Physics-Guided Reflection Separation from a Pair of Unpolarized and Polarized Images

Youwei Lyu[*‡], Zhaopeng Cui[*], *Member, IEEE,* Si Li,
Marc Pollefeys, *Fellow, IEEE,* and Boxin Shi[†], *Senior Member, IEEE*

Abstract—Undesirable reflections contained in photos taken in front of glass windows or doors often degrade visual quality of the image. Separating two layers apart benefits both human and machine perception. The polarization status of the light changes after refraction or reflection, providing more observations of the scene, which can benefit the reflection separation. Different from previous works that take three or more polarization images as input, we propose to exploit physical constraints from a pair of unpolarized and polarized images to separate reflection and transmission layers in this paper. Due to the simplified capturing setup, the system is more under-determined compared to the existing polarization-based works. In order to solve this problem, we propose to estimate the semi-reflector orientation first to make the physical image formation well-posed, and then learn to reliably separate two layers using additional networks based on both physical and numerical analysis. In addition, a motion estimation network is introduced to handle the misalignment of paired input. Quantitative and qualitative experimental results show our approach performs favorably over existing polarization and single image based solutions.

Index Terms—Reflection separation, polarization, deep learning.

1 Introduction

Semi-reflectors like glass windows reflect light from the side of photographers, and the taken photos are often mixtures of two layers of the scene: the layer transmitted through the glass and the other layer reflected by the glass. Separating the reflection and transmission layers enhances the visual quality of images and also benefits downstream computer vision tasks, such as image classification and semantic segmentation. However, it is not an easy task because recovering two images from a single blended image is highly ill-posed and the number of unknowns is twice as many as that of given measurements.

To tackle this challenging task, researchers tend to make the layer separation constrained by introducing assumptions and priors. Strong priors crafted from image formation models or natural image statistics, *e.g.*, gradient sparsity [1], different blur levels of estimated layers [2], [3] and ghost cues [4], assist in solving the problem if the assumed priors are well observed in the input. However, these approaches are likely to produce unsatisfactory results, when they are applied to images that obey disparate priors. As deep learning flourishes in recent years, researchers adopt deep convolutional neural networks to address the limitation of handcrafted priors. For example, CoRRN [5] introduces a concurrent model to tackle this problem in a cooperative and unified framework. Perceptual loss [6], the alignment-invariant loss [7] and other effective objective functions are applied to this task to generate compelling results.

The problem naturally becomes less ill-posed if multiple images are captured from different viewpoints (*e.g.*, five images in [8]) or different polarization angles (*e.g.*, at least three images in [9], [10]). The motions between the layers present in multiple images provide a strong and effective constraint, but aligning multiple-view images contaminated by reflections is not a trivial task [8]. Although additional information is extracted from three polarization images [10], [11], the separation is still under-determined and more priors are required for the solution. Assuming uniform polarization properties across the image, the separated images generated from pixel-wise calculation [9] or iterative optimization [10] are often affected by calculation artifacts and may deteriorate in disparate lighting conditions. The learning-based method [12] considers the intensity difference and pixel information of three input images, but hardly exploits the physical model. Besides, rotating a polarizer to capture multiple images doesn't suffer from the alignment issue [10], [12], but it requires skillful operations and the polarized images always filter part of the incoming light.

In this paper, we propose to separate reflection and transmission layers using a pair of unpolarized and polarized images. Such a setup takes fewer images than existing polarization-based methods [10], [12], [14], [15], [16], and keeps an unpolarized image to maintain sufficient luminous flux (*i.e.*, perceived power of light), which always requires careful consideration in designing of a practical camera system, especially for mobile platforms. Given a pair of (un)polarized images as input (Figure 1-Input), separating

[*] *Authors contributed equally to this work.* [‡] *Part of this was finished while working as a visiting student at Peking University.* [†] *Corresponding author.*

- *Y. Lyu and S. Li are with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. Email: {youweilv, lisi}@bupt.edu.cn.*
- *Z. Cui is with the State Key Lab of CAD&CG and the College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China. E-mail: zhpcui@zju.edu.cn.*
- *M. Pollefeys is with Department of Computer Science, ETH Zürich, CH-8092 Zürich, Switzerland. Email: marc.pollefeys@inf.ethz.ch.*
- *B. Shi is with the National Engineering Research Center of Visual Technology, School of Computer Science and Institute for Artificial Intelligence, Peking University, Beijing 100871, China. Email: shiboxin@pku.edu.cn.*

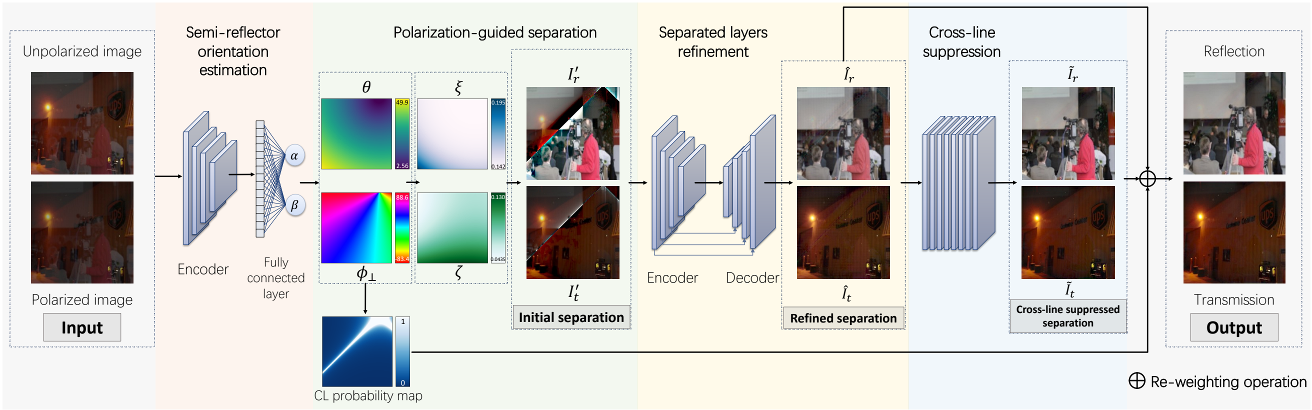


Fig. 1: The overall view of our framework. Our network takes a cascaded architecture with four modules: semi-reflector orientation estimation, polarization-guided separation, separated layers refinement, and cross-line suppression. Given a pair of unpolarized and polarized images as input, an initial separation guided by the polarization image formation model suffers from regional artifacts like "cross lines"; these artifacts can be partially suppressed by a refinement network [13]; by introducing a physics-guided CL probability map and a cross-line suppression module, we can obtain further refined results with less artifacts.

the reflection and transmission layers is still an ill-posed problem, but we find that semi-reflector surface normal encodes essential polarization information of the scenes and facilitates the solving process. By assuming the semi-reflector is mostly planar, only two planar parameters can determine the complete physical image formation model that encodes the solution to layer separation. Based on these observations, we propose an end-to-end physics-guided deep neural network for reflection separation using two (un)polarized images.

The proposed network takes the cascaded architecture consisting of four parts based on the physical and mathematical analysis. At first, we design a semi-reflector orientation estimation module to predict the two crucial variables for a well-posed physical image formation model. Then, according to the physical image formation model, we design a polarization-guided separation module which could generate the initial separation (Figure 1-Initial separation) based on the estimated coefficients and input images. It is noted that the polarization-guided separation is purely based on the physical model which has no parameters to be learned.

Due to the error in coefficient estimation and the nonlinear effect in real scenes, the initially separated layers may not be satisfying as expectation. Moreover, owing to physical limitations of the polarization based image formation model, the results generated from the area with low zenith angles tend to be downgraded and unstable, which is well known in shape from polarization [17], and the initial results might be also affected by regional artifacts, distributed like "cross lines" (Figure 1-Initial separation).

So we further design a refinement module and a cross-line suppression module to improve the initial layer separation. The refinement module is built based on the encoder-decoder architecture, aiming to enhance visual quality. For leveraging information in different aspects, we employ perceptual loss [18] for overall integrity and MSE loss as a pixel-level constraint in the training stage. The majority of artifacts are eliminated after the refinement, but there still remains residual artifact in the output separation (Figure 1-Refined separation). These residual artifacts actually correspond to special polarization angles or low zenith angles of the incident light rays. Based on the analysis of the cross-line effect, in the end, we propose to automatically annotate the cross-line area by a CL probability map, and design a cross-line suppression module to tackle such regional artifacts (Figure 1-Cross-line suppressed separation).

We show that our separation framework grounded on the imaging model exploits physical information effectively and enhances the overall performance (Figure 1-Output). We compare our method with the state-of-the-art methods on both synthetic data and real-world images. On the basis of physical derivations, the unified framework works well on wild images. By further introducing a motion estimation network, our model can handle the small misalignment of the (un)polarized images. The main contributions of this paper can be summarized as follows:

- We propose to solve reflection separation using a pair of unpolarized and polarized images, which integrates polarization cues with a simpler and light-efficient setup;
- We propose a unified end-to-end deep-learning framework with well-designed modules based on both the physical image formation mode and numerical analysis;
- We demonstrate that our method separates the reflection better than the state-of-the-arts, and can handle regional artifacts inherited in theoretical models and small misalignment in practical setups.

A preliminary version of this work appeared in [13], and in this paper we extend it in three aspects. First, we analyse the cross-line formation and propose a CL probability map for annotating cross-line regions and introduce a cross-line suppression module for mitigating the cross-line effect in the local regions. Second, we enhance the network by improving the architecture and introducing the perceptual loss [18] to improve visual quality of final results. Third, we study the influence of the misalignment of the (un)polarized images and show that our method could handle this using an off-the-shelf optical flow network as a pre-processing step.

Additional experiments and ablation studies are added to demonstrate the improvement over [13].

The remainder of this paper is organized as follows. In Section 2, we start with an introduction of existing works relevant to polarization methods and the image separation. Then Section 3 and Section 4 introduce the physical image formation model and our proposed method, respectively. Ablation studies, experimental results, and discussions on misaligned data are presented in Section 5. Finally, we conclude the paper in Section 6.

2 Related Work

2.1 Reflection Separation

In terms of input, reflection separation can take a single image or multiple images. The single image problem has the most relaxed requirement, since it only needs an image captured by an ordinary camera in the wild. But such a problem is also highly ill-posed, and priors formulated using hand-crafted priors [1], [2], [3], [4], [19], [20] or features learned from large-scale training data [6], [21], [22], [23], [24] are explored to facilitate the separation. By taking multiple images from different viewpoints, the difference of projected motion from reflection and transmission layers due to the visual parallax provides useful cues to the separation [25], [26], [27]. By taking multiple images under different polarization angles, the differently polarized images provide "independent" representations of reflection and transmission layers based on the physical image formation model to leverage the separation using independent component analysis [28], [29], [30], closed form expressions [9], [10], or deep learning [12]. Multiple images usually bring more promising separation quality than relying on only a single image, but request more complicated and careful image capturing operations.

In terms of solutions, reflection separation can be solved by non-learning based methods or learning methods. Adopted priors of reflection and transmission layers by non-learning based methods include the sparse gradient prior [1], [19], blur level differences between two layers [2], the ghosting effect due to thick glass [4], [31], and the Laplacian data fidelity term [20]. These gradient cues guiding separation can be extracted from manual annotation [1] and defocus disparity of dual pixel sensors [32]. Recently, the symmetry of reflection is utilized to remove reflection artifacts in the panorama [33]. Such handcrafted priors may get violated in various real scenarios when expected properties are weakly observed. Learning based methods are benefited by the comprehensive modeling ability of deep neural networks. It can be solved by learning the gradient inference and image restoration sequentially [21], [34] or concurrently [22], by incorporating perceptual losses [6] or by considering bidirectional constraints [23]. The laborious data collection process of paired images hinders the generalization ability of learning methods, which encourages researchers to propose the alignment-invariant loss function [7] and develop the weakly-supervised framework [35] exploiting misaligned training data. With differently polarized images available, a simple encoder-decoder architecture is shown to be effective for separating two layers using physics-based image formation model [12] or polarization information such as the degree of polarization and the angle of polarization [15].

Our work belongs to the learning based approach using multiple images and physical constraints. Different from previous works exploring polarization cues [9], [10], [12], [15] that require at least three images with different polarization angles, we take a pair of unpolarized and polarized images and learn to solve a more under-determined system.

2.2 Applications of Polarization

The polarization state of incident light provides cues to the reflection surface and transparent medium, which has been widely used in 3D vision. The surface orientation can be predicted by exploiting polarization properties of reflected light (Shape-from-polarization, SfP). The SfP cues providing the phase angle and degree of polarization, however, introduce ambiguities of surface normal either. Boundary constraints [36], [37] and convexity assumptions [38] are proposed to disambiguate the surface normal. Further, a number of methods combine the polarimetric information with additional constraints, *e.g.*, multi-spectral measurements [39], depth maps obtained by an RGBD camera [40], [41], and shape-from-shading information [42], [43], [44], in which ambiguities are resolved and more accurate estimation of surfaces can be obtained. In multi-view stereo, polarization methods help in enabling transparent surface modeling [45] and recovery of surface shape in featureless regions [46]. Stereo polarization cues have been used for depth estimation [47] and dense SLAM reconstruction [48]. Recent works propose to estimate epipolar geometry by phase information [49] or geometric information available from polarization cameras [50].

In computational photography domains, polarization is used in special imaging systems. For example, scattering of sunlight in the atmosphere creates a characteristic polarization pattern in the sky, which motivates researchers to design the visual compass system for estimating sun direction and its covariance [51]. The polarization cues also enable non-line-of-sight (NLOS) imaging by improving the conditioning of the light transport matrix [52]. Polarization methods are also used in imaging applications, *e.g.*, image dehazing [53], image mosaicing, panoramic stitching [54], and reflection separation [9], [10], [15].

3 Physical Image Formation Model

Given a pair of unpolarized and polarized images captured at the same view, we aim to separate the reflection layer and the transmission layer. In this section, we will first review the reflection and transmission model, and describe the relationship between polarization properties and semi-reflector surface geometry. By assuming the medium is planar, we prove that the separation tightly relies on only two parameters of the plane. Then we analyse the cross-line artifacts that might appear in the initial separation.

3.1 Reflection and Transmission Image Formation

Suppose $I_t(x)$, the intensity of light from the transmission scene, and $I_r(x)$, the intensity of light from the reflection

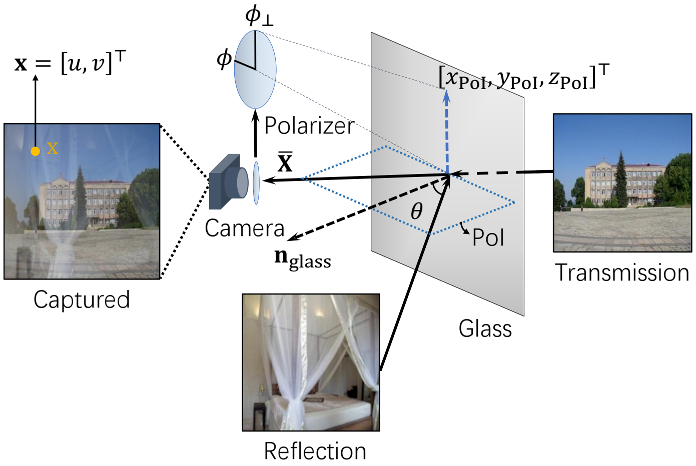

$\mathbf{x} = [u, v]^\top$
$\phi_\perp$
ϕ
Polarizer
$[x_{\text{PoI}}, y_{\text{PoI}}, z_{\text{PoI}}]^\top$
$\overline{\mathbf{X}}$
Camera
θ
PoI
$\mathbf{n}_{\text{glass}}$
Captured
Transmission
Glass
Reflection

Fig. 2: Illustration of physical image formation model.

scene, are both unpolarized. After being reflected or transmitted, the intensity of light observed at pixel x changes depending on $\theta(x)$, the angle of incidence (AoI) at the reflected point corresponding to pixel x, as the following [10]. For conciseness, we omit the reference of pixel x in the rest of the paper.

$$I_{\text{unpol}} = \frac{R_\perp(\theta) + R_\parallel(\theta)}{2} I_r + \frac{T_\perp(\theta) + T_\parallel(\theta)}{2} I_t, \tag{1}$$

where R represents the relative strength of light reflected off the glass surface, T represents the relative strength of light transmitted through the glass, and subscripts $\perp$ and $\parallel$ correspond to the polarized components perpendicular and parallel to the plane of incidence (PoI), respectively.

When we place a linear polarizer with a polarization angle ϕ in front of the camera, according to Malus' law [55], the received light intensity is

$$\begin{aligned} I_{\text{pol}} =& \frac{R_\perp(\theta)\cos^2(\phi - \phi_\perp) + R_\parallel(\theta)\sin^2(\phi - \phi_\perp)}{2} I_r + \\ & \frac{T_\perp(\theta)\cos^2(\phi - \phi_\perp) + T_\parallel(\theta)\sin^2(\phi - \phi_\perp)}{2} I_t, \end{aligned} \tag{2}$$

where $\phi_\perp$ is the orientation of the polarizer for the best transmission of the component perpendicular to the PoI. For simplicity, we denote

$$\xi = R_\perp(\theta) + R_\parallel(\theta), \tag{3}$$

$$\zeta = R_\perp(\theta)\cos^2(\phi - \phi_\perp) + R_\parallel(\theta)\sin^2(\phi - \phi_\perp). \tag{4}$$

The glass can be considered as a double-surfaced semi-reflector, and we have $R_\perp(\theta) + T_\perp(\theta) = 1$ and $R_\parallel(\theta) + T_\parallel(\theta) = 1$ for each pixel x approximately [10]. Under the double-surface assumption, $R_\perp(\theta)$ and $R_\parallel(\theta)$ are given by

$$R_\perp(\theta) = \frac{2\sin^2(\theta - \theta_t(\theta,\kappa))}{\sin^2(\theta - \theta_t(\theta,\kappa)) + \sin^2(\theta + \theta_t(\theta,\kappa))}, \tag{5}$$

$$R_\parallel(\theta) = \frac{2\tan^2(\theta - \theta_t(\theta,\kappa))}{\tan^2(\theta - \theta_t(\theta,\kappa)) + \tan^2(\theta + \theta_t(\theta,\kappa))}, \tag{6}$$

in which κ is the refractive index set to be 1.474 for regular glass, and $\theta_t(\theta,\kappa) = \arcsin(\frac{1}{\kappa}\sin\theta)$ according to Snell's law [55]. Then Equation (1) and Equation (2) can be rewritten as

$$I_{\text{unpol}} = \frac{\xi}{2} I_r + \frac{2-\xi}{2} I_t, \tag{7}$$

$$I_{\text{pol}} = \frac{\zeta}{2} I_r + \frac{1-\zeta}{2} I_t, \tag{8}$$

where $\xi \in (0, 2)$ and $\zeta \in (0, 1)$. Given the value of ξ and ζ, the reflection layer and the transmission layer can be computed by

$$I_r = 2\frac{(2-\xi)I_{\text{pol}} - (1-\zeta)I_{\text{unpol}}}{2\zeta - \xi}, \tag{9}$$

$$I_t = 2\frac{\zeta I_{\text{unpol}} - \xi I_{\text{pol}}}{2\zeta - \xi}, \tag{10}$$

except for $2\zeta = \xi$ where $\phi - \phi_\perp = \pm 45^\circ$ or $\pm 135^\circ$. The angle of a polarizer ϕ can be measured by calibration. Associated with surface geometry of semi-reflector, $\phi_\perp$ is not constant but spatially varying over the whole image plane. There may exist trivial $\phi - \phi_\perp$ corresponding to a few pixels, which can produce cross-line artifacts in the initial separation. We will discuss this issue in Section 3.3.

In short, the reflection layer I_r and the transmission layer I_t are determined by ξ and ζ when a pair of unpolarized and polarized images are given.

3.2 Semi-reflector Surface Geometry

In order to recover the reflection layer I_r and the transmission layer I_t, we first have to solve ξ and ζ according to Equations (7) and (8), which can be further computed by θ and $\phi - \phi_\perp$ according to Equations (3) and (4). In this section, we will describe how we compute θ and $\phi - \phi_\perp$ for each pixel given the surface normal of the semi-reflector and camera parameters.

We assume the semi-reflector has a planar surface, and the camera coordinate coincides with the world coordinate in $x-$ and $y-$axis. Then the semi-reflector plane can be expressed as

$$\sin\alpha \cdot x - \cos\alpha\sin\beta \cdot y + \cos\alpha\cos\beta(z - z_0) = 0, \tag{11}$$

where α represents the rotation angle around y-axis, β represents the angle around x-axis, and z_0 denotes the distance between the imaging plane and the glass. The plane normal is thus given by

$$\mathbf{n}_{\text{glass}} = \begin{bmatrix} \sin\alpha & -\cos\alpha\sin\beta & \cos\alpha\cos\beta \end{bmatrix}^\top. \tag{12}$$

Let f be the focal length of the camera, and (p_x, p_y) be the coordinate of the principal point. For the pixel x located at (u, v) on the image plane, we can easily compute the direction vector of its corresponding 3D point as follows:

$$\mathbf{X} = \begin{bmatrix} u - p_x & v - p_y & f \end{bmatrix}. \tag{13}$$

Let $\overline{\mathbf{X}} = \mathbf{X}/\|\mathbf{X}\|$, then the AoI corresponding to pixel x can be calculated as

$$\theta = \arccos\left|\mathbf{n}_{\text{glass}} \cdot \overline{\mathbf{X}}\right|. \tag{14}$$

We calculate the absolute value for the above term since $\theta \in [0, 90^\circ)$. The normal of PoI $\mathbf{n}_{\text{PoI}} = (x_{\text{PoI}}, y_{\text{PoI}}, z_{\text{PoI}})^\top$ is then calculated as

$$\mathbf{n}_{\text{PoI}} = \mathbf{n}_{\text{glass}} \times \overline{\mathbf{X}}, \tag{15}$$

and the projection of $\mathbf{n}_{\text{PoI}}$ on the imaging plane is $(x_{\text{PoI}}, y_{\text{PoI}})^\top$ denoting the orientation of $\phi_\perp$. For $\phi_\perp \in [-180^\circ, 180^\circ)$, we have

$$\phi_\perp = \arctan\frac{y_{\text{PoI}}}{x_{\text{PoI}}}. \tag{16}$$

We combine the reflection and transmission image formation and semi-reflector surface geometry to compute $\phi_\perp$ and θ for each pixel. Note that they are not affected by z_0, because physically the transparent plane can be projected to parallel plane with arbitrary intercept about z-axis and mathematically before computing $\arctan$ and $\arccos$, z_0 has been eliminated according to Equation (16).

In short, it is the normal of glass that matters, and we only need to estimate coefficients α and β to determine the semi-reflector plane. Different from other numerical expressions in this paper, coefficients α and β are constant at each pixel of the same image, respectively.

3.3 "Cross Line" Artifacts

In Section 3.1, when we calculate the reflection and transmission layers by Equations (9) and (10), cross-line-distributed artifacts may affect the initial separation results as shown in Figure 3a. To figure out how it happens, we rewrite Equations (9) and (10) as

$$I_r = I_{\text{unpol}} + \Theta_r(\theta)\Phi(\phi_\perp)\left(2I_{\text{pol}} - I_{\text{unpol}}\right), \tag{17}$$

$$I_t = I_{\text{unpol}} - \Theta_t(\theta)\Phi(\phi_\perp)\left(2I_{\text{pol}} - I_{\text{unpol}}\right), \tag{18}$$

in which Θ_r and Θ_t denote two functions of θ, corresponding to the reflection and transmission layer respectively, and Φ denotes the function of $\phi_\perp$, as follows:

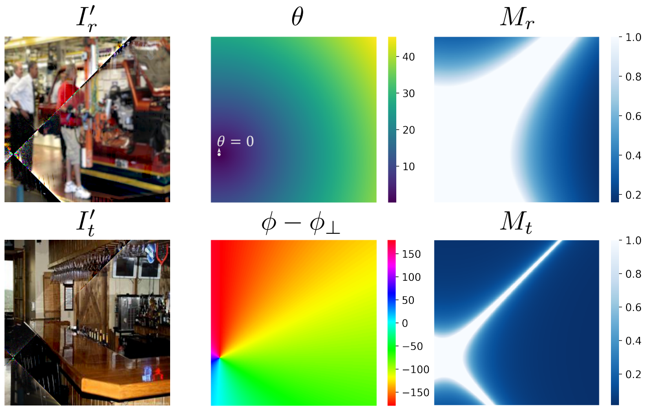
$$\Theta_r(\theta) = \frac{2 - \left(R_\perp(\theta) + R_\parallel(\theta)\right)}{R_\perp(\theta) - R_\parallel(\theta)}, \tag{19}$$

$$\Theta_t(\theta) = \frac{R_\perp(\theta) + R_\parallel(\theta)}{R_\perp(\theta) - R_\parallel(\theta)}, \tag{20}$$

$$\Phi(\phi_\perp) = \frac{1}{\cos 2\left(\phi - \phi_\perp\right)}. \tag{21}$$

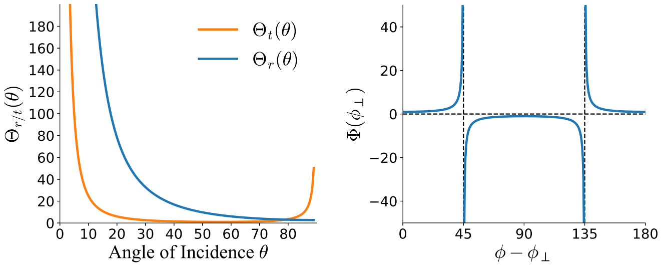
Given the two observed images I_{unpol} and I_{pol} and the estimated physical coefficients θ and $\phi_\perp$, we can compute reflection and transmission layers. To analyse how physical information influences the calculated layers, we plot the magnitude curves of the above three functions in Figure 4. For the terms $\Theta_{r/t}(\theta)$ and $\Phi(\phi_\perp)$, their values increase when the variables approach some trivial points, *e.g.*, $\theta = \{0°, 90°\}$ and $\phi - \phi_\perp = \{\pm 45°, \pm 135°\}$, greatly magnifying the fractional noise and quantization error in mixture images and making part of the separated layer deteriorate. Based on the physical model, $\phi - \phi_\perp$ is always radially distributed and centered at the point $\theta = 0°$ as shown in Figure 3b. Thus, artifacts are magnified to obviously form cross lines over the image.

From the perspective of imaging, revisiting Equations (1) and (2), we find the intensities of the unpolarized image I_{unpol} are just twice as those of the polarized image I_{pol} on pixels $\theta = 0°$ or $\phi - \phi_\perp = \{\pm 45°, \pm 135°\}$, in which the two input images provide the same information about the mixture scene. Lack of necessary polarization cues, the calculation results degrade around these pixels.



(a) Initial separation (b) Parameter map (c) CL probability map

Fig. 3: (a) Initially separated layers I'_r and I'_t are calculated by Equation (9) and (10). Due to the numerical problem, regional artifacts in the form of "cross lines" can be obviously observed in separated layers. (b) An example of the value distribution of physical parameters: angle of incidence θ and $\phi - \phi_\perp$. $\phi - \phi_\perp$ is radially distributed and its value changes around the point $\theta = 0$. So the artifacts in pixels $\phi - \phi_\perp = \{\pm 45°, \pm 135°\}$ are distributed as "cross lines", and the intersection of cross lines lies in the point $\theta = 0$. (c) The CL probability map (stretched and scaled for better visualization) stemmed from physical parameters. For pixels in the initially separated images, larger values of the map represent higher probabilities of being affected by cross-line artifacts.



(a) (b)

Fig. 4: The magnitude curves of the functions (a) $\Theta_{r/t}$ and (b) Φ regarding θ and $\phi - \phi_\perp$. At some trivial point, the values of these two functions increase dramatically, which magnify subtle noise and produce unreliable separated layers.

We try to handle the cross-line issue by adding a regularization term or with the linear constraints that the intensity values are within $(0, 1)$ using the alternating direction method of multipliers (ADMM) [56], but they fail to solve the problem well. Please refer to Appendix A for more details. To alleviate the cross-line effect, we further propose to enhance the network with a cross-line suppression module. Based on the observation that the area affected by "cross line" tends to be with large product terms (*i.e.*, $\Theta_{r/t}(\theta)\Phi(\phi_\perp)$) in Equation (17) and Equation (18), we present the CL probability map, shown in Figure 3c, for annotating the unstable regions in initially separated layers, which is denoted as

$$M_r = 2\text{sigmoid}\left(|\Theta_r(\theta)\Phi(\phi_\perp)|\right) - 1, \tag{22}$$

$$M_t = 2\text{sigmoid}\left(|\Theta_t(\theta)\Phi(\phi_\perp)|\right) - 1, \tag{23}$$

of which larger values represent higher probabilities that the pixels in initial separation are affected by the cross-line artifacts.

With the CL probability map annotating the cross-line area, the cross-line suppression module is designed for further improvement of separation results and we will introduce it in Section 4.1.

4 Reflection Separation Network

In this section, we introduce the proposed reflection separation network which makes use of the physical model discussed in Section 3, and details about loss functions and network training.

4.1 Network Architecture

As shown in Figure 1, our network takes a cascaded architecture which consists of four modules: semi-reflector orientation estimation, polarization-guided separation, separated layers refinement, and cross-line suppression.

Taking a pair of unpolarized and polarized images, the semi-reflector orientation module aims to predict coefficients of the glass plane, *i.e.*, α and β. As we only need to estimate two parameters, the pose estimation module is pretty light and consists of seven convolutional layers followed by two fully connected layers.

The polarization-guided separation module takes α and β as inputs, and computes the reflection layer I_r' and transmission layer I_t'. This module only relies on the physical image formation model in Section 3 using analytic equations, so we do not have any parameters to learn here.

The separated layers using equations may not be satisfactory due to the gap between the physical model and real data. Part of the transmission layers may be still affected by reflection artifacts due to inevitable estimation error of physical parameters. The numerical problem also occurs when the denominators in Equation (9) and Equation (10) approach zero, and the computed results are degenerated as mentioned in Section 3.3. Fortunately, this happens only for a few of pixels and the remaining non-degenerated calculations can guide a refinement network to enhance separation results. We therefore further feed I_r' and I_t' with original input images and ξ, ζ into the separated layers refinement module to improve the initial estimation. The refinement module has a widely adopted encoder-decoder architecture. In detail, the encoder consists of eight convolutional layers and the decoder consists of five deconvolutional layers. We denote the refinement network as

$$\hat{I}_r, \hat{I}_t = \mathcal{F}_{\textbf{RF}}\left(I_{\text{unpol}}, I_{\text{pol}}, I_r', I_t', \xi, \zeta\right). \tag{24}$$

In contrast to [13], we replace 7×7, 5×5 convolutional kernels to 3×3 ones, and add a 1×1 convolutional layer at the end of the deconvolutional blocks.

As shown in Figure 1-Refined separation, fed with the initial separation $\{I_r^{'}, I_t^{'}\}$ that are affected by cross-line artifacts, the refinement module has removed most of the contamination globally, but cross lines obviously remain in intermediate separation, especially in the reflection layer. To eliminate the regional artifacts in the reflection and transmission layers and further improve the visual quality of separated layers, we present the cross-line suppression module by taking the concatenated input (un)polarized images and the estimated layer from the refinement module. Aiming at improving affected regions, the cross-line suppression network is shallow and consists of four ResNet [57] blocks with the output:

$$\widetilde{I}_r, \widetilde{I}_t = \mathcal{F}_{\textbf{CLS}}\left(I_{\text{unpol}}, I_{\text{pol}}, \hat{I}_r, \hat{I}_t\right). \tag{25}$$

Finally, we linearly combine $\hat{I}_r, \hat{I}_t$ and $\widetilde{I}_r, \widetilde{I}_t$, and produce the final separation as the following:

$$I_r^* = \hat{I}_r(1-W) + \widetilde{I}_r W, \tag{26}$$

$$I_t^* = \hat{I}_t(1-W) + \widetilde{I}_t W. \tag{27}$$

In practice, we adopt the CL probability maps of transmission layers M_t as weighting parameters W, since M_t and M_r corresponding to the same image pair have the identical distribution, and only vary in magnitude, as shown in Figure 3c. Empirically, we find that M_t reflects the affected regions better than M_r.

4.2 Loss Functions

Pixel loss. In image reconstruction, pixel-wise loss is one of the popular and efficient objective functions that supervise the network producing results close to the references. We adopt mean squared error (MSE), denoted as $\mathcal{L}_{\text{MSE}}$, to measure the distance between estimated images and ground truth:

$$\mathcal{L}_{\text{pixel}} = \mathcal{L}_{\text{MSE}}\left(I_r^*, I_r\right) + \mathcal{L}_{\text{MSE}}\left(I_t^*, I_t\right). \tag{28}$$

Perceptual loss. The perceptual loss [18] has been proved effective in image decomposition tasks [6], and it helps to preserve details and enhance perceptual quality of output images. We use the pre-trained AlexNet for generating the activation layer:

$$\mathcal{L}_{\text{LPIPS}} = \left\|\Psi(I_r^*) - \Psi(I_r)\right\|_2 + \left\|\Psi(I_t^*) - \Psi(I_t)\right\|_2. \tag{29}$$

For semi-reflector estimation training, MSE is adopted as the error metric. We hope the reconstruction network to preserve details in the original image as many as possible, so we define the loss function for refinement network and cross-line suppression module as:

$$\mathcal{L}_{\textbf{total}} = \lambda_1 \mathcal{L}_{\text{pixel}} + \lambda_2 \mathcal{L}_{\text{LPIPS}}, \tag{30}$$

where $\lambda_{1,2}$ are the weighting parameters of the pixel loss and the learning perceptual loss, respectively.

4.3 Implementation Details

We implemented our model using PyTorch deep-learning framework [58] and used the Adam [59] solver with default parameters. To improve the performance of our model, the multi-stage method was adopted for training. We first trained the orientation estimation network with an initial learning rate of 0.001 for 30 epochs until convergence. Given the predicted planar coefficients, we successively trained the refinement network and the cross-line suppression module for 50 epochs with the same strategy: the learning rate was initially set to 4×10^{-4} and halved every 15 epochs. For the final separation results, we fine-tuned the refinement module with the cross-line suppression module based on Equation (26) and Equation (27) for 40 epochs. The learning rate was 1×10^{-4} and halved every 10 epochs. Additionally, hyperparameters $\lambda_{1,2}$ were set to be $1, 0.05$ in training, respectively.

4.4 Training Data Generation

The deep-learning method tends to be data-hungry, but it is laborious to obtain pairwise reflection and transmission images with both polarized and unpolarized observations at a large scale. It is possible to directly use Equation (1) and Equation (2) to generate the synthetic data, but it is expected that the network trained with such data may not generalize well on real scenarios. Therefore, we propose an effective data generation pipeline to better match images of real-world scenes.

At the first step, we randomly pick two images from PLACE2 dataset [60] as original reflection and transmission layers. Based on a commonly adopted assumption that people take photos focusing on the background scene (the transmission layer) so the reflection layer is likely to be blurry [21], a Gaussian smoothing kernel with a random kernel size in the range of 3 to 5 pixels is applied to a portion of reflection images. We also need to simulate the coefficients α and β of the semi-reflector plane. We assume people rarely take photos in front of the glass that inclines by a weird angle, *e.g.*, glass nearly orthogonal to the image plane, so we set $\alpha \in (-60°, 60°)$ and $\beta \in (-60°, 60°)$. We render a total of $55,000$ sets of synthetic images, where $50,000$ sets are used for training, and $5,000$ sets used for testing.

For the virtual camera, we set the focal length as 1.4 times as long as the image width, and the image resolution as 256×256. By fixing these factors, the normal of glass is specified, θ and $\phi_{\perp}$ can be derived from Equation (14) and Equation (16), respectively. ϕ can be an arbitrary value in the range of $[0, 2\pi)$, as long as the polarization images are captured under the same polarizer angle. In our experiment, we set ϕ to be 0. Additionally, real-world scenes are generally high-dynamic-range (HDR), so we apply dynamic range manipulation as conducted in [12] to simulate the appearance of reflections in a more realistic manner. Finally, the synthetic unpolarized image I_{unpol} and the polarized image I_{pol} can be obtained by Equation (7) and Equation (8).

5 Experimental Results

We evaluate our method on both synthetic and real data with extensive experiments including the comparison with related work and the ablation study. Besides, we simulate misalignment between the (un)polarized images, which may happen in real cases, and test the proposed model on the misaligned pairs. For all quantitative evaluations, both the peak-signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) [61] are used to evaluate the quality of separated images.

5.1 Ablation Study

To understand the impact of each module and loss function on the final performance, we conduct a comprehensive ablation study by disabling each component respectively. Results are shown in Table 1 and Figure 5. We first verify the contribution of the polarization-guided separation module by directly estimating $\hat{I}_r$ and $\hat{I}_t$ from the refinement network (without inferring α, β and the initially separated layers $\{I_r^{'}, I_t^{'}\}$ first). In other words, we also use an encoder-decoder architecture to estimate separation directly from a given pair of (un)polarized images. SSIM and PSNR averaged over $5,000$ validation images are shown in "w/o ori. est." row of Table 1. We can see that, more prior knowledge encoded in the network facilitates the image prediction, and the orientation estimation with only two parameters is easier to learn and also better than directly estimating ξ and ζ for each pixel. To verify the effectiveness of our pipeline, we quantitatively evaluate our intermediate steps, *i.e.*, "cross-line suppression module", "refinement module", "polarization-guided module", and the results listed in Table 1 show the effectiveness of the physical guidance and boost from cascaded networks. The intermediate separation is shown in Figure 5. We omit results of the cross-line suppression module, since they are similar to the final output visually. In the polarization-guided module, rough transmission and reflection layers are generated according to estimated plane coefficients, although they are affected by regional artifacts. Next, the refinement network eliminates most of the contamination and improves the visual quality of the separation. At the final step, the residual of cross lines (especially in the reflection) is attenuated by the cross-line suppression module, and the final re-weighting method boosts the overall performance and generates compelling results by integrating outputs of the two modules.

TABLE 1: Quantitative evaluation results in ablation study.

	Transmission		Reflection	
	SSIM	PSNR	SSIM	PSNR
Ours	0.9812	33.32	0.9526	30.16
Cross-line suppression module	0.9799	32.22	0.9498	29.79
Refinement module	0.9718	31.56	0.9183	28.60
Polarization-guided module	0.8558	21.00	0.6497	15.88
Refinement module w/o LPIPS	0.9695	31.66	0.9062	28.11
Refinement module w/o ori. est.	0.9668	31.31	0.8925	27.31
Ours pre. [13]	0.9659	31.12	0.9010	27.81
Ours on 8-bit dataset	0.9599	30.56	0.6869	20.28
Ours pre. [13] on 8-bit dataset	0.9708	28.23	0.8953	20.92

We further evaluate different loss functions, and train our network without the perceptual loss. The results are listed in "w/o LPIPS" row of Table 1. We find the perceptual loss is particularly useful in improving the visual quality of the layer estimation, though the evaluation indexes are close to the baseline of Refinement Module.

Our preliminary work [13] adopts the gradient loss instead of perceptual loss and produces similar results as Refinement module w/o LPIPS, as listed in "Ours pre. [13]" row of Table 1. Our new framework benefits from supervision with high-level perceptual loss and enhancement of the cross-line suppression module, and shows better results than [13].

We also evaluate the effect of quantization error in polarization-guided separation, since the initially separated results stem from pixel-wise calculation. Note that our model is trained on 16-bit-format dataset. In this experiment, we feed our model with 8-bit-format input images. As shown

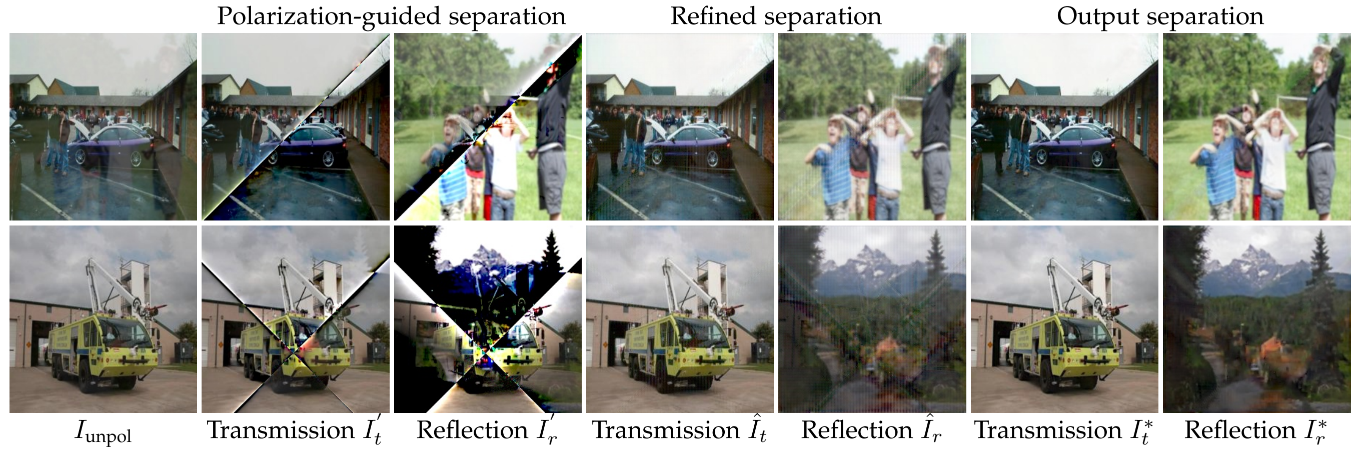


Fig. 5: The separation results in intermediate steps and the final stage. Note that the polarization-guided module coarsely separates reflections and transmissions, but suffers from the cross-line artifacts. Refinement part further improves the results, but there still remain small artifacts in the cross-line regions. In the final stage, most of the artifacts are suppressed and cleaner separation results are generated.

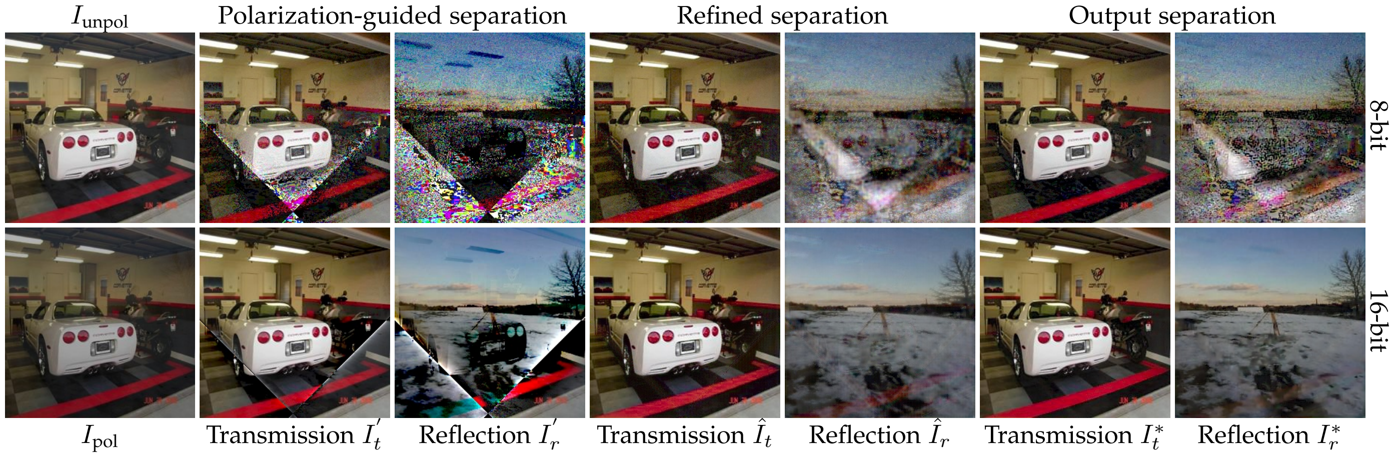


Fig. 6: Visual comparison between 8-bit and 16-bit (un)polarized images. The results generated from 8-bit images are affected by more artifacts than those taking 16-bit images as input owing to the calculation error in the polarization-guided step.

in Figure 6 and Table 1, the reflections are greatly affected by the cross-line effect and the quantitative indexes decline largely. As we discussed in Section 3.3, the value of $\Theta_r(\theta)$ corresponding to the reflection layer is several times as large as the value of $\Theta_t(\theta)$ in the range of $(0, 78°]$, which makes the calculation of reflection layers become more susceptible to small noise. While the transmissions are more robust to the quantization error. When the previous model [13] is trained and tested on 8-bit dataset, PSNR of the results is also worse than that of the model trained on 16-bit-format images, as shown in "Ours pre. [13] on 8-bit dataset" row of Table 1.

5.2 Evaluation on Synthetic Data

We use $5,000$ pairs of images from our synthetic validation dataset with ground truth reflection and transmission layers to quantitatively compare our method with state-of-the-art approaches. ReflectNet [12] is a learning based method using three polarized images; Zhang *et al*. [6], CoRRN [5], ERRNet [7] are deep learning based solutions using a single image. To test the performance of ReflectNet [12], we generated two additional polarization images for each pair of (un)polarized images in our dataset, and fine-tuned ReflectNet using the Adam solver with a learning rate of 0.005 for 5 epochs.

The experimental results are shown in Figure 7 and Figure 8, and the quantitative evaluation is listed in Table 2. We can see that, in contrast to all the single-image based methods, our method has much better performance, which shows the advantage of the additional polarized image. We only compare the transmission results of our method with transmissions of other single-image-based methods, due to their bad performance on reflection layers.

Our method also outperforms ReflectNet [12] which requires three polarized images as input, especially in terms of the quality of the reflection layer, although our method only needs one polarized image in addition to an unpolarized image. Moreover, our method performs the best in suppressing undesired reflection in transmission layers and recovers high-quality reflection layers as well, as indicated by corresponding SSIM and PSNR values under images in Figure 8. In order to compare our model with ReflectNet thoroughly, we retrain ReflectNet on our dataset with the same training strategy. Under this setup, the quantitative results of reflection and transmission are listed in "ReflectNet retrained" row of Table 2. We can see

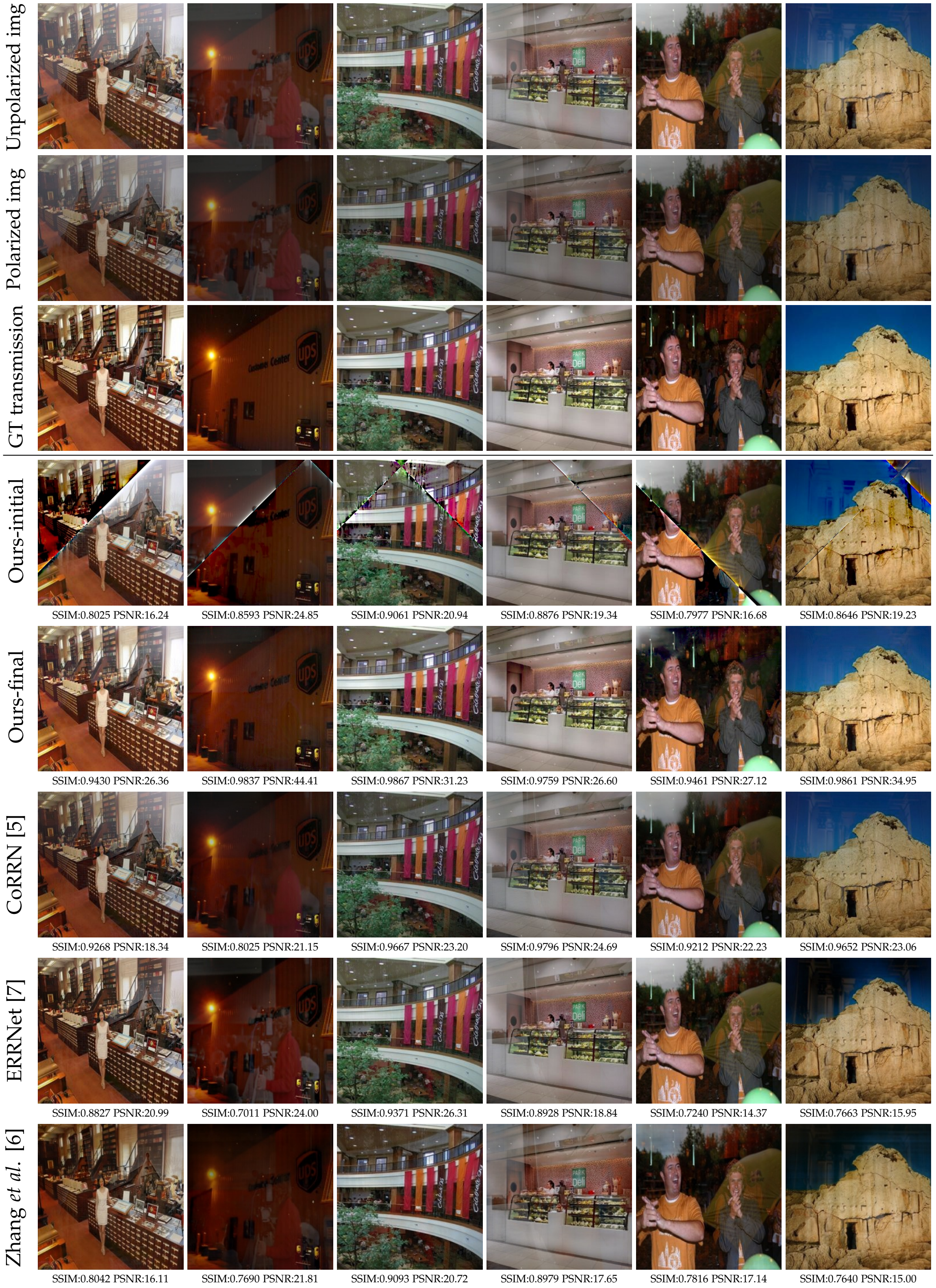


Fig. 7: Quantitative and qualitative evaluation on synthetic data, compared with single-image methods including CoRRN [5], ERRNet [7], and Zhang *et al.* [6].

that even with this retrained ReflectNet still performs worse than our refinement module. We also evaluate our initial polarization-guided separation $I_r^{'}$ and $I_t^{'}$ ("Ours-initial") in Table 2, and we can see that the initial separation is effective, and both the refinement network and cross-line suppression module facilitate attenuating the artifact and noise caused by rough estimation of ξ and ζ. At last, we test our method against Gaussian noise added to images with different standard deviations. The results are shown in Table 2. We can see that the reflection results downgrade a lot since the initially separated reflections are more sensitive to noise referring to Section 3.3. In contrast to reflection results,

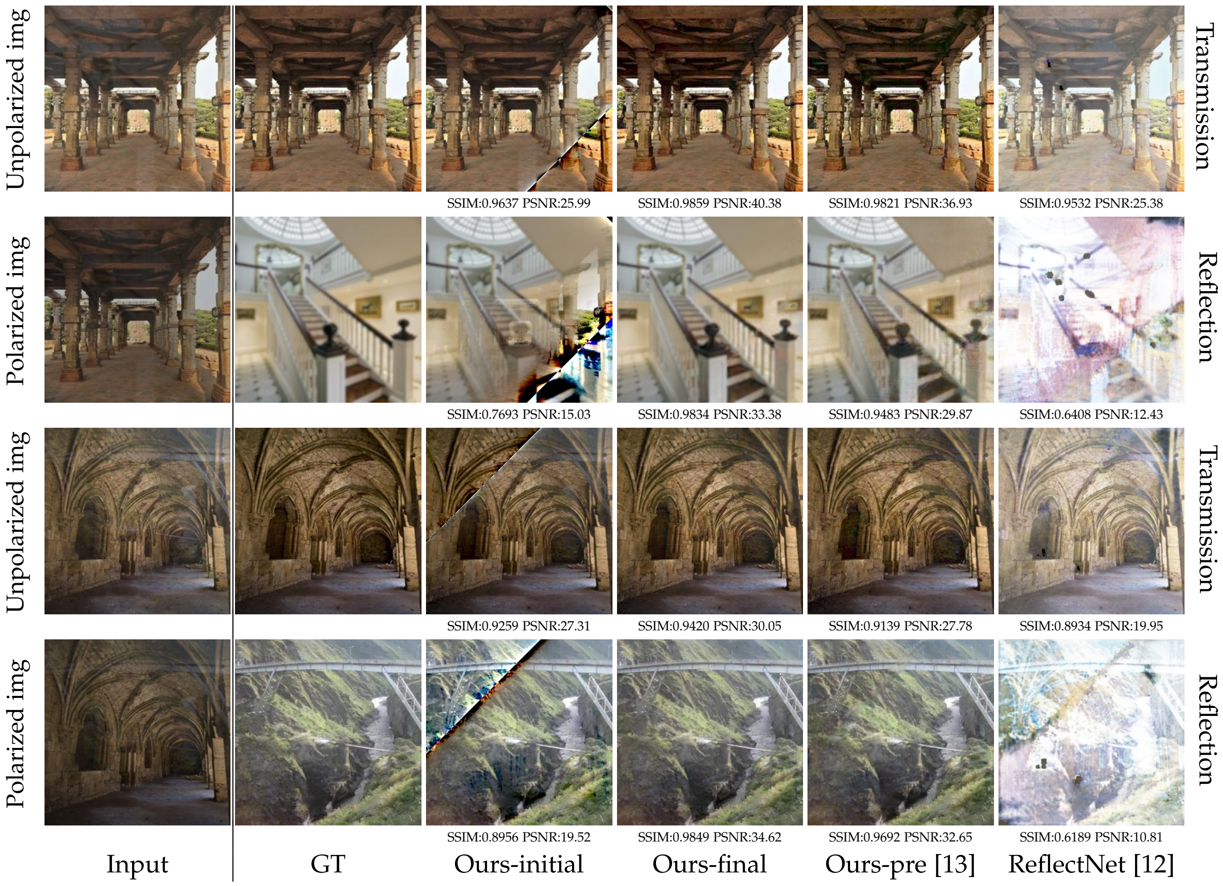


Fig. 8: Quantitative and qualitative evaluation on synthetic data, compared with methods taking multiple polarized images as input, *e.g.*, ReflectNet [12] fine-tuned on our dataset and the preliminary version of this work [13].

produced transmissions are more robust to the additive Gaussian noise due to the high transmission-reflection ratio in the unpolarized and polarized images. This phenomenon is similar to that we feed our model with 8-bit-format (un)polarized images.

5.3 Evaluation on Real Data

We use the Lucid Vision Phoenix[1] (grayscale, 16-bit format) and Triton[2] (RGB, 8-bit format) polarization cameras to capture the real dataset. The polarization camera can take four images with different polarizer angles at a single shot. We use three of them as input images to ReflectNet [12], four of them as input to [15] and [16], and one of them as the polarized image to our method. The unpolarized input image is calculated by summing two polarized images captured with orthogonal polarizer angles [55]. We conduct qualitative comparisons on the real-world data, as displayed in Figure 9, Figure 10, and Figure 11[3]. More results on the real-world data are provided in Appendix C. These scenes contain strong reflections with complex textures, and all the single-image based methods fail to recover the transmissions while removing the reflections. Thanks to the polarimetric cues, ReflectNet [12], Lei *et al*. [15], Li *et al*. [16] and our method all show obvious advantage over the single-image based methods. ReflectNet [12] and our method produce better separations. Compared to ReflectNet [12], our method additionally exploits the semi-reflector geometry and the physics-based imaging model to produce reliable initial separations to the network rather than letting the network conduct separation almost from scratch [12]. Due to this encoded physical knowledge as well as the dedicated network architecture, our framework is able to produce compelling separation results by taking fewer input images (with one of them unpolarized to capture more light) than other polarization-based methods.

TABLE 2: Quantitative evaluation results on synthetic data.

	Transmission		Reflection	
	SSIM	PSNR	SSIM	PSNR
Ours	0.9812	33.32	0.9526	30.16
Ours-initial	0.8558	21.00	0.6497	15.88
ReflectNet finetuned	0.8988	25.91	0.7512	19.64
ReflectNet retrained	0.8659	25.71	0.7412	20.88
Ours-1%noise	0.9319	28.95	0.6308	20.00
Ours-initial 1%noise	0.7600	18.73	0.3585	12.29
Ours-2%noise	0.9119	28.06	0.5236	17.87
Ours-initial 2%noise	0.6845	17.47	0.2429	10.48

5.4 Dealing with Misaligned Inputs

When we take the polarized images with a single-view camera, rotating, mounting, and unmounting the polarizer may cause movement of the camera, and the camera parameters might be slightly different in the two shots. Hence, the input

[1] https://thinklucid.com/product/phoenix-5-0-mp-polarized-model/
[2] https://thinklucid.com/product/triton-5-mp-polarization-camera/
[3] To test the grayscale images, we stack the single channel images into 3 channels and convert the output images back to grayscale.

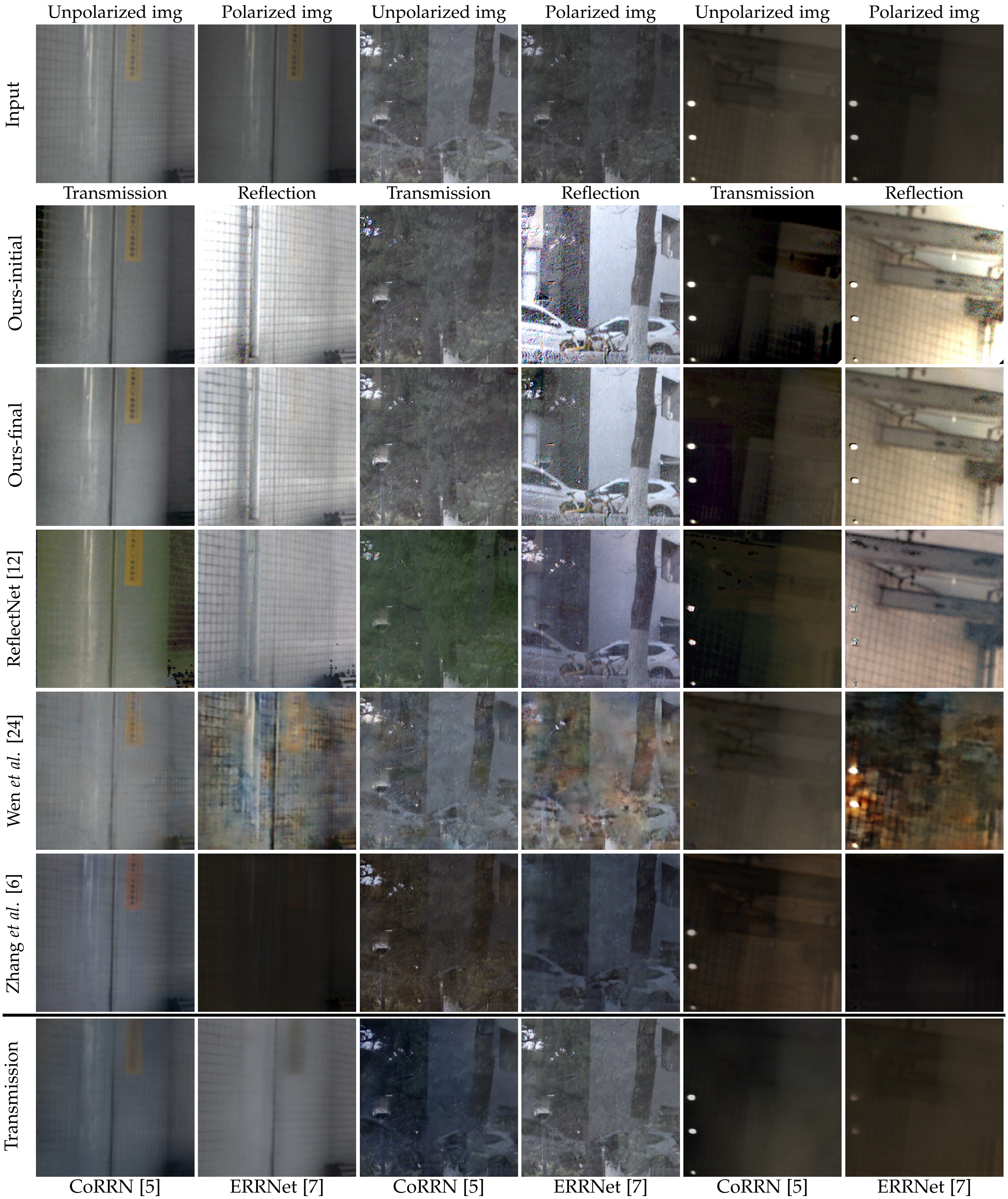


Fig. 9: Qualitative comparisons with ReflectNet [12], Wen *et al.* [24], Zhang *et al.* [6], CoRRN [5], and ERRNet [7], evaluated on real-world images (8-bit format) taken by a Lucid Vision Triton polarization camera. We only show transmission results of CoRRN [5] and ERRNet [7], since they are designed for extracting transmission scenes only.

(un)polarized images are misaligned in practice. We further improve our method to tackle the slight misalignment[1] in the

[1] Our method could handle this slight misalignment but may not work on the dual-view misaligned images with relative large baselines, which is elaborated in Appendix B.

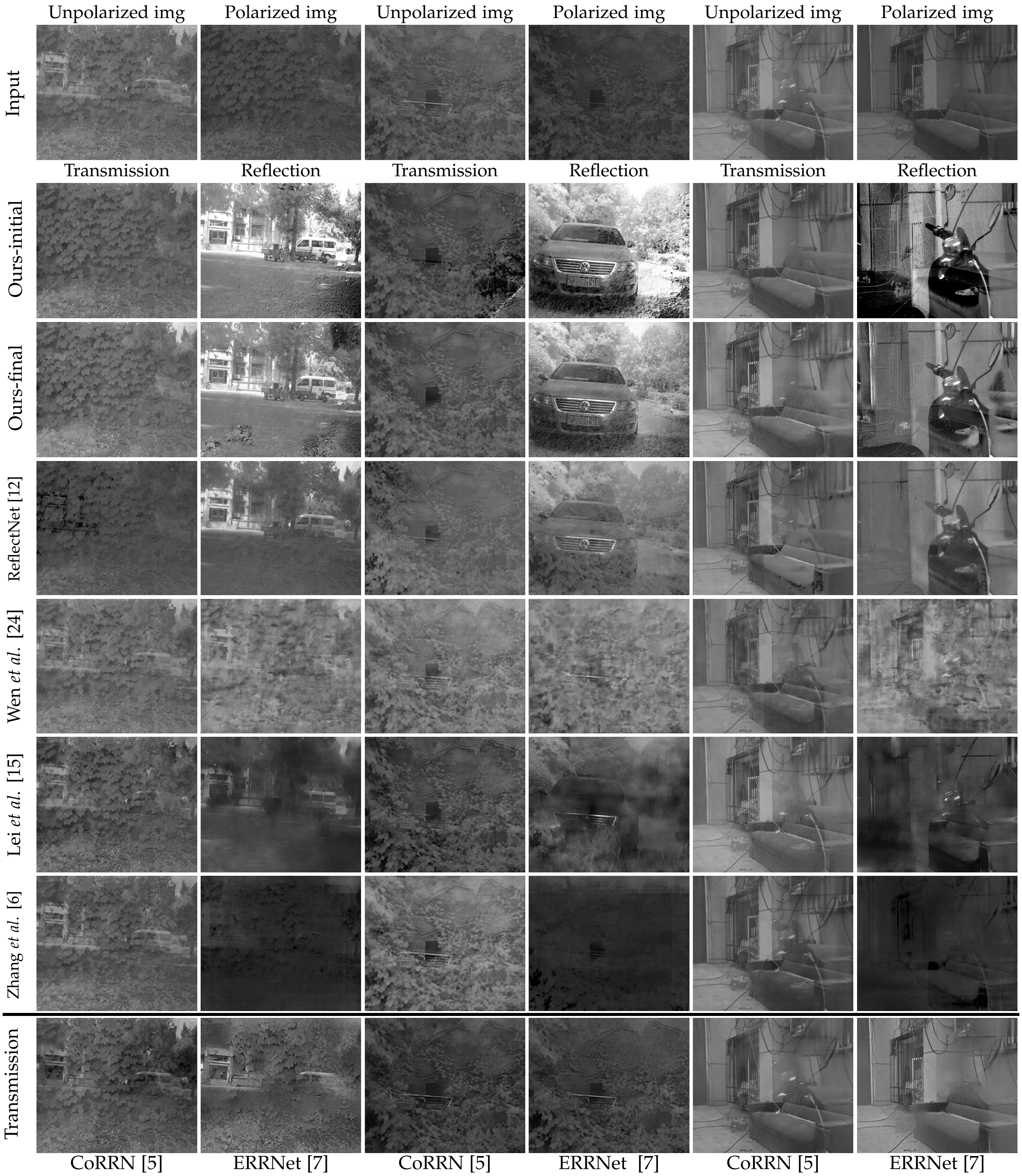


Fig. 10: Qualitative comparisons with ReflectNet [12], Wen *et al.* [24], Lei *et al.* [15], Zhang *et al.* [6], CoRRN [5], and ERRNet [7], evaluated on real-world images (16-bit format) taken by a Lucid Vision Phoenix polarization camera. For better visualization, the minimum and maximum intensity values of different algorithms are stretched in a consistent range.

paired inputs and test it on the synthetic data. Specifically, we assume the camera positions of the two shots are close enough that most of the pixels can be registered and the occlusions caused by different views only cover a small part in captured images. When generating the synthetic images, we set the camera focal length as a normally distributed random variable $f \sim \mathcal{N}(1.4, 0.03)$ for data augmentation. Changes in exposure and gain parameters are considered as well: We re-scale the polarized images with a coefficient k to simulate the exposure difference, where $k \sim \mathcal{U}(0.8, 1.1)$. Moreover, to simulate the parallax motion between two views, we generate a random global shift within 5 pixels

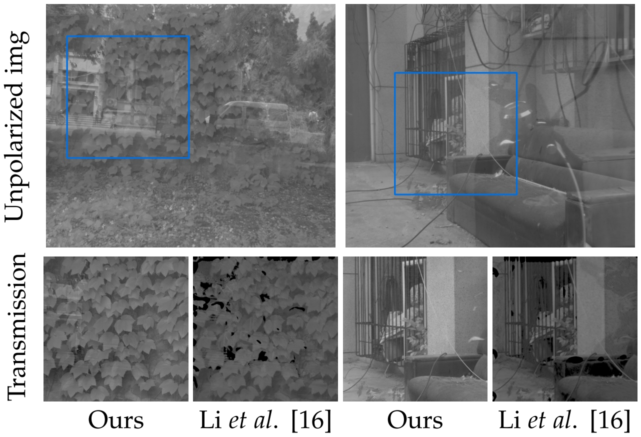


Fig. 11: Qualitative comparison with Li *et al.* [16], evaluated on images in the wild. Our method generates more reliable results for the recovered transmission images.

TABLE 3: Quantitative comparison between the proposed pipeline with the alignment module and the pipeline without alignment module, which is evaluated on the the misaligned inputs.

	Transmission		Reflection	
	SSIM	PSNR	SSIM	PSNR
Ours-initial w/o align	0.3190	11.60	0.1284	6.566
Ours-initial w/ align	0.8106	18.17	0.5229	12.27
Ours w/o align	0.8191	22.27	0.2444	10.19
Ours w/ align	0.9449	27.81	0.7115	17.49

and apply a Gaussian deformation on its grid, based on which we warp the images to produce misaligned polarized images.

Misalignment can degrade the performance of our separation results, since our model relies on the pixel-wise calculations in the initial separation. Thus, we introduce the motion estimation network to predict the optical flow between the polarized and unpolarized images, and to align the images for further calculation. Specifically, we build the optical flow model based on PWC-Net [62], and fine-tune it on our synthetic dataset. Note that the warped polarized images are hardly to be registered perfectly, which might suffer from occlusions and pixel mismatch. To alleviate the ghost effect caused by misalignment, we connect the proposed method with the optical flow model in [62] as a unified framework, and then fine-tune it on the misaligned data until convergence. To simulate the slight misalignment of the paired images, we use the polarization camera to capture the two images with a manual displacement within 1 centimeter. The quantitative results on synthetic data are listed in Table 3, and qualitative results on the synthetic and the real-world images are shown in Figure 12. Fed with the misaligned data, the polarization-guided module performs badly, and our method fails to separate the reflections. After adding the alignment module, our model still works well on the misaligned images, thanks to the additional optical flow estimation. Such handling of misalignment has the potential to extend our method using multi-lens camera phones if the registered raw data is accessible, with one of the lens equipped with a polarizer. Moreover, we can design a new micro-polarizer sensor array to capture both polarized and unpolarized images and avoid the alignment issue, which is also interesting while beyond the scope of this paper. We will consider these as future work.

6 Conclusion

We solve the problem of integrating polarimetric constraints from a pair of unpolarized and polarized images to separate reflection and transmission layers. To deal with the ill-posedness introduced by using fewer polarized images, we derive the semi-reflector orientation constraint to make the physical image formation for layer separation valid given our setup, and train a well-designed neural network to refine the separated layers and eliminate the cross-line effect, showing state-of-the-art performance. Our simple yet unique capturing setup explores polarimetric constraints for separating reflection and transmission layers as reliably as existing approaches using three or more polarized images. Besides, the newly added optical estimation module enables registration of misaligned input, allowing our framework to be potentially integrated into multi-lens camera phones, if the registered raw data is accessible.

Our model assumes the semi-reflector approximately has a planar shape. When it becomes a curved shape such as the windshield in a car, our semi-reflector orientation estimation module will fail, and thus the performance of our method will deteriorate, and we will consider this as our future work.

Acknowledgments

This project is supported in National Natural Science Foundation of China under Grant No. 62136001, 61872012 and 62102356.

References

[1] A. Levin and Y. Weiss, “User assisted separation of reflections from a single image using a sparsity prior,” *IEEE TPAMI*, 2007.

[2] Y. Li and M. S. Brown, “Single image layer separation using relative smoothness,” in *Proc. CVPR*, 2014.

[3] R. Wan, B. Shi, A. H. Tan, and A. C. Kot, “Depth of field guided reflection removal,” in *Proc. ICIP*, 2016.

[4] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, “Reflection removal using ghosting cues,” in *Proc. CVPR*, 2015.

[5] R. Wan, B. Shi, H. Li, L.-Y. Duan, A.-H. Tan, and A. Chichung, “CoRRN: Cooperative reflection removal network,” *IEEE TPAMI*, 2019.

[6] X. Zhang, R. Ng, and Q. Chen, “Single image reflection separation with perceptual losses,” in *Proc. CVPR*, 2018.

[7] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang, “Single image reflection removal exploiting misaligned training data and network enhancements,” in *Proc. CVPR*, 2019.

[8] Y. Li and M. S. Brown, “Exploiting reflection change for automatic reflection removal,” in *Proc. ICCV*, 2013.

[9] Y. Y. Schechner, J. Shamir, and N. Kiryati, “Polarization and statistical analysis of scenes containing a semireflector,” *Journal of the Optical Society of America*, 2000.

[10] N. Kong, Y. Tai, and J. S. Shin, “A physically-based approach to reflection separation: from physical modeling to constrained optimization,” *IEEE TPAMI*, 2014.

[11] Y. Y. Schechner, J. Shamir, and N. Kiryati, “Polarization-based decorrelation of transparent layers: The inclination angle of an invisible surface,” in *Proc. ICCV*, 1999.

[12] P. Wieschollek, O. Gallo, J. Gu, and J. Kautz, “Separating reflection and transmission images in the wild,” in *Proc. ECCV*, 2018.

[13] Y. Lyu, Z. Cui, S. Li, M. Pollefeys, and B. Shi, “Reflection separation using a pair of unpolarized and polarized images,” in *Proc. NeurIPS*, 2019.

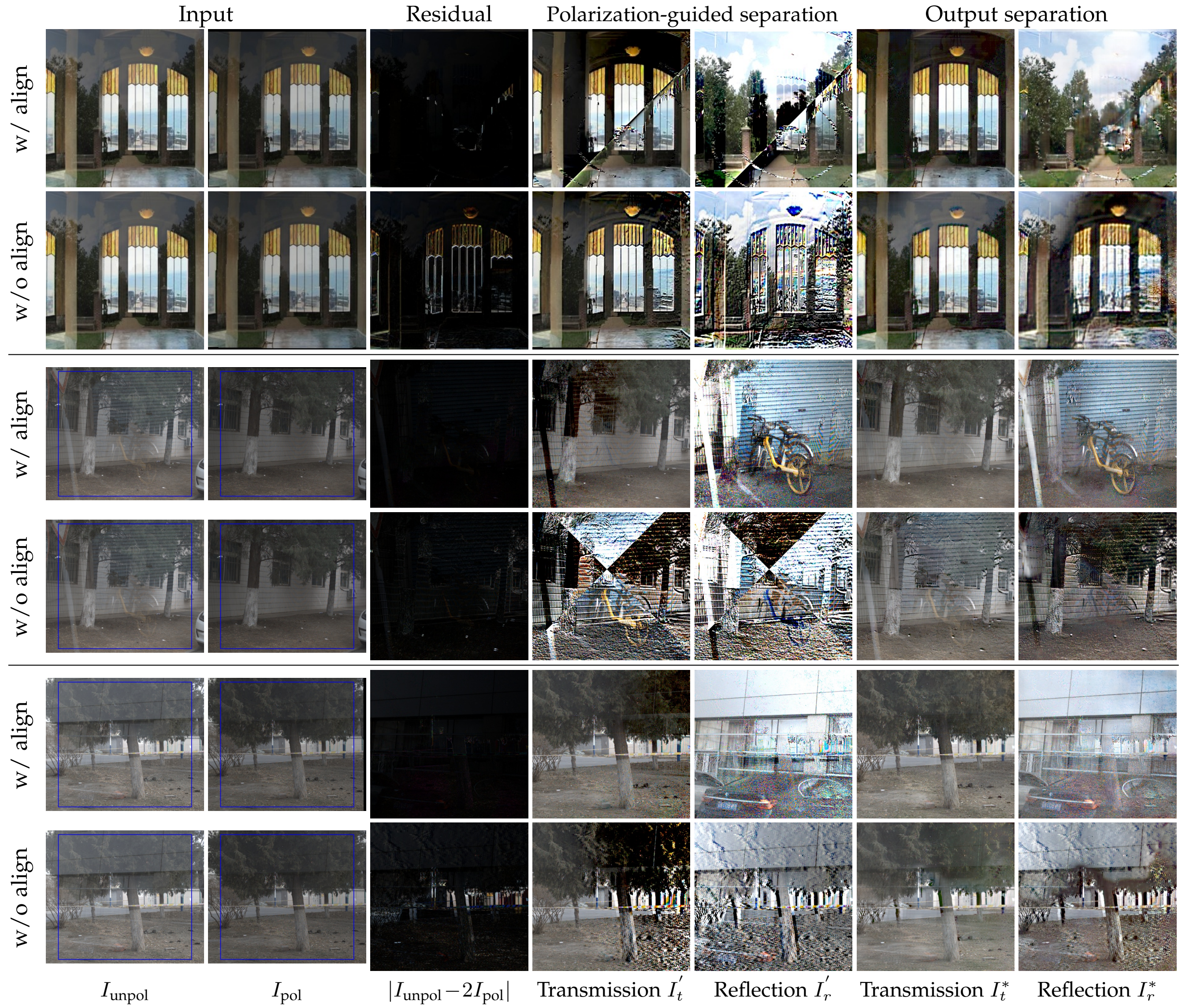


Fig. 12: Qualitative results of misaligned pairs of (un)polarized images, evaluated on synthetic data (the top two rows) and real-world images (the other rows). We first align the input paired images and then conduct separations on the center regions of the aligned images (marked by the blue bounding box). Misalignment of paired images makes the polarization-guided module fail to conduct pixel-wise separation, while the warping module registers the two input images and produces reasonable results.

[14] Y. Y. Schechner, N. Kiryati, and R. Basri, “Separation of transparent layers using focus,” *Springer IJCV*, 2000.

[15] C. Lei, X. Huang, M. Zhang, Q. Yan, W. Sun, and Q. Chen, “Polarized reflection removal with perfect alignment in the wild,” in *Proc. CVPR*, 2020.

[16] R. Li, S. Qiu, G. Zang, and W. Heidrich, “Reflection separation via multi-bounce polarization state tracing,” in *Proc. ECCV*, 2020.

[17] L. B. Wolff and T. E. Boult, “Constraining object features using a polarization reflectance model,” *IEEE TPAMI*, 1991.

[18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proc. CVPR*, 2018.

[19] A. Levin and Y. Weiss, “User assisted separation of reflections from a single image using a sparsity prior,” in *Proc. ECCV*, 2004.

[20] N. Arvanitopoulos, R. Achanta, and S. Süsstrunk, “Single image reflection suppression,” in *Proc. CVPR*, 2017.

[21] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, “A generic deep architecture for single image reflection removal and image smoothing,” in *Proc. ICCV*, 2017.

[22] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, “CRRN: Multi-scale guided concurrent reflection removal network,” in *Proc. CVPR*, 2018.

[23] J. Yang, D. Gong, L. Liu, and Q. Shi, “Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal,” in *Proc. ECCV*, 2018.

[24] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han, and S. He, “Single image reflection removal beyond linearity,” in *Proc. CVPR*, 2019.

[25] E. Be’Ery and A. Yeredor, “Blind separation of superimposed shifted images using parameterized joint diagonalization,” *IEEE TIP*, 2008.

[26] K. Gai, Z. Shi, and C. Zhang, “Blind separation of superimposed moving images using image statistics,” *IEEE TPAMI*, 2012.

[27] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, “A computational approach for obstruction-free photography,” *ACM TOG*, 2015.

[28] H. Farid and E. H. Adelson, “Separating reflections and lighting using independent components analysis,” in *Proc. CVPR*, 1999.

[29] Hermanto, A. K. D. B. Filho, T. Yamamura, and N. Ohnishi, “Separating virtual and real objects using independent component analysis,” *IEICE Transactions on Information and Systems*, 2001.

[30] A. M. Bronstein, M. M. Bronstein, M. Zibulevsky, and Y. Y. Zeevi, “Sparse ICA for blind separation of transmitted and reflected images,” *International Journal of Imaging Systems and Technology*,

2005.
[31] Y. Diamant and Y. Y. Schechner, "Overcoming visual reverberations," in *Proc. CVPR*, 2008.
[32] A. Punnappurath and M. S. Brown, "Reflection removal using a dual-pixel sensor," in *Proc. CVPR*, 2019.
[33] J.-S. Yun and J.-Y. Sim, "Reflection removal for large-scale 3D point clouds," in *Proc. CVPR*, 2018.
[34] P. Chandramouli, M. Noroozi, and P. Favaro, "Convnet-based depth estimation, reflection separation and deblurring of plenoptic images," in *Proc. ACCV*, 2016.
[35] D. Ma, R. Wan, B. Shi, A. C. Kot, and L.-Y. Duan, "Learning to jointly generate and separate reflections," in *Proc. ICCV*, 2019.
[36] Miyazaki, Tan, Hara, and Ikeuchi, "Polarization-based inverse rendering from a single view," in *Proc. ICCV*, 2003.
[37] E. Hancock and G. Atkinson, "Recovery of surface orientation from diffuse polarization," *IEEE TPAMI*, 2006.
[38] C. P. Huynh, A. Robles-Kelly, and E. R. Hancock, "Shape and refractive index recovery from single-view polarisation images," in *Proc. CVPR*, 2010.
[39] ——, "Shape and refractive index from single-view spectro-polarimetric images," *IJCV*, 2013.
[40] A. Kadambi, V. Taamazyan, B. Shi, and R. Raskar, "Polarized 3D: High-quality depth sensing with polarization cues," in *Proc. ICCV*, 2015.
[41] ——, "Depth sensing using geometrically constrained polarization normals," *Springer IJCV*, 2017.
[42] T. Ngo Thanh, H. Nagahara, and R.-i. Taniguchi, "Shape and light directions from shading and polarization," in *Proc. CVPR*, 2015.
[43] W. Smith, R. Ramamoorthi, and S. Tozza, "Linear depth estimation from an uncalibrated, monocular polarisation image," in *Proc. ECCV*, 2016.
[44] S. Tozza, W. A. P. Smith, D. Zhu, R. Ramamoorthi, and E. R. Hancock, "Linear differential constraints for photo-polarimetric height estimation," in *Proc. ICCV*, 2017.
[45] D. Miyazaki, M. Kagesawa, and K. Ikeuchi, "Transparent surface modeling from a pair of polarization images," *IEEE TPAMI*, 2004.
[46] Z. Cui, J. Gu, B. Shi, P. Tan, and J. Kautz, "Polarimetric multi-view stereo," in *Proc. CVPR*, 2017.
[47] D. Zhu and W. A. P. Smith, "Depth from a polarisation + RGB stereo pair," in *Proc. CVPR*, 2019.
[48] L. Yang, F. Tan, A. Li, Z. Cui, Y. Furukawa, and P. Tan, "Polarimetric dense monocular slam," in *Proc. CVPR*, 2018.
[49] L. Chen, Y. Zheng, A. Subpa-asa, and I. Sato, "Polarimetric three-view geometry," in *Proc. ECCV*, 2018.
[50] Z. Cui, V. Larsson, and M. Pollefeys, "Polarimetric relative pose estimation," in *Proc. ICCV*, 2019.
[51] W. Sturzl, "A lightweight single-camera polarization compass with covariance estimation," in *Proc. ICCV*, 2017.
[52] K. Tanaka, Y. Mukaigawa, and A. Kadambi, "Polarized non-line-of-sight imaging," in *CVPR*, 2020.
[53] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Instant dehazing of images using polarization," in *Proc. CVPR*, 2001.
[54] Y. Y. Schechner and S. K. Nayar, "Generalized mosaicing: polarization panorama," *IEEE TPAMI*, 2005.
[55] E. Hecht, *Optics*, ser. Pearson education. Addison-Wesley, 2002.
[56] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, 2011.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.
[58] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NeurIPS Workshops*, 2017.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
[60] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE TPAMI*, 2017.
[61] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, 2004.
[62] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. CVPR*, 2018.

Youwei Lyu received his B.S. degree from Beijing University of Posts and Telecommunications in 2019. He is currently studying at Beijing University of Posts and Telecommunications. His research interests are centered around computational photography and physics-based vision..

Zhaopeng Cui received the Ph.D. degree from Simon Fraser University in 2017. He was a senior researcher at ETH Zurich. He is currently a research professor with the State Key Lab of CAD&CG and the College of Computer Science and Technology, Zhejiang University. His research interests include 3D mapping and localization, 3D scene understanding, image and video editing.

Si Li received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2012. She is currently an Associate Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. Her research interests include multimodal artificial intelligence and machine learning.

Marc Pollefeys received the PhD degree from the KU Leuven, Belgium, in 1999. He is Director of the Microsoft Mixed Reality and AI Zurich lab and a full professor in the Department of Computer Science, ETH Zurich. From 2002 to 2007 he was on the faculty with the University of North Carolina, Chapel Hill. His main area of research is computer vision, but he is also active in robotics, machine learning, and computer graphics. He has received several prizes for his research, including a Marr prize, a NSF CAREER award, a Packard Fellowship and a European Research Council Grant. He is the author or co-author of more than 300 peer-reviewed publications. He was the general chair of ICCV 2019 in Seoul and ECCV 2014 in Zurich, and the program chair of CVPR 2009. He is the president of the European Computer Vision Association and is a fellow of the IEEE.

Boxin Shi 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 postdoctoral research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University from 2013 to 2016, and worked as a researcher in the National Institute of Advanced Industrial Science and Technology from 2016 to 2017. His papers were awarded as Best Paper Runner-Up at International Conference on Computational Photography 2015 and selected as Best Papers from ICCV 2015 for IJCV Special Issue. He has served as an editorial board member of IJCV and an area chair of CVPR/ICCV. He is a senior member of IEEE.

Physics-Guided Reflection Separation from a Pair of Unpolarized and Polarized Images

Youwei Lyu*‡, Zhaopeng Cui*, *Member, IEEE,* Si Li,
Marc Pollefeys, *Fellow, IEEE,* and Boxin Shi†, *Senior Member, IEEE*

Appendix A
Using Least Squares with Regularization and ADMM in Initial Separations

In stead of directly solving Equation (7) and Equation (8) in the main paper, we try to formulate the initial separation as a linear least-squares problem and add $\mathcal{L}_2$ regularization terms to handle cross-line artifacts. Specifically speaking, the linear equations,

$$I_{\text{unpol}} = \frac{\xi}{2} I_r + \frac{2-\xi}{2} I_t, \tag{1}$$

$$I_{\text{pol}} = \frac{\zeta}{2} I_r + \frac{1-\zeta}{2} I_t, \tag{2}$$

can be converted into a matrix form,

$$\begin{bmatrix} I_{\text{unpol}} \\ I_{\text{pol}} \end{bmatrix} = \begin{bmatrix} \frac{\xi}{2} & \frac{2-\xi}{2} \\ \frac{\zeta}{2} & \frac{1-\zeta}{2} \end{bmatrix} \begin{bmatrix} I_r \\ I_t \end{bmatrix}. \tag{3}$$

For conciseness, we denote Equation (3) as

$$\mathbf{b} = \mathbf{A}\mathbf{x}. \tag{4}$$

In order to solve the reflection and the transmission ($\mathbf{x}$), we formulate the separation as a least-squares problem:

$$\min \left\| \mathbf{A}\mathbf{x} - \mathbf{b} \right\|_2^2. \tag{5}$$

The cross-line artifacts are attributed to the zero division when directly solving Equation (1) and Equation (2). We apply an $\mathcal{L}_2$ regularization term to penalize large values in reflection and transmission layers, and derive the formulation LS-regul.:

$$\min \left\| \mathbf{A}\mathbf{x} - \mathbf{b} \right\|_2^2 + \lambda \left\| \mathbf{x} \right\|_2^2, \tag{6}$$

where λ denotes the weighting parameter of the regularization term. Then we solve Equation (6) by the least-squares method, obtain the separation vector $\mathbf{x}$, and convert it into the output images. We experimentally set $\lambda = 0.002$ after the grid search. We also try to use the regularization term with an offset $a = 0.5$, *i.e.*, $\min \left\| \mathbf{A}\mathbf{x} - \mathbf{b} \right\|_2^2 + \lambda \left\| \mathbf{x} - a \right\|_2^2$, and obtain the separation similar to that of LS-regul.

In addition, we employ the alternating direction method of multipliers (ADMM) [1] to solve the reflection and transmission components with the linear constraints that the intensity values of reflection and transmission images are within $(0, 1)$. Considering the linear constraint, we formulate the objective function as

$$\min \left\| \mathbf{A}\mathbf{x} - \mathbf{b} \right\|_2^2, \quad s.t.\ 0 \leq x \leq 1, x \in \mathbf{x}, \tag{7}$$

and employ the ADMM solver[1] to obtain the separation vector $\mathbf{x}$. We simultaneously optimize all the pixels in the separation image by the ADMM solver and conduct enormous experiments to search for appropriate hyper-parameters in the optimization. For better convergence, the reflection and transmission vector $\mathbf{x}$ is initialized with the unpolarized values[2] of the same pixel, and the penalty parameter ρ is set to be 0.01. We observe the optimization takes about 100 iterations to reach convergence.

5,000 pairs of images from our synthetic validation dataset with ground truth reflection and transmission are used for comparisons between the least-squares method (LS-regul.), ADMM, and the closed-form solution. The qualitative results are shown in Figure 1, and quantitative evaluation is listed in Table 1.

Despite that the least-squares method and the ADMM solver perform better on recovering transmissions, they produce much worse reflection layers compared to the closed-form solution. The output of LS-regul. and ADMM is free of cross-line issue, but the recovered reflection layers are still affected by black pixels (LS-regul.) or transmission layers (ADMM). LS-regul. and ADMM also fail to locally separate the reflection and transmission (*i.e.*, the helicopter in the transmission in the second sample of Figure 1), introducing

Authors contributed equally to this work. ‡Part of this was finished while working as a visiting student at Peking University. †Corresponding author.

- *Y. Lyu and S. Li are with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. Email: {youweilv, lisi}@bupt.edu.cn.*
- *Z. Cui is with the State Key Lab of CAD&CG and the College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China. E-mail: zhpcui@zju.edu.cn.*
- *M. Pollefeys is with Department of Computer Science, ETH Zürich, CH-8092 Zürich, Switzerland. Email: marc.pollefeys@inf.ethz.ch.*
- *B. Shi is with the National Engineering Research Center of Visual Technology, School of Computer Science and Institute for Artificial Intelligence, Peking University, Beijing 100871, China. Email: shiboxin@pku.edu.cn.*

[1] https://web.stanford.edu/~boyd/papers/admm/quadprog/quadprog.html

[2] We also try to initialize $\mathbf{x}$ with the closed-form solution and obtain the similar separation results.

TABLE 1: Quantitative comparisons among the least squares with $\mathcal{L}_2$ regularization, the optimization method using the ADMM solver, and our closed-form solution.

	Transmission		Reflection		Running time
	SSIM	PSNR	SSIM	PSNR	sec/(100 samples)
Ours-initial	0.8557	21.00	0.6497	15.88	0.019
LS-regul.	0.8997	25.07	0.4356	12.16	0.714
ADMM	0.9197	24.77	0.5941	15.59	8.654

ripple-like artifacts in reflection images. In contrast, our method retains reflection information, and our refinement and cross-line suppression network can well solve the cross-line artifacts, as shown in our experiments. For ADMM, it's hard to select an appropriate penalty parameter to well recover the reflection and transmission simultaneously. Moreover, we conduct experiments to compare the computational costs of these methods on a single NVIDIA GeForce RTX 3090, and the results are listed in Table 1. ADMM is about 450 times slower than our method to converge to a high accuracy, which is infeasible in real-time applications. Our method directly obtains a closed-form solution instead of having to calculate the inverse of a matrix, which has less computational burden compared to LS-regul. Based on all these points, the direct closed-form solution is more suitable for our end-to-end framework.

Appendix B
Misalignment Caused by the Baseline

We try to capture the paired (un)polarized images with two parallel polarization cameras but find that this setup does not work for our method. We notice that the baseline between two cameras is too large, and it causes alignment issues for reflection separations, even if we place the two cameras side by side. It would be hard to simultaneously align the reflection and transmission layers of the two captured images, due to the baseline between two cameras and different depths in reflection and transmission scenes. The aligned pixels in the paired images are mixed from different points of the scene, which violates the assumption of our method and makes the per-pixel computation fail.

We illustrate this issue in Figure 2. As shown in Figure 2a, when the transmission component dominates the mixture image, the alignment module tends to register image pixels corresponding to the same points in the transmission scene, and an aligned pixel may correspond to different points in the reflection scene. We provide a numerical analysis to demonstrate the disparity in the reflection scenes, as listed in Table 2a. We assume the binocular system consists of two cameras with a baseline of 5 cm, which is denoted as b. The distance from the transmission scene to the glass d_t is about 500 cm, the distance from the reflection scene to the glass d_r is about 1,000 cm, and the distance from Cam_1 to the glass is 40 cm. We assume the light ray is received by Cam_1 with a reflection angle of θ_1 ($\theta_1 = 30°$). According to the principle of light propagation, we can compute the disparity Δx_r in the reflection scene is about 5.3 cm^3. The disparity will be enlarged as the baseline increases. Under this large disparity, the same point of the transmission scene in the two views may be mixed with different reflections, which violates our assumption of the simultaneous alignment of reflection and transmission. Our performance will degrade on such a parallel camera rig. When capturing the real misaligned data shown in Figure 12 of the main paper, we move the camera within 1 cm between two shots, and the simulated reflection disparity is less than 1.1 cm. This slight disparity may bring small changes in reflection values and can be handled by our pipeline.

TABLE 2: Computed disparities in the reflection (transmission) scenes when aligning the pixels in the transmission (reflection) scenes. The baseline between Cam_1 and Cam_2 is b. The light ray is received by Cam_1 with a reflection angle of θ_1, and the rotation angle of the glass is represented as φ. The distance from the glass to Cam_1 is denoted as d_c, the distance from the glass to the reflection scene is denoted as d_r, and the distance from the glass to the transmission scene is denoted as d_t. Δx_r and Δx_t are disparities in the reflection scene and in the transmission scene, respectively.

(a) Disparities in the reflection scene, when transmission components dominate.

b	θ_1	φ	d_c	d_t	d_r	Δx_r
5.0	30.0°	30°	40	500	1,000	5.3
1.0	30.0°	30°	40	500	1,000	1.1

(b) Disparities in the transmission scene, when reflection components dominate.

b	θ_1	φ	d_c	d_t	d_r	Δx_t
5.0	30.0°	30°	40	1,000	500	5.3
1.0	30.0°	30°	40	1,000	500	1.1

In brief, our method could tolerate small misalignment caused by rotating or (un)mounting the polarizer in the front of the camera between two shots, but it is still challenging to apply our method to the dual-camera rig with a relative large baseline.

Appendix C
Additional Qualitative Comparisons on the Real-world Data

We show more separation results on the grayscale images (16-bit format) compared with ReflectNet [2], Wen *et al.* [3], Lei *et al.* [4], Zhang *et al.* [5], CoRRN [6] and ERRNet [7], as displayed in Figure 3.

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, 2011.

[2] P. Wieschollek, O. Gallo, J. Gu, and J. Kautz, "Separating reflection and transmission images in the wild," in *Proc. ECCV*, 2018.

[3] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han, and S. He, "Single image reflection removal beyond linearity," in *Proc. CVPR*, 2019.

[4] C. Lei, X. Huang, M. Zhang, Q. Yan, W. Sun, and Q. Chen, "Polarized reflection removal with perfect alignment in the wild," in *Proc. CVPR*, 2020.

[5] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *Proc. CVPR*, 2018.

[3] Similarly, we could calculate the disparity in the transmission scenes when we align pixels with dominant reflection components, as listed in Table 2b.

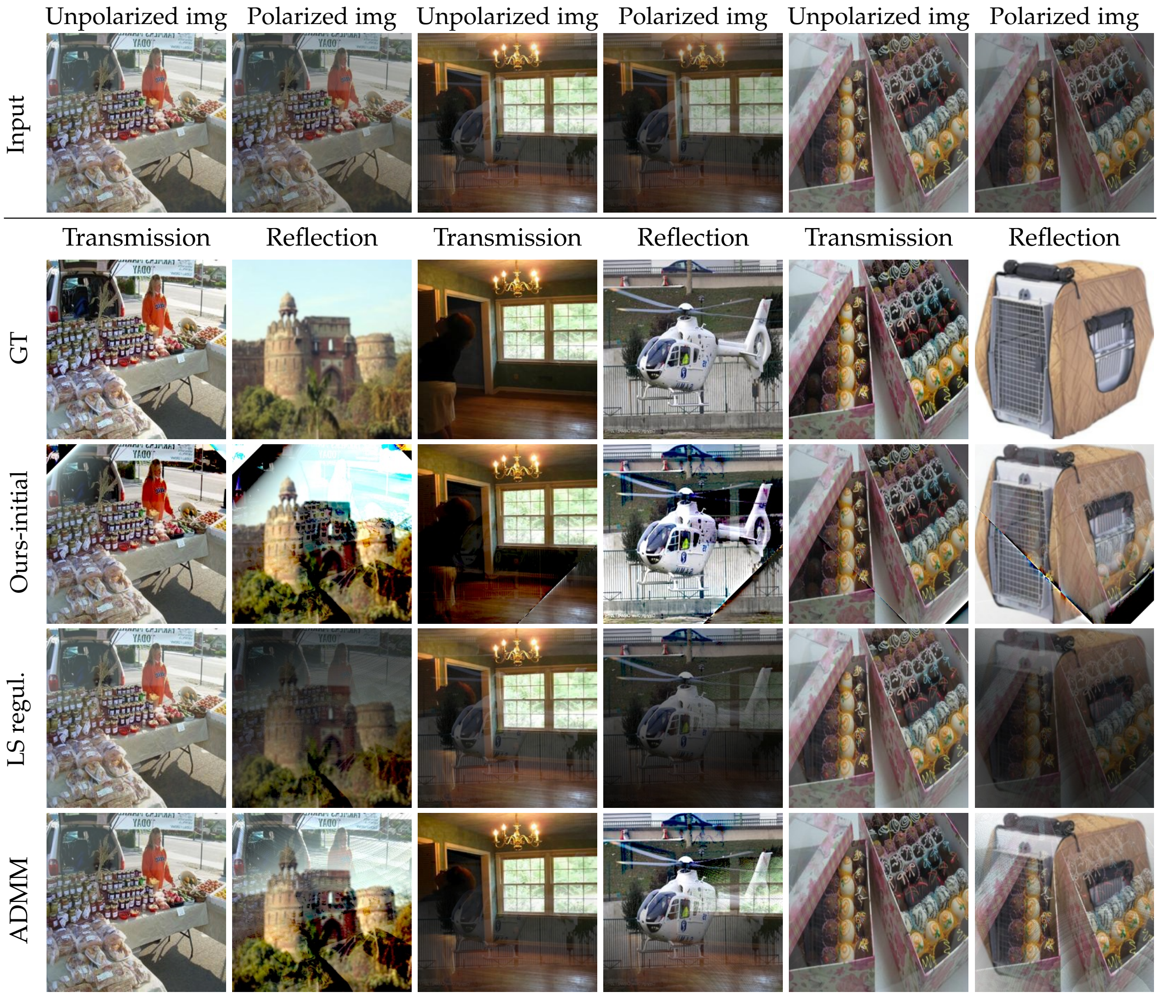
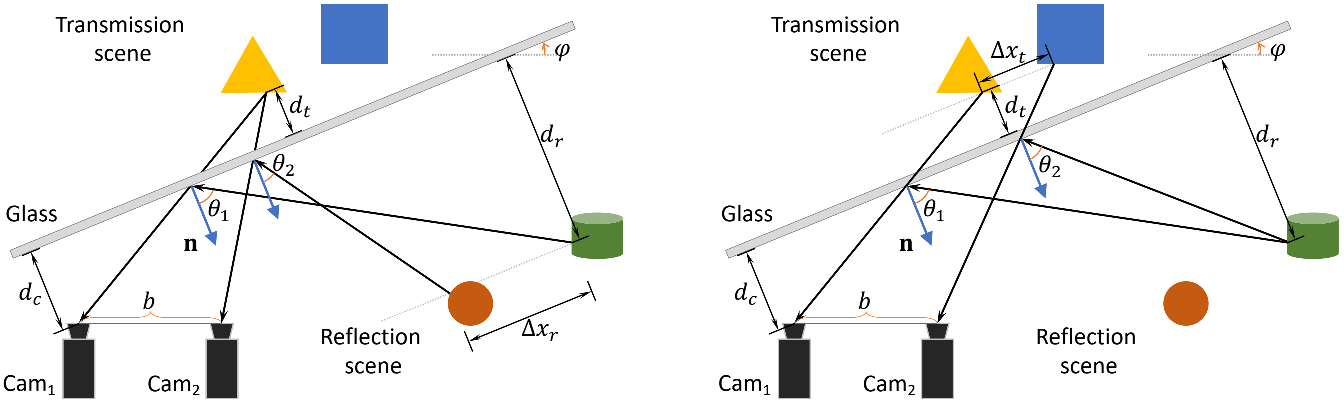


Fig. 1: Qualitative comparisons between our closed-form solution and the results of the least squares and ADMM, evaluated on synthetic data. The least-squares method with regularization and ADMM fail to separate some part of the image and introduce ripple-like artifacts in the reflection images.



(a) When the transmission component is dominant in the mixture image, the model tends to align the pixels corresponding to the same point in the transmission scene, and the misalignment of the reflection scene may occur.

(b) When the reflection component dominates the mixture image, the model tends to align the pixels corresponding to the same point in the reflection scene, and the misalignment of the transmission scene may occur.

Fig. 2: Illustration of the misalignment of reflection scenes and transmission scenes in a binocular camera system.

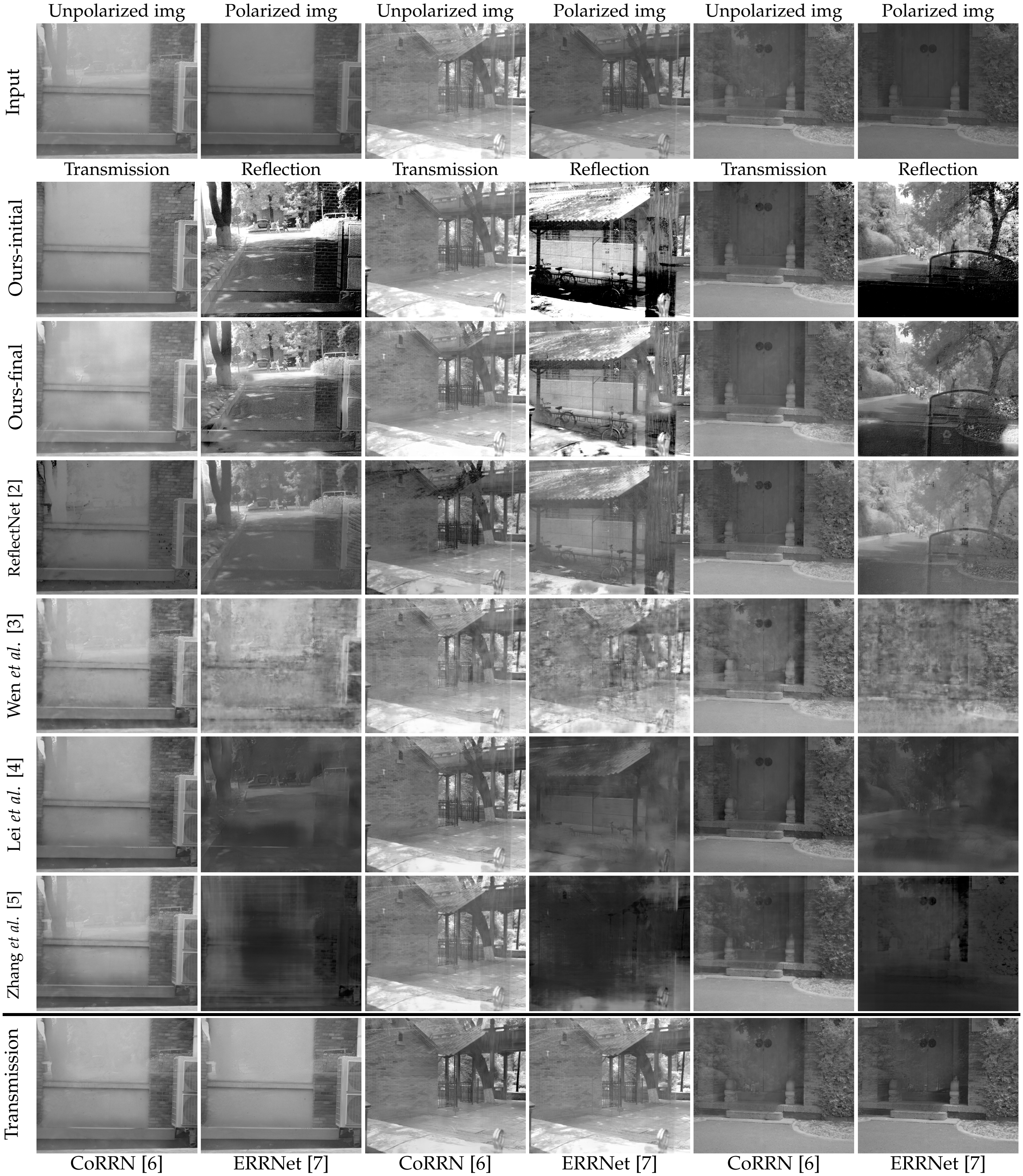


Fig. 3: More qualitative comparisons with ReflectNet [2], Wen *et al*. [3], Lei *et al*. [4], Zhang *et al*. [5], CoRRN [6] and ERRNet [7], evaluated on images in the wild. Our method works well on reflection separations, with less input images compared to other polarization-based methods.

[6] R. Wan, B. Shi, H. Li, L.-Y. Duan, A.-H. Tan, and A. Chichung, "CoRRN: Cooperative reflection removal network," *IEEE TPAMI*, 2019.

[7] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang, "Single image reflection removal exploiting misaligned training data and network enhancements," in *Proc. CVPR*, 2019.