Polarized Reflection Removal with Perfect Alignment in the Wild

Chenyang Lei\textsuperscript{1} Xuhua Huang\textsuperscript{1} Mengdi Zhang\textsuperscript{1} Qiong Yan\textsuperscript{2} Wenxiu Sun\textsuperscript{2} Qifeng Chen\textsuperscript{1}

\textsuperscript{1}HKUST \textsuperscript{2}Sensetime

Abstract

We present a novel formulation to removing reflection from polarized images in the wild. We first identify the misalignment issues of existing reflection removal datasets where the collected reflection-free images are not perfectly aligned with input mixed images due to glass refraction. Then we build a new dataset with more than 100 types of glass in which obtained transmission images are perfectly aligned with input mixed images. Second, capitalizing on the special relationship between reflection and polarized light, we propose a polarized reflection removal model with a two-stage architecture. In addition, we design a novel perceptual NCC loss that can improve the performance of reflection removal and general image decomposition tasks. We conduct extensive experiments, and results suggest that our model outperforms state-of-the-art methods on reflection removal.

1. Introduction

It is often desirable to remove glass reflection as it may contaminate the visual quality of a photograph. Reflection separation is also arguably important for robots to work robustly in the real world as the content in reflection usually does not exist in the viewing frustum of a camera. One intriguing property of reflection is that reflected light is often polarized, which may facilitate reflection removal. In this paper, we study reflection removal with polarized sensors by designing a customized deep learning framework.

An image with reflection is a mixture of reflection and transmission, as shown in Fig. 1. In raw data space, the mixed image $M$ can be formulated as

$$M = T + R,$$

where $T$ and $R$ are transmission and reflection, respectively. We name the light behind glass as background $B$ and the light that passes through glass as transmission $T$. Although most prior work treats $B$ as the same as $T$ \cite{34, 32, 7, 31, 24}, we argue that $T$ and $B$ are different. $T$ is darker than $B$ as some light is reflected or absorbed by glass, and there is a spatial shift between $T$ and $B$ due to refraction.

A common issue of many existing reflection removal methods \cite{34, 32, 7, 31, 24} is that strict assumptions are imposed on reflection. These assumptions make previous methods work well in special cases but fail in many others. For example, many works assume reflection images are out of focus \cite{7, 34}. As a result, these approaches may not remove reflection properly when the reflection is sharp and strong. Another prior assumption is on ghost cues \cite{24} that result from multiple reflections inside a thick glass. However, ghost cues do not exist in thin glass.

The lack of diverse and high quality real-world data is another challenging issue. Zhang et al. \cite{34} and Wei et al. \cite{28} have collected a small set of real-world data where only background images (in Fig. 1 and Fig. 4) are captured as the ground-truth transmission images. However, background images are not perfectly aligned with the mixed image $M$ due to refraction and also have the problem of intensity decay ($T$ appears darker than $B$) and color distortion (colored glass). Misalignment introduces great challenges in training a machine learning model \cite{28} and the intensity difference makes it even more difficult. Moreover, since the type of reflection depends on the glass type and only one type of glass is used to collect data, the models trained on these data cannot generalize well to other types of glass.

To be able to relax the assumptions about the appear-
of reflection, we leverage polarization that inherently exists in almost all reflected light. Fig. 2 shows an example polarized image. Existing works based on polarization often impose strict assumptions. A common one is that all light sources are unpolarized [3], which is easily violated in the real world because reflection happens in different types of surfaces in addition to glass and polarized or partially polarized light source exists commonly, such as the LED light. As can be seen in Fig. 2, polarization exists both inside and outside the glass. We cannot solely rely on this information. To rule out the case that polarization also happens in transmission image, our work in this paper removes this assumption. Therefore, our method is more general and applicable to more scenarios.

To ensure the diversity and quality of real-world data, we propose a new data collection pipeline called M-R based on the principle that raw image space is linear. We capture $M$ and $R$ only and obtain the transmission through $T = M - R$. Note that we capture the raw sensor data so that Eq. 1 holds. Our formulation is physically faithful to image formation and eases the process of data collection. We show that with our novel M-R pipeline, it is easy to capture reflection caused by a diverse set of glass. We use the M-R pipeline to build a real-world polarization dataset collected by a novel polarization sensor for reflection removal.

With the collected dataset, we propose a two-stage framework for reflection removal from polarized images. Our approach firstly estimates reflection, with which it infers the transmission image secondly. Our PNCC (perceptual NCC) loss is used to minimize the similarity between the output reflection and transmission. Experiments demonstrate that our method achieves state-of-the-art performances on various metrics. The ablation study shows that our approach benefits from polarized data, PNCC, and the two-stage framework design. Our contributions are summarized as follows:

- We observe two important factors for the task of reflection removal: 1) the difference between transmission $T$ and background $B$ is noticeable. 2) the linearity from reflection to mixed image holds perfectly on raw data.

- We design a new data collection pipeline called M-R, which helps us collect diverse real-world data with perfect alignment by utilizing glass in the real world.

- We propose a deep learning method for reflection removal based on polarization data. Our method does not impose any assumption on the appearance of reflection. A two-stage framework is adopted to get better performance. We design a PNCC loss, which can be applied to many image decomposition tasks. Experiments show that our method outperforms all state-of-the-art methods and has better generalization.

2. Related Work

**Single image reflection removal.** Most single image reflection removal methods [7, 34, 31, 32] rely on various assumptions. Considering image gradients, Arvanitopoulos et al. [2] propose the idea of suppressing the reflection, and Yang et al. [32] propose a faster method based on convex optimization. These methods fail to remove sharp reflection. Under the assumption that transmission is always in focus, Punnappurath et al. [20] design a method based on dual-pixel camera input. For most deep learning based approaches, training data is critical for good performance. CEILNet [7], Zhang et al. [34] and BDN [31] assume reflection is out of focus and synthesize images to train their neural networks. CEILNet [7] estimates target edges first and uses it as guidance to predict the transmission layer. Zhang et al. [34] use perceptual and adversarial losses to capture the difference between reflection and transmission. BDN [31] estimates the reflection image, which is then used to estimate the transmission layer. These methods [34, 7, 31] work well when reflection is more defocused than transmission but fail otherwise. To break the limitation of using solely synthetic data, Zhang et al. [34] and Wei et al. [28] collect real-world datasets for training. However, their datasets have misalignment issues and do not contain sufficient diversity. Wei et al. [28] propose to use high-level features that are less sensitive to small misalignment to calculate losses. To obtain more realistic and diverse data, Wen et al. [29] and Ma et al. [17] propose methods to synthesize data using a deep neural network and achieve better performance and generalization. Though the data is more perceptually appealing, physical authenticity remains a doubt.

**Polarization-based reflection removal.** Polarization is known to be useful in image reflection removal since decades ago [33, 8]. Schechner et al. [33] and Bronstein et al. [4] utilize independent component analysis to separate reflection and transmission images. With the assumption of unpolarized light sources, Kong et al. [14] proposed an optimization method to automatically find the optimal separation of the reflection and transmission layer. Wieschollek et al. [19] combine deep learning with a polarization-based
reflection removal method. Different from previous works, they eliminate a number of assumptions (e.g., the glass must be perfectly flat) and propose a pipeline to synthesize data with polarization information from regular RGB images. However, all light sources are still assumed to be unpolarized.

**Multi-image reflection removal** Polarization-based reflection removal methods are a special category of multi-image approaches. Agrawal et al. [1] use a pair of flash/no-flash images. Many works [26, 22, 21, 16, 10, 11, 30] move the camera to exploit the relative motion between reflection and transmission for reflection removal, while most works assume that motion of the reflection layer is larger than that of the transmission layer. Sarel and Irani [21, 22] assume that both reflection and transmission should be static. Li et al. [16] use SIFT-flow to align the images to make a pixel-wise comparison under the assumption that the background dominates in the mixed image. Xue et al. [30] also require that objects in reflection and transmission are roughly static. Han et al. [11] require the transmission to be more dominant than the reflected scenes.

### 3. M-R Dataset

Real-world reflection removal datasets [34, 28] are limited in quantity and diversity because of the complicated data collection procedure and the difficulty of acquiring ground-truth reflection and transmission. We propose a new method named M-R to collect paired data for reflection removal. A triple \{M, R, T\} is collected for each scene where M, R, T are the mixed image, the reflection image, and the transmission image, respectively.

#### Polarization information

We use the PHX050S-P polarization camera, which is equipped with an IMX250MZR CMOS. This sensor captures an image with four different polarizer angles in one single shot. Each polarization pixel consists of \(2 \times 2\) units with four sub-pixels corresponding to the polarization angles 0°, 45°, 90°, 135°. The light intensity passing through a polarizer follows Malus’ law [12]:

\[
I_{out} = I_{in} \cos^2(|\phi - \theta|),
\]

where \(\theta\) is the angle of polarizer, and \(\phi\) is the polarization angle of incoming light. Note that the equations related to polarization hold only for raw data that is linear to light intensity, and thus we adopt the RAW format in our dataset. The resolution of each captured RAW image is \(2048 \times 2448\). We extract sub-pixels with the same polarization angle to form an image, and we can get four images with resolution \(1024 \times 1224\). The value range of each pixel is from 0 to 4095. Let \(I\) be the light intensity, and let \(I_{np}, I_p\) be the intensity for unpolarized light and linear polarized light. The degree of polarization \(\rho\) equals to \(I_p/(I_p + I_{np})\). Then we define \(I_1, I_2, I_3, I_4\) as the light intensity passed through 4 angles 0°, 45°, 90°, 135°. According to the properties of polarization, we have:

\[
I = (I_1 + I_3 + I_2 + I_4)/2, \quad (3)
\]

\[
\rho = \sqrt{((I_1 - I_3)^2 + (I_2 - I_4)^2)/I}, \quad (4)
\]

\[
\phi = \arctan\frac{I_2 - I_4}{I_1 - I_3}. \quad (5)
\]

**Data collection pipeline** Fig. 4 shows the comparison between our pipeline and previous [34, 28, 27]. Previous methods take a photo in front of glass as a mixed image \(M\)
shows the comparison between
shows the difference between RGB
mission
17
35
27
17
3
3
41
154
154
1
50
-456
transmission
T
Eq. (1)
while prior work \cite{29, 17} claims that the combination
and then remove the glass to take another one as the trans-
mission \( T \) so that the difference between background \( B \) and
transmission \( T \) is ignored. As mentioned before, \( M \) is the
sum of \( R \) and \( T \) (not \( B + R \)). Therefore, inferring \( T \) is
believed to be easier than \( B \). However, it is relatively dif-
cult to capture \( T \) directly because all the reflection must
be blocked. Therefore, we capture \( M \) and \( R \) only and then obtain \( T = M - R \).

While prior work \cite{29, 17} claims that the combination of
reflection and transmission is beyond linearity, we argue
that the non-linearity is introduced by ISP pipeline when
operating in RGB space. On the other hand, there is no such
problem for raw data since the voltage on the sensor
is linearly correlated with the intensity of light. Therefore,
Eq. (1) holds, and we can obtain a transmission image \( T \)
directly by \( M - R \). Fig. 3 shows the difference between RGB
data and raw data. It is clear that our formulation conforms
with reality, and the direct subtraction removes reflection
perfectly. To the best of our knowledge, we are the first to
use \( M - R \) as ground truth on raw data.

To ensure perfect alignment between \( M \) and \( R \), we use
a tripod to fix the camera and take the polarized images re-
moderately controlled by a computer. We first use a piece
of black cloth to cover the back of the glass to block all trans-
mision \( T \) to obtain a clear reflection \( R \). Then we remove
the cloth to collect the mixed image \( M \). To ensure the in-
tensity of reflection are the same in \( M \) and \( R \), we set the
camera to manual model with relatively long exposure time
to avoid noise.

**Analysis of M-R** Table 1 shows the comparison between
our dataset and previous datasets. Compared with previous
methods, M-R has the following advantages:

a). **More diversity.** Previous methods require the glass
to be thin, non-colored, removable, and flat. As long as the
transmission is clear, we do not make such assumptions on
the glass. Therefore, it is possible for us to utilize numerous
glass in our daily life, such as glass doors and windows. The
glass can be flat or curved, thin or thick, colored, or non-
colored. We are even able to record dynamic scenes if the
reflection is static.

b). **Simplified task.** Since \( B \) might be different with
\( T \) in color, intensity, and position, using \( B \) as ground truth
introduces extra problems in reflection removal. Estimating
\( T \) is an equally useful and simplified task. Our dataset has
provided perfectly aligned pairing data.

c). **Improved simulation.** Even if we use our method,
collecting paired data is time-consuming. Since previous
methods have the misalignment problem, they can not cor-
correctly obtain \( R \) by \( M - B \). Besides, they use RGB images
instead of raw images, so non-linearity in intensity is intro-
duced. Derived from the linearity discussed above, we can
use \( M = a \ast R + b \ast T \) directly to simulate various realistic
data where \( a \) and \( b \) varies from 0 to 1 with unpaired \( R \) and \( T \).

**Data cleaning** To improve the quality of our dataset, we
calculate the mean intensity ratio for each pair of \( R \) and \( T \),
and discard the pairs if the ratio is greater than 10 or smaller
than 0.1. As in this situation, either \( R \) or \( T \) is perceptually
invisible. Negative values after subtraction, due to noise, are
set to zero. If there is more than one layer of glass, we crop
the image to keep only the part with a single layer. Polar-
ization can be calculated correctly only if each polarization
image is correct. Hence, we need to pay special attention to

<table>
<thead>
<tr>
<th></th>
<th>Glass type</th>
<th>Data format</th>
<th>Scene</th>
<th>Alignment</th>
<th>Intensity decay</th>
<th>Raw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang et al. \cite{34}</td>
<td>1</td>
<td>( M, B )</td>
<td>110</td>
<td>Misalignment (calibrated)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Wei et al. \cite{28}</td>
<td>1</td>
<td>( M, B )</td>
<td>450</td>
<td>Misalignment</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>SIR benchmark \cite{27}</td>
<td>3</td>
<td>( M, R, B )</td>
<td>100+20+20</td>
<td>Misalignment (calibrated)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Ours</td>
<td>&gt;100</td>
<td>( M, R, T )</td>
<td>807</td>
<td>Perfect Alignment</td>
<td>Small</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 1. Comparison between our collected dataset and others. Our dataset has more diverse glass types, perfect alignment, and little intensity decay. Besides, since we provide raw data, we can synthesize new \( \{ M, R, T \} \) triples faithful to the real image formation process.

Figure 4. Comparison between our data collection pipeline and prior methods. Previous methods use removable, relatively thin, and non-colored glass to avoid misalignment and color distortion. Since we do not enforce these constraints, we can utilize a diverse set of glass types that exist in our daily life.

and then remove the glass to take another one as the trans-
mision \( T \) so that the difference between background \( B \) and
transmission \( T \) is ignored. As mentioned before, \( M \) is the
sum of \( R \) and \( T \) (not \( B + R \)). Therefore, inferring \( T \) is
believed to be easier than \( B \). However, it is relatively dif-
cult to capture \( T \) directly because all the reflection must
be blocked. Therefore, we capture \( M \) and \( R \) only and then obtain \( T = M - R \).

While prior work \cite{29, 17} claims that the combination of
reflection and transmission is beyond linearity, we argue
that the non-linearity is introduced by ISP pipeline when
operating in RGB space. On the other hand, there is no such
problem for raw data since the voltage on the sensor
is linearly correlated with the intensity of light. Therefore,
Eq. (1) holds, and we can obtain a transmission image \( T \)
directly by \( M - R \). Fig. 3 shows the difference between RGB
data and raw data. It is clear that our formulation conforms
with reality, and the direct subtraction removes reflection
perfectly. To the best of our knowledge, we are the first to
use \( M - R \) as ground truth on raw data.

To ensure perfect alignment between \( M \) and \( R \), we use
a tripod to fix the camera and take the polarized images re-
overexposed areas. We calculate an overexposure mask $O$ based on the intensity of $I_1, I_2, I_3, I_4$.

$$O(x) = \begin{cases} 0, & \max\{I_1(x), I_2(x), I_3(x), I_4(x)\} > \delta \\ 1, & \text{otherwise} \end{cases} \quad (6)$$

where $\delta$ is a threshold and we use $\delta = 0.98$ here.

4. Method

4.1. Reflection-Based Framework

Unpolarized light reflected from the glass surface or passed through the glass becomes partially polarized. The degree of polarization, $\rho$, depends on the property of glass and the angle of incidence. For a specific type of glass with refractive index $n = 1.7$. Fig. 6 shows how the degree of polarization changes. Based on this fact, Kong et al. [14] and Wieschollek [19] proposes two methods for reflection removal. However, in the real world, unpolarized light sources assumption doesn’t hold well because partially polarized light sources exist commonly, and reflection exists not only through glass surfaces. These methods would then fail [19, 14]. Different from Wieschollek et al. [19] and Kong et al. [14], we do not assume all light sources are unpolarized. We utilize the fact that the $\rho$ of transmission is quite different. Hence we propose to use deep learning based and two-stage method to catch the differences between reflection and transmission and separate them.

Fig. 5 shows an overview of our framework. Our method takes a multi-channel image as input. The first 4 channels, $I_1, I_2, I_3, I_4$ are extracted from mixed image $M$ for each polarization angle. The next 4 channels are $I, \rho, \phi, O$, calculated from Eq. 3, 4, 5, 6. The final network output is a one-channel image, the recovered transmission $\hat{T}$, the same size as $I_1$, that is half of $M$ in width and height.

There are two stages in our process. The first stage is dedicated to estimating reflection $\hat{R}$, and the second is for transmission $\hat{T}$ with estimated $\hat{R}$. We use a two-stage design for two reasons. Firstly, reflection contributes a lot to mix image and has a strict relationship on RAW space (Eq. 1). Furthermore, as discussed above, reflection and transmission are quite different in terms of polarization. The separated decoders for them are helpful to learn specific features. BDN [31] also observes the importance of reflection and improves performance by training a bidirectional network. However, their performance relies on an assumption to make $R$ and $T$ more different: the reflection is blurry. Undoubtedly, their model cannot distinguish $R$ and $T$ well when reflection is sharp. Note that if without polarization, such design may deteriorate the performance as the difference between them becomes subtle in regular image data.

4.2. Loss function

PNCC loss In general, reflection and transmission images would be different on most pixels. We propose a perceptual normalized cross-correlation (PNCC) loss to minimize the...
correlation between estimated reflection and transmission on different feature maps. Our PNCC loss is defined on different feature maps of VGG-19 [25]. Given two images $I_A$ and $I_B$, we try to calculate the NCC of their feature maps. In practice, the monotonicity is not right in extreme cases where the intensity between $R$ and $T$ has a big difference. Therefore, we normalize $I_A$, $I_B$ to $[0, 1]$, denoted as $\tilde{I}_A$, $\tilde{I}_B$. The PNCC loss is defined as follows:

$$L_{PNCC}(I_A, I_B) = \sum_{l=1}^{n} NCC(v_l(\tilde{I}_A), v_l(\tilde{I}_B)), \quad (7)$$

where $v_l$ denotes the $l$-th layer feature maps of VGG-19 [25]. In practice, we use three layers 'conv2_2', 'conv3_2', 'conv4_2', PNCC can also be applied using another pre-trained neural network.

Fig. 7 shows the monotonicity of PNCC and the impact of normalization. We choose 100 pairs of images randomly from the dataset used in [34]. For each pair ($R, T$) we generate synthetic data $I_A$ and $I_B$ by:

$$I_A = T + (1 - \alpha) \ast R, I_B = \alpha \ast R, \quad (8)$$

where $\alpha$ is sampled from 0.01 to 1. When $\alpha = 1$, $I_A$ and $I_B$ are completely two different images, PNCC is the lowest. When $\alpha = 0.01$, $I_A$ contains most part of $I_B$, PNCC is the largest, but the non-normalized version is not. Our PNCC loss can also be applied to other image decomposition tasks. More results are demonstrated in experiments.

**Perceptual loss** The perceptual loss [13] has been proved effective on various computer vision tasks [15, 34, 5]. In our task, we modify it to account for the overexposed area. Given the overexposure mask $O$, the perceptual loss is defined as:

$$L_p(T, \hat{T}) = \sum_{l=1}^{n} \beta_l |v_l(O \ast T) - v_l(O \ast \hat{T})|_1. \quad (9)$$

$\beta_l$ is the weight for the $l$-th layer. Following Chen et al. [5], we initialize $\beta_l$ based on the number of parameters in each layer and we adopt 6 layers 'conv1_1', 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2', and 'conv5_2'.

In total, the loss function we optimize is the sum of PNCC loss between $\hat{R}$ and $\hat{T}$ and perceptual loss.

### 4.3. Implementation

To improve the performance of our model, we augment the input to the network with the hypercolumn features extracted from the VGG-19 network [25]. In particular, we extract 'conv1_2' from the VGG-19 network for $I_1$, $I_2$, $I_3$, $I_4$, $I$ and upsample the layers bilinearly to match the resolution of the input image. Since our data is in RAW format and pre-trained VGG-19 [25] was trained on ImageNet dataset [6] in RGB space, we first apply a gamma correction to the raw input and then feed them into the network. We adopt U-Net [18] as our network architecture for both $f$ and $g$. We modify the kernel size of the first layer to $1 \times 1$ and use it to reduce the dimensionality of the augmented input [5]. At the training, we first train $f$ and $g$ together for 200 epochs using Adam optimizer and learning rate 0.0001. Then we decay the learning rate to 0.00001 and train for 50 more epochs.

### 5. Experiments

#### 5.1. Experimental procedure

**Baselines.** We compare our method with several state-of-the-art reflection removal approaches, including both deep learning and traditional methods. Specifically, in the deep learning track, we choose Zhang et al. [34], Wei et al. [28], BDN [31], Wieschollek et al. [19], and Wen et al. [29]. For fairness, we re-train models on our M-R dataset using official source codes for Zhang et al. [34] and Wei et al. [28]. For BDN [31] and Wieschollek et al. [19], we directly use the available pre-trained models since no training codes are available. For Wen et al. [29], as their training requires additional alpha matting masks that are not available in our task, we also use their pre-trained model.

For polarization based methods, we choose Kong et al. [14], Schechner et al. [23] and Fraid et al. [8]. Third-party implementations by Wieschollek et al. [19] are used. We also evaluate the convex optimization based method by Yang et al. [32] using their official source codes.

DoubleDIP [9] is an unsupervised image decomposition model, but it fails in our setting. The possible reason is that DoubleDIP holds a simple assumption that a mixed image is composed of two images with spatial-invariant coefficients, but real-world data break the assumption.

**Experimental setup** The experiments are mainly conducted on our M-R dataset since it is the only available raw
Table 2. Quantitative results on our M-R dataset. Our method outperforms all others in PSNR and SSIM. Note methods tagged with * are evaluated with pre-trained models and tag ** stands for third-party implementation. To compare fairly with other methods [8, 33, 14, 19] which use three polarization images, we use $I_1 + I_3 - I_2$ to represent $I_4$ as input as a ‘3 inputs’ version.

<table>
<thead>
<tr>
<th>Method</th>
<th>Transmission PSNR</th>
<th>Transmission SSIM</th>
<th>Reflection PSNR</th>
<th>Reflection SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraid et al.** [8]</td>
<td>21.99</td>
<td>0.714</td>
<td>6.48</td>
<td>0.241</td>
</tr>
<tr>
<td>Schechner et al.** [23]</td>
<td>23.42</td>
<td>0.655</td>
<td>12.40</td>
<td>0.247</td>
</tr>
<tr>
<td>Kong et al.** [14]</td>
<td>18.76</td>
<td>0.402</td>
<td>12.96</td>
<td>0.271</td>
</tr>
<tr>
<td>Yang et al. [32]</td>
<td>25.42</td>
<td>0.780</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wieschollek et al.* [19]</td>
<td>22.15</td>
<td>0.711</td>
<td>15.93</td>
<td>0.462</td>
</tr>
<tr>
<td>BDN* [31]</td>
<td>24.49</td>
<td>0.805</td>
<td>12.34</td>
<td>0.377</td>
</tr>
<tr>
<td>Wen et al.* [29]</td>
<td>26.62</td>
<td>0.827</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wei et al. [28]</td>
<td>30.13</td>
<td>0.899</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Zhang et al. [34]</td>
<td>31.91</td>
<td>0.903</td>
<td>32.02</td>
<td>0.88</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>34.62</strong></td>
<td><strong>0.936</strong></td>
<td><strong>33.88</strong></td>
<td><strong>0.907</strong></td>
</tr>
<tr>
<td>Ours (3 inputs)</td>
<td>33.91</td>
<td>0.930</td>
<td>33.53</td>
<td>0.903</td>
</tr>
</tbody>
</table>

Table 2. Quantitative results on our M-R dataset. Our method outperforms all others in PSNR and SSIM. Note methods tagged with * are evaluated with pre-trained models and tag ** stands for third-party implementation. To compare fairly with other methods [8, 33, 14, 19] which use three polarization images, we use $I_1 + I_3 - I_2$ to represent $I_4$ as input as a ‘3 inputs’ version.

5.2. Comparisons with baselines

Quantitative evaluation Table 2 summarizes the evaluation results on our dataset. Our method presents a new state-of-the-art performance. Performance of traditional polarization-based methods [23, 14, 8] rank low since their assumption that all light sources are unpolarized is oversimplified for real-world data. An interesting phenomenon is that BDN [31] scores badly in reflection despite its bi-directional network design. After analysis, we find out that BDN confuses between transmission and reflection in many cases, which affects the performance significantly. Scores of Zhang et al. [34] and Wei et al. [28] are the closest to ours. In addition to being retrained on our dataset, another common characteristic of the two methods is that they are designed for not only synthetic data but also real data.
Figure 10. Without the polarization information, the network cannot distinguish objects from $R$ or $T$ in many cases, especially for sharp reflection.

Figure 11. Our method achieves satisfying results on curved glass.

Figure 12. Our performance is reasonable on non-ideal data [19]

Qualitative evaluation Fig. 8 shows several samples by different methods in different situations. We choose the best two single image models and the best polarization method for perceptual comparisons. As seen in Fig. 8, our method can handle different types of reflection well and remove the reflection pretty well without introducing artifacts. Wieschollek et al. [19] can also remove different types of reflection based on polarization, but their results have visible artifacts, and it even amplifies the reflection for the third case. For Zhang et al. [34] and Wei et al. [28], the results have visible residual reflection left. Fig. 9 shows a hard case where the mixed image is a bit blurry. Previous methods [32, 31, 28, 29] assuming the reflection is blurry perform poorly and tend to remove too much content. Our result shows better generalization without such an assumption. Our model can also achieve good performance on curved glass and non-ideal data collected by Wieschollek et al. [19], as shown in Fig. 11 and Fig. 12.

5.3. Ablation study

To study the influence of polarization information, we replace the input channels $I_1$, $I_2$, $I_3$, $I_4$, $\rho$, $\phi$ all with $I$ and keep the network structure the same. To study the effect of our two-stage structure, we remove the loss on $R$. Finally, we conduct an experiment with the setting without PNCC. The results are shown in Table 3. Polarization information improves the performance most. Fig. 10 shows a sample. The model predicts $T$ as $R$ without the support of polarization information. The two-stage design also boosts the performance of our model by a large margin. Our proposed PNCC can further increase the performance of our model on reflection removal.

As an additional evaluation, we compare PNCC with the exclusion loss proposed by Zhang et al. [34]. The experiment is conducted in DoubleDIP [9] framework, which adopts exclusion loss to decompose images. By replacing the exclusion loss with our PNCC, we get the evaluation results in Table 4. Our approach outperforms their official implementation easily and still performs better after tuning the hyperparameters for them.

6. Discussion

We propose a two-stage polarized reflection removal model with perfect alignment of input-output image pairs. With a new reflection formulation to bypass the misalignment problem between the background and mixed images, we build a polarized reflection removal dataset that covers more than 100 types of glass in the real world. A general decomposition loss called PNCC is proposed to minimize the correlation of two images at different feature levels. We have conducted thorough experiments to demonstrate the effectiveness of our model. We hope our novel model formulation and the M-R dataset can inspire research in reflection removal in the future.

Acknowledgement

We thank SenseTime Group Limited for supporting this research project.
References


[34] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In CVPR, 2018.