Panoramic Image Reflection Removal

Yuchen Hong[1#] Qian Zheng[2#] Lingran Zhao[1] Xudong Jiang[2] Alex C. Kot[2] Boxin Shi[1,3,4*]
[1] NELVT, Department of Computer Science and Technology, Peking University, Beijing, China
[2] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[3] Institute for Artificial Intelligence, Peking University, Beijing, China [4] Peng Cheng Laboratory, Shenzhen, China
yuchenhong.cn@gmail.com, {zhengqian, exdjiang, eackot}@ntu.edu.sg, {calvinzhao, shiboxin}@pku.edu.cn

Abstract

This paper studies the problem of panoramic image reflection removal, aiming at reliving the content ambiguity between reflection and transmission scenes. Although a partial view of the reflection scene is included in the panoramic image, it cannot be utilized directly due to its misalignment with the reflection-contaminated image. We propose a two-step approach to solve this problem, by first accomplishing geometric and photometric alignment for the reflection scene via a coarse-to-fine strategy, and then restoring the transmission scene via a recovery network. The proposed method is trained with a synthetic dataset and verified quantitatively with a real panoramic image dataset. The effectiveness of the proposed method is validated by the significant performance advantage over single image-based reflection removal methods and generalization capacity to limited-FoV scenarios captured by conventional camera or mobile phone users.

1. Introduction

Single-image reflection removal addresses a severely ill-posed problem of recovering the transmission $\mathbf{T}$ from a reflection-contaminated or mixture image $\mathbf{M}$. A general image formation model of $\mathbf{M}$ is formulated as [11]

$$\mathbf{M} = \mathbf{\Omega} \odot \mathbf{T} + \mathbf{\Phi} \odot \mathbf{R}, \quad (1)$$

where $\odot$ is the element-wise multiplication operator, $\mathbf{\Omega}$ and $\mathbf{\Phi}$ are the refractive and reflective amplitude coefficient map, and $\mathbf{R}$ is the reflection scene [37]. The major challenge of this problem is that both $\mathbf{T}$ and $\mathbf{R}$ are part of different natural scenes, arousing the difficulty to differentiate the dominant content for $\mathbf{M}$. We call it *content ambiguity* in this paper. Early methods address it through content-free priors, *e.g.*, sparse distribution of reflection gradients [17] or ghosting cues [27], while state-of-the-art methods leverage both content and content-free priors from a large scale of training data, *e.g.*, LBCLN [14] and Kim *et al.* [9] (denoted as 'KH20' for brevity). Unfortunately, different distributions of $\mathbf{T}$ and $\mathbf{R}$ modeled by low-level or deep priors are not always observed in real scenarios, especially for strong reflections with sharp edges. Figure 1 displays an example where state-of-the-art methods fail to remove strong reflections.

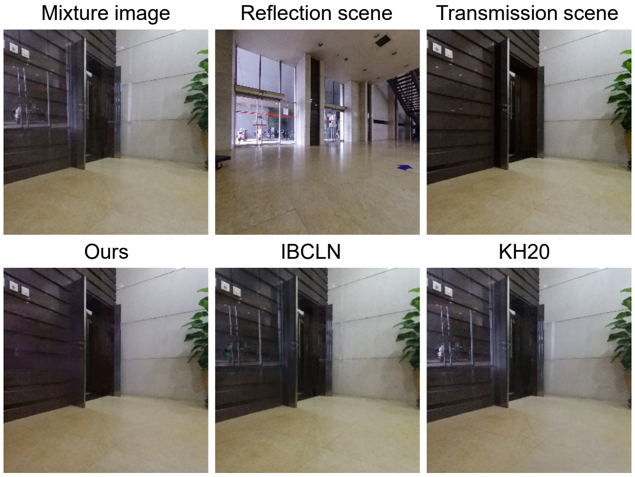

Figure 1. An example of our testing data and reflection removal results from our method, IBCLN [14], and KH20 [9].

The content ambiguity could be significantly relieved if we can (partially) capture the reflection scene. Fortunately, with the development of image stitching technology (*e.g.*, [6]), capturing panoramic images (also called 'panorama') becomes handily available, *i.e.*, by either off-the-shelf panoramic cameras for professionals (*e.g.*, Ricoh Theta series and Insta360 Pro series *etc.*) or camera phones for casual users (*e.g.*, panorama photography is a standard function for almost all smartphones nowadays such as Google Pixel and Apple iPhone *etc.*). A panoramic image has 360° field-of-view (FoV) and naturally contains a partial view of the reflection scene within a single shot, as

[#] Equal contribution. [*] Corresponding author.

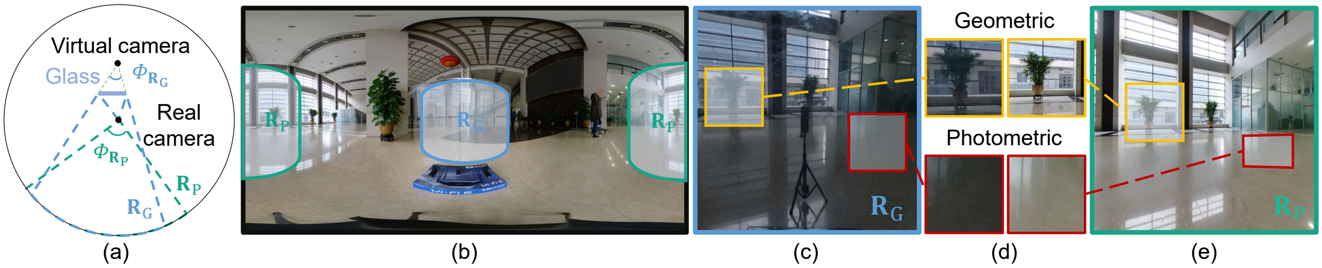
Virtual camera
Glass
$\Phi_{\mathbf{R}_{\mathrm{G}}}$
Real camera
$\Phi_{\mathbf{R}_{\mathrm{P}}}$
$\mathbf{R}_{\mathrm{P}}$
$\mathbf{R}_{\mathrm{G}}$
Geometric
Photometric
(a)
(b)
(c)
(d)
(e)

Figure 2. (a) Camera model of capturing a scene containing a glass plate by a panoramic camera. (b) Captured panoramic image. (c) Glass-reflected reflection image $\mathbf{R}_{\mathrm{G}}$, which is 'captured' by the virtual camera. (d) Illustration of the geometric and photometric misalignment. (e) Panoramic reflection scene $\mathbf{R}_{\mathrm{P}}$, which is captured by the real camera.

shown in Figure 2 (b). This motivates us to relieve the content ambiguity of reflection removal with a panoramic image.

Given a panoramic image, it seems to be straightforward to solve Equation (1) and remove reflections, since $\mathbf{R}$ has been 'captured' and $\mathbf{\Omega}$ and $\mathbf{\Phi}$ can be further simplified (*e.g.*, by assuming they are uniform across each image, as commonly adopted by previous works [38]). However, as shown in Figure 2 (a) and (b), the panoramic reflection scene captured by the real camera (*i.e.*, $\mathbf{R}_{\mathrm{P}}$) is not the glass-reflected reflection image 'captured' by the virtual camera (*i.e.*, $\mathbf{R}_{\mathrm{G}} = \mathbf{\Phi} \odot \mathbf{R}$). There exists geometric and photometric misalignment between the panoramic view $\mathbf{R}_{\mathrm{P}}$ and the glass-reflected view $\mathbf{R}_{\mathrm{G}}$, as shown in Figure 2 (c)−(e). The geometric misalignment is mainly caused by different positions of the real camera and the virtual one formed by glass, while the photometric misalignment is aroused from light attenuation when interacting with glass.

In this paper, we consider reflection removal using a single panoramic image. We solve this problem with a two-step solution including reflection alignment and transmission recovery. The first step adopts a coarse-to-fine strategy to align $\mathbf{R}_{\mathrm{P}}$ to $\mathbf{R}_{\mathrm{G}}$. The coarse alignment is achieved by a pre-processing procedure that explicitly considers misalignment factors (Section 3.2), while the fine-grained one is accomplished by a reflection refinement network which imposes the mutual information between $\mathbf{R}_{\mathrm{P}}$ and $\mathbf{M}$ (Section 3.3). With a precisely aligned $\mathbf{R}_{\mathrm{G}}$, the second step utilizes a transmission recovery network to restore $\mathbf{T}$ from $\mathbf{M}$ with the guidance of $\mathbf{R}_{\mathrm{G}}$ (Section 3.4). Our contributions can be summarized as follows:

- We present the first work to explicitly relieve the content ambiguity for reflection removal using a panoramic image.
- We solve the geometric and photometric misalignment between reflection scenes in panoramic and glass-reflected views, accompanying with high-fidelity transmission recovery after the alignment.
- We show that our method not only achieves superior performance advantage over single-image methods but also generalizes well to casual users without panoramic cameras.

2. Related Work

A panoramic image is generated by stitching multiple images from different viewpoints, but it cannot provide motion or parallax cues as inputs of multi-image reflection removal methods [11, 42, 21, 25, 20], since the overlap and correspondence information across different viewpoints have been lost after merging the panoramic image. 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.

Reflection removal. Existing methods for single-image reflection removal rely on the assumption of different distributions of transmission and reflection images, *i.e.*, reflection images are likely to be more blurry and with lower intensity compared with transmission images. Traditional methods formulate this assumption in their optimization pipeline, *e.g.*, image gradient sparsity priors [13], image gradient smoothness priors [16], ghosting cues [27], image content [34], and penalty on the gradient of recovered transmission images [1, 46]. For learning-based methods, they are developed to generalize the knowledge learnt from training data. They consider the assumption in the procedure of training data synthesizing, *e.g.*, blur natural images as reflection images with a Gaussian kernel [3, 49, 44, 15], directly capture the out-of-focus reflection images by placing black cloth behind a piece of glass [36, 38, 50], or render the out-of-focus reflection image [9]. Both optimization-based methods and learning-based methods rely on the assumption of different distributions of transmission and reflection images. However, they could fail to deal with scenarios where such assumption violates, *e.g.*, when there are strong reflections with sharp edges or content of two images are easily confused.

Applications of panoramic images. Thanks to the full FoV, panoramic images are useful in various computer vi-

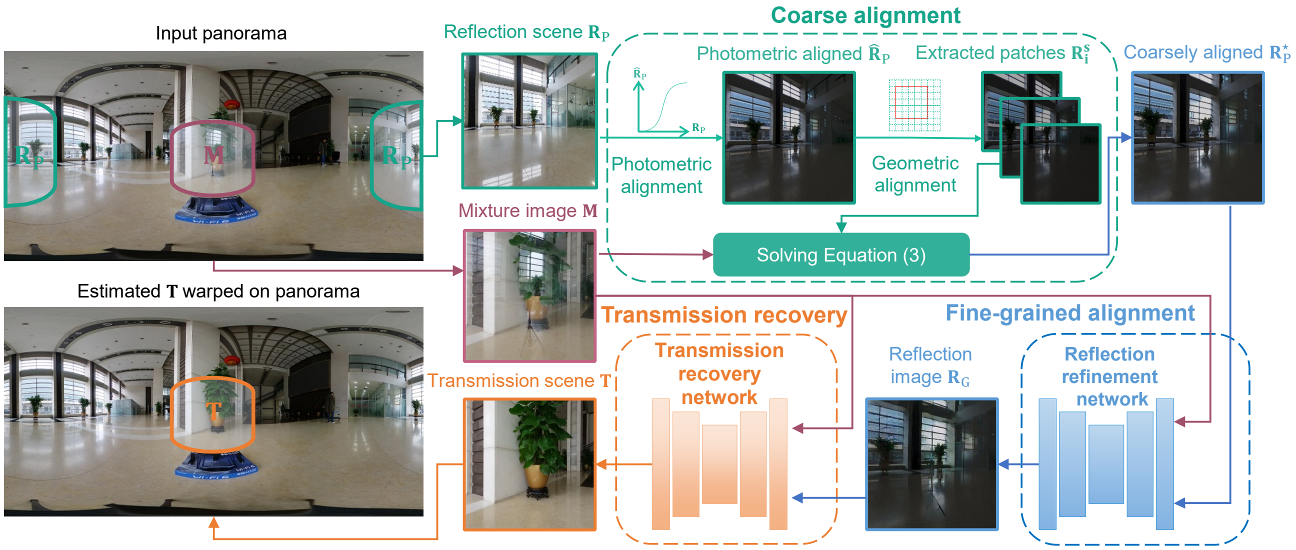
Input panorama
$\mathbf{R}_\mathrm{P}$
$\mathbf{M}$
$\mathbf{R}_\mathrm{P}$
Reflection scene $\mathbf{R}_\mathrm{P}$
Coarse alignment
Photometric aligned $\hat{\mathbf{R}}_\mathrm{P}$
Extracted patches $\mathbf{R}_\mathrm{i}^\mathrm{s}$
Coarsely aligned $\mathbf{R}_\mathrm{P}^*$
$\hat{\mathbf{R}}_\mathrm{P}$
$\mathbf{R}_\mathrm{P}$
Photometric alignment
Geometric alignment
Mixture image $\mathbf{M}$
Solving Equation (3)
Estimated $\mathbf{T}$ warped on panorama
Transmission recovery
Fine-grained alignment
Transmission recovery network
Reflection image $\mathbf{R}_\mathrm{G}$
Reflection refinement network
Transmission scene $\mathbf{T}$
$\mathbf{T}$

Figure 3. Given a panoramic image containing glass reflections with the mixture image $\mathbf{M}$ according to user-provided RoI, the panoramic reflection scene $\mathbf{R}_\mathrm{P}$ is automatically cropped. Our method settles the geometric and photometric misalignment between the panoramic reflection scene and the glass-reflected reflection image with a coarse-to-fine strategy. For the coarse alignment, $\mathbf{R}_\mathrm{P}$ is roughly aligned to the glass-reflected reflection image $\mathbf{R}_\mathrm{G}$ via a pre-processing procedure with the help of $\mathbf{M}$. Then a reflection refinement network for the fine-grained alignment takes the coarsely aligned $\mathbf{R}_\mathrm{P}^*$ and $\mathbf{M}$ as inputs, outputting an aligned reflection image $\mathbf{R}_\mathrm{G}$. The transmission recovery network estimates the transmission scene $\mathbf{T}$ from $\mathbf{M}$ with the guidance of $\mathbf{R}_\mathrm{G}$ and the estimated $\mathbf{T}$ is finally warped back to the panoramic image.

sion 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 [45, 32], indoor depth reconstruction [33], semantic segmentation, and vehicle detection [12]. 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 [31], lighting represented by an HDR panoramic image for either indoor [29, 18] or outdoor [48, 7] scenarios. In this paper, we further investigate how partial views of the reflection scene in a single panoramic image could be utilized to relieve the content ambiguity of reflection removal.

3. Proposed Method

Given a single panoramic image partially contaminated by reflections, the proposed method focuses on how to exploit content cues from the reflection scene to recover the transmission scene behind glass. Without losing generality, we ask the user to define a Region-of-Interest (RoI), which usually contains the reflection-contaminated area. The RoI is then automatically rectified to obtain the mixture image $\mathbf{M}$ for reflection removal. We then roughly extract and rectify the region of the panoramic reflection scene $\mathbf{R}_\mathrm{P}$ from the panoramic image based on glass orientation, which can be estimated according to the ratio of its two vertical edges (assuming the glass to be planar and orthogonal to the ground). We set the cropped regions with a wide FoV (*i.e.*, 90°) to avoid the omission of useful content information as shown in Figure 2 (a). Both of $\mathbf{M}$ and $\mathbf{R}_\mathrm{P}$ are resized to $h \times w$ for computation purpose. The reflection-removed results (transmission scene $\mathbf{T}$) could be visualized as directly or optionally warped back to the panoramic image.

In this section, we first analyze factors that impact the geometric and photometric misalignment between $\mathbf{R}_\mathrm{P}$ and $\mathbf{R}_\mathrm{G}$ in Section 3.1. We then propose a coarse-to-fine strategy to align $\mathbf{R}_\mathrm{P}$ to $\mathbf{R}_\mathrm{G}$ in Section 3.2 and Section 3.3. Finally, we introduce our transmission recovery network to recover $\mathbf{T}$ from $\mathbf{M}$ with the guidance of $\mathbf{R}_\mathrm{G}$ in Section 3.4. The pipeline of our method is displayed in Figure 3.

3.1. Misalignment Issues

Geometric misalignment. We define the geometric misalignment as the pixel-wise spatial discrepancy caused by different viewpoints of the real camera and the virtual one, as illustrated in Figure 2 (a). Besides, as $\mathbf{R}_\mathrm{P}$ is roughly cropped with a wide FoV, it contains a large proportion of content that cannot be found from $\mathbf{R}_\mathrm{G}$, which leaves additional problems for geometric alignment to solve.

Photometric misalignment. Given $\mathbf{R}_\mathrm{P}$ which has been well aligned to $\mathbf{R}_\mathrm{G}$ regarding geometry, we define the photometric misalignment as their pixel-wise difference [37].

Such misalignment comes from the light attenuation caused by glass, which is described by the reflective amplitude coefficient map $\mathbf{\Phi}$ [50]. We then represent the relationship between $\mathbf{R}_\text{P}$ and $\mathbf{R}_\text{G}$ as

$$\mathbf{R}_\text{G} \quad = \quad \mathbf{\Phi} \odot \mathbf{R}_\text{P}. \tag{2}$$

As existing alignment approaches (*e.g.*, RANSAC-Flow [26]) usually fail to handle above misalignment issues due to the impact of glass[1], we propose a coarse-to-fine alignment strategy to achieve geometric and photometric consistency for the relief of the content ambiguity.

3.2. Coarse Alignment

Our coarse alignment is achieved by a pre-processing procedure, where the geometric and photometric alignment are explicitly considered. We assume the glass orientation is not too large (*i.e.*, $< \ 30°$), based on our observation that people usually photograph in front of the glass plate rather than sideways if they intend to capture the transmission scene.

3.2.1 Geometric Alignment

We mainly consider the scale discrepancy and the spatial translation, while leaving the refinement of parallax to our fine-grained alignment. Because the distance from the camera to the glass plate is generally much smaller than that to the reflection scene, resulting significant scale discrepancy and spatial translation with slight parallax.

We employ an ergodic searching and matching method to deal with the scale discrepancy and spatial translation. To be specific, we use a sliding window to address spatial translation. That is, for all patches from $\mathbf{R}_\text{P}$ with different spatial positions, we aim at finding the one which best matches with $\mathbf{R}_\text{G}$. We use different sizes of sliding windows to consider the scale discrepancy, *i.e.*, each patch of $\mathbf{R}_\text{P}$ is resized according to a scalar s before matching. Denote each patch as $\mathbf{R}_i^s$, where $i \in P$ represents the position of the patch on $\mathbf{R}_\text{P}$, and $s \in S$ determines the patch size (*i.e.*, $sh \times sw$). The geometric alignment is achieved by finding the best matched patch $\mathbf{R}_\text{P}^\star$ among $\mathbf{R}_i^s$:

$$\mathbf{R}_\text{P}^\star = \arg\min_{\mathbf{R}_i^s} \{\text{D}(\mathbf{R}_i^s, \mathbf{R}_\text{G}) | \forall i \in P, \forall s \in S\}, \tag{3}$$

where $\mathbf{R}_i^s$ is scaled to be the same size as $\mathbf{R}_\text{G}$, $\text{D}(\cdot)$ measures the similarity of $\mathbf{R}_i^s$ and $\mathbf{R}_\text{G}$. We highlight the similarity regarding the global and local image structure:

$$\text{D}(\mathbf{R}_i^s, \mathbf{R}_\text{G}) = \Psi(\nabla \mathbf{R}_i^s, \nabla \mathbf{R}_\text{G}) + \frac{1}{K} \sum_{j=1}^{K} \Psi(\triangle_j \mathbf{R}_i^s, \triangle_j \mathbf{R}_\text{G}), \tag{4}$$

where ∇ is the operator to calculate the image gradient, $\triangle_j$ is the operation to extract the j-th small patch, and $\Psi(\cdot)$ is a function that measures the correlation between two images. In our experiment, we use the normalization cross correlation (NCC) [47] as $\Psi(\cdot)$ and set $K = 64$.

Though $\mathbf{R}_\text{G}$ is not accessible from a single panoramic image, $\mathbf{M}$ is a proper alternate with useful cues for our alignment algorithm as $\mathbf{R}_\text{G}$ is a component of $\mathbf{M}$. We also perform photometric alignment (introduced in Section 3.2.2) for $\mathbf{R}_i^s$ before the geometric alignment to highlight the role of strong reflections. We set s to be in the interval of $[0.45, 0.85]$[2] and sample s with the step of 0.05 for the above ergodic searching and matching procedure.

3.2.2 Photometric Alignment

Photometric alignment aims at producing $\hat{\mathbf{R}}_P$ from $\mathbf{R}_\text{P}$ to make it show the same photometric distribution as $\mathbf{R}_\text{G}$. Directly multiplying $\mathbf{R}_\text{P}$ with $\mathbf{\Phi}$ based on Equation (2) cannot output desired results owing to the non-linear in-camera image processing pipeline, especially for regions with strong reflections. That is, Equation (2) only holds for scene radiance, while it is not valid after the mapping of in-camera pipeline. Specifically, the dynamic range clipping (due to saturation) and non-linear mapping (due to radiometric response functions) [19] of in-camera image processing pipeline make intensities of $\mathbf{R}_\text{P}$ and $\mathbf{R}_\text{G}$ to be comparative for regions with large values, though those from $\mathbf{R}_\text{P}$ should be much larger than those from $\mathbf{R}_\text{G}$ in real scenes.

To this end, we use a paired dataset from [37] with geometrically aligned $\mathbf{R}_\text{P}$ and $\mathbf{R}_\text{G}$ to account for the in-camera image processing pipeline and the light attenuation together. The approximated relationship between $\mathbf{R}_\text{P}$ and $\mathbf{R}_\text{G}$ is built by estimating the non-linear function using fifth-order polynomial fitting as it is a common choice to model the camera's radiometric response function [22]. We then apply this relationship to process $\mathbf{R}_\text{P}$ and obtain a reflection scene that is more similar to $\mathbf{R}_\text{G}$, in terms of the photometric distribution, as shown in Figure 3.

3.3. Reflection Refinement

The coarse alignment addresses the majority of the misalignment between $\mathbf{R}_\text{P}$ and $\mathbf{R}_\text{G}$. However, the slight parallax still exists and the approximation for photometric alignment may not generalize well to all scenarios. Inspired by the fact that $\mathbf{M}$ provides useful cues for geometric and photometric alignment, we further perform the fine-grained alignment by employing a reflection refinement network, aiming to align the output of the coarsely aligned $\mathbf{R}_\text{P}^\star$ to $\mathbf{R}_\text{G}$. As illustrated in Figure 4, the reflection refinement network is composed of three modules, *i.e.*, feature extraction, feature fusion, and reflection generation. The re-

[1] Examples can be found in the supplementary material.

[2] More details about this setup are in the supplementary material.

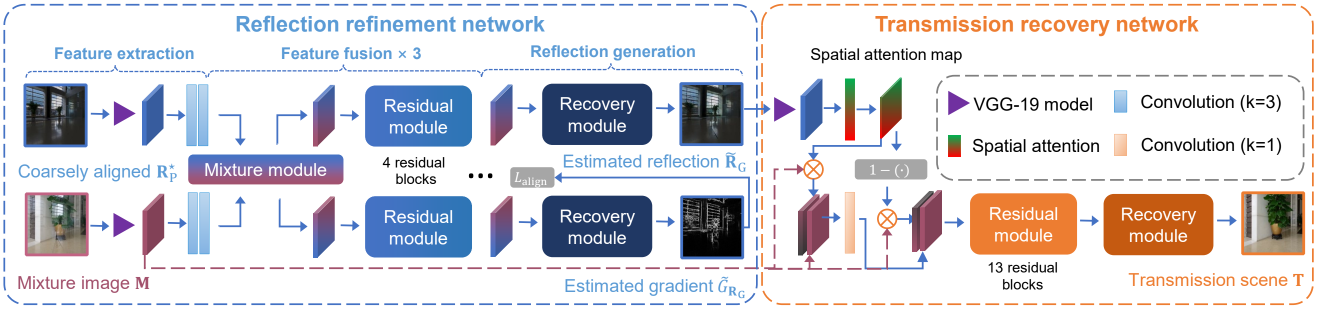
Reflection refinement network
Feature extraction
Feature fusion × 3
Reflection generation
Residual module
Recovery module
Coarsely aligned $\mathbf{R}_{\mathrm{P}}^{\star}$
Mixture module
4 residual blocks
$\mathcal{L}_{\text{align}}$
Estimated reflection $\widetilde{\mathbf{R}}_{\mathrm{G}}$
Residual module
Recovery module
Mixture image $\mathbf{M}$
Estimated gradient $\widetilde{G}_{\mathbf{R}_{\mathrm{G}}}$
Transmission recovery network
Spatial attention map
VGG-19 model
Convolution (k=3)
Spatial attention
Convolution (k=1)
1 − (·)
Residual module
Recovery module
13 residual blocks
Transmission scene $\mathbf{T}$

Figure 4. Illustration of detailed architectures of the reflection refinement network and the transmission recovery network. The reflection refinement network takes the coarsely aligned $\mathbf{R}_{\mathrm{P}}^{\star}$ and $\mathbf{M}$ as inputs and outputs an aligned reflection scene $\widetilde{\mathbf{R}}_{\mathrm{G}}$. The transmission recovery network estimates the transmission scene $\mathbf{T}$ by taking $\mathbf{M}$ and $\widetilde{\mathbf{R}}_{\mathrm{G}}$ as inputs.

finement network is a two-branch neural network, each of which takes the input of $\mathbf{M}$ and $\mathbf{R}_{\mathrm{P}}^{\star}$, respectively.

Feature extraction. Features of inputs are first extracted by the multi-level image feature pyramids based on the widely-used VGG-19 network [28], with the last four layers (*i.e.*, three fully connected layers and a Softmax layer) being removed to adapt our target of image-to-image translation. The extracted feature pyramids are then transformed into hypercolumn features, as they have been proved to be effective to learn semantic cues for the problem of reflection removal [49, 40]. To balance the efficiency and effectiveness, hypercolumn features are condensed by two convolutional blocks, respectively, each of which is composed of a convolution layer with kernel size 1×1 and an activation layer with the ReLU activation function [23].

Feature fusion. This stage contains three repetitive components, each of which consists of a mixture module for feature exchange and two parallel streams for feature learning. In the mixture module, features of $\mathbf{R}_{\mathrm{P}}^{\star}$ are firstly used to generate a spatial attention map via the spatial attention module [43] to highlight the spatial information of remarkable reflections. Features of $\mathbf{M}$ are then multiplied to the spatial attention map, concatenated with the original features, and condensed via a convolutional block with kernel size 1×1, to obtain features of content cues for reflections. For features of $\mathbf{R}_{\mathrm{P}}^{\star}$, they are concatenated with the condensed features of $\mathbf{M}$ to produce more content information regarding the reflection scene while isolating that from the transmission scene. After condensed by another convolutional block (kernel size 1×1), features of $\mathbf{R}_{\mathrm{P}}^{\star}$ for alignment can be acquired. The parallel streams then learn features that mutually exchange their content information. We repeat above procedures three times to extract discriminative features for reflection generation.

Reflection generation. This stage contains two generation streams for refined features of $\mathbf{M}$ and $\mathbf{R}_{\mathrm{P}}^{\star}$, respectively. Each stream generates features of reflections, which is composed of a transposed convolutional block for up-sampling and two convolutional blocks with a pyramid pooling module [40]. The top branch generates the estimated reflection image denoted as $\widetilde{\mathbf{R}}_{\mathrm{G}}$, while the bottom one obtains an estimated gradient map as $\widetilde{G}_{\mathbf{R}_{\mathrm{G}}}$, which is utilized for loss calculation to better constrain the fine-grained alignment.

Loss functions. Considering pixel-wise similarity and human perceptions jointly, we train our reflection refinement network with the following loss function:

$$\mathcal{L}_{\text{total}} = \omega_1 \mathcal{L}_{\text{pixel}} + \omega_2 \mathcal{L}_{\text{ssim}} + \omega_3 \mathcal{L}_{\text{feat}} + \omega_4 \mathcal{L}_{\text{align}}. \tag{5}$$

The weights are empirically set as $\omega_1 = 1$, $\omega_2 = 1$, $\omega_3 = 0.1$, and $\omega_4 = 0.5$ throughout our experiments. The Pixel loss function $\mathcal{L}_{\text{pixel}}$ is defined as the mean square error (MSE) between the estimated $\widetilde{\mathbf{R}}_{\mathrm{G}}$ and its ground truth $\mathbf{R}_{\mathrm{G}}$:

$$\mathcal{L}_{\text{pixel}}(\widetilde{\mathbf{R}}_{\mathrm{G}}, \mathbf{R}_{\mathrm{G}}) \quad = \quad \left\| \widetilde{\mathbf{R}}_{\mathrm{G}} - \mathbf{R}_{\mathrm{G}} \right\|_2^2. \tag{6}$$

The structural similarity loss function $\mathcal{L}_{\text{ssim}}$ tackles the blurry regions caused by the pixel loss and it is defined as:

$$\mathcal{L}_{\text{ssim}}(\widetilde{\mathbf{R}}_{\mathrm{G}}, \mathbf{R}_{\mathrm{G}}) = 1 - \mathrm{SSIM}(\widetilde{\mathbf{R}}_{\mathrm{G}}, \mathbf{R}_{\mathrm{G}}), \tag{7}$$

where the structural similarity index (SSIM) [39] measures the similarity of the illuminance, contrast, and structure between two images. The feature loss function $\mathcal{L}_{\text{feat}}$ is designed to measure the discrepancy of $\widetilde{\mathbf{R}}_{\mathrm{G}}$ and $\mathbf{R}_{\mathrm{G}}$ in feature space, which is similar to that in [49, 40]:

$$\mathcal{L}_{\text{feat}}(\widetilde{\mathbf{R}}_{\mathrm{G}}, \mathbf{R}_{\mathrm{G}}) = \sum\nolimits_i \lambda_i \left\| \Phi_i(\widetilde{\mathbf{R}}_{\mathrm{G}}) - \Phi_i(\mathbf{R}_{\mathrm{G}}) \right\|_1, \tag{8}$$

where $\{\lambda_i\}$ are the weights for equilibrium of multi-stage feature differences, and Φ_i represents the i-th convolutional layer in the VGG-19 model [28]. The alignment loss function $\mathcal{L}_{\text{align}}$ is designed to ensure the guidance via spatial attention maps from refined features of $\mathbf{R}_{\mathrm{P}}^{\star}$ to be reliable. It is accomplished by diminishing the pixel difference between the estimated gradient of the aligned reflection scene $\widetilde{G}_{\mathbf{R}_{\mathrm{G}}}$ and the gradient of the ground truth reflection image $G_{\mathbf{R}_{\mathrm{G}}} = \nabla \mathbf{R}_{\mathrm{G}}$:

$$\mathcal{L}_{\text{align}}(\widetilde{G}_{\mathbf{R}_{\mathrm{G}}}, G_{\mathbf{R}_{\mathrm{G}}}) \quad = \quad \left\| \widetilde{G}_{\mathbf{R}_{\mathrm{G}}} - G_{\mathbf{R}_{\mathrm{G}}} \right\|_2^2. \tag{9}$$

3.4. Transmission Recovery

As shown in Figure 4, our transmission recovery network recovers the transmission scene with the help of the estimated reflection image $\widetilde{\mathbf{R}}_{\mathrm{G}}$. As $\widetilde{\mathbf{R}}_{\mathrm{G}}$ is expected to contain the geometrically and photometrically aligned content information with the glass-reflected image, the content ambiguity in our transmission recovery is supposed to be relieved. Our network uses a similar architecture of [40], except that content cues of reflection scenes are embedded. Note that this architecture can also be replaced by other reflection separation methods such as CoRRN [38]. Compared with the architecture in [40], we add a branch to exploit helpful content information from $\widetilde{\mathbf{R}}_{\mathrm{G}}$. The additional branch extracts hypercolumn features from $\widetilde{\mathbf{R}}_{\mathrm{G}}$ and generates a spatial attention map to indicate the reflection region. Then features of $\mathbf{M}$ are multiplied with the attention map and concatenated to their product. A convolutional block condenses the fused features after the concatenation. The network repeats above operations once again except for the attention map, we subtract it by one to avoid the omission of certain transmission scenes. The fused features will be processed by the following residual blocks and up-sampling components of [40], ultimately recovering the distinct and clean transmission scene. We use the same training loss as that in [40] to train our transmission recovery network.

3.5. Data Preparation

Training data. Our method is trained using a synthetic dataset, which contains 5000 sets of data with transmission scenes, mixture images, reflection images, and unaligned reflection scenes. Reflection images are generated from reflection scenes with manually added photometric and geometric misalignment and blended with transmission scenes multiplied with a randomly scalar to simulate the light attenuation[3]. Images for synthesis are selected from an indoor image dataset SUN RGB-D [30] and an outdoor dataset Cityscapes [2], to cover various real scenarios.

Testing data. We collect two groups of real panoramic images for evaluation, including 30 sets as PORTABLE and 10 sets as NATURAL dataset. Images in PORTABLE are used for quantitative evaluation and visual quality comparison, which are captured by putting a portable glass in the scene. The corresponding transmission and reflection scenes are captured in the same way as in [35, 49]. Images in NATURAL are used for visual quality comparison, which are captured with glass found in different natural scenarios, such as office buildings. Samples from these two sets are collected by a single-shot panorama camera Ricoh Theta Z1. We further collect a real dataset named PHONE, which only contains mixture images and reflection scenes collected by casual users with a Huawei P40 Pro+ smartphone, for validation of the generalization capability of the proposed method.

Table 1. Comparisons of quantitative results in terms of PSNR [8], SSIM [39], NCC [47], and LMSE [4] on our PORTABLE dataset. ↑ (↓) indicates larger (smaller) values are better. Bold numbers indicate the best performing results.

Method	Error Metrics			
	PSNR↑	SSIM↑	NCC↑	LMSE↓
Ours	**23.986**	**0.749**	**0.926**	**0.021**
IBCLN [15]	20.636	0.709	0.862	0.031
KH20 [9]	20.443	0.711	0.849	0.035
CoRRN [38]	20.539	0.696	0.865	0.033
ERRNet [40]	21.444	0.701	0.87	0.029

Implementation details. The reflection refinement network and the transmission recovery network are implemented with PyTorch, a widely-used deep learning framework [24]. These networks are both trained 40 epochs with Adam [10] optimizer. The weights are initialized as in [5]. The learning rate is set to 10^{-4} initially and decreases to 10^{-5} at epoch 30.

4. Experiments

4.1. Comparison with Single-image Methods

We compare our method with four sate-of-the-art methods for single-image reflection removal[4], including IBCLN [15], KH20 [9], CoRRN [38], and ERRNet [41]. Following the evaluation for existing reflection removal methods [35, 40], we utilize PSNR [8], SSIM [39], NCC [47], and LMSE [4] as error metrics.

As can be found from Table 1, our method achieves much better quantitative performance regarding all error metrics compared with state-of-the-art single-image methods, *e.g.*, 0.749 over 0.709 regarding SSIM [39]. As the visual quality comparison on PORTABLE dataset in Figure 5 shown, all of these single-image methods fail to address the content ambiguity, *i.e.*, incorrectly enhance the image content from reflections due to sharp edges (first row) and fail to remove strong reflections caused by light sources (second row). Our method successfully suppresses strong reflections with sharp edges and produces much more faithful recovery of transmission images. We further conduct experiments on NATURAL dataset as illustrated in Figure 6[5], which further demonstrates the effectiveness of the proposed method using a panoramic view.

[3] More details about the synthetic data generation are in the supplementary material.

[4] The evaluation on the estimation of reflection images can be found in the supplementary material.

[5] A demo video displaying reflection removal in panoramic images with user interaction can be found in the supplementary material.

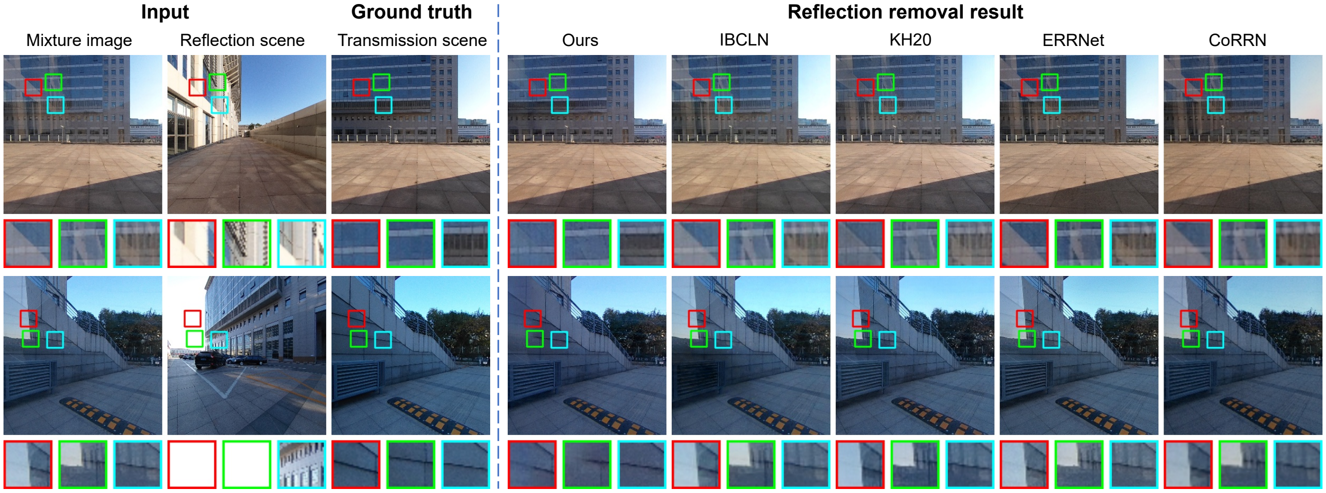

Figure 5. Examples of reflection removal results on PORTABLE dataset, compared with IBCLN [14], KH20 [9], ERRNet [40], and CoRRN [38]. Close-up views are displayed at the bottom of each image. Zoom in for better details.

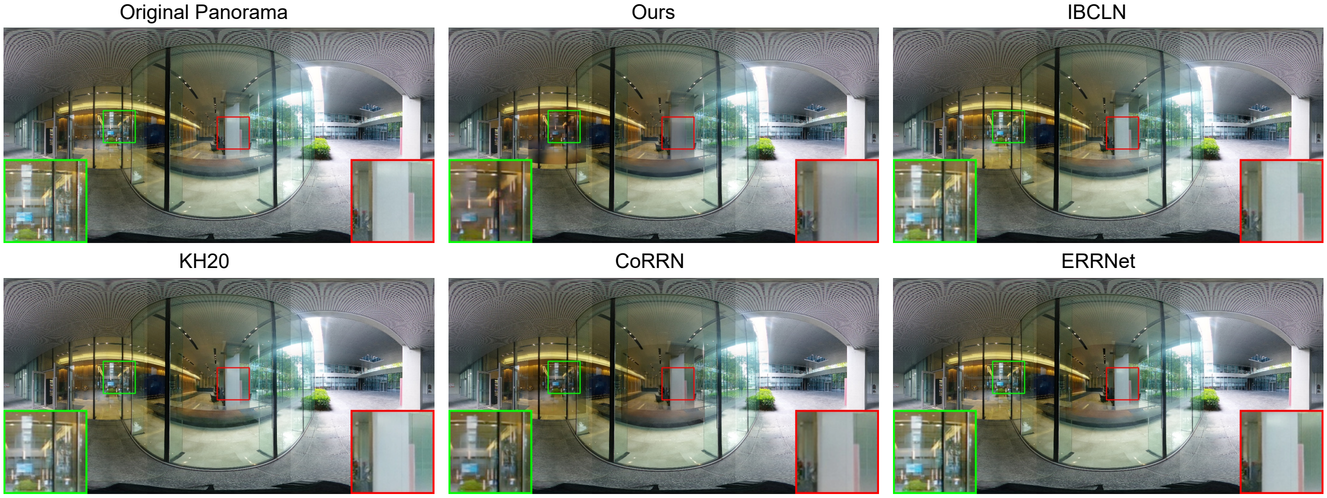

Figure 6. Visual quality comparison with state-of-the-art single-image reflection removal methods on NATURAL dataset. Close-up views are displayed in each image. Zoom in for better details.

4.2. Validation for Coarse-to-fine Alignment

We evaluate the effectiveness of our method in comparison with three variants: 1) 'coarse only' method that skips the fine-grained procedure, 2) 'fine-grained only' method that skips the pre-processing procedure, and 3) 'no alignment' method that directly takes $\mathbf{R}_\text{P}$ and $\mathbf{M}$ as inputs of the transmission recovery network.

Table 2. Comparisons of quantitative results in terms of PSNR [8], SSIM [39], NCC [47], and LMSE [4] on our PORTABLE dataset. Bold numbers indicate the best performing results.

Method	Error Metrics			
	PSNR↑	SSIM↑	NCC↑	LMSE↓
Ours	**23.986**	**0.749**	**0.926**	**0.021**
No alignment	22.539	0.724	0.895	0.027
Fine-grained only	23.473	0.738	0.909	0.024
Coarse only	23.288	0.737	0.907	0.025

Table 2 reports the quantitative comparison. As can be observed, both the coarse and fine-grained alignment play important roles in our method. Compared with the performance of single-image methods in Table 1, the no alignment method still achieves much better performance, indicating the effectiveness of our transmission recovery network and the setup of panoramic image reflection removal[6]. As can be observed from the visual comparison results in Figure 7,

[6] The ablation study on the transmission recovery network can be found in the supplementary material.

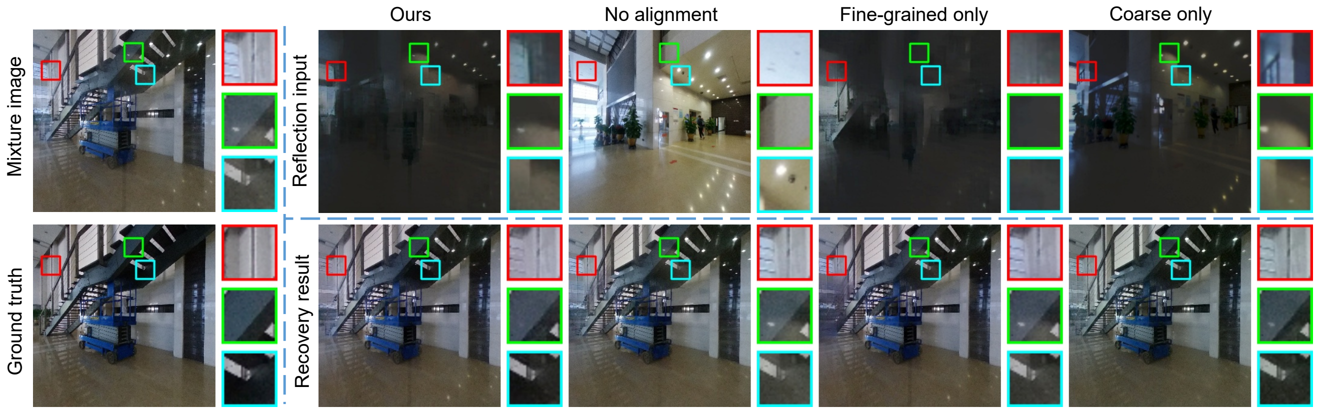

Figure 7. Visual quality comparison of different variants of our methods. From left to right: the mixture image/the ground truth of the transmission scene, estimated reflection images/recovered transmission scenes from our method, the no alignment method, the fine-grained only method, and the coarse only method. Close-up views are displayed at the right side of each image. Zoom in for better details.

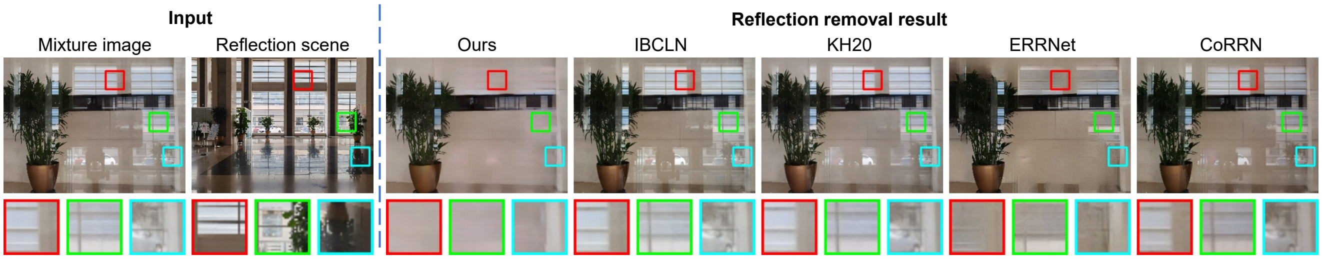

Figure 8. An example of reflection removal results on PHONE dataset, compared with IBCLN [14], KH20 [9], ERRNet [40], and CoRRN [38]. Close-up views are displayed at the bottom of each image. Zoom in for better details.

our method produces a more accurate estimation of the reflection image as compared with its variants (first row) and thereby recovers the transmission scene with better details and fewer reflection artifacts (second row).

4.3. Without using Panoramic Cameras

This section considers a more practical case for casual users who does not have panoramic cameras but use conventional cameras or mobile phones with limited FoV. After capturing a mixture image, the reflection scene can be obtained by turning over the camera for about 180°, while the constraints on $\mathbf{R}_{\mathrm{P}}$ and $\mathbf{R}_{\mathrm{G}}$ is not the same as that in a panoramic image, bringing different challenges for reflection alignment and transmission recovery. To evaluate the generalization capacity, we conduct experiments on PHONE dataset using our method and other single-image methods. As can be observed from Figure 8, our method achieves much better results and suppresses most of reflection artifacts. Single-image methods as IBCLN [14] and KH20 [9] incorrectly enhance reflection artifacts, with ERRNet [40] and CoRRN [38] only suppressing partial reflections. From the promising results, it can be verified that our two-step pipeline is well generalized for limited-FoV images, accompanying with the prominent advantage on relieving the content ambiguity for reflection removal.

5. Conclusion

We consider relieving the content ambiguity by taking a panoramic image as the input for reflection removal. We show that the major challenge of this problem is the geometric and photometric misalignment between the panoramic reflection scene and the glass-reflected reflection image, based on which a two-step solution composing of reflection alignment and transmission recovery is proposed. Experimental results demonstrate that our method not only achieves a significant performance advantage over single-image methods by relieving the content ambiguity, but also generalizes well to casual users. Though our method relieves the content ambiguity for most scenarios, it fails to inpaint extremely strong reflections with missing image content knowledge. Failure cases are shown in our supplementary material.

Acknowledgement

This work is supported by National Natural Science Foundation of China under Grant No. 61872012, 62088102, and Beijing Academy of Artificial Intelligence (BAAI). Qian Zheng is supported by the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore.

References

[1] Nikolaos Arvanitopoulos, Radhakrishna Achanta, and Sabine Süsstrunk. Single image reflection suppression. In *CVPR*, pages 1752–1760, 2017. 2

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 6

[3] 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 *ICCV*, pages 3258–3267, 2017. 2

[4] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009. 6, 7

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 6

[6] Charles Herrmann, Chen Wang, Richard Strong Bowen, Emil Keyder, Michael Krainin, Ce Liu, and Ramin Zabih. Robust image stitching with multiple registrations. In *ECCV*, 2018. 1

[7] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *CVPR*, pages 6927–6935, 2019. 3

[8] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 6, 7

[9] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single image reflection removal with physically-based training images. In *CVPR*, June 2020. 1, 2, 6, 7, 8

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[11] Naejin Kong, Yu-Wing Tai, and Joseph S Shin. A physically-based approach to reflection separation: from physical modeling to constrained optimization. *IEEE TPAMI*, 36(2):209–221, 2013. 1, 2

[12] Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In *CVPR*, pages 9181–9189, 2019. 3

[13] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE TPAMI*, 29(9):1647–1654, 2007. 2

[14] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E. Hopcroft. Single image reflection removal through cascaded refinement. In *CVPR*, June 2020. 1, 7, 8

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

[16] Yu Li and Michael S Brown. Exploiting reflection change for automatic reflection removal. In *ICCV*, pages 2432–2439, 2013. 2

[17] Yu Li and Michael S Brown. Single image layer separation using relative smoothness. In *CVPR*, pages 2752–2759, 2014. 1

[18] 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 *CVPR*, June 2020. 3

[19] 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 *CVPR*, pages 1651–1660, 2020. 4

[20] YuLun Liu, WeiSheng Lai, MingHsuan Yang, YungYu Chuang, and Jia-Bin Huang. Learning to see through obstructions. In *CVPR*, 2020. 2

[21] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Reflection separation using a pair of unpolarized and polarized images. In *NeurIPS*, pages 14532–14542, 2019. 2

[22] Zhipeng Mo, Boxin Shi, Sai-Kit Yeung, and Yasuyuki Matsushita. Ambiguity-free radiometric calibration for internet photo collections. *IEEE TPAMI*, 42(7):1670–1684, 2019. 4

[23] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 5

[24] 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 *NeurIPS*, 2019. 6

[25] Abhijith Punnappurath and Michael S Brown. Reflection removal using a dual-pixel sensor. In *CVPR*, pages 1556–1565, 2019. 2

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

[27] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman. Reflection removal using ghosting cues. In *CVPR*, pages 3193–3201, 2015. 1, 2

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[29] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *CVPR*, pages 6918–6926, 2019. 3

[30] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 2015. 6

[31] 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 *CVPR*, pages 3847–3856, 2018. 3

[32] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *CVPR*, pages 1047–1056, 2019. 3

[33] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *ECCV*, pages 707–722, 2018. 3

[34] 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 TIP*, 27(6):2927–2941, 2018. 2

[35] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *ICCV*, 2017. 6

[36] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. CRRN: Multi-scale guided concurrent reflection removal network. In *CVPR*, pages 4777–4785, 2018. 2

[37] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, and Alex C. Kot. Reflection scene separation from a single image. In *CVPR*, 2020. 1, 3, 4

[38] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, Ah-Hwee Tan, and Alex Kot Chichung. CoRRN: Cooperative reflection removal network. *IEEE TPAMI*, 2019. 2, 6, 7, 8

[39] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 5, 6, 7

[40] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *CVPR*, pages 8178–8187, 2019. 5, 6, 7, 8

[41] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *CVPR*, pages 3771–3779, 2019. 6

[42] Patrick Wieschollek, Orazio Gallo, Jinwei Gu, and Jan Kautz. Separating reflection and transmission images in the wild. In *ECCV*, pages 89–104, 2018. 2

[43] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, 2018. 5

[44] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *ECCV*, pages 654–669, 2018. 2

[45] 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 *CVPR*, pages 3363–3372, 2019. 3

[46] Yang Yang, Wenye Ma, Yin Zheng, Jian-Feng Cai, and Weiyu Xu. Fast single image reflection suppression via convex optimization. In *CVPR*, pages 8141–8149, 2019. 2

[47] Jae-Chern Yoo and Tae Hee Han. Fast normalized cross-correlation. *Circuits, systems and signal processing*, 28(6):819, 2009. 4, 6, 7

[48] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *CVPR*, pages 10158–10166, 2019. 3

[49] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, pages 4786–4794, 2018. 2, 5, 6

[50] 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 *CVPR*, 2020. 2, 4