Reflection Separation via Multi-bounce Polarization State Tracing

Rui Li*, Simeng Qiu*, Guangming Zang, and Wolfgang Heidrich

King Abdullah University of Science and Technology, Thuwal, SA
{rui.li, simeng.qiu, guangming.zang, wolfgang.heidrich}@kaust.edu.sa

Abstract. Reflection removal from photographs is an important task in computational photography, but also for computer vision tasks that involve imaging through windows and similar settings. Traditionally, the problem is approached as a single reflection removal problem under very controlled scenarios. In this paper we aim to generalize the reflection removal to real-world scenarios with more complicated light interactions. To this end, we propose a simple yet efficient learning framework for supervised image reflection separation with a polarization-guided ray-tracing model and loss function design. Instead of a conventional image sensor, we use a polarization sensor that instantaneously captures four linearly polarized photos of the scene in the same image. Through a combination of a new polarization-guided image formation model and a novel supervised learning framework for the interpretation of a ray-tracing image formation model, a general method is obtained to tackle general image reflection removal problems. We demonstrate our method with extensive experiments on both real and synthetic data and demonstrate the unprecedented quality of image reconstructions.

Keywords: Reflection Removal, Polarization Simulation Engine, ray-tracing, Polarization Tracing

1 Introduction

There are a number of circumstances in photography as well as scientific and computer vision imaging systems in which it is unavoidable to capture images through glass windows or transparent enclosures. In these scenarios, the captured image is a mixture of transmitted and reflected light paths, which degrades both the visual quality of the scene in photography applications, as well as the performance of computer vision algorithms. Therefore, reflection removal and the separation of reflected and transmitted images are topics of considerable interest in both computational photography and computer vision.

For traditional photography, a range of different techniques have been developed to suppress reflections. These range from controlling the lighting on the near side of the window (the darker the better), or using a linear polarizer while

* Jointly first author.

shooting the image at an angle close to the Brewster angle of the reflective surface. However, even in such controlled environments, complete reflection removal is surprisingly hard in practice – for example, the angle to the glass surface varies spatially over the image plane, especially for wide-angle photography. Moreover, in many situations "in the wild" such constrained setups are simply not possible and we need to contend with bright reflections at arbitrary angles. As a result, there has been considerable interest in recovering the transmitted image by utilizing a single image as input [1–4], assuming that the reflected and transmitted scenes are independent in terms of low-level features, high-level semantics or even the motion of views. The independence assumption, however, is not always reliable or discriminating enough to solve this highly ill-posed inverse problem. Moreover, transparent windows often cause weak ghost images due to multiple reflections in the glass (Fig. 1 b), which are not modeled by most prior work.

In this work, we propose a full multi-bounce reflection model for the interaction of reflected and transmitted light with the glass window. At every surface bounce, we model the full change of the polarization state due to the Fresnel equations. The spatial shift introduced by each bounce also creates weaker ghost images, especially of high-intensity light sources. The change in each bounce is influenced by unknown parameters such as the material of the transparent medium and the incident angle. The weak polarization of light from these light-surface interactions acts as an additional cue for separating the reflected and the transmitted component, and to reconstruct a clean image for both components. We design a simple yet effective deep network architecture to perform this separation and image restoration.

To separate the two independent components (i.e., reflected, transmitted) from captured polarization images efficiently, we propose a comprehensive polarization image formation model by taking into account the Stokes vector and Mueller matrices conversion, coordinate between the image sensor and transparent medium, multiple reflectances as well as transmittance, and medium thickness caused ghosts by solving an simple iterative parameter searching problem. Considering a high-quality polarization image dataset is lacking in this area, our experiment dataset includes both unpolarized and polarized cases. This carefully captured high-resolution dataset includes different polarization scenarios that are required to be solved in the real world, we will make it public for this research area. In particular, the main contributions of our works are:

- We propose a differentiable polarization simulation engine to accurately trace the polarization state of light in the transparent medium, combining it with a deep learning framework for end-to-end training.
- We design a mixed network architecture that considers a learnable forward inference model to separate reflected and transmitted scene, and a physically backward simulation loss to further verify and refine network output.
- We also release a real scene polarized image dataset with and without medium reflection, it contains a pair of clear polarized images and several reflected scenes with different scene parameters.

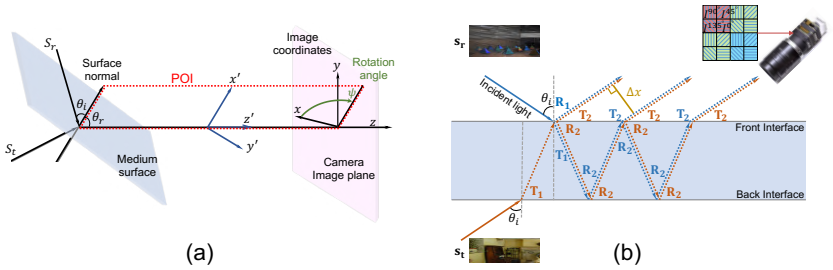

Fig. 1: Light reflection and transmission model: Incident angle θ_i, reflection angle θ_r, transmission angle θ_t, and POI. (a) Object surface and camera image plane are in different coordinate (b) Multiple light paths for reflected the (blue) and transmitted (red) scenes. The top-right is a a color polarization camera with micropolarizer at four different angles and Bayer filter layout.

2 Related Work

Single image reflection separation is a well-known ill-posed problem. There are two general approaches to this. One is based on handcrafted priors [5–11], while the other is based on deep learning methods [1–3]. The handcrafted priors that are based on the observations from specific natural images. Properties like gradient sparsity [8, 5, 11], relative smoothness [12–14], and ghosting cues [10], are leveraged in the literature. Although reasonable performance can be observed when these assumptions hold, high-level understanding for the specific input data is required for the prior based approaches. In deep learning based methods, deep convolutional neural networks (CNN) are applied to solve this inverse problem. Fan et al. [1] propose an image learning network to predict the background layer in an end-to-end approach. Yang et al. [15] estimate background and reflection in an alternating fashion, to improve the accuracy of reflection removal. Jin et al. [16] propose a neural network approach with a focus on handling color ambiguity and saturation. Wan et al. [17] applies a multi-scale strategy on the learning network to improve the target details. Wen et al. [18] synthesizes and remove reflection with a non-linear model. Perceptual loss functions have been adopted in [4]. An alignment-invariant loss is introduced in [19] to improve the performance under misaligned data.

Multiple image reflection removal is also an active research area, where some measurement diversity is introduced in the form of motion or rotating polarizers etc. By estimating the motion between the transmitted and reflected images with different strategies [20–22], researchers manage to separate them for reflection removal. Recently, there is an emerging interest in polarization guided image reflection separation [23–25]. With multiple images captured for the same scene at different polarization angles, the reflections from glasses are separated by applying independent components analysis [26]. Kong et al. [25] propose a multi-scale strategy to find the reflection separation by investigating the properties of different polarization angles. However, they failed to consider the thickness of the

transparent medium, which introduced ghost images. By leveraging the properties of polarized light, the reflection and transmission images are separated with more impressive results in recent approaches [24]. Lyu et al. [23] takes unpolarized and polarized images for reflection separation. However, to the best of our knowledge, all these methods assume a simplified model with only one transmitted light path, whereas multiple reflections between the two glass surfaces are ignored. In contrast, we proposed a comprehensive polarization engine, to jointly describe polarization states for each reflection in a physical plausible way, which significantly improves our results compared to the state of the art.

3 Physically-Based Image Formation Model

When photographing in front of a double-surfaced transparent planar medium, people always capture a transmitted image with the unwanted reflected images. Therefore, we propose a comprehensive polarization-based image formation model for separating the reflection layer and transmission layer of blending scenes. Specifically, we separate the image into a reflected and a transmitted component, while taking into account multiple bounces in the glass surface and the associated ghost images (Fig. 1b).

3.1 Polarization Image Formation Model

Considering a local coordinate frame of a light ray hitting a transparent surface (Fig. 1a), a plane of incidence (POI) subsumes the transmission angle θ_t and the reflection angle θ_r, which is equal to the incident angle θ_i. The angles are related to the refractive indices via Snell's law: $n_0 \sin\theta_i = n \sin\theta_t$, where the transparent medium has refractive index n, and n_0 is the refractive index of the ambient medium (e.g. $n_0 \approx 1$ for air). An incident light passing through or reflected off a transparent media is partially polarized and consists of two orthogonal polarized components that perpendicular and parallel to the POI. This relationship is guided by the Fresnel equations, which we briefly summarize in the following. We define reflectance R and transmittance T as the intensity ratio of reflected light and transmitted light to incident light, respectively. The subscripts $\parallel$ and $\perp$ represent the polarized components parallel and perpendicular to the POI. R is derived from two orthogonal polarized elements of reflectance, $R = (R_\parallel + R_\perp)/2$. Likewise, $T = (T_\parallel + T_\perp)/2$.

We adopt Mueller calculus to represent the polarization state of light. The full polarization state is described by a 4D Stokes vectors $\mathbf{s} = (s_0, s_1, s_2, s_3)^T$. However, in our case we only consider linear polarization, in which case we only require a 3D vector, corresponding to the first 3 components of the full Stokes vector. In the coordinate frame of the camera sensor we perform polarization measurements with four different linear polarizer angles resulting in four images I^{0°, I^{45°, I^{90° and I^{135°, which are acquired simultaneously with a polarization image sensor. Given the four linear polarizer images our Stokes vector in the

image plane can be computed as

$$\mathbf{s} = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(I^{0^\circ} + I^{45^\circ} + I^{90^\circ} + I^{135^\circ}) \\ I^{0^\circ} - I^{90^\circ} \\ I^{45^\circ} - I^{135^\circ} \end{bmatrix}, \tag{1}$$

The change of polarization states as we propagate the light through the transparent surface can be described by Mueller matrices, which operate on the Stokes vectors. First, it is necessary to transform the vector between the local coordinate frame of the transparent surface and the coordinate frame of the camera sensor, which we also take to be the global coordinate frame for simplicity. This is achieved with a rotation Mueller matrix:

$$\mathbf{C}(\psi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos 2\psi & -\sin 2\psi \\ 0 & \sin 2\psi & \cos 2\psi \end{bmatrix}, \tag{2}$$

which maps a Stokes vector $\hat{\mathbf{s}}$ in local coordinates to a Stokes vector $\mathbf{s}$ in global coordinates. $\mathbf{C}^{-1}(\psi) = \mathbf{C}(-\psi)$ is used for the reverse mapping.

Next, we need to model the Mueller matrices for the individual reflection and transmission operations along the light path (Fig. 1a,b). There are two pairs of Mueller matrices: $\mathbf{R}_1$ and $\mathbf{T}_1$ describe the reflection and transmission as the light travels from the outside (optically thinner medium) to the transparent object (optically thicker medium). When the light travels from the inside (optically thicker) to the outside (optically thinner), reflection and transmission are respectively described by $\mathbf{R}_2$ and $\mathbf{T}_2$. Please see the supplemental materials or [27, 28] for the definition of these four matrices.

The contribution of the reflected scene consists of the direct reflection on the surface, as well as all possible ghost images, with the latter being characterized as all light paths transmitted into the medium, followed by an odd number of reflections, and transmission out of the medium. In the local coordinate frame of the transparent object, this can be described as

$$\hat{\mathbf{s}}_r(x) = \mathbf{R}_1\hat{\mathbf{s}}_r^0(x) + \sum_{i=0}^{\infty} \mathbf{T}_2\mathbf{R}_2^{2i+1}\mathbf{T}_1\hat{\mathbf{s}}_r^0(x - i \cdot \Delta x), \tag{3}$$

where $\Delta x = 2d \tan\theta_t \sin\theta_t$ is the spatial offset between ghost images (see Fig. 1b), and $\hat{\mathbf{s}}_r^0 = \mathbf{C}(-\psi)\mathbf{s}_r^0$ describes the polarization state of the reflected scene before interacting with the transparent object.

Likewise, the total contributions of the transmitted scene are given by a transmission into the object, followed by an even number of internal reflections (possibly zero), and a transmission out of the object:

$$\hat{\mathbf{s}}_t(x) = \sum_{i=0}^{\infty} \mathbf{T}_2\mathbf{R}_2^{2i}\mathbf{T}_1\hat{\mathbf{s}}_t^0(x - i \cdot \Delta x), \tag{4}$$

with $\hat{\mathbf{s}}_t^0 = \mathbf{C}(-\psi)\mathbf{s}_t^0$ being the initial polarization state. The total light imaged by the camera can then be described in global/camera coordinates as

$$\mathbf{s} = \mathbf{C}(\psi)(\hat{\mathbf{s}}_r(x) + \hat{\mathbf{s}}_t(x)). \tag{5}$$

3.2 Polarization Simulation Engine

To accurately simulate the polarization state for a reflected I_r and transmitted I_t scene, our forward simulator takes RGB image pairs (I_r, I_t) that correspond to "clean" images without interaction with a transparent surface. From these inputs, the simulator generates blended images of reflected and transmitted scenes with full polarization state, i.e., simulated polarized images as described above. The initial polarization states $\mathbf{s}^0_{\{r,t\}}$ can either simply be set to unpolarized states or some other manually selected value, or it could be estimated from the reflected and transmitted scenes.Thus, our simulator can be represented as,

$$I^{\phi} = \mathcal{S}(I_r, I_t, \Theta), \tag{6}$$

where I^{ϕ} are a set of linearly polarized images corresponding to simulated sensor images. The scene parameters Θ correspond to the thickness of the transparent object d, light incident angles θ_i, and refractive index n. In our setup, the glass object has a constant thickness refractive index. The glass thickness d and incident angle θ_i, together with the corresponding transmission angle θ_t, mainly affect the spatial shift Δx in Eqn. (3) of multi-bounce images (maximal bounce number is 10). Therefore, we only need to estimate the spatial shift Δx to obtain the light incident angle θ_i. The energy loss within the glass was considered, but can actually be neglected in our setting: the attenuation coefficient in glass is approximately $\alpha \approx 0.5db/km$, corresponding to an energy loss of around 12% per km. This amounts to only a reduction by a factor of $2.4 \times 10^{-4}\%$ per bounce for a typical glass thickness. We will show that the simulation results can verify and refine the output of recovered reflected and transmitted scenes.

4 Proposed Method

Given a set of linearly polarized images $I^{\phi} = \{I^{\phi_i} \mid \phi_i = 0, \cdots, N\}$, our approach first independently decomposes each polarized image into a transmitted scene $\hat{I}_t^{\phi} = \{I_t^{\phi_i} \mid \phi_i = 0, \cdots, N\}$ and a reflected scene $\hat{I}_r^{\phi} = \{I_r^{\phi_i} \mid \phi_i = 0, \cdots, N\}$ via the proposed PolarNet, then FusionNet takes all pairs of proposed separations $\hat{I}_r^{\phi}$, $\hat{I}_t^{\phi}$ and polarized image set I^{ϕ} as input to generate a refined final $\hat{I}_r$ and $\hat{I}_t$.

4.1 Network Architecture

Our pipeline mainly contains two cascade networks, PolarNet and FusionNet, for processing single polarized images and combining multiple separated results for refinement, and one polarization simulation engine (PSE) takes refined $\hat{I}_r$ and $\hat{I}_t$ as input to recover polarized image set I^{ϕ} by physically simulating light traveling in transparent medium. The PolarNet first decomposes each polarized image I^{ϕ_i} into two independent reflected and transmitted images $\hat{I}_r^{\phi_i}$ and $\hat{I}_t^{\phi_i}$. The encoder uses a pre-trained VGG-19 network as feature extractor with fixed parameters, and the decoder concatenates feature maps in the selected downsampling layers and upsampling layers (see Fig. 2).

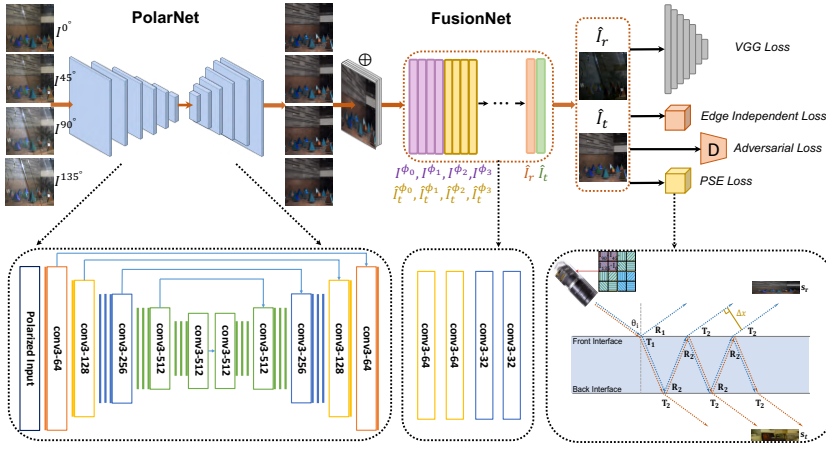

Fig. 2: Overview of system architecture.

FusionNet combines all the pairs of predicted $(\hat{I}_r^{\phi_i}, \hat{I}_t^{\phi_i})$ to have a final refined $\hat{I}_r$ and $\hat{I}_t$. Our PSE automatically estimate scene parameters and accurately simulate the polarization state of light when traveling thought transparent medium, to match the input of polarized image set I^ϕ by $\hat{I}_r$ and $\hat{I}_t$. For example, light travels through transparent medium will have multiple bounces between two surface of glass, for each bounce, the polarization state will change as detailed in Section 3. PolarNet, FusionNet, and the PSE are connected together to form a loop with an end-to-end training process.

4.2 Perceptual and Simulation-based Loss Function

We assume that the real reflected scene I_r and transmitted scene I_t are perceptually different, and the recovered $\hat{I}_r$ and $\hat{I}_t$ can physically reconstruct the input polarized images I^ϕ by given the estimated scene parameter (e.g., incident angle). Our pipeline integrates those two assumptions into the loss function and network architecture design. Our loss function contains 5 terms: a pixel-wise loss term $\mathcal{L}_d$ measures the per pixel difference between estimated $\hat{I}_r$, $\hat{I}_t$ and synthetic ground truth I_r and I_t. For real scene polarized data acquisition, I_r is hard to capture in general, so $\mathcal{L}_d$ only measures $\hat{I}_t$ and I_t for training. The perceptual loss $\mathcal{L}_p$ measures the perceptual independence of reflection and transmission, while the PSE loss $\mathcal{L}_s$ forces the output to accurately reconstruct polarized inputs. The edge independent loss $\mathcal{L}_e$ encourages the gradient of two scene to be independent, and an adversarial loss $\mathcal{L}_a$ encourages production of realistic images. Therefore, our overall loss function is,

$$\mathcal{L} = \mathcal{L}_d + \lambda_p \mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_e \mathcal{L}_e + \lambda_a \mathcal{L}_a. \tag{7}$$

Pixelwise Loss $\mathcal{L}_d$. $\mathcal{L}_d$ compares estimated polarized images $\{\hat{I}_t^{\phi_i}\}$, $\{\hat{I}_r^{\phi_i}\}$, $\hat{I}_r$ and $\hat{I}_t$ with synthetic ground truth $\{I_r^{\phi_i}\}$, $\{I_t^{\phi_i}\}$ or camera captured scenes $\{I_t^{\phi_i}\}$, in

terms of feature space,

$$\mathcal{L}_d = \|I_t - \hat{I}_t\|_2 + \|I_r - \hat{I}_r\|_2 + \sum_{i=0}^{N-1} \|I_r^{\phi_i} - \hat{I}_r^{\phi_i}\|_2 + \|I_t^{\phi_i} - \hat{I}_t^{\phi_i}\|_2, \tag{8}$$

where $N = 4$ in our system setting indicates 4 polarized images, the target I_t or I_r could be calculated via canonical Stokes vector s_0 from Eqn. (1).

Perceptual Loss $\mathcal{L}_p$. The reflected and transmitted scenes are perceptually independent, therefore, they should be semantically different. As shown in many successful applications (e.g. [4]), pre-trained networks can be directly utilized as high-level feature extractors. To combine high-level loss, we use a pre-trained VGG-19 network [29] to measure the difference between recovered layers and ground-truth layers (reflected and transmitted scenes). The perceptual loss function can be defined as the concatenation of a selected layers of VGG-19 network, and we compute the L_1 loss to measure distance in feature space as,

$$\mathcal{L}_p = \sum_l \lambda_l (\|\Phi_l(I_r) - \Phi_l(\hat{I}_r)\|_1 + \|\Phi_l(I_t) - \Phi_l(\hat{I}_t)\|_1), \tag{9}$$

where Φ_l is the output of VGG-19, we stack 5 layers output as features: 'conv1_2', 'conv2_2', 'conv3_4', 'conv4_4', 'conv5_4', and λ_l is the weight of the layer.

PSE Loss $\mathcal{L}_s$. A good recovery of reflected or transmitted scenes could ideally recover the input polarized images using our polarization simulation engine, if scene parameters (e.g., incident angle, thickness of glass, etc.) are given or can easily be estimated. For normal reflection removal cases, the reflected and transmitted light mainly interact with double-parallel surface of transparent glass. Since the physical properties of glass are within a narrow range, the major scene parameters could be estimated by several simple line search to obtain reasonable values for the thickness and incident light angle. $\mathcal{S}(\hat{I}_r, \hat{I}_t, \Theta)$ is our polarization simulation model, it takes a predicted $\hat{I}_r$, $\hat{I}_t$ as well as scene parameters Θ as input, and traces each scene light's polarization state and spatial shift inside the two surface of transparent glass. We adopt an L_1 loss to measure the difference between ground-truth polarized images and the simulator output as,

$$\mathcal{L}_s = \sum_i \|I^{\phi_i} - \mathcal{S}(\hat{I}_r, \hat{I}_t, \Theta)\|_1. \tag{10}$$

We implement $\mathcal{S}$ by pytorch tensor data structure, all the simulation can be computed by interior tensor operations, therefore, $\mathcal{L}_s$ can be optimized by autograd as a normal deep network training pipeline.

Edge Independent Loss $\mathcal{L}_e$. Two independent scenes are unlikely to have overlapping gradients or edges. Based on this observation, we proposed an edge independent loss to penalize those overlapping edge to recover better scenes. We formulate the $\mathcal{L}_e$ as the normalized downsampled gradient difference as,

$$\mathcal{L}_e = \sum_n \|f_n^{\downarrow} \nabla I_r - f_n^{\downarrow} \nabla \hat{I}_r\|_1 + \|f_n^{\downarrow} \nabla I_t - f_n^{\downarrow} \nabla \hat{I}_t\|_1, \tag{11}$$

where $f_n^{\downarrow}$ is the downsampling operator with a factor of 2^{n-1}.

Adversarial Loss $\mathcal{L}_a$. To encourage realistic images, we apply normal conditional GAN's adversarial loss to avoid potential artifacts in recovered images, such as, black holes or color errors. The loss for conditional the GAN discriminator $\mathcal{D}$ is,

$$\sum_{i=0}^{N} \log D(I^{\phi_i}, I_t) - \log D(I^{\phi_i}, \hat{I}_t), \tag{12}$$

then the adversarial loss $\mathcal{L}_a$ can be formed as,

$$\mathcal{L}_a = \sum_{i=0}^{N} - \log D(I^{\phi_i}, \hat{I}_r). \tag{13}$$

5 Experiments

To further illustrate the superiority of this polarization-based model, we compare our proposed method with five recently published baseline reflection separation techniques. The first one is an alignment-invariant loss which is introduced in [19] (*ERRnet*) to improve the performance under misaligned data. The second one is perceptual loss functions proposed by [4] (*Zhang et al.*). Wan et al. [17] (*Wen et al.*) apply a multi-scale strategy on the learning network to improve the target details. We also compared our results with polarization guided reflection separation methods [23, 24](*ReflectNet and Lyu et al.*), which are proposed most recently. They assume a simplified model with only one transmitted light path, whereas multiple reflections between the two glass surfaces are ignored.

5.1 Datasets

Synthetic Data. We generate synthetic data using our polarization simulation engine. Any pair of reflected and transmitted scene can generate a set of blended images by changing the scene parameters, e.g., incident angle, glass thickness, and refractive index. We use a widely-adopted single RGB image dataset for scene blending [4], and a proposed polarization dataset [28]. Our PSE can simulate several usual cases of reflection, e.g, strong/weak reflection, polarized incoming light, mirror reflection, etc.

Real Data. For the real data, we use a color polarization camera PHX050S-Q from Lucid Vision Labs ⋆. The polarization sensor is shipped with Sony IMX250MYR CMOS with 2048×2448 pixels. Each pixel size is $3.45\mu m \times 3.45\mu m$, and every 4×4 pixels sample polarization at angles of $\{0°, 45°, 90°, 135°\}$ as well as color filter array for jointly capturing polarization and color information pixel layout shown in Fig. 1b. We use a 16mm C-mount lens for the results. To capture

⋆ https://thinklucid.com/phoenix-machine-vision/

a clear polarized scene, we carefully select static scenes, setup camera parameters and adjust focus, and then take the clear scene as ground truth of the current scene. To capture the blended polarized image, we set up a glass planar in the front of the lens, and adjust the angle and position of glass to capture reflected scenes with various scenarios. Our dataset contains 26 carefully captured scenes, where for each scene we captured one real ground-truth polarized transmission image and $6-12$ blended images. The resolution of the raw sensor is 2448×2048, while the resolution for each angle of the polarized image is 1224×1024. To capture high quality polarized data, we fine-tune the exposure time of polarization camera within a range of $[1 \times 10^4, 1 \times 10^5]\mu s$.

Training Data and Parameters. To train our proposed network, we generated 10,000 pairs of synthetic blended image from a widely-used RGB image dataset [4], with synthetic parameters $d = 10$, incident angle $\theta = [10, 40, 60, 70, 80, 85]$, refractive index $n = 1.5$, and 4 polarization angles matching the camera we use. For the training parameters, we use the Adam optimizer[30] with a learning rate of 1×10^{-4}, $\lambda_p = 0.1$, $\lambda_s = 2$, $\lambda_e = 0.1$, $\lambda_a = 0.1$. To augment the real scene dataset, we randomly sample multiple rectangle regions and resize it to 512×512 for each training iteration with proper flip and downsampling operations. We use 13 real scene and synthetic data for training, and other 13 scene for testing.

5.2 Visual Comparison in Synthetic and Real Scene

We compare our method with state of the art methods [19, 18, 4, 23, 24], including single or polarization-based approaches. The implementation is based on both polarization-guided synthetic data and our delicately captured experiment data. We first show a visual comparison ([19, 18, 4]) for synthetic polarized dataset (Fig. 3). Here, our results clearly remove the reflection as well as the ghosting effect induced by the thick glass. In the second part, we compare the previous methods along with our approach on the experiment dataset (Fig. 4), which are 'pyramid', 'wood', 'DNA', and 'bird' from top to the bottom. Our proposed method yields superior results in suppressing complex reflections (row 1), or high intensity and strongly polarized reflections (row 2). It also manages to rectify color when the reflected light distorts the color of the transmitted scene (row 3), and to remove glossy reflection (row 4). In addition, we compared our results with polarization-based approaches [24, 23] from real dataset (Fig. 5), which are 'violin', 'paint', 'library', and 'tea'. *ReflectNet* changes the background color of the transmitted images slightly, and *Lyu et al.* only focus on the gray scale images that captured by a monochrome polarization camera. Again, our real experiment results provide a set of distinctive transmission images.

Special Cases. Next, we test several difficult illumination scenarios, with results shown in Fig. (6). Specifically, we simulate scenes that are weakly polarized or with partially polarized illumination, which contains an LCD screen. We also simulate very bright and high contrast (HDR) scenes. The weakly polarized scene contains diffuse transmitted light with a weakly polarized reflected scene. All the

Fig. 3: Visual comparison for synthetic polarized dataset. From left to right are: polarized images (*Polarized*), transmitted scene (*GT*), *ERRnet*[19], *Wen et al.* [18], *Zhang et al.*[4], and our transmitted results (*Ours*), respectively.

competitors achieve reasonable results, and our proposed method outputs a clear transmitted scene in line with the state-of-the-art. For the fully polarized object, we show a scene with an LCD screen in the transmission image. This scenario is more challenging for the competing methods, and results in incomplete removal of the reflections, whereas out approach recovers a clean transmission image. High light levels will generally lead to overexposure, and our polarized input can suppress strongly polarized reflection and recover details. Finally, the HDR scene contains non-semantic highlights, which is challenging for our competitors due to the lack of extra cues to identify the two scenes.

5.3 Quantitative Evaluation

In Tab. (1) we quantitatively compare our method against other state-of-the-art methods using both PSNR and SSIM. Since our approach fully utilizes polarization information, it achieves best results in most of the challenging cases. We compare *Lyu et al.* with converged gray-scale images, and our PSNR/SSIM results in Tab. (2) perform significantly better in each cases.

Table 1: PSNR/SSIM measurements for three approaches [19, 18, 4].

	eccv	bird	violin	pyramid	wood	DNA
ERRnet	19.25 / 0.823	24.36 / 0.843	26.77 / 0.885	21.67 / 0.689	11.80 / 0.650	20.66 / 0.314
Wen et al.	22.77 / 0.791	23.40 / 0.823	24.14 / 0.840	19.37 / 0.699	11.49 / 0.620	18.87 / 0.321
Zhang et al.	18.83 / 0.931	25.11 / 0.854	24.33 / 0.879	19.38 / 0.660	16.89 / 0.690	20.79 / 0.290
Ours	**26.14** / **0.83**	**26.78** / **0.802**	**30.34** / **0.877**	**27.9** / **0.824**	**26.79** / **0.79**	**28.72** / **0.788**

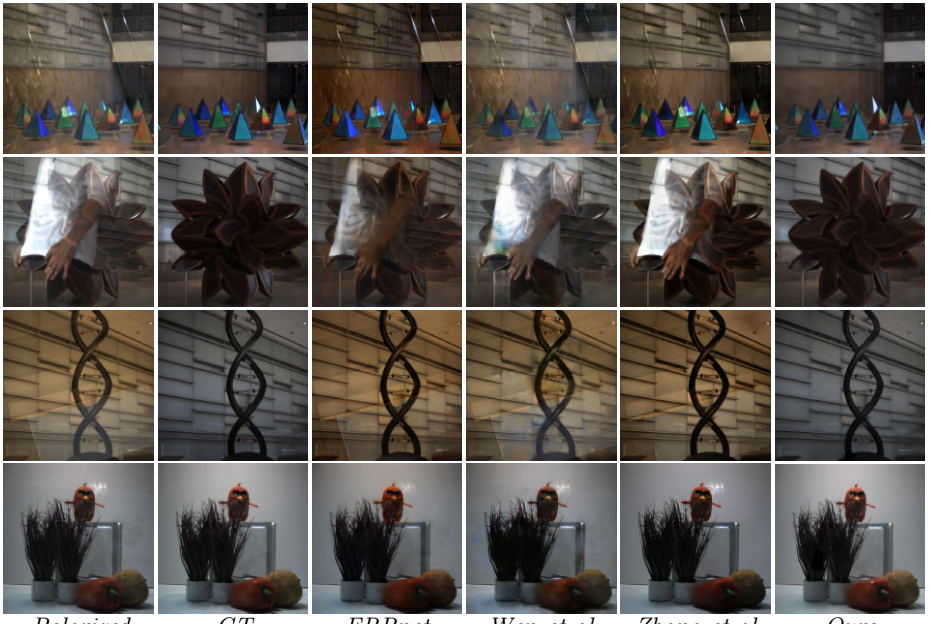

Fig. 4: Visual comparison for our real scene dataset. From left to right are: polarized images (*Polarized*), scene without glass (*GT*), *ERRnet*[19], *Wen et al.*[18], *Zhang et al.*[4], and our transmitted results (*Ours*), respectively.

Table 2: PSNR/SSIM measurements for polarization-based approaches [24, 23].

	violin	paint	library	tea
Lyu et al.	10.56 / 0.519	10.31 / 0.267	11.10 / 0.425	10.33 / 0.450
ReflectNet	14.16 / 0.458	13.28 / 0.504	14.09 / 0.592	12.10 / 0.479
Ours	**31.08** / **0.911**	**25.73** / **0.817**	**29.16** / **0.907**	**22.91** / **0.696**

5.4 Ablation Study

In Fig. 7, we conduct a comprehensive ablation study to show the behavior of each loss term. To cancel the individual terms in the main loss function, we set the related weight parameter to 0 and re-train the model for around 100 epochs.

In Fig. 7(k), we remove our PSE loss, and observed significant artifacts in the reflection mixtures. In Fig. 7(l), the lack of $\mathcal{L}_e$ produces an overlapping gradient in the transmitted scene. In Fig. 7(m), omission of $\mathcal{L}_a$ leads to a noticeable black spot in the image. Fig. 7(n) shows that replacing the VGG-19 feature map with a single L_2 loss of raw pixel values produces an over-smoothed result. We also replaced our encoder/decoder architecture with skip connections with a single multi-layer convolutional network. Fig. 7(o) that these results are in poor performance and lack of local details. Our complete pipeline shows the best performance and clear details in the reconstruction. Tab. (3) shows a quantitative

Fig. 5: Real scene comparison with polarization approaches. From left to right are: 2×2 polarized images at four angles (*Polarized* I^{ϕ}), 2×2 ground truth images (GT), *ReflectNet* [24] with reflected scene on the left and transmitted on the left, *Lyu et al.* [23], reflected and transmitted scene (*Ours*), respectively.

ablation study conducted by muting one of the loss terms and refining other terms weights. We compare generated polarized images with GT by using PSNR and SSIM measurements. Our GT are captured by the same polarization camera without glass while keeping lighting conditions and scenes fixed.

Table 3: PSNR/SSIM measurements for ablation study by muting loss terms.

Loss	all	-$\mathcal{L}_p$	-$\mathcal{L}_s$	-$\mathcal{L}_e$	-$\mathcal{L}_a$
Ablation	28.3/0.889	24.1/0.797	26.4/0.842	26.6/0.809	28.1/0.878

6 Conclusion and Future Work

In this work, we have presented polarization guided image reflection separation. With a new image formation model where polarization information is leveraged for physically plausible measurements, we use the captured image pair as input for our designed supervised deep learning framework. Due to the natural properties of polarization for separating the reflections as well as the elegant network for training, an unprecedented quality can be achieved from our approach, which is demonstrated by the extensive experiments conducted on both synthetic data and real world captures. Although the polarized-based reflection removal performs well in general, it can fail in extremely cases such as unpolarized and metallic reflection, fully polarized illumination. Among these specific scenarios, our polarized-based approach has less benefit for the reflection separation.

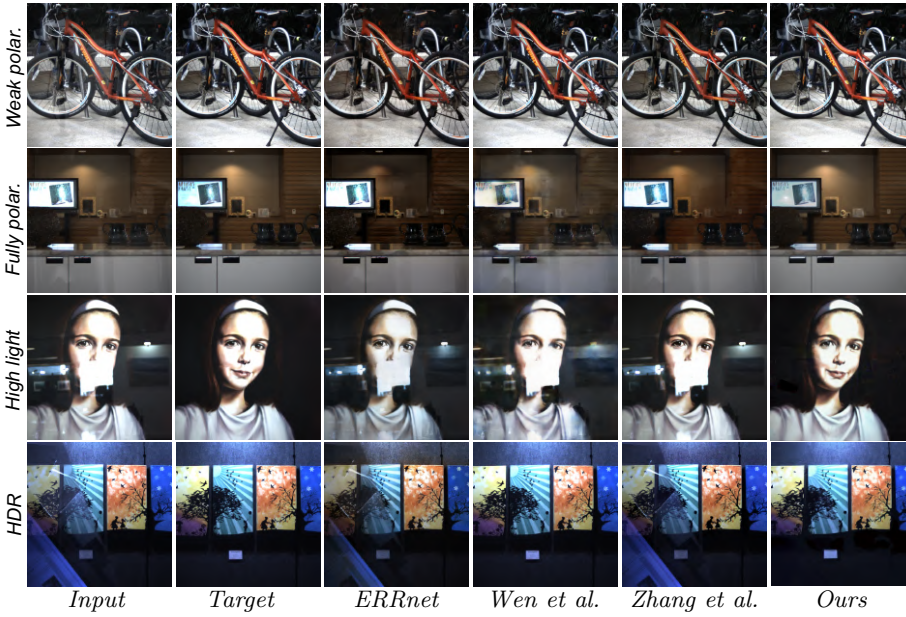

Fig. 6: Visual comparison for special cases. From top to bottom: weak polarized diffuse scene (*Weak polar.*), fully polarized light (*Fully polar.*), high light reflection (*High light*), and dark scene with high dynamic range. From left to right: *Input*, reference image, *ERRnet* [19], *Wen et al.* [18], *Zhang et al.* [4], and *Ours*.

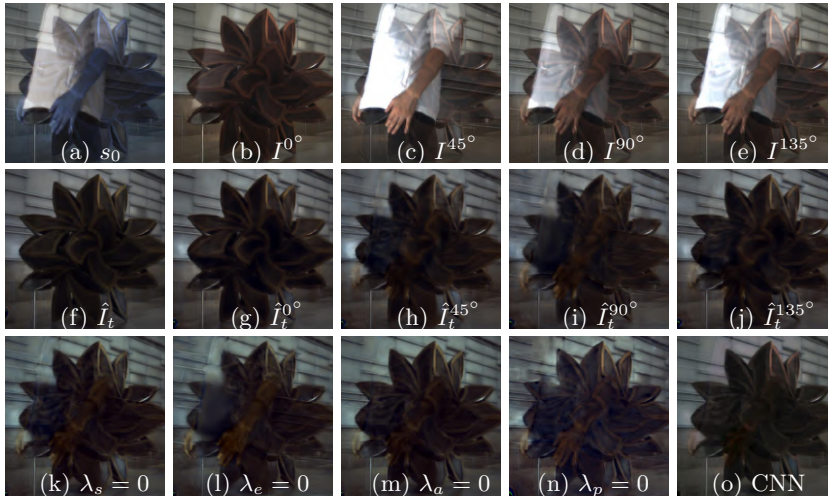

Fig. 7: Ablation study on five individual loss functions for visual comparison. First row: total intensity, 4 polarized images with polarizer angle of $\{0°, 45°, 90°, 135°\}$. Second row: recovered $\hat{I}_t$, $\hat{I}_t^{0°}$, $\hat{I}_t^{45°}$, $\hat{I}_t^{90°}$, and $\hat{I}_t^{135°}$. Third row: without $\mathcal{L}_s$, without $\mathcal{L}_e$, without $\mathcal{L}_a$, without $\mathcal{L}_p$, and replace PolarNet with a simple multi-layer CNN, respectively.

References

1. Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: A generic deep architecture for single image reflection removal and image smoothing. In: CVPR. (2017)
2. Fan, Q., Yang, J., Wipf, D., Chen, B., Tong, X.: Image smoothing via unsupervised learning. TOG (2019)
3. Gandelsman, Y., Shocher, A., Irani, M.: Double-dip: Unsupervised image decomposition via coupled deep-image-priors. In: CVPR. (2019)
4. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: CVPR. (2018)
5. Levin, A., Weiss, Y.: User assisted separation of reflections from a single image using a sparsity prior. PAMI (2007)
6. Li, Y., Brown, M.S.: Exploiting reflection change for automatic reflection removal. In: ICCV. (2013)
7. Levin, A., Zomet, A., Weiss, Y.: Learning to perceive transparency from the statistics of natural scenes. In: NeurIPS. (2003)
8. Levin, A., Zomet, A., Weiss, Y.: Separating reflections from a single image using local features. In: CVPR. (2004)
9. Li, Y., Brown, M.S.: Single image layer separation using relative smoothness. In: CVPR. (2014)
10. Shih, Y., Krishnan, D., Durand, F., Freeman, W.T.: Reflection removal using ghosting cues. In: CVPR. (2015)
11. Arvanitopoulos, N., Achanta, R., Susstrunk, S.: Single image reflection suppression. In: CVPR. (2017)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015)
13. Xu, L., Lu, C., Xu, Y., Jia, J.: Image smoothing via l 0 gradient minimization. In: TOG. (2011)
14. Wan, R., Shi, B., Hwee, T.A., Kot, A.C.: Depth of field guided reflection removal. In: ICIP. (2016)
15. Yang, J., Gong, D., Liu, L., Shi, Q.: Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In: ECCV. (2018)
16. Jin, M., Süsstrunk, S., Favaro, P.: Learning to see through reflections. In: ICCP. (2018)
17. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: Crrn: Multi-scale guided concurrent reflection removal network. In: CVPR. (2018)
18. Wen, Q., Tan, Y., Qin, J., Liu, W., Han, G., He, S.: Single image reflection removal beyond linearity. In: CVPR. (2019)
19. Wei, K., Yang, J., Fu, Y., David, W., Huang, H.: Single image reflection removal exploiting misaligned training data and network enhancements. In: CVPR. (2019)
20. Xue, T., Rubinstein, M., Liu, C., Freeman, W.T.: A computational approach for obstruction-free photography. TOG (2015)
21. Guo, X., Cao, X., Ma, Y.: Robust separation of reflection from multiple images. In: CVPR. (2014)
22. Han, B.J., Sim, J.Y.: Reflection removal using low-rank matrix completion. In: CVPR. (2017)
23. Lyu, Y., Cui, Z., Li, S., Pollefeys, M., Shi, B.: Reflection separation using a pair of unpolarized and polarized images. In: NeurIPS. (2019)
24. Wieschollek, P., Gallo, O., Gu, J., Kautz, J.: Separating reflection and transmission images in the wild. In: ECCV. (2018)

25. Kong, N., Tai, Y., Shin, J.S.: A physically-based approach to reflection separation: From physical modeling to constrained optimization. PAMI (2014)
26. Farid, H., Adelson, E.H.: Separating reflections and lighting using independent components analysis. In: CVPR. (1999)
27. Miyazaki, D., Ikeuchi, K.: Inverse polarization raytracing: estimating surface shapes of transparent objects. In: CVPR. (2005)
28. Qiu, S., Fu, Q., Wang, C., Heidrich, W.: Polarization Demosaicking for Monochrome and Color Polarization Focal Plane Arrays. In: Vision, Modeling and Visualization. (2019)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR. (2014)