# Reflection Scene Separation From a Single Image

Renjie Wan[†,*] Boxin Shi[★,◇,*] Haoliang Li[†] Ling-Yu Duan[★,◇] Alex C. Kot[†]

[†]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[★]National Engineering Laboratory for Video Technology, Department of CS, Peking University, China
[◇]Peng Cheng Laboratory, Shenzhen, China
{rjwan, lihaoliang, eackot}@ntu.edu.sg, {shiboxin,lingyu}@pku.edu.cn

## Abstract

*For images taken through glass, existing methods focus on the restoration of the background scene by regarding the reflection components as noise. However, the scene reflected by glass surface also contains important information to be recovered, especially for the surveillance or criminal investigations. In this paper, instead of removing reflection components from the mixture image, we aim at recovering reflection scenes from the mixture image. We first propose a strategy to obtain such ground truth and its corresponding input images. Then, we propose a two-stage framework to obtain the visible reflection scene from the mixture image. Specifically, we train the network with a shift-invariant loss which is robust to misalignment between the input and output images. The experimental results show that our proposed method achieves promising results.*

## 1. Introduction

When taking a photo through glass, the camera sensors always receives a mixture of the light emitted by the background scene behind glass and the reflection scene in front of glass. By regarding the reflection components as the noise to be removed, previous reflection removal methods [20, 29] aim at improving the visibility of the background scene by removing the interference from the reflections.

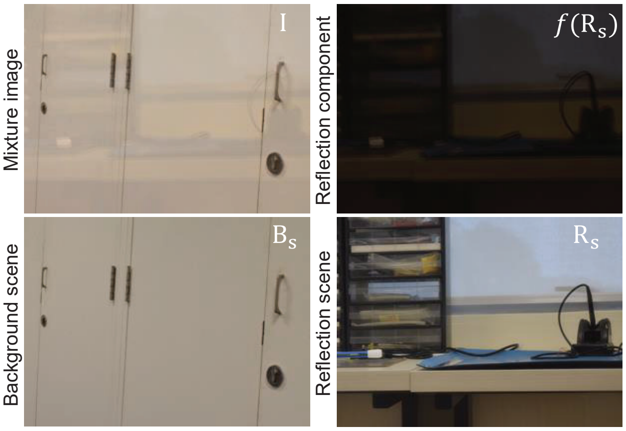

Figure 1. Given a scene image contaminated by reflections, reflection removal recovers $\mathbf{B}_s$ (and sometimes $f(\mathbf{R}_s)$ as a byproduct), while reflection scene separation aims at further extracting $\mathbf{R}_s$ from $f(\mathbf{R}_s)$.

However, glass is not only a noise-generator but also a semi-reflector recording the surrounding world. Important information can be recovered by approximating the light emitted by the reflection scenes. For criminal investigations, the investigators may feel interested in the suspicious objects or persons appeared on glass surface of an image [10]. By analyzing the oncoming vehicles at junctions from their reflections appeared on the parked cars or windows, the self-driving cars or the robotics may use reflections as alternatives to perceive the surrounding world [23].

Considering the degradation for both the background light and reflection light, the formation of a mixture image can be generally expressed as:

$$\mathbf{I} = g(\mathbf{B}_s) + f(\mathbf{R}_s), \tag{1}$$

where $\mathbf{B}_s$ and $\mathbf{R}_s$ denote the light emitted by the background scenes and reflection scenes, respectively; $g(\cdot)$ and $f(\cdot)$ denote the various degradation for $\mathbf{B}_s$ and $\mathbf{R}_s$ during the light transmission, respectively; $g(\mathbf{B}_s)$ and $f(\mathbf{R}_s)$ de-

[*]Corresponding authors. This research is supported by the National Research Foundation, Prime Minister's Office, Singapore, under the NRF-NSFC grant NRF2016NRF-NSFC001-098 and NTU-PKU Joint Research Institute with donation from Ng Teng Fong Charitable Foundation. The research work was done at the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore. This research is in part supported by the National Natural Science Foundation of China under Grants U1611461 and 61872012, National Key R&D Program of China (2019YFF0302902), Shenzhen Municipal Science and Technology Program under Grant JCYJ20170818141146428, and Beijing Academy of Artificial Intelligence (BAAI).

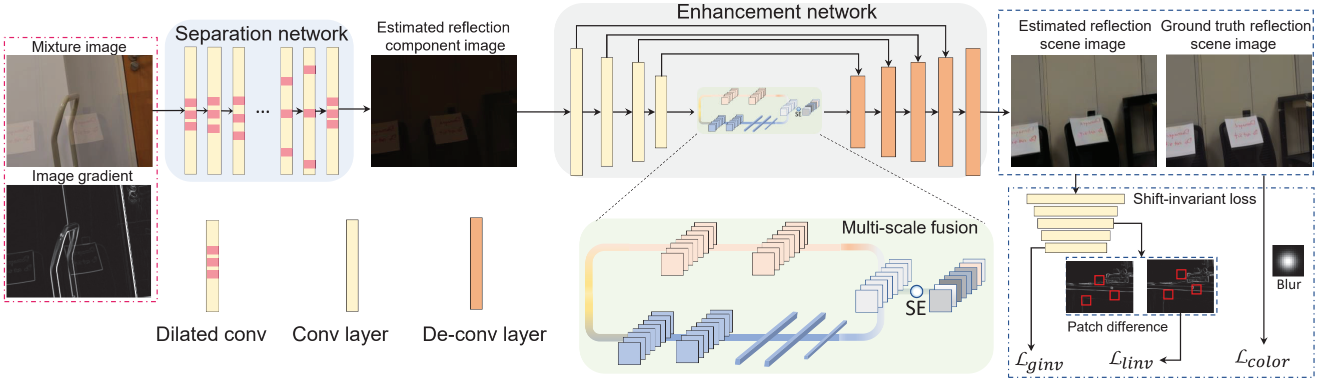

Figure 2. The framework of our proposed approach. The separation network separates the reflection component from the mixture image. The enhancement network estimates the reflection scene image by recovering the visibility of the reflection scene. Specifically, we design a shift-invariant loss by considering the global, local, and color consistency to optimize the whole network with the non-aligned information.

note the background and reflection irradiance that finally reach the camera sensor, respectively; and a mixture of them forms $\mathbf{I}$.

The degradation functions $g(\cdot)$ and $f(\cdot)$ are caused by complicated factors when light travels from a scene to the sensor, which is hard to be written in analytic form. For example, the light $\mathbf{B}_s$ from the background scene may be absorbed when it transmits through glass, which leads to $g(\mathbf{B}_s) \approx \mathbf{B}_s$. However, as shown in Figure 1, due to the high transmittance rate of glass [2], $g(\mathbf{B}_s)$ is almost a copy of $\mathbf{B}_s$, which makes the background component dominates the mixture image. However, for the light $\mathbf{R}_s$ emitted by the reflection scenes, only a limited amount of $\mathbf{R}_s$ can be reflected by glass [2] and received by the camera, which leads to $f(\mathbf{R}_s) < \mathbf{R}_s$.

A popular research topic dealing with such degradation is reflection removal [5], which aims at recovering $\mathbf{B}_s$ by treating $f(\mathbf{R}_s)$ as noise. Though some reflection removal methods [20, 30] also estimate the $f(\mathbf{R}_s)$, their results are always with obvious residue edges from the background, which is difficult for further processing. Moreover, as shown in Figure 1, $f(\mathbf{R}_s)$ contains little meaningfully visual information, since it appears rather dark in most cases. Thus, further extracting $\mathbf{R}_s$ from $f(\mathbf{R}_s)$ with enhanced visibility is still a challenging problem.

We define this problem as the reflection scene separation. This problem aims at obtaining a clear reflection scene image from the mixture image $\mathbf{I}$. To address the above two challenges, we divide the whole solutions into the *separation* and *enhancement* stage. As shown in Figure 2, in the separation stage, we leverage advantages from the progress in the reflection removal [20, 30, 12] to specifically focus on the reflection component separation. In the enhancement stage, to extract $\mathbf{R}_s$ from $f(\mathbf{R}_s)$, we consider the global and local features to improve the visibility of the reflection scenes. At last, since we use the data-driven approach to solve this problem, we propose a strategy to prepare the dataset for training and evaluating this unique task. By putting a planar mirror behind glass, our dataset contains the ground truth reflection scene image, which facilitates the network optimization with real data. Correspondingly, we introduce a new shift-invariant loss function to handle the misalignment in the dataset. Our major contributions are summarized as follows:

- We define the reflection scene separation problem which aims at recovering clear and meaningful reflection information from the mixture image.
- We propose a strategy to obtain the training and evaluation dataset for this problem. By putting a mirror behind glass, our dataset successfully captures the light directly emitted by the reflection scenes as the ground truth.
- We propose a two-stage framework to solve this problem and a shift-invariant loss to facilitate the network optimization for real data where misalignment often occurs.

## 2. Related work

**Reflection removal.** Reflection removal aims at approximating the light emitted by the background scenes. Recent methods mainly adopt the deep learning framework to solve this problem. For example, Fan *et al.* [5] proposed a two-stage deep learning approach to learn the mapping between the mixture images and the estimated clean images. Recently, Wan *et al.* [19] proposed a concurrent model to better preserve the background details. They further proposed a cooperative model [20] to better utilize the intermediate information. Zhang *et al.* [30] also proposed a perceptual reflection removal method based on the generative adversarial network. Recently, Wei *et al.* [22] proposed another method to solve the reflection removal problem with the non-aligned information. Since the background always dominates the mixture image, these methods have achieved

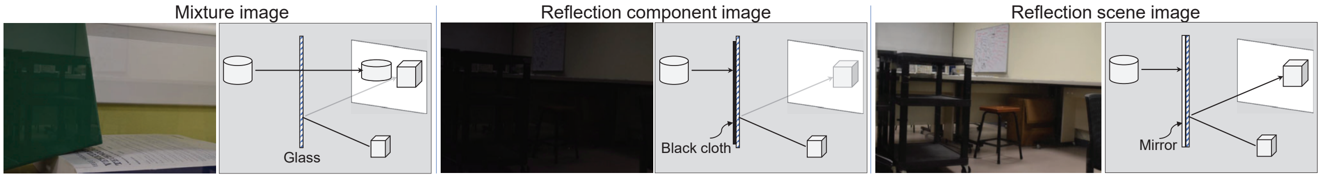

Figure 3. Examples of images in the image triplet and the corresponding capturing setups. From left to right, the mixture image is taken through glass; the reflection component image is taken by putting a piece of black cloth behind glass; the reflection scene image is taken by putting a mirror behind glass.

promising results. Different from these reflection removal methods, our method aims at approximating the light emitted by the reflection scenes.

**The usage of reflections.** The reflection scene separation problem is not specifically discussed by previous methods. The most related one is proposed by Yano *et al.* [25]. However, different from our problem where the background and reflection are highly mixed, the reflection component dominates the mixture image and the background is almost invisible, which cannot be directly used for our problem. Besides, the recent method [10] extracted faces of bystanders from corneal reflections, and proposed its possible application in criminal investigations. Wu *et al.* [23] also proposed another method to locate the objects from the reflections. Nishino *et al.* [17, 16] estimated lighting from corneal reflections for the relighting applications.

## 3. Dataset preparation

Since we use a data-driven approach to solve this problem, the appropriate dataset becomes necessary to learn the inherent image properties. The reflection removal methods [18] proposed to capture an image triplet $(\mathbf{I}, \mathbf{B}, \mathbf{R})$, where $\mathbf{I}$ is the mixture image, $\mathbf{B}$ is the background component image, and $\mathbf{R}$ is the reflection component image recording $f(\mathbf{R}_s)$, the light reflected by the glass. It provides a reasonable way to obtain the ground truth for the reflection component image. By putting a piece of black cloth behind glass, only the reflection light reflected by the glass can be captured by the camera. However, for the reflection scene separation problem, since the final goal is to estimate the light emitted by the reflection scenes, $\mathbf{R}$ alone is not enough for the whole estimation process. To approximate the intensity of light emitted by the reflection scenes, we put a piece of mirror closely behind glass, so that almost all light emitted by the reflection scenes can be reflected by the mirror and captured by the camera.

We introduce a new image triplet $(\mathbf{I}, \mathbf{R}, \mathbf{R}_s)$ for the reflection scene separation problem, where $\mathbf{I}$ and $\mathbf{R}$ are with similar meanings to previous definitions and $\mathbf{R}_s$ denotes the reflection scene image. As shown in Figure 3, the new image triplet is captured as follows: 1) Taking the mixture image $\mathbf{I}$ through glass; 2) capturing a reflection component image $\mathbf{R}$ by putting a piece of black cloth behind glass; 3) capturing a reflection scene image $\mathbf{R}_s$ by putting a piece of mirror closely behind glass.

During the data capture, we use a DSLR camera with an 18-55mm zoom lens to collect the image triplet. The resolution of each image in the triplet is $4000 \times 6000$. Specifically, we use a tripod and remote control to keep the camera stable and choose a small aperture size to minimize the depth of field difference.

**Training and evaluation dataset.** We capture $80$ image triplets in the real world from indoor and outdoor scenarios. $30$ triplets among them are used as the evaluation dataset for the proof-of-concept purpose. The rest $50$ triplets are used for training. We then crop each image for training with original high resolution into approximately $40$ smaller images. By flipping these smaller images, our training dataset contains $2000$ image triplets all from the real world. To increase the diversity of the background scene, we generate another $500$ image triplets with the synthetic mixture image by adding the reflection component images with another image from the public dataset (*e.g.*, COCO [14]). In total, our training dataset is with 2500 image triplets and the evaluation dataset is with 30 image triplets.

**Misalignment analysis.** The tripod and the remote control used during the data capture can well attenuate the misalignment between the mixture image and the reflection component image. However, since the mirror is closely attached behind glass, the misalignment between the reflection component image and its corresponding reflection scene image can hardly be avoided. This misalignment is inherently caused by the touch of the glass and mirror and usually causes 20-70 pixel shifts.

## 4. Proposed method

In this section, we describe the design methodology of the proposed network and the implementation details. From the physical model in Equation (1), we adopt the two-stage framework to address the difficulty separately. In the first stage, we aim at separating the reflection components $f(\mathbf{R}_s)$ from the mixture image. In the second stage, we enhance the visibility of the reflection scene by optimizing the whole network with the non-aligned information.

### 4.1. Separation network

Instead of regarding the reflection components in the mixture image as a byproduct, our separation network

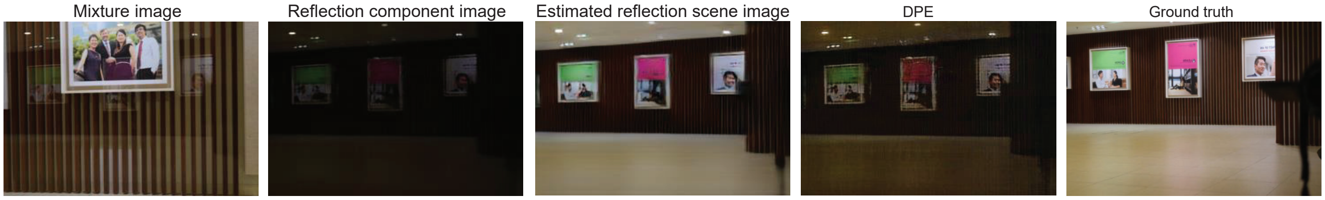

Figure 4. Examples of the mixture image, the reflection component image estimated by the separation network, the estimated reflection scene image by our method and a recently proposed low-light image enhancement method DPE [4], and the ground truth reflection scene image.

aims at separating the reflection components by leveraging advantages from the existing reflection removal methods [30, 20]. Our separation network is similar to [30], which is also used in other image restoration tasks. To adapt to our problems, we first remove the layers for the background component estimation. Then, since the reflection components and background components are more easily to be differentiated in the gradient domain [12], we concatenate the input image with its corresponding gradient image as the input to the separation network. The separation network separates the reflection components from the mixture image as follows:

$$f(\mathbf{R}_s) = \mathcal{G}_S([\mathbf{I}, \nabla\mathbf{I}]), \tag{2}$$

where $\mathcal{G}_S$ denotes the reflection separation network, $\mathbf{I}$ denotes the mixture image, $\nabla\mathbf{I}$ denotes the gradient image of $\mathbf{I}$, $f(\mathbf{R}_s)$ denotes the separated reflection components, and $[\cdot,\cdot]$ denotes the concatenation operation. $\mathcal{G}_s$ contains the dilated convolutional layers, which increase the receptive field size. For simplicity, we denote $f(\mathbf{R}_s)$ as $\mathbf{R}$.

We employ the Wasserstein GAN [1] with the gradient penalty term to separate the reflection components from the mixture image:

$$\mathcal{L}_{\mathbf{adv}}(z, z^*) = \min_{\mathcal{G}_s} \max_{D_s \in \mathcal{D}} E_{z^* \sim \mathbf{P}_g}[D_s(z^*)] - E_{z \sim \mathbf{P}_r}[D_s(z)], \tag{3}$$

where $D_S$ is the discriminator network, $\mathcal{D}$ is the set of 1-Lipschitzonger, $z$ is the estimation of the network, $z^*$ denotes the ground truth of $z$, and $\mathbf{P}_r$ and $\mathbf{P}_g$ are the corresponding data distributions.

We also use a feature loss to measure the difference between the predicted reflection layer and the reference reflection layer in the feature space as:

$$\mathcal{L}_{\mathbf{feat}} = \sum_l \delta_l \left\|\Phi_l(\mathbf{R}) - \Phi_l(\mathbf{R}^*)\right\|_1, \tag{4}$$

where $\Phi_l$ denotes the $l$-layer in the VGG-19 network and $\delta_l$ is used to balance the different terms. We selecte the 'conv1_2', 'conv2_2', 'conv3_2', and 'conv4_2' for layers in Equation (4).

We adopt the classical pixel-wise $\mathcal{L}_1$ loss to increase the robustness of the final estimated results. Combining the above terms, the loss functions for the reflection separation network are as follows:

$$\mathcal{L}_{SN} = \lambda_1 \mathcal{L}_{\mathbf{adv}}(\mathbf{R}, \mathbf{R}^*) + \lambda_2 \mathcal{L}_1(\mathbf{R}, \mathbf{R}^*) + \lambda_3 \mathcal{L}_{\mathbf{feat}}(\mathbf{R}, \mathbf{R}^*), \tag{5}$$

where $\lambda_1 = 0.001$, $\lambda_2 = 1$, and $\lambda_3 = 0.3$ are the weights to balance different terms.

An example of the separated reflection component image is shown in Figure 4.

### 4.2. Enhancement network

As discussed in Section 1 and the examples shown in Figure 4, due to the interference from $f(\cdot)$, the separated reflection component image $f(\mathbf{R}_s)$ obtained in Equation (2) is still with residue edges and dark appearance. To extract $\mathbf{R}_s$ from $f(\mathbf{R}_s)$, an alternative solution is to adopt exposure correction methods, due to the similar appearance between the separated reflection component image and the underexposed image [11, 21, 27]. However, the cause of dark appearance in the underexposed image is mainly due to the low environmental illumination and noise in those images is mainly generated during the camera post-processing process. This is different from our problem, where the dark appearance is due to the low reflectance of glass and the artifacts are mainly generated during the light transmission process. Moreover, the majority of previous exposure correction methods [3, 21] are optimized on the aligned information, which is not applicable to our problems with the non-aligned information. From the results shown in Figure 4, though the results generated by the exposure correction methods become brighter, they are still with obvious distortions, which affect human perception.

To effectively enhance the visibility of the scenes in the separated reflection image, we adopt the U-Net as the backbone to build the enhancement network as follow:

$$\mathbf{R}_s = \mathcal{G}_E(\mathbf{R}), \tag{6}$$

where $\mathbf{S}$ denotes the estimated reflection scene images and $\mathcal{G}_E$ denotes the enhancement network.

The convolutional layers in U-Net only computes the local image features, where the features like the average intensity affecting the overall visibility are not considered. To solve this problem, as shown in Figure 2, instead of predicting the results purely based on the local image features, we

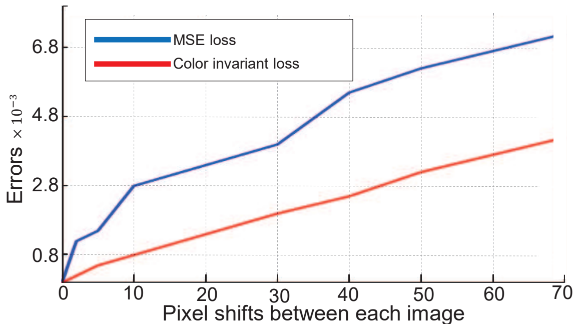

Figure 5. Comparison between the pixel-wise loss and color invariant loss as a function of the magnitude of pixel shifts.

apply a multi-scale fusion scheme for the features from the encoder as follows:

$$\mathcal{F} = SE(\mathbf{F}_g + \mathbf{F}_l). \tag{7}$$

From Equation (7), the fusion scheme splits the intermediate layers between the decoder and encoder into a local path and a global path. The local path contains two convolutional layers to extract the semantic features based on the features from the decoder. The global path contains two convolutional layers and three full-connected layers. The global features obtained by the full-connected layers contain the features related to the average intensity of the entire image [6]. Then, the features are fused and fed into a squeeze-and-excitation (SE) block [7]. Instead of concatenating $\mathbf{F}_g$ and $\mathbf{F}_l$, the two features are added together, where the global feature acts as a prior to regularize the features estimated by the local path [6]. SE block computes normalized weights for each channel and selectively emphasizes the features by multiplying the weights learned by the SE block [7].

**Shift-invariant loss.** Since the misalignment is unavoidable during the data capture, we need to design a loss function applicable to image pairs with moderate misalignment. Loss functions like CXLoss [15] are proposed recently to facilitate network optimization with the misaligned information. However, CXLoss [15] is computationally expensive [22] and simply relying on the information from the feature domain also cannot achieve the best results [29] due to the new distortions in other spaces (*e.g.*, color). We propose a shift-invariant loss by considering the consistency of the color space and the content space.

An effective way to preserve the color consistency is to directly adopt the pixel-wise loss functions in the RGB domain. However, due to the misalignment between the estimated image and the ground truth, the pixel-wise loss functions may introduce new artifacts. To solve this problem, previous methods [8, 9] measure the pixel-wise difference in the Euclidean space by first applying Gaussian Blur for the estimated image and its ground truth. The Gaussian Blur removes high-frequencies and makes color comparison easier [8] especially for small distortions ($\leq$ 2 pixels). However, for our problems where the pixel shifts range from 20 pixels to 70 pixels, this strategy is also not applicable since the blurring operation cannot effectively attenuate the influence from the pixel shifts. Since we focus on the color space similarity, instead of relying on the measurements in the Euclidean space, we measure the color difference based on the cosine distance as follows:

$$\mathcal{L}_{color} = 1 - \mathbf{mean}\left(\frac{z_b \cdot z_b^*}{\max(\|z_b\|_2 \cdot \|z_b^*\|_2, \epsilon)}\right), \tag{8}$$

where $z_b$ denotes the blurred estimated image, $z_b^*$ denotes the blurred ground truth, $\epsilon$ is a small value to avoid division by zero, and $\mathbf{mean}(\cdot)$ denotes the mean operation. As shown in Figure 5, by measuring the angle between two vectors, the color invariant loss in Equation (8) is less insensitive to the pixel shifts than the $\mathcal{L}_1$ and $\mathcal{L}_2$ based loss functions.

For the consistency in the content space, we first introduce the local invariant loss based on the Markov Random Field (MRF) loss [13]. Instead of measuring the pixel-wise feature differences, the local invariant loss measures the patch-based differences. For a patch from the estimations, it finds its similar patch from the corresponding ground truth as follows:

$$\mathcal{L}_{linv} = \sum_{i=1}^{m} \|\Psi_i(\Phi(z)) - \Psi_{NN(i)}(\Phi(z^*))\|_1, \tag{9}$$

where $m$ is the cardinality of $\Psi(\Phi(\mathbf{x}))$, $\Psi_i(\Phi(z))$ denotes a patch from $\phi(z)$, $\Psi_{NN(i)}(\Phi(z^*))$ denotes the similar patch of $\Psi_i(\Phi(z))$ from $\Phi(z^*)$, and $NN(i)$ denotes the patch differences between the two patches.

In the original definition of the MRF loss [13], the similar patch is found from the entire feature map and $NN(i)$ is measured by the feature differences between the two patches. However, since the pixel shifts in our image triplets only range from 20 to 70 pixels, the similar patch search from the whole images become less meaningful and also occupy extra computational resources. To accelerate the computational speed and increase patch matching accuracy, for each patch $\Psi_i(\Phi(\mathbf{x}))$, instead of only relying on the cosine similarity between the image features, we integrate the spatial pixel coordinates into the original formulation as follow:

$$NN(i) = \text{argmax}_{j=1,\ldots,m_s}(\mathbb{D}_f(p_i, q_j) + \omega_s \mathbb{D}(p_i, q_j)), \tag{10}$$

where $\mathbb{D}(p_i, q_j)$ denotes the spatial distances between $p_i$ centered in position $i$ and $q_j$ centered in position $j$ and $\mathbb{D}_f(p_i, q_j) = \langle \frac{p_i}{\|p_i\|}, \frac{q_j}{\|q_j\|} \rangle$ denotes the cross-correlation [26] between features $p_i$ and $q_j$, where the matching process can be efficiently executed by an additional convolutional layer [13, 26].

For the global content consistency, we directly use the differences between the image features from the VGG19 network described in Equation (4) as the global invariant loss. Since the deeper feature may be more insensitive to the misalignment [22], we simply adopt the 'conv5_2' layer as the features for the global invariant loss.

By combining the loss functions in Equation (8), Equation (9), and the global invariant loss, the shift-invariant loss is formulated as follows:

$$\mathcal{L}_{sinv} = \omega_g \mathcal{L}_{ginv} + \omega_l \mathcal{L}_{linv} + \omega_c \mathcal{L}_{color}, \qquad (11)$$

where $\mathcal{L}_{ginv}$ denotes the global invariant loss; $\omega_g = 0.3$, $\omega_l = 0.4$, and $\omega_c = 0.3$ are the weighting coefficients to balance the influence of the three terms.

By incorporating the adversarial loss, the loss functions for the enhancement network is as follows:

$$\mathcal{L}_{EN} = \mathcal{L}_{sinv} + \omega_a \mathcal{L}_{adv}, \qquad (12)$$

$\omega_a = 1$ is the weighting coefficients.

### 4.3. Implementation and training details

We have implemented our model using PyTorch and TensorFlow. The whole training process of our network can be divided into two stages. In the first stage, we first train the separation network to convergence. In the second stage, the separation network is fixed and connected with the enhancement network. We then train the whole network to convergence. The learning rate for the separation network and whole network training is all set to $1 \times 10^{-4}$.

## 5. Experiments

Due to the lack of directly related methods, we evaluate the performances of the separation network and enhancement network separately. We first compare with three reflection removal methods (CoRRN [20], Zhang *et al.* [30], and Yang *et al.* [24]) to evaluate the performances of our separation network. The three methods all estimate the reflection component as a byproduct and we train them on our dataset. Then, for the enhancement network, we choose several image translation and exposure correction methods as the baseline for comparisons: DPE [4], CycleGAN [31], and EnlightenGAN [11]. Since these methods are not designed for our problem, for the fair comparison, we use the reflection component image estimated by our separation network as the input of the baseline methods. EnlightenGAN [11] and CycleGAN [31] are all trained on our dataset due to their unsupervised or weakly supervised learning strategy. For DPE [4], we directly run their released codes and pre-trained model on our evaluation dataset.

We adopt the recently proposed learned perceptual metric LPIPS [28] as the error metric. It measures perceptual image similarity using a pre-trained deep network. The lower LPIPS values indicate better performances. Without loss of generality, we also choose the SSIM and PSNR metrics to do the comparisons for references. Though there is moderate misalignment in the input-output image pairs, this misalignment exists across all methods and thus the comparisons are fair [29].

Table 1. Quantitative evaluations for the reflection component separation using three different error metrics, and compared with CoRRN [20], Zhang *et al.* [30], and Yang *et al.* [24].

| | LPIPS | SSIM | PSNR |
|---|---|---|---|
| Ours | **0.194** | **0.942** | **35.085** |
| CoRRN [20] | 0.491 | 0.801 | 21.718 |
| Zhang *et al.* [30] | 0.606 | 0.854 | 25.625 |
| Yang *et al.* [24] | 0.502 | 0.422 | 7.2894 |

Table 2. Quantitative evaluations for the reflection scene enhancement using three different error metrics, and compared with CycleGAN [31], EnlightenGAN [11], and DPE [4].

| | LPIPS | SSIM | PSNR |
|---|---|---|---|
| Ours | **0.151** | **0.716** | **16.886** |
| CycleGAN [31] | 0.202 | 0.583 | 15.581 |
| EnlightenGAN [11] | 0.315 | 0.595 | 12.956 |
| DPE [4] | 0.331 | 0.479 | 10.998 |

### 5.1. Experimental results

**Quantitative evaluations for reflection separation** Though we leverage advantages from previous reflection removal methods, the results in Table 1 show that our concise setup in this stage achieves better results than other reflection removal methods, which estimates the reflection components as the byproduct. CoRRN [20] assumes the strict linear additive relationship between the reflection and background. Since the relationship may not exist, it may introduce the residue background edges to the estimated reflection component images. Though Zhang *et al.*'s method [30] well estimates the background components, the estimated reflection components are always with residue edges from the background, which undermine its overall performances.

**Quantitative evaluations for the scene enhancement** From the results shown in Table 2, our method achieves the best values among all other methods, which demonstrates that the proposed method better enhances the visibility of the reflection scenes under the same separated reflection components. Due to the existence of the misalignment, the SSIM and PSNR values are not as good as the values usually observed in other image restoration tasks (*e.g.*, image denoising and deraining). However, our method also achieves better performances than other compared methods. Moreover, the higher LPIPS values demonstrate that the proposed method provides better visual pleasing results.

**Qualitative evaluations for the scene enhancement** The qualitative comparisons are shown in Figure 6. Our proposed method not only enhances the visibility of the reflection scenes than other methods but also preserves some de-

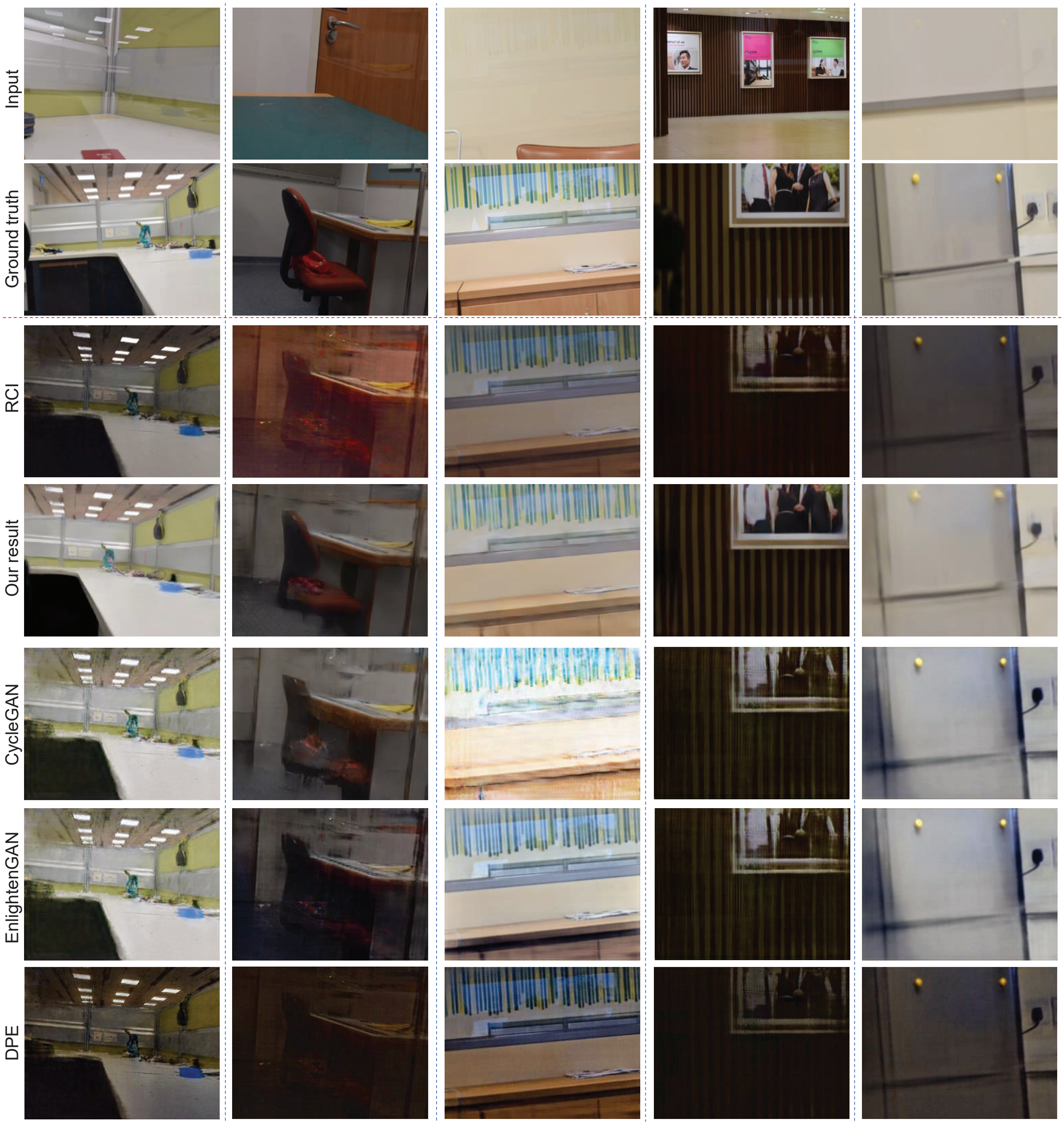

Figure 6. Examples of enhancement results on the evaluation dataset. From top to down, the input image, the ground truth for the reflection scene, the reflection component image (RCI) estimated by our separation network, the estimated reflection scene images of our method, CycleGAN [31], EnlightenGAN [11], and DPE [4]. The RCIs are multiplied by 5.

tails of the reflection scenes. Though CycleGAN also makes the estimated images brighter than the source images, it cannot handle the residue edges existed in the separated reflection images (*e.g.*, the first and fourth column of Figure 6). For the exposure correction methods, if the reflection component image is with fewer artifacts, they can achieve acceptable results (*e.g.*, the results of EnlightenGAN [11] and DPE [4] in the third column of Figure 6). As we discussed before, since the cause for the dark appearance is different between the under-exposed image and the reflection component image, the exposure correction generally fails on our examples, where the separated reflection images are with

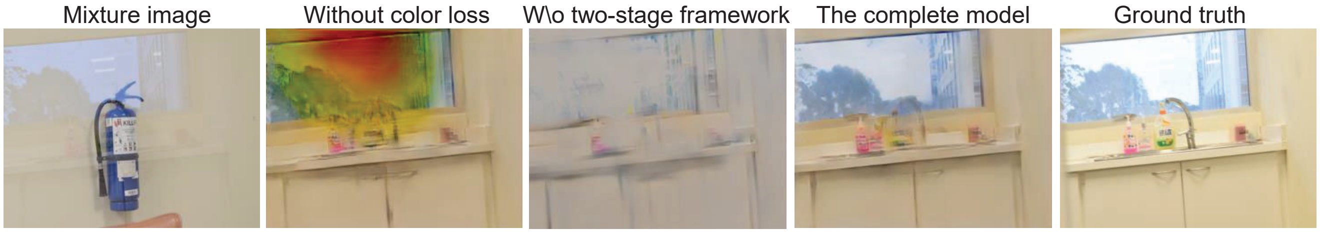

Figure 7. Examples of the result without the two-stage framework, the result without the color loss, the result of our complete model, and the corresponding ground truth reflection scene image.

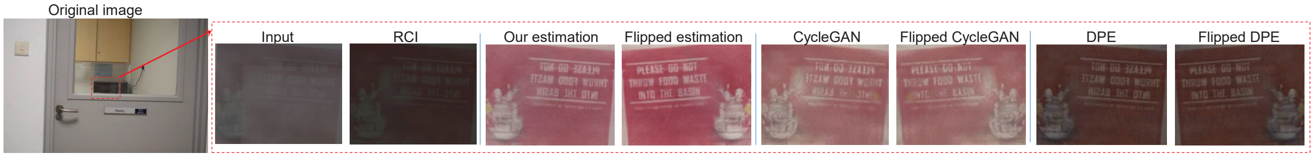

Figure 8. One example taken by a camera in front of glass in the real scenario. We show the estimation of our method, CycleGAN [31], DPE [4], and their flipped version. The sentence in the right example: PLEASE DO NOT THROW FOOD WASTE INTO THE BASIN.

Table 3. Quantitative evaluation results for the model without the proposed complete model, the model without the two-stage framework and the model without the color loss.

| | LPIPS | SSIM | PSNR |
|---|---|---|---|
| Ours | **0.151** | **0.716** | **16.886** |
| W/o two-stage framework | 0.401 | 0.584 | 12.856 |
| W/o color loss | 0.198 | 0.675 | 14.875 |

various artifacts. For example, the image estimated by EnlightenGAN [11] in the first and fourth column still contains obvious artifacts. The performances of DPE [4] are not as good as others.

### 5.2. Ablation study

Our network consists of two parts: the separation network and the enhancement network. In this section, we conduct several experiments to evaluate the contributions of the two parts and the effectiveness of the shift-invariant loss.

We first remove the separation network and feed the enhancement network with the mixture image directly. From the results shown in Figure 7, the enhancement network alone cannot effectively recover the reflection scenes from the mixture image. The quantitative values in Figure 7 also prove the effectiveness of the two-stage framework.

We then remove the color loss in the shift-invariant loss. The results in Figure 7 show that the final estimated images without the color loss are with color distortions. Since the results without the color invariant loss may introduce more artifacts, its error metric values in Table 3 are not as good as the model without the local invariant loss.

### 5.3. Application scenarios

Besides images shown in Figure 6 taken by the standard setup, we show one example in Figure 8 causally taken by a camera under the wild and uncontrolled scenarios. With the appropriate training data and loss functions, our method generates results with more clear and visible details. Our method successfully recovers the sentences from the separated reflection components. The information in the other two results is still far from being visible. However, due to the complicated environments in the real world, the estimations in the right part still look blurred.

## 6. Conclusions

We propose a method to solve the reflection scene separation problem, which aims at recovering clear and meaningful information of the reflection scenes from a single mixture image. We first propose a strategy to obtain the dataset for this unique task. Taking the misaligned data as the input, we further introduce a shift-invariant loss to increase the robustness. Our experiments show promising results for this problem. The results in Figure 8 show that our method could be further used for criminal investigations and surveillance purposes.

**Limitation.** Though our method achieves acceptable results, it may still fail in the separation stage, especially when the reflection and background are correlated or with similar properties. Besides, for some images with too weak reflections, we leverage some advantages from image inpainting techniques to optimize the network with training images from specific scenarios, which may undermine the generalization ability of our method for some "unseen" images. At last, since our evaluation dataset is only for the proof-of-concept purpose, its limited size makes it difficult to judge the availability of our method in more general scenarios. In the future, we will propose a more effective separation network by focusing on the special properties of the reflections. Then, how to improve the generalization ability and increase the dataset diversity will be further investigated.

## References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 4

[2] H. Bach and N. Neuroth. *The properties of optical glass*. Springer Science & Business Media, 2012. 2

[3] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to see in the dark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[4] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4, 6, 7, 8

[5] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf. A generic deep architecture for single image reflection removal and image smoothing. *arXiv preprint arXiv:1708.03474*, 2017. 2

[6] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*. 5

[7] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5

[8] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool. DSLR-quality photos on mobile devices with deep convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision*, 2017. 5

[9] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool. WESPE: weakly supervised photo enhancer for digital cameras. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 5

[10] R. Jenkins and C. Kerr. Identifiable images of bystanders extracted from corneal reflections. *PloS one*, 8(12):e83325, 2013. 1, 3

[11] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang. Enlightengan: Deep light enhancement without paired supervision. *arXiv preprint arXiv:1906.06972*, 2019. 4, 6, 7, 8

[12] A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007. 2, 4

[13] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European conference on computer vision*, 2014. 3

[15] R. Mechrez, I. Talmi, and L. Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of European Conference on Computer Vision*, 2018. 5

[16] K. Nishino, P. N. Belhumeur, and S. K. Nayar. Using eye reflections for face recognition under varying illumination. In *Proceedings of International Conference on Computer Vision*, 2005. 3

[17] K. Nishino and S. K. Nayar. Eyes for relighting. *ACM Transactions on Graphics (TOG)*, 2004. 3

[18] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of International Conference on Computer Vision*, 2017. 3

[19] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot. CRRN: Multi-scale guided concurrent reflection removal network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[20] R. Wan, B. Shi, H. Li, L.-Y. Duan, A.-H. Tan, and A. C. Kot. CoRRN: Cooperative reflection removal network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 2, 4, 6

[21] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[22] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 5, 6

[23] J. Wu and Z. Ji. Seeing the unseen: Locating objects from reflections. In *Proceedings of Annual Conference Towards Autonomous Robotic Systems*, 2018. 1, 3

[24] J. Yang, D. Gong, L. Liu, and Q. Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of European Conference on Computer Vision*, 2018. 6

[25] T. Yano, M. Shimizu, and M. Okutomi. Image restoration and disparity estimation from an uncalibrated multi-layered image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 3

[26] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5

[27] Q. Zhang, G. Yuan, C. Xiao, L. Zhu, and W.-S. Zheng. High-quality exposure correction of underexposed photos. In *Proceedings of ACM Multimedia Conference on Multimedia Conference*, 2018. 4

[28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 6

[29] X. Zhang, Q. Chen, R. Ng, and V. Koltun. Zoom to learn, learn to zoom. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 5, 6

[30] X. Zhang, R. Ng, and Q. Chen. Single image reflection separation with perceptual losses. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 4, 6

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of International conference on computer vision*, 2017. 6, 7, 8