mathbf{F}_{\mathbf{R}_\text{S}}$) and the initially decomposed features of the reflection layer (*i.e.*, $\mathbf{F}^{\text{ini}}_{\mathbf{R}_\text{L}}$). To narrow the searching space for corresponding points and reduce the computational cost, we apply the correspondence module on patches instead of the whole features, whose mechanism is shown in Fig. 8. $\mathbf{F}_{\mathbf{R}_\text{S}}$ and $\mathbf{F}^{\text{ini}}_{\mathbf{R}_\text{L}}$ are firstly divided into $N$ overlapped square patches, and the patches are denoted as $\mathbf{P}^i_{\mathbf{R}_\text{S}}$, $\mathbf{P}^i_{\mathbf{R}_\text{L}} \in \mathbb{R}^{h \times h \times C}$, respectively, where $1 \leqslant i \leqslant N$ represents the index of the current patch, $h$ is the side length of patches, and $C$ is the number of channels. The overlapped area between two adjacent patches is set as $h^2/2$ (*i.e.*, the overlapped region is a rectangle of size $h \times \frac{h}{2}$). According to Eqn. (2) and Fig. 5 in Sec. 3.1, the maximum pixel shift between corresponding points is $0.032W$ (*i.e.*, $0.064H$, where $W$, $H$ are the width and height of the panoramic image, respectively) under the assumtion that the camera-scene distance is much larger than the camera-glass distance, so we set the side length of patches as $h = 0.4H$ to cover the potential pixels shift. Before calculating the pixel-wise correspondence, $\mathbf{P}^i_{\mathbf{R}_\text{S}}$ and $\mathbf{P}^i_{\mathbf{R}_\text{L}}$ are processed by convolutional blocks with $1 \times 1$ kernels for the alignment in the feature domain [63] and converted into vectors $\mathbf{V}^i_{\mathbf{R}_\text{S}}$, $\mathbf{V}^i_{\mathbf{R}_\text{L}} \in \mathbb{R}^{h^2 \times C}$, respectively. Then a correspondence matrix $\mathbf{\Psi}^i \in \mathbb{R}^{h^2 \times h^2}$ is calculated as follows:

$$\mathbf{\Psi}^i(u, v) = \frac{\hat{\mathbf{V}}^i_{\mathbf{R}_\text{L}}(u) \hat{\mathbf{V}}^i_{\mathbf{R}_\text{S}}(v)^\top}{\|\hat{\mathbf{V}}^i_{\mathbf{R}_\text{L}}(u)\| \|\hat{\mathbf{V}}^i_{\mathbf{R}_\text{S}}(v)\|}, \tag{5}$$

where $\hat{\mathbf{V}}^i_{\mathbf{R}_\text{L}}(u)$ and $\hat{\mathbf{V}}^i_{\mathbf{R}_\text{S}}(v) \in \mathbb{R}^C$ denote centralized vectors of $\mathbf{V}^i_{\mathbf{R}_\text{L}}$ and $\mathbf{V}^i_{\mathbf{R}_\text{S}}$ at position $u$ and $v$ (a position in the vector corresponds to a pixel in the patch), respectively.

Though the correspondence matrix $\mathbf{\Psi}^i$ integrates correspondences between $\mathbf{V}^i_{\mathbf{R}_\text{L}}$ and $\mathbf{V}^i_{\mathbf{R}_\text{S}}$, it cannot be utilized directly. Intuitively, since the reflection scenes and layer observe the same scene from different viewpoints, correspondences between $\mathbf{V}^i_{\mathbf{R}_\text{L}}$ and $\mathbf{V}^i_{\mathbf{R}_\text{S}}$ should be sparse, *i.e.*, for each position in $\mathbf{V}^i_{\mathbf{R}_\text{L}}$ (corresponding to each pixel in $\mathbf{P}^i_{\mathbf{R}_\text{L}}$), there should be only a few positions in $\mathbf{V}^i_{\mathbf{R}_\text{S}}$ as its correspondences (considering the defocus blur and ghosting effects). However, $\mathbf{\Psi}^i$ describes the correspondence between each position in $\mathbf{V}^i_{\mathbf{R}_\text{L}}$ and all positions in $\mathbf{V}^i_{\mathbf{R}_\text{S}}$ (*i.e.*, all pixels

in the image patch $\mathbf{P}^i_{\mathbf{R}_\mathrm{S}}$), which generates dense while inaccurate correspondences. Therefore, we decompose $\mathbf{\Psi}^i$ into a geometric correspondence matrix $\mathbf{\Psi}^i_{\mathrm{geo}} \in \mathbb{R}^{h^2 \times h^2}$ and a photometric attenuation matrix $\mathbf{\Psi}^i_{\mathrm{pho}} \in \mathbb{R}^{h^2 \times 1}$:

$$\mathbf{\Psi}^i_{\mathrm{geo}}(u,v) = \underset{u}{\mathrm{softmax}}(\mathbf{\Psi}^i(u,v)/\tau), \tag{6}$$

$$\mathbf{\Psi}^i_{\mathrm{pho}}(v) = \max_u(\mathbf{\Psi}^i(u,v)), \tag{7}$$

where $\tau$ is set as $0.01$ to ensure values of the row vector $\mathbf{\Psi}^i_{\mathrm{geo}}(u,\cdot)$ to be sparse, corresponding to the aforementioned sparse correspondences between $\mathbf{V}^i_{\mathbf{R}_\mathrm{L}}$ and $\mathbf{V}^i_{\mathbf{R}_\mathrm{S}}$. Then the geometric and photometric alignments can be conducted through:

$$\tilde{\mathbf{V}}^i_{\mathbf{R}_\mathrm{L}}(u) = \sum_v \mathbf{\Psi}^i_{\mathrm{pho}}(v) \cdot \mathbf{\Psi}^i_{\mathrm{geo}}(u,v) \cdot \mathbf{V}^i_{\mathbf{R}_\mathrm{S}}(v), \tag{8}$$

which can also be written as:

$$\tilde{\mathbf{V}}^i_{\mathbf{R}_\mathrm{L}} = \mathbf{\Psi}^i_{\mathrm{pho}} \odot (\mathbf{\Psi}^i_{\mathrm{geo}} \otimes \mathbf{V}^i_{\mathbf{R}_\mathrm{S}}), \tag{9}$$

where $\odot$ and $\otimes$ denote the element-wise and matrix multiplication, respectively.

Coincidentally, if considering $\mathbf{\Psi}^i_{\mathrm{pho}}$ as $\mathbf{\Phi}$ and $\mathbf{\Psi}^i_{\mathrm{geo}}$ as $\mathcal{F}_{S\to L}$, then Eqn. (9) can be regarded as an approximation of Eqn. (4), which shows the consistency between the employed correspondence module and the reflection formulation process. Finally, the transformed vectors $\tilde{\mathbf{V}}^i_{\mathbf{R}_\mathrm{L}}$ are reshaped to patches and merged (with the overlapped regions being averaged) to obtain the final aligned features of the reflection layer (denoted as $\mathbf{F}^{\mathrm{align}}_{\mathbf{R}_\mathrm{L}}$), which will facilitate the following recovery of the reflection layer.

#### 4.2.4 Reflection refinement

The reflection correspondence stage above conducts photometric and geometric alignments in the feature domain, exploits auxiliary information in reflection scenes, and generates aligned features of the reflection layer $\mathbf{F}^{\mathrm{align}}_{\mathbf{R}_\mathrm{L}}$. For the recovery of the reflection layer $\mathbf{R}_\mathrm{L}$, the initially decomposed features of the reflection layer (*i.e.*, $\mathbf{F}^{\mathrm{ini}}_{\mathbf{R}_\mathrm{L}}$) is also needed to avoid discarding contextual details. Utilizing $\mathbf{F}^{\mathrm{align}}_{\mathbf{R}_\mathrm{L}}$ and $\mathbf{F}^{\mathrm{ini}}_{\mathbf{R}_\mathrm{L}}$, this stage recovers $\mathbf{R}_\mathrm{L}$ by feature fusion and refinement.

Aligned features $\mathbf{F}^{\mathrm{align}}_{\mathbf{R}_\mathrm{L}}$ and initially decomposed features $\mathbf{F}^{\mathrm{ini}}_{\mathbf{R}_\mathrm{L}}$ are firstly merged by the channel concatenation operation, and a convolutional block with a $1 \times 1$ kernel is utilized to condense and fuse the concatenated features. Then the fused features are fed into a refinement module containing six successively connected residual blocks [9], and we denote the output feature as $\mathbf{F}_{\mathbf{R}_\mathrm{L}}$. Finally, the reflection layer $\mathbf{R}_\mathrm{L}$ is recovered by a recovery module which contains a transposed convolutional block for up-sampling, a pyramid pooling module [54] for the aggregation of multi-scale contextual global information, and a convolutional block with a $1 \times 1$ kernel for the transformation from the feature domain to the image domain.

#### 4.2.5 Transmission recovery

With the assistance from the refined features of the reflection layer (*i.e.*, $\mathbf{F}_{\mathbf{R}_\mathrm{L}}$) to relieve content ambiguity, this stage recovers the transmission scene $\mathbf{T}_\mathrm{S}$ from the initially decomposed features (*i.e.*, $\mathbf{F}^{\mathrm{ini}}_{\mathbf{T}_\mathrm{S}}$). According to Eqn. (1), if subtracting $\mathbf{F}_{\mathbf{R}_\mathrm{L}}$ from features of the mixture image (*i.e.*, $\mathbf{F}_\mathbf{M}$), the features containing contextual information of the transmission scene $\mathbf{T}_\mathrm{S}$ can be obtained, which we denote as $\mathbf{F}^{\mathrm{sub}}_{\mathbf{T}_\mathrm{S}}$. For the preparation of recovering the transmission scene $\mathbf{T}_\mathrm{S}$, $\mathbf{F}^{\mathrm{sub}}_{\mathbf{T}_\mathrm{S}}$ and $\mathbf{F}^{\mathrm{ini}}_{\mathbf{T}_\mathrm{S}}$ are merged by channel concatenation and condensed by a $1 \times 1$ convolutional block to generate $\mathbf{F}_{\mathbf{T}_\mathrm{S}}$. However, $\mathbf{F}_{\mathbf{T}_\mathrm{S}}$ suffers from the intensity attenuation caused by the refraction of the glass, *i.e.*, the coefficient map $\mathbf{\Omega}$ in Eqn. (1).

To tackle the above issue, we feed $\mathbf{F}_{\mathbf{T}_\mathrm{S}}$ into a $1 \times 1$ convolutional block to generate an attention-like weighted map $\mathbf{W}$, then element-wise multiplication is conducted on $\mathbf{F}_{\mathbf{T}_\mathrm{S}}$ and $\mathbf{W}$ for intensity enhancement. Note that we use $\mathbf{W}$ to approximate the inverse of the attenuation coefficient map (*i.e.*, $1/\mathbf{\Omega}$) for counteracting the influence of the intensity attenuation from the refractive effect, and the reliability of $\mathbf{W}$ is ensured by a reconstruction loss which will be introduced in Sec. 4.3. Finally, by feeding the enhanced features into a refinement module and a recovery module which are with the same structure as in Sec. 4.2.4, a clean transmission scene $\mathbf{T}_\mathrm{S}$ without the reflection contamination is recovered.

### 4.3 Loss functions

For the high-fidelity recovery of the two components, *i.e.*, reflection layers and transmission scenes, the proposed method is optimized with several loss functions which conduct supervision on the visual quality of estimated images (the pixel, structural similarity, and perceptual loss) or exploit the inherent relationship in compositing mixture images (the reconstruction loss). We denote the estimated reflection layers and transmission scenes as $\mathbf{R}^{\mathrm{est}}_\mathrm{L}$ and $\mathbf{T}^{\mathrm{est}}_\mathrm{S}$, with the corresponding ground truths as $\mathbf{R}_\mathrm{L}$ and $\mathbf{T}_\mathrm{S}$. It must be noted that mixture images, reflection layers, and transmission scenes are all masked by user interaction, so loss functions are only calculated on the masked regions. Detailed formulations of loss functions are as follows.

**Pixel loss**. We apply the $l_1$ distance to penalize the pixel-wise discrepancy between estimated images with their ground truths, which is formulated as:

$$\mathcal{L}_{\mathrm{pixel}} = \left\|\mathbf{R}_\mathrm{L} - \mathbf{R}^{\mathrm{est}}_\mathrm{L}\right\|_1 + \left\|\mathbf{T}_\mathrm{S} - \mathbf{T}^{\mathrm{est}}_\mathrm{S}\right\|_1. \tag{10}$$

**Structural similarity loss**. Simply utilizing pixel loss results in low-frequency artifacts [32] and degrades the image quality. Thus the structural similarity index (SSIM) [52] is introduced to form a loss function, which conforms to human perception closely and measures the similarity of the luminance, contrast, and structure between a pair of images $\{\mathbf{X}, \mathbf{Y}\}$. The SSIM index is defined as follows:

$$\mathrm{SSIM}(\mathbf{X},\mathbf{Y}) = \frac{(2\mu_\mathbf{X}\mu_\mathbf{Y} + c_1)(2\sigma_{\mathbf{XY}} + c_2)}{(\mu^2_\mathbf{X} + \mu^2_\mathbf{Y} + c_1)(\sigma^2_\mathbf{X} + \sigma^2_\mathbf{Y} + c_2)}, \tag{11}$$

where $c_1$ and $c_2$ are regularization constants, $\mu_\mathbf{X}$ and $\mu_\mathbf{Y}$ are the means of $\mathbf{X}$ and $\mathbf{Y}$, $\sigma_\mathbf{X}$ and $\sigma_\mathbf{Y}$ are the variances of $\mathbf{X}$ and $\mathbf{Y}$, and $\sigma_{\mathbf{XY}}$ represents their covariance. Considering the common setting of loss functions for network training, the structural similarity loss is defined as:

$$\mathcal{L}_{\mathrm{ssim}} = 2 - [\mathrm{SSIM}(\mathbf{R}_\mathrm{L}, \mathbf{R}^{\mathrm{est}}_\mathrm{L}) + \mathrm{SSIM}(\mathbf{T}_\mathrm{S}, \mathbf{T}^{\mathrm{est}}_\mathrm{S})]. \tag{12}$$

**Feature loss**. To measure the multi-level discrepancy between estimated images and their ground truths in the feature domain, a feature loss [64] is employed. We use the VGG-19 model [38] trained on the ImageNet dataset [36] to extract features which contain both low-level and high-level contextual information, and combine them to calculate the feature loss, which is defined as follows:

$$\mathcal{L}_{\text{feat}} = \sum\nolimits_i \lambda_i [\mathcal{L}^i_{\text{VGG}}(\mathbf{R}_{\text{L}}, \mathbf{R}^{\text{est}}_{\text{L}}) + \mathcal{L}^i_{\text{VGG}}(\mathbf{T}_{\text{S}}, \mathbf{T}^{\text{est}}_{\text{S}})], \quad (13)$$

where $\{\lambda_i\}$ are the weights for equilibrium of multi-level feature differences, and $\mathcal{L}^i_{\text{VGG}}$ presents the $l_1$ distance between features from the $i$-th convolutional layer in the VGG-19 model [38]. Similar to [64], the layers as 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2', and 'conv5_2' are selected in our experiments.

**Reconstruction loss**. In the stage of transmission recovery (Sec. 4.2.5), after obtaining $\mathbf{R}^{\text{est}}_{\text{L}}$, $\mathbf{T}^{\text{est}}_{\text{S}}$ and the coefficient map $\mathbf{W}$, it is intuitive to follow Eqn. (1) and reconstruct a mixture image $\mathbf{M}^{\text{est}}$ via recombining above estimated components:

$$\mathbf{M}^{\text{est}} \quad = \quad \frac{1}{\mathbf{W}} \odot \mathbf{T}^{\text{est}}_{\text{S}} + \mathbf{R}^{\text{est}}_{\text{L}}, \quad (14)$$

where $\mathbf{W}$ is designed to approach the inverse of the refractive map (*i.e.*, $1/\mathbf{\Omega}$). If the network is trained well, the reconstructed mixture image $\widetilde{\mathbf{M}}$ is supposed to approximate the original mixture image $\mathbf{M}$. Therefore, we also apply the $l_1$ distance to supervise the reconstruction quality:

$$\mathcal{L}_{\text{recon}} = \left\| \mathbf{M} - \mathbf{M}^{\text{est}} \right\|_1. \quad (15)$$

Overall, we train the proposed network with the following loss function:

$$\mathcal{L}_{\text{total}} = \omega_1 \mathcal{L}_{\text{pixel}} + \omega_2 \mathcal{L}_{\text{ssim}} + \omega_3 \mathcal{L}_{\text{feat}} + \omega_4 \mathcal{L}_{\text{recon}}. \quad (16)$$

Following previous methods [4], [47], [54], [64], the weights are empirically set as $\omega_1 = 1$, $\omega_2 = 1$, $\omega_3 = 0.1$, and $\omega_4 = 1$ throughout our experiments.

We implement the proposed method with PyTorch [33] on two Nvidia GeForce RTX 3090 GPUs. The model is trained in an end-to-end manner for 40 epochs with Adam [16] optimizer to update learnable parameters. Weights are initialized as in [8]. The learning rate is set to $10^{-4}$ initially and decreases to $10^{-5}$ at epoch 30.

## 5 DATASET

### 5.1 Synthetic data

Performances of learning-based reflection removal methods heavily rely on synthetic training data [4], [5], [22], [54], [64], since capturing real mixtures images with ground truths of transmission scenes and reflection layers is complicated and time-consuming. Currently, prevalent synthetic data generation methods mainly concentrate on simulating the perceptual properties of mixture images by applying linear addition and regional subtraction operations (*e.g.*, Fan *et al.* [5] and Zhang *et al.* [64], denoted as 'FY17' and 'ZN18', respectively), or by employing spatially-varying blending masks generated from specially designed networks (*e.g.*, Wen *et al.* [55], denoted as 'WT19'). By taking the physical formation process of glass reflections into consideration, Zheng *et al.* [67] (denoted as 'ZC20') synthesize mixture images using spatially-varying reflective and refractive coefficient maps, which are independent from image contents [17]. Generally, to synthesize reflection layers, intensities of selected reflection scenes are often decreased to simulate the attenuation of lights when they reflect from the glass surface. However, due to the low dynamic range (LDR) of commonly-used images (resulted from the dynamic range clipping [26] in the camera pipeline), selected reflection scenes may contain saturated regions which suffer from the missing of contextual information. As shown in red boxes of Fig. 9, saturation in LDR reflection scenes causes reflection layers generated from them to be unrealistic in the corresponding regions. We call such regions in synthetic reflection layers as *pseudo valid* regions, since they lack valid information despite not being saturated in intensities.

To tackle the weaknesses of existing synthetic data generation methods, we propose to synthesize more realistic mixture images based on the physics-based and high dynamic range (HDR) image formation pipeline to prevent the pseudo valid regions. Considering the misalignment between reflection scenes and layers in the setup of panoramic image reflection removal, the image formulation model of mixture images [17], [28], [67] can be written as:

$$\hat{\mathbf{M}} = (1 - \mathbf{\Theta}) \odot \hat{\mathbf{T}}_{\text{S}} + \mathbf{\Theta} \odot \mathcal{F}_{S \to L}(\hat{\mathbf{R}}_{\text{S}}), \quad (17)$$

where $\mathbf{\Theta}$ is the reflective amplitude coefficient map, $\mathcal{F}_{S \to L}(\cdot)$ represents the pixel-wise geometric transformation operation as Eqn. (4), $\hat{\mathbf{M}}$, $\hat{\mathbf{T}}_{\text{S}}$, and $\hat{\mathbf{R}}_{\text{S}}$ represent the mixture image, transmission scene, and reflection scene in the linear image space, respectively. It has to be noted that, different from $\mathbf{\Omega}$ and $\mathbf{\Phi}$ (in Sec. 3) which conduct intensity attenuation in the non-linear image space, $\mathbf{\Theta}$ is content-free and only concerned with the angle of incidence and the refractive index of the glass [17], [28], [67].

For the blending of reflection and transmission scenes, reflective amplitude coefficient maps $\mathbf{\Theta}$ are first generated randomly as in ZC20 [67]. To prevent the missing of scene information caused by the dynamic range clipping [26], we utilize HDR images collected from Poly Heaven[3] as $\hat{\mathbf{R}}_{\text{S}}$. The geometric transformation operation $\mathcal{F}_{S \to L}(\cdot)$ includes random translation and spatial scaling, which simulates the geometric misalignment between reflection scenes and layers. Then reflection layers in the linear image space (denoted as $\hat{\mathbf{R}}_{\text{L}}$) can be obtained by $\hat{\mathbf{R}}_{\text{L}} = \mathbf{\Theta} \odot \mathcal{F}_{S \to L}(\hat{\mathbf{R}}_{\text{S}})$. We randomly select a proportion of generated reflection layers to apply Gaussian smoothing kernels with kernel sizes in the range of 3 to 7 pixels to simulate the situation that reflections are out of focus [5]. To obtain transmission scenes $\hat{\mathbf{T}}_{\text{S}}$ in the linear image space, we multiply panoramic images in the Structured3D dataset [66] with randomly generated masks which indicate glass regions to obtain $\mathbf{T}_{\text{S}}$ first, and then conduct inverse gamma correction $\Gamma^{-1}(\cdot)$ as previous methods [4], [64] to transform them into linear space: $\hat{\mathbf{T}}_{\text{S}} = \Gamma^{-1}(\mathbf{T}_{\text{S}})$. Afterward, mixture images $\hat{\mathbf{M}}$ in the linear image space can be generated according to Eqn. (17). Finally, LDR mixture images $\mathbf{M}$ is obtained by the dynamic range clipping $\mathcal{C}(\mathbf{M}) = \min(\mathbf{M}, 1)$ and non-linear mapping (here we use gamma correction function $\Gamma(\cdot)$ as in [4], [64]): $\mathbf{M} = \Gamma[\mathcal{C}(\hat{\mathbf{M}})]$. Similarly, LDR reflection scenes $\mathbf{R}_{\text{S}}$,

3. https://polyhaven.com/hdris

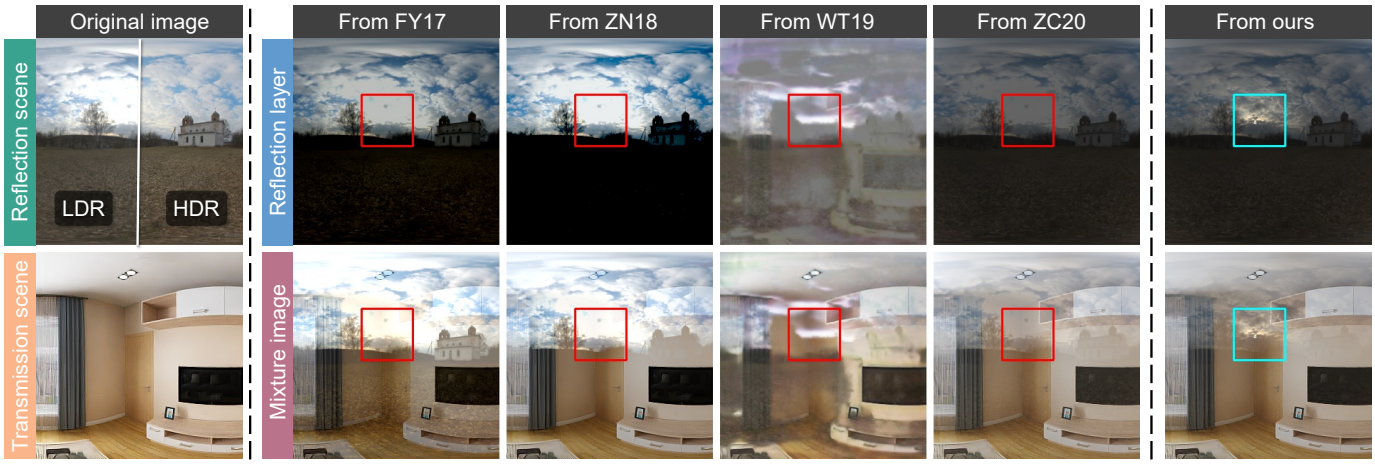


Fig. 9: Visual quality comparison of synthetic data from existing data generation methods (*i.e.*, FY17 [5], ZN18 [64], WT19 [55], and ZC20 [67]) and our method. Note that previous methods utilize LDR images as reflection scenes, generating data with pseudo valid regions (red boxes). Our method employs HDR images (tone mapped here for visualization) to generate LDR refection scenes, reflection layers, and mixture images, which avoids the problem of pseudo valid regions (blue boxes).

reflection layers $\mathbf{R}_{\mathrm{L}}$, and transmission scenes $\mathbf{T}_{\mathrm{S}}$ can also be obtained by conducting the dynamic range clipping and non-linear mapping to their corresponding images (*i.e.*, $\hat{\mathbf{R}}_{\mathrm{S}}$, $\hat{\mathbf{R}}_{\mathrm{L}}$, and $\hat{\mathbf{T}}_{\mathrm{S}}$) in the linear image space respectively. As shown in blue boxes of Fig. 9, our data generation method tackles the problem of pseudo valid regions compared with previous methods [5], [55], [64], [67], which attributes to the utilization of HDR images. In total, we generate 10000 sets of synthetic data (*i.e.*, $\{\mathbf{M}, \mathbf{R}_{\mathrm{S}}, \mathbf{R}_{\mathrm{L}}, \mathbf{T}_{\mathrm{S}}\}$) for the training of the proposed method[4].

### 5.2 Real data

Due to the insufficiency of available datasets for panoramic image reflection removal, we collect two groups of real panoramic images for evaluation, including 30 sets as PORTABLE dataset and 10 sets as NATURAL dataset. Images in PORTABLE dataset are used for both quantitative evaluation and visual quality comparison, which are captured by putting a portable glass in the scene. Corresponding transmission scenes and reflection layers are captured in the same way as SIR$^2$ dataset [45]. Images in NATURAL dataset are used for visual quality comparison (due to the lack of ground truths), which are captured with glass found in different natural scenarios, such as office buildings. Samples from the above two datasets are collected by a single-shot panoramic camera, *i.e.*, Ricoh Theta Z1. To validate the generalization capability of the proposed method for casual users, we further collect a real dataset named PHONE, which only contains mixture images and reflection scenes (no ground truths) collected by a Huawei P40 Pro+ smartphone.

## 6 EXPERIMENTS

### 6.1 Evaluation on real data

To evaluate the performance of the proposed method (PAR$^2$Net), we conduct quantitative and qualitative experiments on our real data, *i.e.*, the PORTABLE and NATURAL datasets. We compare the proposed method with our preliminary work [12] (denoted as 'HZ21') and five state-of-the-art single-image reflection removal methods, including CGDNet [65], Dong *et al.* [4] (denoted as 'DX21'), IBCLN [22], Kim *et al.* [15] (denoted as 'KH20'), and CoRRN [47]. For fair comparisons, we finetune the above-mentioned methods on our training data if their training codes are provided. Since source codes of CGDNet [65] have not been released, we sent our testing data to the authors and obtained results of CGDNet [65].

TABLE 1: Comparison of quantitative results in terms of PSNR [14] and SSIM [53] on the PORTABLE dataset for evaluating the recovery of both transmission scenes and reflection layers. ↑ (↓) indicates larger (smaller) values are better. Bold numbers indicate the best performing results.

| Method | Transmission | | Reflection | |
|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| PAR$^2$Net | **26.189** | **0.857** | **22.159** | **0.743** |
| HZ21 [12] | 24.673 | 0.830 | 20.486 | 0.697 |
| CGDNet [65] | 21.178 | 0.772 | 16.573 | 0.535 |
| DX21 [4] | 21.445 | 0.793 | 17.091 | 0.603 |
| IBCLN [22] | 21.657 | 0.778 | 16.687 | 0.496 |
| KH20 [15] | 21.235 | 0.782 | 16.547 | 0.526 |
| CoRRN [47] | 20.959 | 0.769 | 16.667 | 0.492 |

As for the input setting, it has to be noted that our PAR$^2$Net can process panoramic images with masks which can have arbitrary shapes for labeling glass regions, while HZ21 [12] and single-image methods [4], [15], [22], [47] need to input with rectangular images. Therefore, in our experiments, HZ21 [12] and single-image methods [4], [15], [22], [47], [65] are input with extracted patches (of resolution $320 \times 320$) from panoramic images. For PAR$^2$Net, we directly input panoramic images (of resolution $512 \times 1024$) with non-rectangular masks and extract patches from results for both quantitative and qualitative comparisons.

4. Note that HDR images are only required in data synthesis to diminish the domain gap between training and testing data, while inputs for PAR$^2$Net are all LDR images.

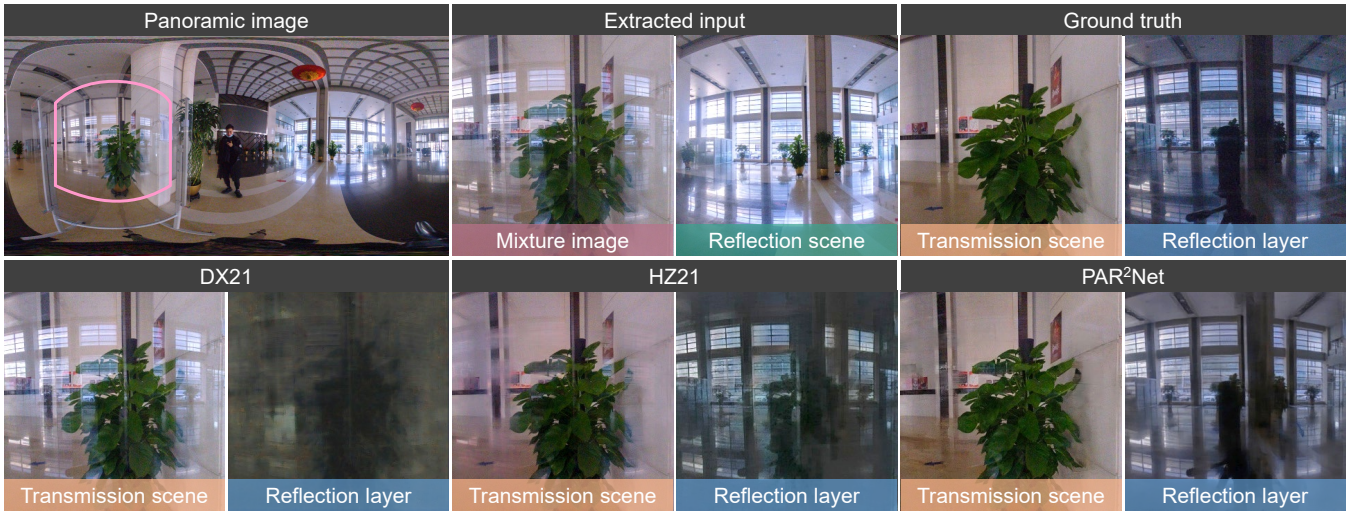


Fig. 10: Qualitative results on the PORTABLE dataset. First row (from left to right): the panoramic image with the non-rectangular mask (labeled by pink curves) as the input of the proposed method (PAR$^2$Net), rectangular images extracted from panoramic images as inputs of HZ21 [12] and the single-image method DX21 [4] (which only uses mixture images), and the corresponding ground truths. Second row: results in the rectangular format, where results of PAR$^2$Net are extracted from its panoramic result (not shown here) for comparison. Please zoom in for details.

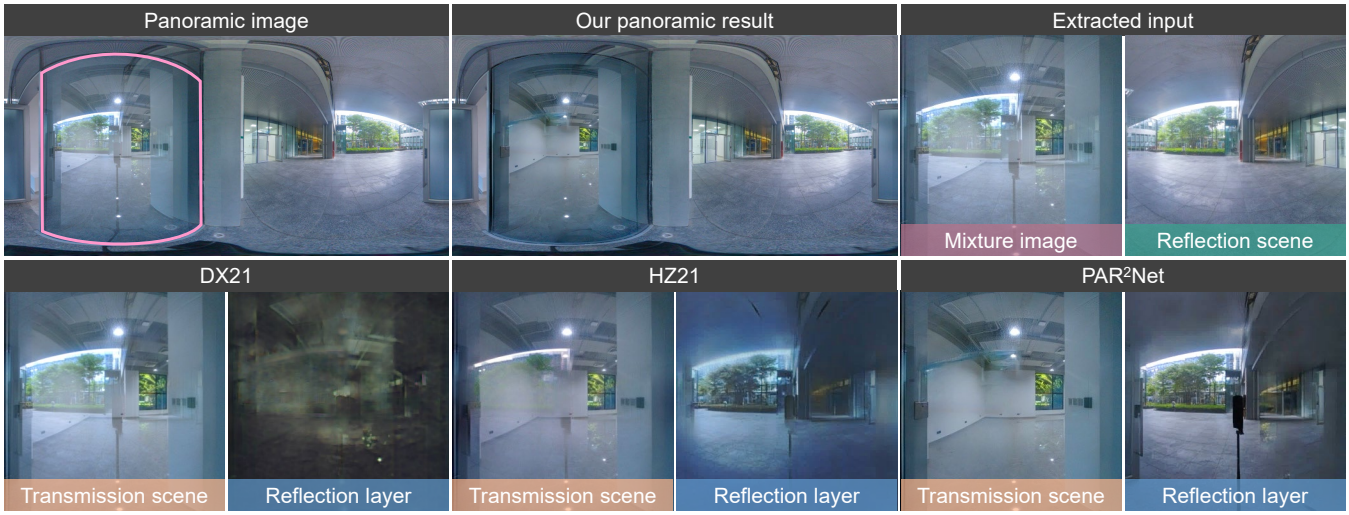


Fig. 11: Qualitative results on the NATURAL dataset. First row (from left to right): the panoramic image with the non-rectangular mask (labeled by pink curves) as the input of the proposed method (PAR$^2$Net), the panoramic result from PAR$^2$Net, and rectangular images extracted from panoramic images as inputs of HZ21 [12] and the single-image method DX21 [4] (which only uses mixture images). Second row: results in the rectangular format, where results of PAR$^2$Net are extracted from its panoramic result for comparison. Please zoom in for details.

In the quantitative comparison, we follow the setting of existing reflection removal methods [4], [28] to utilize PSNR [14] and SSIM [53] as error metrics for evaluating the recovery of both transmission scenes and reflection layers. Quantitative results are shown in Table 1. Comparing to state-of-the-art single-image methods [4], [15], [22], [47], [65], the proposed PAR$^2$Net achieves much better performance on both transmission and reflection recovery in all error metrics, indicating that introducing other views of reflection scenes in panoramic images successfully provides auxiliary information for the relief of content ambiguity. HZ21 [12] outperforms single-methods [4], [15], [22], [47], [65] since it also utilizes additional reflections scenes, while it performs worse than PAR$^2$Net especially in recovering reflection layers, which demonstrates the effectiveness of the network design of PAR$^2$Net, *i.e.*, the initial decomposition and the reflection correspondence stage to better handle misalignment issues between reflection scenes and layers.

Visual quality comparisons are conducted on the PORTABLE and NATURAL datasets with results shown in Fig. 10 and Fig. 11, respectively[5]. We compare PAR$^2$Net with HZ21 [12] and a single-image method DX21 [4] 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). In both Fig. 10 and Fig. 11, it can be observed that due to

5. Additional qualitative results are provided in Appendix A.

the lack of auxiliary information, the single-image method DX21 [4] fails to address regions with content ambiguity, *i.e.*, fails to distinguish transmission scenes and reflection layers when reflections are strong or with sharp edges. HZ21 [12] achieves acceptable results thanks to the exploitation of reflection scenes. However, its photometric alignment which is based on fixed polynomial fitting results in chromatic aberration in estimated reflection layers and further affects the transmission recovery to generate degraded results. Besides, errors in the matching-based geometric alignment accumulate through the multi-stage framework of HZ21 [12], causing strong reflections still left in recovered transmission scenes. Comparing to HZ21 [12], PAR$^2$Net is able to process larger regions in panoramic images and achieves more precise transmission recovery in terms of image content and color, which attributes to the new setup that inputs panoramic images with arbitrary masks and the reflection correspondence stage that tackles the misalignment issues. In addition, PAR$^2$Net outperforms all compared methods in recovering reflection layers, which demonstrates the effectiveness of our data generation method to synthesize realistic reflections.

### 6.2 Ablation study

In this section, we implement an unsupervised version[6] of the proposed method inspired by Han *et al.* [6] to compare the supervised learning and unsupervised learning strategy on the panoramic image reflection removal task. Besides, we conduct several ablation studies to investigate the influence of the additional reflection scene, the YTMT blocks [13], the adaptive reflection correspondence module, and the loss function. Note that the majority of loss functions employed for training the proposed method are commonly used and validated to be effective in reflection removal [4], [22], [47], so we only evaluate the influence of the reconstruction loss ($\mathcal{L}_{\text{recon}}$). In addition, we validate the influence of the proposed data generation method by comparing to variants trained with data which is generated from existing methods [5], [55], [64], [67] (mentioned in Sec. 5.1). In general, we evaluate the effectiveness of the proposed method with the following variants: 'PAR$^2$Net-U' that trains the proposed method in an unsupervised manner, 'w/o $\mathbf{R}_{\text{S}}$' that only inputs with mixture images while lacking auxiliary reflection scenes, 'w/o YTMT' that replaces the YTMT blocks [13] for initial decomposition with a simple two-stream module composed of residual blocks [9], 'w/o correspondence' that disables the reflection correspondence module and directly uses features of $\mathbf{R}_{\text{S}}$, 'w/o $\mathcal{L}_{\text{recon}}$' that trains the model without the reconstruction loss, 'with FY17 data', 'with ZN18 data', 'with WT19 data', and 'with ZC20 data' that trains the model with data synthesized from data generation methods of FY17 [5] (image processing-based), ZN18 [64] (image processing-based), WT19 [55] (learning-based), and ZC20 [67] (physics-based), respectively.

Table 2 reports the quantitative results on the PORTABLE dataset. As can be observed, though inferior to the model trained with the supervised strategy, the unsupervised version 'PAR$^2$Net-U' surpasses the variant 'w/o $\mathbf{R}_{\text{S}}$' due to the auxiliary contextual information in reflection scenes,

6. Details of the unsupervised version are provided in Appendix B.

TABLE 2: Quantitative results of the ablation study, in terms of PSNR [14] and SSIM [53] on our PORTABLE dataset. ↑ (↓) indicates larger (smaller) values are better. Bold numbers indicate the best performing results.

| Method | Transmission | | Reflection | |
|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| PAR$^2$Net | **26.189** | **0.857** | **22.159** | **0.743** |
| PAR$^2$Net-U | 22.759 | 0.798 | 18.070 | 0.641 |
| w/o $\mathbf{R}_{\text{S}}$ | 21.793 | 0.789 | 17.102 | 0.606 |
| w/o YTMT | 25.332 | 0.851 | 21.178 | 0.731 |
| w/o correspondence | 24.521 | 0.815 | 20.008 | 0.672 |
| w/o $\mathcal{L}_{\text{recon}}$ | 25.768 | 0.849 | 21.565 | 0.724 |
| with FY17 [5] data | 24.692 | 0.832 | 20.922 | 0.706 |
| with ZN18 [64] data | 24.735 | 0.833 | 20.862 | 0.707 |
| with WT19 [55] data | 23.432 | 0.801 | 18.195 | 0.632 |
| with ZC20 [67] data | 25.554 | 0.843 | 21.452 | 0.715 |

while the performance of 'w/o $\mathbf{R}_{\text{S}}$' is still comparable to state-of-the-art single-image methods in Table 1, which is benefited from our network design. The variant 'w/o YTMT' performs worse than our complete model, since the two branches in the substituted two-stream module are independent, which lacks the capability of exploiting the complementary information in transmission scenes and reflection layers comparing to YTMT blocks [13]. For the variant 'w/o correspondence', since the reflection correspondence module is utilized for diminishing misalignment issues between reflection scenes and layers, its absence also affects the performance of the proposed method, especially affecting the recovery of reflection layers. Discarding the reconstruction loss (w/o $\mathcal{L}_{\text{recon}}$) also has impacts on the performance, as the missing of constraints on the relationship between transmission scenes and reflection layers intuitively influences their recovery.

As for the ablation study on training data, due to the unstable learning-based synthesis, the variant 'with WT19 data' performs worst among the four variants. Performances of the rest three variants, *i.e.*, 'with FY17 data', 'with ZN18 data', and 'with ZC20 data' increase successively, while the model trained on our synthetic data exceeds all of them, indicating the necessity to consider the physics-based and high dynamic range image formation model in data generation. Overall, the complete version of the proposed method (*i.e.*, PAR$^2$Net) outperforms all variants, demonstrating the effectiveness of the network architecture and the proposed data generation method.

### 6.3 Without using panoramic cameras

This section considers more practical cases by using conventional cameras with limited FoV instead of panoramic cameras. For casual users using digital cameras or mobile phones, after capturing mixture images, reflection scenes can be obtained by turning over the camera for about 180°. For industrial applications like moving platforms or robots, by equipping with two oppositely-orientated industrial cameras, mixture images and reflection scenes can be captured simultaneously. However, under such cases, constraints on reflection scenes and layers are weakened compared with panoramic images, which brings challenges for

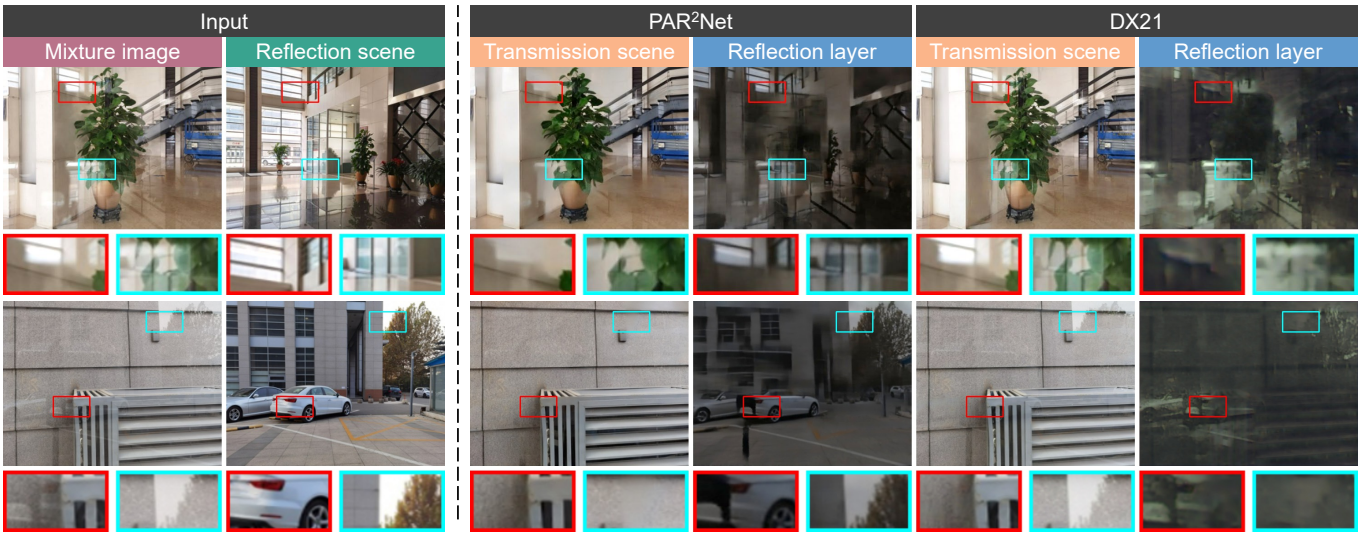


Fig. 12: Qualitative results on the PHONE dataset, compared with the state-of-the-art single-image method DX21 [4]. Close-up views are displayed at the bottom of images. Please zoom in for details.

tackling misalignment issues and recovering transmission scenes and reflection layers. To evaluate the generalization capacity, we conduct experiments on the PHONE dataset by comparing PAR$^2$Net with the typical single-image method DX21 [4][7].

As can be observed from Fig. 12, PAR$^2$Net recovers clear reflection layers and suppresses most of reflection artifacts in transmission scenes, even in regions with strong reflections, *e.g.*, the red box in the first example and the blue box in the second example. The single-image method DX21 [4] fails to suppress reflection artifacts, since it only relies on the deep content priors learned from network training and lacks other reliable auxiliary information to precisely identify reflection regions. From these results, it can be verified that the proposed end-to-end framework is sufficiently capable of dealing with limited-FoV images, and the partial panoramic information could be useful to relieve the content ambiguity for reliable reflection removal, which is potentially applicable to mobile devices and platforms.

## 7 Conclusion

This paper addresses the issue of relieving the content ambiguity in reflection removal by using panoramic images. The main challenge of this problem lies in the geometric and photometric misalignments [12] between the reflection scene and the reflection layer. An end-to-end framework is proposed to adaptively tackle this challenge and achieves the recovery of the transmission scene and the reflection layer with higher fidelity than [12]. Experimental results demonstrate that the proposed method not only achieves a significant performance advantage over single-image methods but also generalizes well to limited-FoV images captured without panoramic cameras.

**Limitations.** The performance of the proposed method may degrade if the image content in some regions of reflection layers cannot find correspondences in reflection scenes. In such regions, the ablation study (specifically, the variant 'w/o $\mathbf{R}_S$') demonstrates that the proposed method can still be comparable with state-of-the-art methods, which also shows our robustness.

7. Additional qualitative results are provided in Appendix A.

## Acknowledgments

This work was supported by the National Key R&D Program of China (2021ZD0109800), the National Natural Science Foundation of China under Grant No. 62136001, 62088102, 61972119, and in part by the Rapid-Rich Object Search (ROSE) Lab of Nanyang Technological University, Singapore.

## References

[1] Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiros Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Pano3d: A holistic benchmark and a solid baseline for 360deg depth estimation. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2021.

[2] Nikolaos Arvanitopoulos, Radhakrishna Achanta, and Sabine Süsstrunk. Single image reflection suppression. In *Proc. Computer Vision and Patter Recognition (CVPR)*, pages 1752–1760, 2017.

[3] Yakun Chang, Cheolkon Jung, Jun Sun, and Fengqiao Wang. Siamese dense network for reflection removal with flash and no-flash image pairs. *International Journal of Computer Vision (IJCV)*, 2020.

[4] 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.

[5] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David P Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proc. International Conference on Computer Vision (ICCV)*, pages 3258–3267, 2017.

[6] 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.

[7] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2015.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. International Conference on Computer Vision (ICCV)*, 2015.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. Computer Vision and Patter Recognition (CVPR)*, pages 770–778, 2016.

[10] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2019.

[11] Yuchen Hong, Youwei Lyu, Si Li, and Boxin Shi. Near-infrared image guided reflection removal. In *Proc. International Conference on Multimedia and expo*, 2020.

[12] 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.

[13] Qiming Hu and Xiaojie Guo. Trash or treasure? an interactive dual-stream strategy for single image reflection separation. *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[14] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[15] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single image reflection removal with physically-based training images. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2020.

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

[17] Naejin Kong, Yu-Wing Tai, and Joseph S Shin. A physically-based approach to reflection separation: from physical modeling to constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(2):209–221, 2013.

[18] Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2019.

[19] Chenyang Lei and Qifeng Chen. Robust reflection removal with reflection-free flash-only cues. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2021.

[20] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2020.

[21] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(9), 2007.

[22] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2020.

[23] Yu Li and Michael S. Brown. Exploiting reflection change for automatic reflection removal. In *Proc. International Conference on Computer Vision (ICCV)*, 2013.

[24] Yu Li and Michael S Brown. Single image layer separation using relative smoothness. In *Proc. Computer Vision and Patter Recognition (CVPR)*, pages 2752–2759, 2014.

[25] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2020.

[26] YuLun Liu, WeiSheng Lai, YuSheng Chen, YiLung Kao, MingHsuan Yang, YungYu Chuang, and JiaBin Huang. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2020.

[27] YuLun Liu, WeiSheng Lai, MingHsuan Yang, YungYu Chuang, and Jia-Bin Huang. Learning to see through obstructions. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2020.

[28] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Reflection separation using a pair of unpolarized and polarized images. In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[29] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Physics-guided reflection separation from a pair of unpolarized and polarized images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[30] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson W.H. Lau. Don't hit me! glass detection in real-world scenes. In *Proc. Computer Vision and Patter Recognition (CVPR)*, June 2020.

[31] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. International Conference on Machine Learning (ICML)*, 2010.

[32] Simon Niklaus, Xuaner Cecilia Zhang, Jonathan T Barron, Neal Wadhwa, Rahul Garg, Feng Liu, and Tianfan Xue. Learned dual-view reflection removal. In *Proc. Winter Conference on Application of Computer Vision (WACV)*, 2021.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[34] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. Slicenet: Deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2021.

[35] Abhijith Punnappurath and Michael S Brown. Reflection removal using a dual-pixel sensor. In *Proc. Computer Vision and Patter Recognition (CVPR)*, pages 1556–1565, 2019.

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[37] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman. Reflection removal using ghosting cues. In *Proc. Computer Vision and Patter Recognition (CVPR)*, pages 3193–3201, 2015.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2019.

[40] Shuran Song, Andy Zeng, Angel X Chang, Manolis Savva, Silvio Savarese, and Thomas Funkhouser. Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2018.

[41] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for 360 recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[42] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2019.

[43] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[44] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, Wen Gao, and Alex C Kot. Region-aware reflection removal with unified content and gradient priors. *IEEE Transactions on Image Processing (TIP)*, 27(6):2927–2941, 2018.

[45] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proc. International Conference on Computer Vision (ICCV)*, 2017.

[46] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. CRRN: Multi-scale guided concurrent reflection removal network. In *Proc. Computer Vision and Patter Recognition (CVPR)*, pages 4777–4785, 2018.

[47] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, Ah-Hwee Tan, and Alex Kot Chichung. CoRRN: Cooperative reflection removal network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[48] Renjie Wan, Boxin Shi, Haoliang Li, Yucheng Hong, Ling-Yu Duan, and Alex C. Kot. Benchmarking single-image reflection removal algorithms. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[49] Ce Wang, Dejia Xu, Renjie Wan, Bin He, Boxin Shi, and Ling-Yu Duan. Background scene recovery from an image looking through colored glass. *IEEE Transactions on Multimedia (TMM)*, 2022.

[50] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Led2-net: Monocular 360deg layout estimation via differentiable depth rendering. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2021.

[51] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2018.

[52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004.

[53] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, 2003.

[54] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proc. Computer Vision and Patter Recognition (CVPR)*, pages 8178–8187, 2019.

[55] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *Proc. Computer Vision and Patter Recognition (CVPR)*, pages 3771–3779, 2019.
[56] Patrick Wieschollek, Orazio Gallo, Jinwei Gu, and Jan Kautz. Separating reflection and transmission images in the wild. In *Proc. European Conference on Computer Vision (ECCV)*, pages 89–104, 2018.
[57] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.
[58] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T. Freeman. A computational approach for obstruction-free photography. *ACM Transactions on Graphics (TOG)*, 2015.
[59] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2019.
[60] Yang Yang, Wenye Ma, Yin Zheng, Jian-Feng Cai, and Weiyu Xu. Fast single image reflection suppression via convex optimization. In *Proc. Computer Vision and Patter Recognition (CVPR)*, pages 8141–8149, 2019.
[61] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2019.
[62] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2019.
[63] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2020.
[64] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proc. Computer Vision and Patter Recognition (CVPR)*, pages 4786–4794, 2018.
[65] Ya-Nan Zhang, Linlin Shen, and Qiufu Li. Content and gradient model-driven deep network for single image reflection removal. In *ACM MM*, 2022.
[66] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.
[67] Qian Zheng, Jinnan Chen, Zhan Lu, Boxin Shi, Xudong Jiang, Kim-Hui Yap, Ling-Yu Duan, and Alex C. Kot. What does plate glass reveal about camera calibration. In *Proc. Computer Vision and Patter Recognition (CVPR)*, 2020.

**Yuchen Hong** received the B.E. degree from Beijing University of Posts and Telecommunications in 2020. He is currently studying at School of Computer Science, Peking University. His research interests include computational photography and computer vision.

**Qian Zheng** is an assistant professor at the College of Computer Science and Technology, Zhejiang University. He received his B.E. and Ph.D. degrees in computer science from Zhejiang University, China, in 2011 and 2017, respectively. From 2018 to 2022, he was a research fellow with the ROSE lab, Nanyang Technological University. His interests include neuromorphic computing and computer vision. He has published several papers in international journals and conferences, such as T-PAMI, T-IP, T-IFS, CVPR, ICCV, AAAI, and IJCAI. He is a guest editor of Frontiers in Neuroscience and a reviewer of T-IP, CVPR, ICCV, ICLR, NeurIPS, and ICML.

**Lingran Zhao** is currently working toward the B.S. degree with the Department of Computer Science and Technology, Peking University, China. His research interests include computer vision, neural network compression, and multi-modality learning.

**Xudong Jiang** (Fellow, IEEE) received the B.E. and M.E. degrees from the University of Electronic Science and Technology of China (UESTC), and the Ph.D. degree from Helmut Schmidt University, Hamburg, Germany. From 1986 to 1993, he was a Lecturer with UESTC. From 1998 to 2004, he was with the Institute for Infocomm Research, A*STAR, Singapore, as a Lead Scientist, and the Head of the Biometrics Laboratory. He joined Nanyang Technological University (NTU), Singapore as a Faculty Member in 2004, where he served as the Director of the Centre for Information Security from 2005 to 2011. He is currently a professor with the School of EEE, NTU and serves as the Director of the Centre for Information Sciences and Systems. He has authored over 200 papers with over 50 papers in the IEEE journals, including 9 papers in T-PAMI and 17 papers in T-IP. He served as IFS TC Member of the IEEE Signal Processing Society and Associate Editors for IEEE SPL and IEEE T-IP. Currently Dr. Jiang is an IEEE Fellow and serves as Senior Area Editor for IEEE T-IP and Editor-in-Chief for IET Biometrics. His current research interests include image processing, pattern recognition, computer vision, machine learning and biometrics.

**Alex C. Kot** (Fellow, IEEE) has been with the Nanyang Technological University, Singapore since 1991. He was Head of the Division of Information Engineering and Vice Dean Research at the School of Electrical and Electronic Engineering. Subsequently, he served as Associate Dean for College of Engineering for eight years. He is currently Professor and Director of Rapid-Rich Object SEarch (ROSE) Lab and NTU-PKU Joint Research Institute. He has published extensively in the areas of signal processing, biometrics, image forensics and security, and computer vision and machine learning. Dr. Kot served as Associate Editor for more than ten journals, mostly for IEEE transactions. He served the IEEE SP Society in various capacities such as the General Co-Chair for the 2004 IEEE International Conference on Image Processing and the Vice-President for the IEEE Signal Processing Society. He received the Best Teacher of the Year Award and is a co-author for several Best Paper Awards including ICPR, IEEE WIFS and IWDW, CVPR Precognition Workshop and VCIP. He was elected as the IEEE Distinguished Lecturer for the Signal Processing Society and the Circuits and Systems Society. He is a Fellow of IEEE, and a Fellow of Academy of Engineering, Singapore.

**Boxin Shi** (Senior Member, IEEE) received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Assistant Professor and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper Runner-Up at ICCP 2015 and selected as Best Papers from ICCV 2015 for IJCV Special Issue. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV.

# 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 $\alpha$ 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.

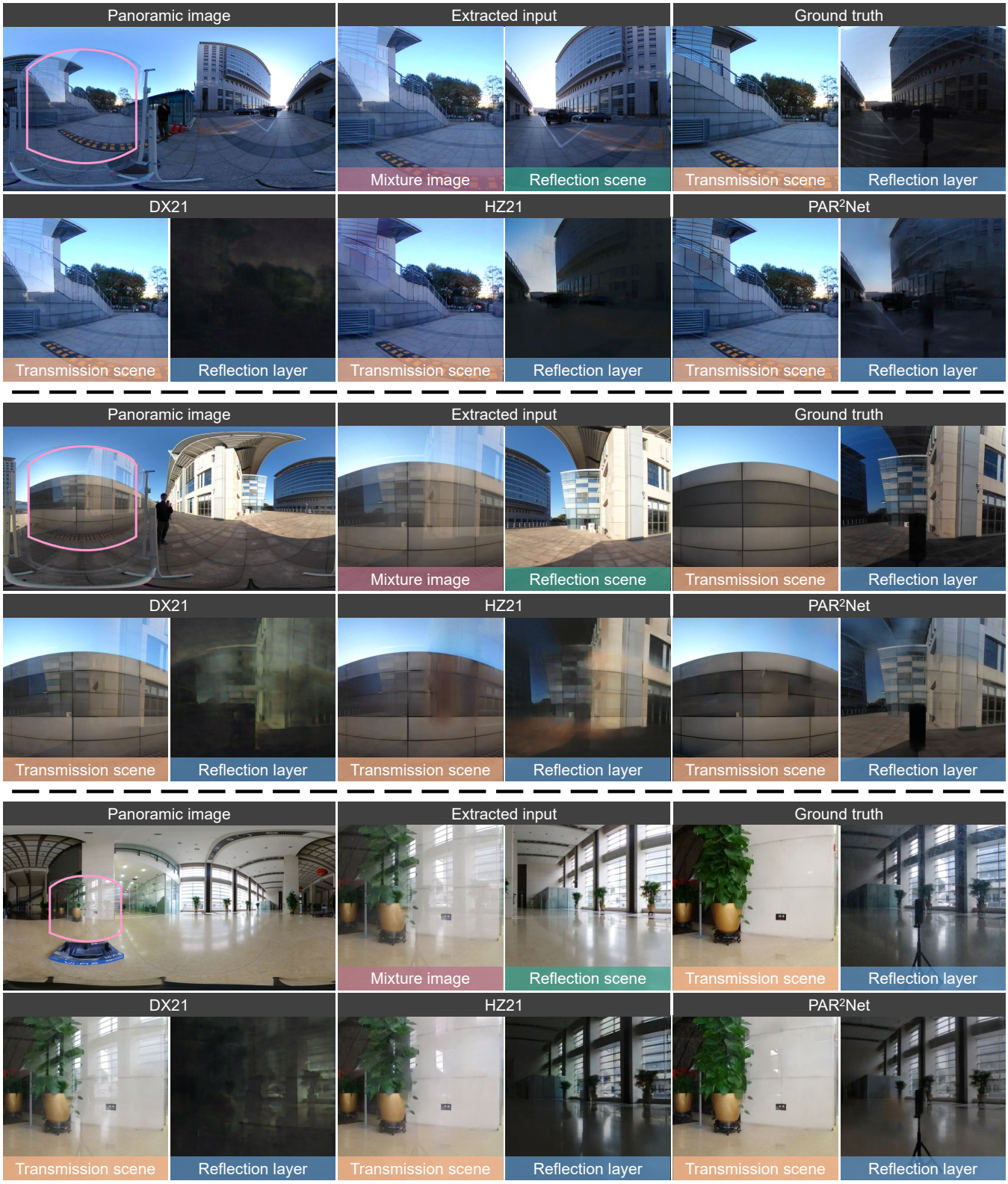


Fig. 13: More qualitative results on the PORTABLE dataset. Inputs and results are shown in the same manner as Fig. 10 of the main paper. Please zoom in for details.


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.

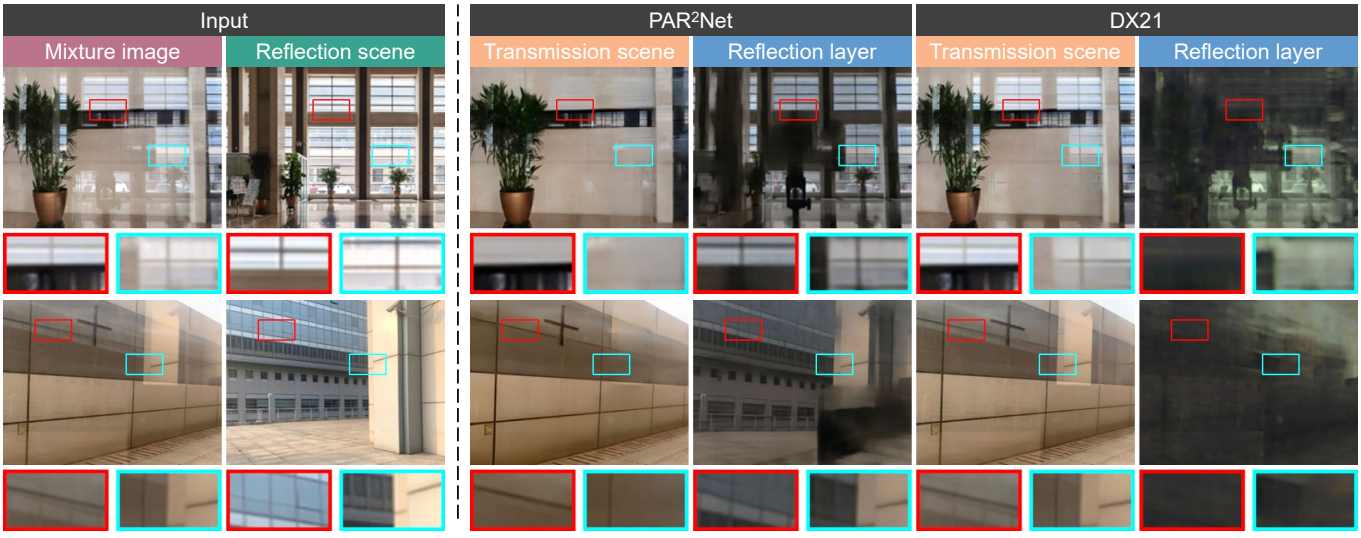


Fig. 15: More qualitative results on the PHONE dataset, compared with the state-of-the-art single-image method DX21 [1]. Close-up views are displayed at the bottom of images. Please zoom in for details.