Learning to Jointly Generate and Separate Reflections

Daiqian Ma[1,2], Renjie Wan[3], Boxin Shi[2,4], Alex C. Kot[3], Ling-Yu Duan[1,2,4*]
The SECE of Shenzhen Graduate School, Peking University, Shenzhen, China[1]
The National Engineering Lab for Video Technology, Peking University, Beijing, China[2]
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore[3]
The Peng Cheng Laboratory, Shenzhen, China[4]
{madaiqian, shiboxin, lingyu}@pku.edu.cn, {rjwan,eackot}@ntu.edu.sg

Abstract

Existing learning-based single image reflection removal methods using paired training data have limitations about the generalization capability of dealing with real-world reflections due to the limited variations in training pairs. In this work, we propose to jointly generate and separate reflections within a weakly-supervised learning framework, aiming to model the reflection image formation more comprehensively with abundant unpaired supervision. By imposing the entanglement and disentanglement mechanisms, the proposed framework elegantly integrates two independent stages of reflection generation and separation into a unified model. For better performance, the image gradient constraint is incorporated into the concurrent training process of the multi-task learning as well. In particular, we built up an unpaired reflection dataset with 4,027 images, which is useful for investigating the problem of reflection removal in the weakly supervised learning manner, and further improving model performance. Extensive experiments on a public benchmark dataset show that our framework performs favorably against state-of-the-art methods and consistently produces visually appealing results.

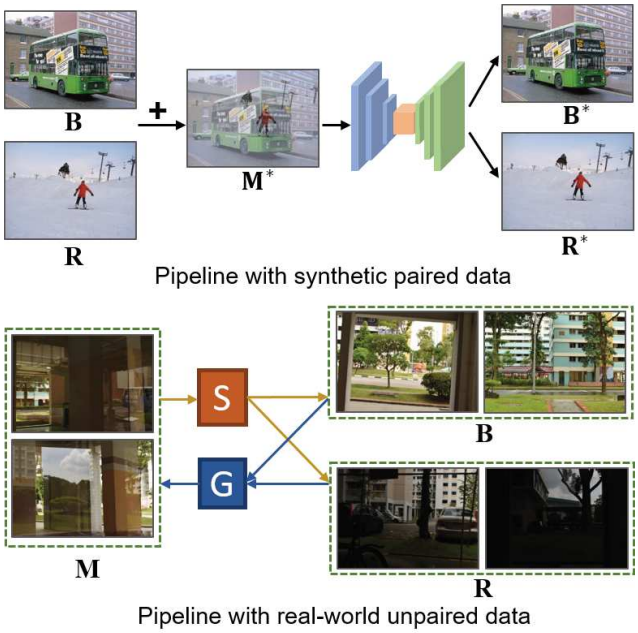

Figure 1. Illustration of two different training pipelines. $\mathbf{S}$ and $\mathbf{G}$ denote the separator and generator in our framework. $\mathbf{M}$, $\mathbf{B}$, and $\mathbf{R}$ represent the mixture image, background and reflection, respectively. * represents the generated images in the training procedure.

1. Introduction

When taking photos through a piece of transparent glass, the presence of reflections accompanied with the background image is undesirable. In addition to the visual quality degradation, the reflections hinder the performance of computer vision systems by obstructing and deforming the background scene behind the glass. The classical representation for image formation with reflections is formulated as,

$$\mathbf{M} = \alpha\mathbf{B} + \beta\mathbf{R}, \tag{1}$$

where $\mathbf{M}$, $\mathbf{B}$, and $\mathbf{R}$ represent observed mixture images with reflections, background, and reflection images, respectively. Here, α and β are the mixing coefficients [5, 27, 26]. Reflection removal aims at removing the reflections $\mathbf{R}$ from $\mathbf{M}$, such that the visibility of the background scenes $\mathbf{B}$ is enhanced. In this scenario, image priors such as different blur levels between the background and reflection [26, 17], ghosting effects [22], and the non-local similarity in the images [25], have been explored. However, these low-level image priors are constrained by limited phenomena causing reflections, which may often be impractical in real-world applications. Moreover, these methods mainly rely on the

*Ling-Yu Duan is the corresponding author.

Figure 2. Examples of our collected unpaired reflection removal dataset (under various scenes and illumination conditions). The mixture images ($\mathbf{M}$), background images ($\mathbf{B}$), and reflection images ($\mathbf{R}$) are shown from left to right.

linear additive formulation in Equation 1 to model the relationship between the mixture image, background, and reflections, which may not well reflect the real interactions.

In practical scenarios, the appearance of real-world reflection is quite complicated, as it is influenced by the interactions of various factors and much beyond the straightforward linear combination. For example, either non-uniform lighting conditions [12] or the non-flat surface of glass [27] may make Equation 1 invalid. As such, a general image formation with reflections is given by,

$$\mathbf{M} = G(\mathbf{B}, \mathbf{R}), \tag{2}$$

where $G(\cdot, \cdot)$ is the mapping function to generate a mixture image. It's not trivial to learn this function accurately.

Recently, deep learning based reflection removal methods [24, 5] with better generalization ability have been proposed to address the limitations arising from the handcrafted image priors. However, most existing methods work in a supervised manner, which requires abundant paired training data, *i.e.*, in the form of a triplet of $\{\mathbf{M}, \mathbf{B}, \mathbf{R}\}$ containing perfectly registered images from the same scene. The recently proposed benchmark dataset [23] is an example. Due to the high cost in capturing the real-world paired data, synthetic mixture images are often applied in accordance with the traditional representation in Equation 1, as shown in Figure 1(a). Clearly, such a strategy ignores various factors in real world image formation process. In particular, the phases of image generation and separation are dealt with as two independent stages, which would degenerate the performance of models by improperly handling the mutual effects of two phases in training models.

In contrast with previous methods [24, 13, 5], that heavily rely on the simplified model in Equation 1 and regard the image generation and separation as two independent stages, the proposed model leverages the mutual benefits of the image generation and separation in a joint learning manner to improve the robustness. It is worthy to note that traditional cycle-consistent network, like CycleGAN [30], cannot be directly applied to reflection removal, as its original setup of one-to-one mapping problem is less comprehensive for modeling the process of reflection generation. Accordingly, we propose to incorporate the entanglement and disentanglement mapping mechanisms between the mixture images and the associated background as well as reflection, which may contribute to more realistic generation results and clearer separation results. Moreover, we introduce the gradient constraints [5, 26] to make the model learning more effective, in which the edge map estimation is elegantly dealt with as an auxiliary task via a multi-task learning structure. We summarize the main contributions as follows:

- We propose to model the real world reflection image formation within a weakly supervised learning framework. Through jointly learning the process of generating and separating reflections, we have achieved encouraging reflection removal performance by leveraging abundant but lower cost unpaired supervision.
- We propose to incorporate the mechanism of entanglement and disentanglement to generate more natural mixture images, and separate clearer backgrounds and reflections, respectively. This also results in a more flexible framework to accommodate auxiliary tasks to further improve the robustness of learning models.
- We have collected a moderate scale reflection image dataset comprising 4,027 unpaired images, which is expected to facilitate the research of removing reflection in a weakly supervised learning manner.

2. Related Work

Reflection Removal. Reflection removal has been widely studied for more than decades. Previous works can be

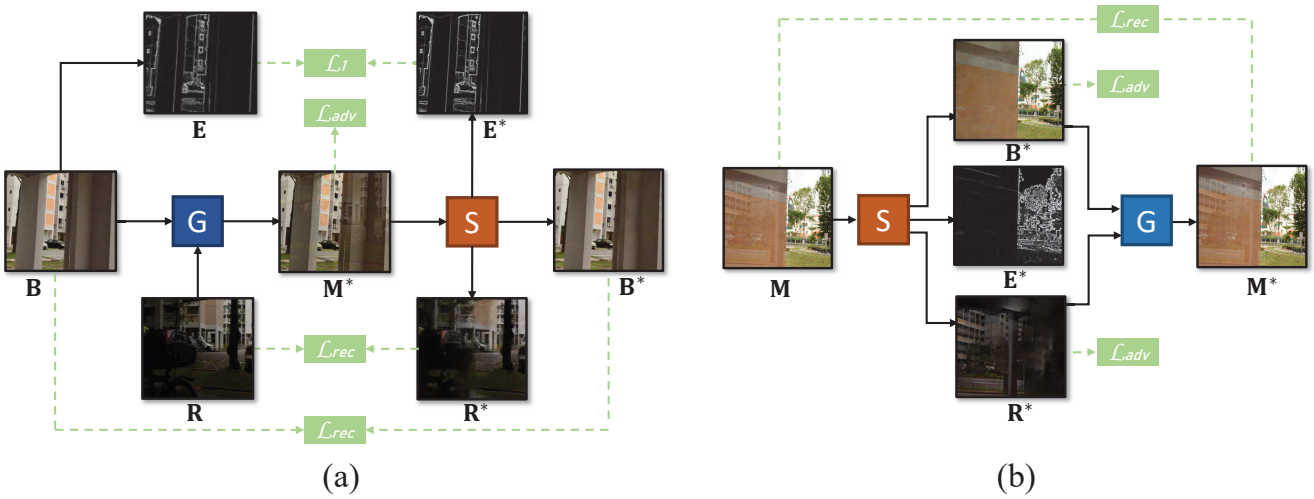

Figure 3. Our framework contains two mapping functions $G : (\mathbf{B}, \mathbf{R}) \rightarrow \mathbf{M}$ and $S : \mathbf{M} \rightarrow (\mathbf{B}, \mathbf{R}, \mathbf{E})$, where $\mathbf{M}$, $\mathbf{B}$, and $\mathbf{R}$ represent the real-world mixture image, background, and reflection, respectively. * represents the generated images in the training procedure. We introduce three reconstruction losses with two cycles (a) and (b). The reconstruction loss for cycle (a) is formulated as $S(G(\mathbf{B}, \mathbf{R})) \approx (\mathbf{B}, \mathbf{R})$, and the reconstruction loss for the mixture image from cycle (b) is formulated as $G(S(\mathbf{M})) \approx \mathbf{M}$. $\mathbf{E}$ (the ground truth edge map is calculated by Sobel operator) is an auxiliary edge map estimation task with $\mathcal{L}_1$ loss and the generation of the intermediate images in these two cycles are guided by the adversarial loss $\mathcal{L}_{adv}$.

classified into two categories. The first category addresses this problem based on hand-crafted priors without learning. Due to the ill-posed nature, different priors have been employed to exploit the properties of the background and reflection layers, including the sparsity prior [15, 14], the blur level differences between the background and reflection layer [17], the ghosting effects [22] and the Laplacian data fidelity term [1]. Other methods in this area remove reflections by virtue of multiple images. By exploiting different image correlation cues [2, 6], the modelling based methods using the multiple images show more reliable results. However, the requirements for specific capturing conditions hinder such methods for practical use, especially for mobile devices or images downloaded from the Internet.

Another category attempts to address this problem in a data-driven learning manner. The comprehensive modeling ability of deep learning has benefited the reflection removal problems and shown very promising results. For example, Chandramouli *et al.* [4] proposed a two-stage deep learning approach to learn the edge features of the reflections with the light field camera. The framework introduced in [5] exploited the edge information when training the whole network to better preserve the image details. Though the deep learning based methods can better capture the image properties, the conventional two-stage framework ignores the intrinsic correlations, which largely limits their performances.

Generative Adversarial Networks (GAN). GAN [7] has become one of the most successful approaches for image-to-image translation problems. In GANs, two networks are adversarially trained simultaneously, where the discriminator is updated to distinguish the real samples from the output of the generator, and the generator is updated to generate fake data to fool the discriminator. For instance, pix2pix GAN [9] learns the translation task in a supervised manner using cGANs[20]. To alleviate the problem of obtaining data pairs, unpaired image-to-image translation frameworks [30, 10, 18] have been proposed. UNIT [18] combines variational autoencoders (VAEs) [11] with CoGAN [19], a GAN framework where two generators share weights to learn the joint distribution of images in cross domains.

It is worthy to mention that some existing mature frameworks like CycleGAN [30] and DiscoGAN [10] are limited in handling reflection removal problem. They are only capable of learning the relationship between two different domains at a time, in which key attributes between the input and the translated images are preserved by utilizing a cycle consistency loss. Undoubtedly, the background is untransferable to the mixture image without the reflection. Unlike the aforementioned approaches, our specifically designed framework for reflection removal attempts to learn the mapping functions amongst three domains, including reflection, background and the mixture.

3. Unpaired Reflection Removal Dataset

In principle, the traditional triplet of $\{\mathbf{M}, \mathbf{B}, \mathbf{R}\}$ (mixture image, background image and reflection image) can be captured in a "remove-and-occlude" manner [23, 28]: 1) Taking a photo of the mixture image through the glass; 2) capturing an image of the background scenes by removing the glass; and 3) capturing a reflection image by putting a

black sheet of the paper behind the glass. However, the perfectly registered triplet sets with "remove-and-occlude" approach is quite time-consuming, which provides limited scalability when much more ground truth data is required by model training. Thanks to the capability of the proposed weakly supervised training framework, paired pixel-wise correspondence is not required when collecting image dataset. So we capture 4000+ images, which allows for a much larger scale than those used in existing methods [23, 28]. Finally, we build a dataset containing 4027 images under various scenes, and example images are shown in Figure 2.

The proposed dataset enriches the diversity and generality over existing datasets in the following aspects:

- **Devices.** Besides using the high-end devices (*e.g.*, the DSLR camera with fully manual control model) like previous methods [23, 28], we also use the cameras on different types of mobile phones (iphone 8, iphone X, *etc.*) to capture images.
- **Illuminations.** We capture images under different illumination conditions. More specifically, we capture reflection images in both cloudy and sunny days, at different time of the day (*e.g.*, morning, afternoon, and night) and indoor scenes with different lighting conditions (*e.g.*, office, living room, *etc.*).
- **Scene.** Our images cover a variety of scenes, *e.g.*, the campus, streets, parks, gardens.

4. Proposed Method

In this section, we first discuss the motivation and the network architecture, following which the loss functions are introduced. Finally, the training strategy is presented.

4.1. Framework of the Proposed Scheme

In contrast to the conventional pipelines [5, 29, 24] that treat the image generation and separation as two independent stages, we come up with a unified model, such that the mutual effects between two stages can benefit the robustness. As shown in Figure 3, our model contains a generator network to generate the mixture images, a separator network to separate the mixture image into background and reflection, and three discriminator networks to produce the adversarial losses. Existing method [5] can be treated as a special instance of our method when the generator is simplified as a linear function.

As shown in Figure 3, our framework involves two cycles of the generator and separator. In particular, each cycle serves as a two-step conversion to convert the generated image back to the original image. There are three reconstruction losses in these two cycles, aiming to incorporate the cycle-consistent constraints to guide the training procedure. Moreover, in contrast with the classical cycle-consistent

Figure 4. Examples of mixture images with different generation methods. The references are the real-world mixture images with similar reflection properties. Please note the similarity of our generated reflections to the captured reference images in the yellow boxes.

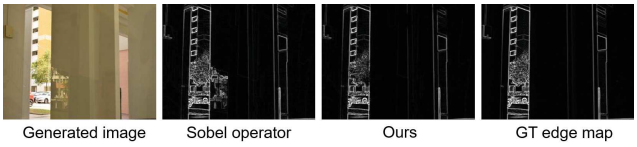

Figure 5. The estimated edge map compared with the ground truth (GT) edge map.

model with the one-to-one framework [30, 10], we propose a joint mapping mechanism based on the additive relationship between the mixture image, the background, and the reflection, such that the whole procedure can be modelled in a better way. The details of our proposed framework are described as follows.

Generator (G). We propose to design an entanglement mechanism in the generator with a general formulation to derive more natural mixture image from the background and reflection.

In general image-to-image translation tasks, generators are mainly designed for the one-to-one mapping conversion. Due to the very nature of the reflection removal problem that the mixture image is a composite of the background and reflection, the traditional one-to-one mapping cannot be directly used to translate the background into the mixture

images due to the lack of reflection. Instead of the one-to-one mapping in previous methods, our generator learn the mapping as $G : (\mathbf{B}, \mathbf{R}) \rightarrow \mathbf{M}$, where the non-linear mappings can produces more realistic reflection appearances (first to third columns in Figure 4 [1]) than previous linear functions [5, 26, 17, 1] with fixed coefficients.

Separator (S). We perform a disentanglement in the separator for the mixture images by leveraging multi-task learning to estimate the background, reflection, and the background edge map (E) concurrently. Instead of one-to-one framework in previous methods [5, 29], our separator learns the mapping function as $S : \mathbf{M} \rightarrow (\mathbf{B}, \mathbf{R}, \mathbf{E})$, where the multi-task learning framework models the image separation process in a more reasonable way, especially the auxiliary task of edge map estimation, that provides useful information to make the separator more efficient.

As shown in Figure 5, compared with the edge map extracted with Sobel operator, our proposed separator successfully removes the gradient information from the reflection and retains the edge map related with the background.

Network architecture. The generator and separator exhibit similar structures: a downsampling unit with two convolutional layers to increase the receptive field size, a feature extraction unit with 9 residual blocks [8] to extract robust features and an upsampling unit at the last stage with two transposed convolutional layers. More specifically, the generator contains two downsampling units to receive the inputs of background and reflection, and the multi-task learning mechanism with three upsampling units (relative to three tasks) is employed in the separator to improve the reflection removal ability. For the discriminator networks, we use 70×70 PatchGANs [9, 16] which can be applied to arbitrarily-sized images in a fully convolutional fashion.

4.2. Loss Functions

The learning of these two mappings are guided by the adversarial losses and reconstruction losses, with training samples $\mathbf{B} = \{b_i\}_{i=1}^{K}$ for the background, $\mathbf{R} = \{r_i\}_{i=1}^{N}$ for the reflection and $\mathbf{M} = \{m_i\}_{i=1}^{L}$ for the mixture.

Adversarial loss. Adversarial loss has been widely used in the image-to-image translation problems. Here, regarding the mapping function $G : (\mathbf{B}, \mathbf{R}) \rightarrow \mathbf{M}$ and its discriminator $D_{\mathbf{M}}$, the objective is given by,

$$\begin{aligned}\mathcal{L}_{adv}^{m} =& E_{m\sim p_{data}(m)}[\log D_{\mathbf{M}}(m)] + \\ & E_{b,r\sim p_{data}(b,r)}[\log(1 - D_{\mathbf{M}}(G(b,r)))],\end{aligned} \tag{3}$$

where G tries to generate images $G(b, r)$ conditioned on both the background and reflection images, while $D_{\mathbf{M}}$ aims to distinguish between the generated fake mixture image $G(b, r)$ and real-world mixture image m. In other words, G aims to minimize this objective against an adversary D that tries to maximize it. Then we introduce two similar adversarial losses for the mapping function $S : \mathbf{M} \rightarrow (\mathbf{B}, \mathbf{R}, \mathbf{E})$ and their discriminators $D_{\mathbf{B}}$ and $D_{\mathbf{R}}$: $\mathcal{L}_{adv}^{b}$ and $\mathcal{L}_{adv}^{r}$.

Reconstruction loss. We employ a reconstruction loss on the pixel and content domain to better preserve both the pixel level and feature level similarity.

Though the minimization of the adversarial loss in Equation 3 can generate images with similar distributions towards the target domain, it does not guarantee that generated images preserve the content of its input images with only the regions covered by reflections changed. To alleviate this problem, we first adopt the pixel reconstruction loss for the generated mixture image $\mathcal{L}_{prec}^{m}$ to preserve the consistency in the pixel domain as follows:

$$\mathcal{L}_{prec}^{m} = E_{m\sim p_{data}(m)}[||m - G(S(m))||_1], \tag{4}$$

where G takes the separated b and r from $S(m)$ and tries to reconstruct the original image m. The pixel reconstruction losses for the background and reflection ($\mathcal{L}_{prec}^{b}$ and $\mathcal{L}_{prec}^{r}$) are with similar scheme but in an inverted order of G and S.

On the other hand, the content reconstruction loss aims to constrain the whole procedure in a high-level feature space. The content reconstruction loss for the generated mixture image $\mathcal{L}_{crec}^{m}$ is defined as:

$$\mathcal{L}_{crec}^{m} = E_{m\sim p_{data}(m)}[\frac{1}{WH}||\phi(m) - \phi(G(S(m)))||_2], \tag{5}$$

where ϕ is the feature map from the $relu4\_3$ layer of a pre-trained VGG-16 network, W and H indicate the dimensions of the $relu4\_3$ layer. The content reconstruction losses for the background and reflection ($\mathcal{L}_{crec}^{b}$ and $\mathcal{L}_{crec}^{r}$) are defined in a similar fashion but in an inverted order of G and S.

Full objective. Finally, the objective functions to optimize G and S are written as:

$$\begin{aligned}\mathcal{L}_G =& \mathcal{L}_{adv}^{m} + \lambda_p \mathcal{L}_{prec}^{m} + \lambda_c \mathcal{L}_{crec}^{m}, \\ \mathcal{L}_S =& \mathcal{L}_{adv}^{b} + \lambda_p \mathcal{L}_{prec}^{b} + \lambda_c \mathcal{L}_{crec}^{b} \\ & + \lambda_r(\mathcal{L}_{adv}^{r} + \lambda_p \mathcal{L}_{prec}^{r} + \lambda_c \mathcal{L}_{crec}^{r}) + \lambda_e \mathcal{L}_e,\end{aligned} \tag{6}$$

where $\mathcal{L}_e$ is the $\mathcal{L}_1$ loss for the edge map estimation, λ_r and λ_e aim to balance the main and auxiliary tasks in the separator; and λ_p and λ_c aim to balance the pixel reconstruction loss and content reconstruction loss.

As such, we aim to solve:

$$G^*, S^* = \arg\min_{G,S} \max_{D_{\mathbf{M}}, D_{\mathbf{R}}, D_{\mathbf{B}}} \mathcal{L}_G + \mathcal{L}_S. \tag{7}$$

In the experiments, λ_p, λ_c, λ_r and λ_e are empirically set as 10, 2, 0.5 and 0.5.

[1] More examples are listed in the supplementary material.

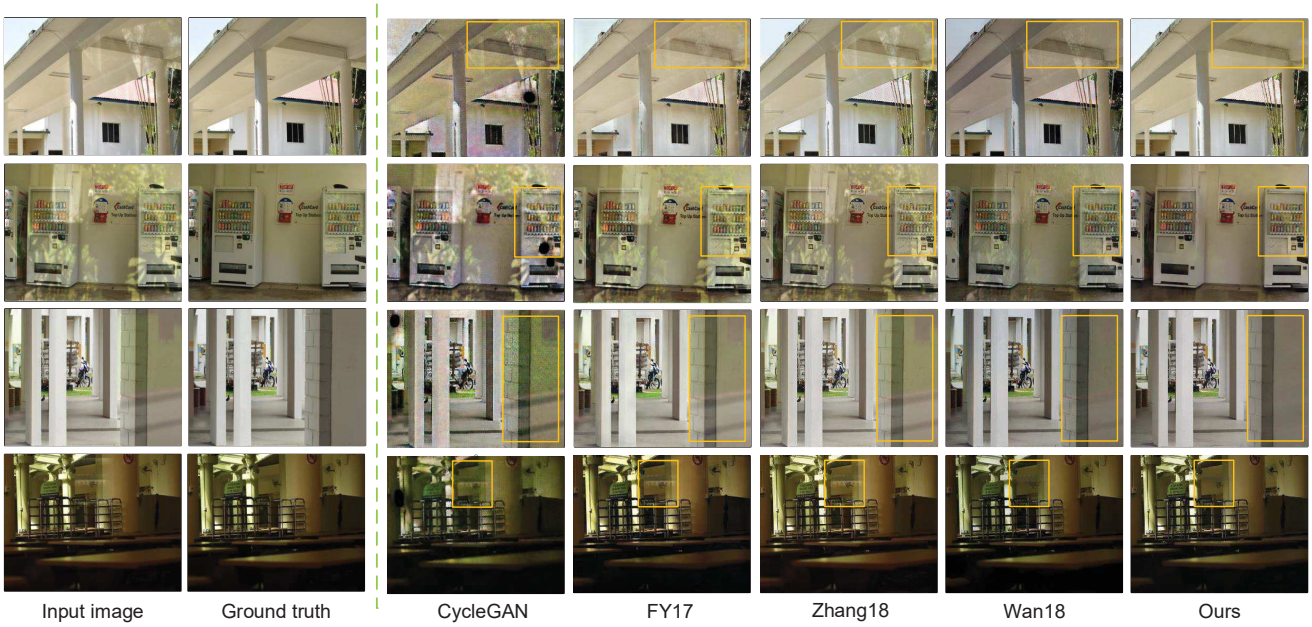

Figure 6. Examples of the reflection removal results on four wild scenes in the SIR^2 dataset. The comparison methods include Wan18 [24], Zhang18 [29], CycleGAN [30], and FY17 [5]. The yellow boxes highlight some noticeable differences.

Table 1. Quantitative evaluation results on SIR^2 with the state-of-the-arts methods using three different error metrics.

	SSIM$_r$	SSIM	PSNR(dB)
LB14 [17]	0.801	0.829	21.77
WS16 [26]	0.833	0.877	22.39
NR17 [1]	0.832	0.882	23.70
FY17 [5]	0.820	0.871	22.51
CycleGAN [30]	0.794	0.813	20.10
Zhang18 [29]	0.842	0.885	24.01
Wan18 [24]	0.854	0.891	24.08
Eq. (1)	0.833	0.880	24.06
Ours	0.858	0.892	24.32
Ours + Eq. (1)	**0.870**	**0.903**	**24.48**

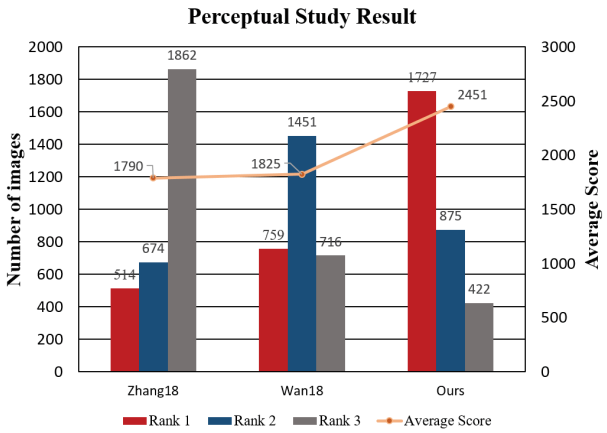

Figure 7. Perceptual study results on the whole SIR^2 dataset for the three best methods (Zhang18 [29], Wan18 [24], and ours) in terms of the quantitative scores in Table 1. The stastistics are obtained by collecting the ranking results from 30 participants and 100 images.

4.3. Training Strategy

The model is implemented with PyTorch[2]. To inject scale-invariance to the network [21], we adopt a multi-size training strategy by feeding images of two sizes: coarse scale 336×252 and fine scale 224×168. The learning rate is set to 2×10^{-4} for the first 100 epochs and we linearly decay it to 0 over the next 100 epochs. We also augment the training data with three different operations: image translation, flipping and cropping. The sizes of mini-batch and momentum are set to 4 and 0.9, respectively.

[2]http://pytorch.org

5. Experiments

To verify the effectiveness of our proposed method, we perform several experiments on the SIR^2 [23] benchmark dataset with state-of-the-art reflection removal methods. All results are evaluated in terms of both quantitative scores and visual quality. Due to the regional properties of the reflection [23], we also adopt SSIMr [21] to assess the quality by focusing on local reflections.

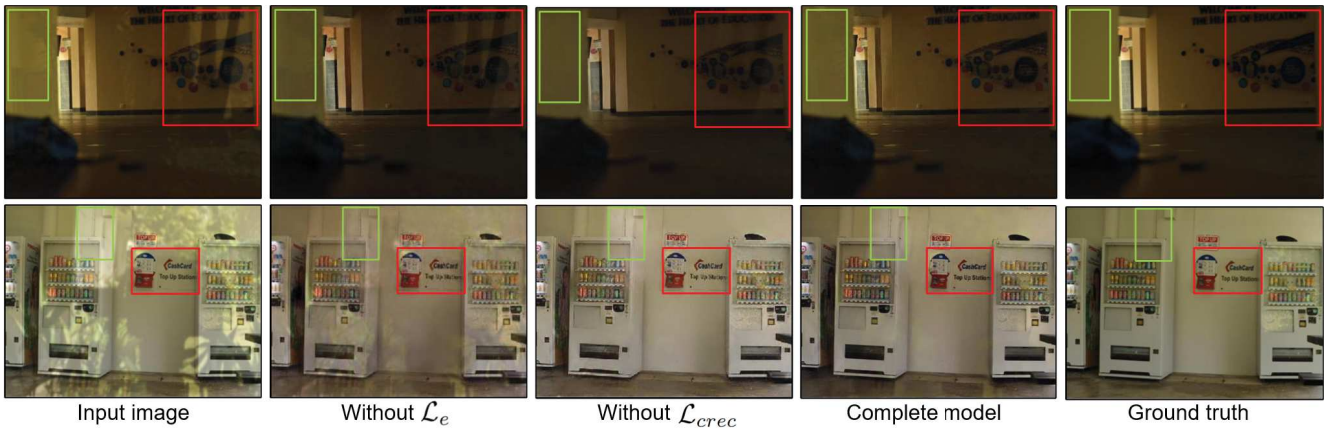

Figure 8. Visual quality comparisons on the ablation study of edge map loss and content reconstruction loss based on the SIR^2 dataset.

Table 2. Numerical comparisons regarding the ablation study of the edge map loss and content reconstruction loss based on the SIR^2 dataset.

	SSIM$_r$	SSIM	PSNR(dB)
w/o $\mathcal{L}_e$	0.826	0.873	23.52
w/o $\mathcal{L}_{prec}$	0.845	0.886	24.31
Complete model	0.858	0.892	24.32

5.1. Comparison with State-of-the-art Methods

The proposed method is compared with seven state-of-the-art single image reflection removal methods, including Wan18 [24], Zhang18 [29], CycleGAN [30], FY17 [5], NR17 [1], WS16 [26], and LB14 [17]. For a fair comparison, we use the codes provided by their authors and set the parameters as suggested in their original papers, and we follow the same training protocol to retrain their networks using our dataset.

Quantitative Comparison. The comparisons with seven state-of-the-art methods are performed with three different error metrics. The results are summarized in Table 1, where the numbers displayed are the mean values over all 100 sets of wild images in the SIR^2 [23] dataset. In particular, Ours + Eq. (1) means that we set a random variable and use the data with probability 0.7 from our generator and probability 0.3 from the Equation 1 to train the separator. Though the left three columns in Figure 4 show that our generator better preserves the backgrounds while highlighting the reflection part, the performance of the generator may drop due to the limited number of our training dataset (see the fourth column in Figure 4). Thus, to increase the stability, we propose to incorporate Equation 1 into the design of our whole framework. As shown in Table 1, our proposed model obviously outperforms other methods in terms of both PSNR and SSIM. The higher objective quality values indicate that our method recovers the background images with better fidelity. Note that almost all images in the SIR^2 dataset are partially reflected images, such that the global changes are small between the recovered background and the original mixture images. To deal with the limitations of global error metrics, we manually label the reflection dominant regions and evaluate the SSIM values in these regions analogously to the evaluation method proposed in [21]. As a result, higher SSIM$_r$ results have been obtained as shown in Table 1, indicating that the proposed method can remove strong reflections more effectively in the regions overlaid with reflections than the state-of-the-art methods.

Note that our framework is inspired by the fact that the mixture images are complicated combinations of reflection and background images in a generative process, and our target is to explicitly model this mechanism in a weakly supervised manner. As shown in Figure 6 and Table 1, CycleGAN shows poor performance in the reflection removal task, because it is rather difficult for CycleGAN to learn the mapping functions between the reflection-contaminated images and reflection-free images directly.

Reflection-Removal Perceptual Study. Recent research [3] pointed out that PSNR and SSIM may not exactly tell the perceptual visual quality. Since there is no suitable error metric specifically developed for the reflection removal task, we conduct a reflection-removal perceptual study and invite 30 subjects to evaluate the quality of 100 images from the SIR^2 dataset. In particular, we focus on the top three methods reported in Table 1 for this perceptual study with the following procedures:

- The participants are well trained with the common reflection images to gain a general sense on this task.
- Each participant is requested to view four images at a time, with the leftmost image showing the input reflection-contaminated image followed by three reflection-removed images generated by differ-

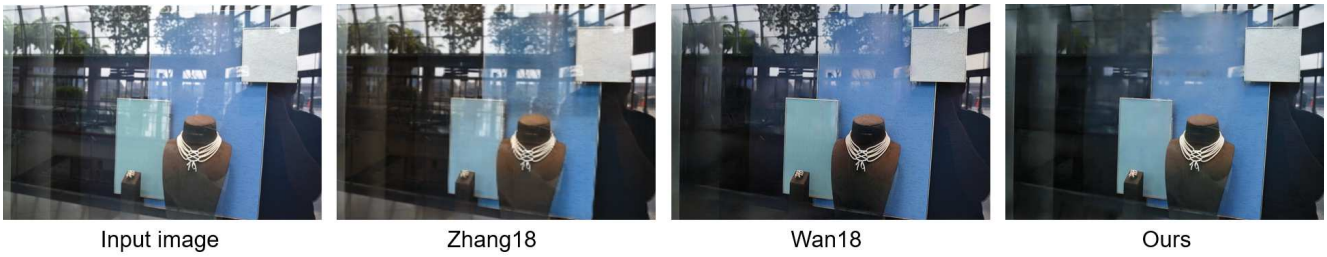

Figure 9. Illustration of an extreme example with the real-world vitrine, compared with Zhang18 [29] and Wan18 [24].

Table 3. Efficiency comparisons with FY17 [5], Zhang18 [29] and Wan18 [24] of an image with size 224×288 on a single Titan XP GPU.

	Framework	Time (s)
FY17 [5]	Torch	0.0705
Zhang18 [29]	Tensorflow	0.0438
Wan18 [24]	PyTorch	0.3488
Ours [29]	PyTorch	0.0214

ent methods displayed in a random order. They rank the reflection removal quality without any time constraint. This test is performed for 100 groups.

- The average score ϕ for a method is calculated from the ranking as $\phi_k = \frac{1}{N}\sum_i \sum_j (N - rank_{i,j,k} + 1)$, where N is the total number of evaluated methods and i, j, k indicate the i-th participant, j-th group of images and k-th method, respectively.

The results in Figure 7 show that the rank-1 number of our method is even higher than the sum of the rest two methods and the rank-3 number of our method is obviously smaller, which demonstrates the superior perceptual quality of our method. Moreover, from the result in Figure 6, we can find that our method removes the reflections more effectively and recovers the details of the background images more clearly. It should be noted that in the third row, our method is able to remove the reflection on the right vending machine, which is even clearer than the ground truth.

5.2. Loss Ablation

Besides the basic cycle consistency with pixel construction loss, we further apply the content reconstruction loss and edge map loss to improve the performance. To analyze how these two loss functions contribute to the final performance, we remove the relative loss in the final objective function and re-train the network. The results are shown in Figure 8 and Table 2. Without the edge map loss, we notice that visible content of the reflection image appears in the background prediction. Moreover, the content reconstruction loss helps to recover cleaner and more natural results (the characters shown in second row). These results demonstrate the necessity in introducing these loss functions.

5.3. Efficiency Analysis

To evaluate the efficiency, we record the average execution time of an image with size 224×288 on a single Titan XP GPU, though these methods are implemented on different deep learning frameworks. The results are shown in Table 3. In particular, the SSIM-guided loss proposed by Wan18 [24] performs well while our method is much more efficient (15 times faster) and achieves higher PSNR value.

6. Conclusions and Discussions

In this paper, we propose a novel approach to jointly generate and separate reflections. Based on the public dataset SIR^2 [23] and the proposed real-world dataset, our method outperforms state-of-the-art methods in terms of both the quantitative and subjective quality.

There remain several open issues for the future work. In some extreme cases like Figure 9, the whole image can be dominated by the reflection, our method cannot remove the reflection completely and the estimated background still remains with some visible residual edges. However, even in this challenging case, our method still removes the majority of reflections and restores the background details, which performs better than other state-of-the-art methods. Moreover, when testing the models across datasets with different collecting protocols (*e.g.*, the dataset of SIR^2 [23] and the dataset of Zhang18 [29]), we have observed that the dataset gap problem is worth further investigating to achieve consistently good performance on diverse real-world scenes. Meanwhile, the proposed framework can be further extended in various ways to facilitate other image restoration tasks (*e.g.*, derain, dehaze, deshadow, *etc.*), which leaves more space for future exploration as well.

Acknowledgement This work was supported by the National Natural Science Foundation of China under Grant 61661146005, Grant U1611461, and 61872012, in part by the Shenzhen Municipal Science and Technology Program under Grant JCYJ20170818141146428, and in part by the National Research Foundation, Prime Minister's Office, Singapore, through the NRF-NSFC Grant, under Grant NRF2016NRF-NSFC001-098. Renjie Wan is supported by the Microsoft Cloud Research Software Fellowships (CRSF) program.

References

[1] Nikolaos Arvanitopoulos, Radhakrishna Achanta, and Sabine Süsstrunk. Single image reflection suppression. In *Proc. CVPR*, 2017.

[2] Erfrat Be'Ery and Arie. Yeredor. Blind separation of superimposed shifted images using parameterized joint diagonalization. *IEEE Transactions on Image Processing*, 17(3):340–353, 2008.

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proc. CVPR*, pages 6228–6237, 2018.

[4] Paramanand Chandramouli, Mehdi Noroozi, and Paolo Favaro. Convnet-based depth estimation, reflection separation and deblurring of plenoptic images. In *Proc. ACCV*, 2016.

[5] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proc. ICCV*, 2017.

[6] Kun Gai, Zhenwei Shi, and Changshui Zhang. Blind separation of superimposed moving images using image statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):19–32, 2012.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[10] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[12] N. Kong, Y. Tai, and J. S. Shin. A physically-based approach to reflection separation: from physical modeling to constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[13] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. Generative single image reflection separation. *arXiv preprint arXiv:1801.04102*, 2018.

[14] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. In *Proc. ECCV*, 2004.

[15] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9), 2007.

[16] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Proc. ECCV*, 2016.

[17] Yu Li and Michael S. Brown. Single image layer separation using relative smoothness. In *Proc. CVPR*, 2014.

[18] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[19] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016.

[20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[21] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Ryson W.H Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proc. CVPR*, 2017.

[22] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T. Freeman. Reflection removal using ghosting cues. In *Proc. CVPR*, 2015.

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

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

[25] Renjie Wan, Boxin Shi, Ah.H Tan, and Alex C. Kot. Sparsity based reflection removal using external patch search. In *Proc. ICME*, 2017.

[26] Renjie Wan, Boxin Shi, Ah Hwee Tan, and Alex C. Kot. Depth of field guided reflection removal. In *Proc. ICIP*, 2016.

[27] Patrick Wieschollek, Orazio Gallo, Jinwei Gu, and Jan Kautz. Separating reflection and transmission images in the wild. In *Proc. ECCV*, 2018.

[28] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T. Freeman. A computational approach for obstruction-free photography. *ACM Transactions on Graphics (TOG)*, 2015.

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

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, 2017.