



## Article

# Transmission Removal from a Single Glass Scene and Its Application in Photographer Identification

Zhen Li <sup>1</sup>, Heng Yao <sup>1,\*</sup> , Ran Shi <sup>2</sup>, Tong Qiao <sup>3</sup>  and Chuan Qin <sup>1</sup>

<sup>1</sup> School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

<sup>2</sup> School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

<sup>3</sup> School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, China

\* Correspondence: hyao@usst.edu.cn

**Abstract:** In daily life, when taking photos of scenes containing glass, the images of the dominant transmission layer and the weak reflection layer are often blended, which are difficult to be uncoupled. Meanwhile, because the reflection layer contains sufficient important information about the surrounding scene and the photographer, the problem of recovering the weak reflection layer from the mixture image is of importance in surveillance investigations. However, most of the current studies mainly focus on extracting the transmission layer while often ignoring the merit of the reflection layer. To fill that gap, in this paper, we propose a network framework that aims to accomplish two tasks: (1) for general scenes, we attempt to recover reflection layer images that are as close as possible to the ground truth ones, and (2) for scenes containing portraits, we recover the basic contour information of the reflection layer while improving the defects of dim portraits in the reflection layer. Through analyzing the performance exhibited by different levels of feature maps, we present the first transmission removal network based on an image-to-image translation architecture incorporating residual structures. The quality of generated reflection layer images is improved via tailored content and style constraints. We also use the patch generative adversarial network to increase the discriminator's ability to perceive the reflection components in the generated images. Meanwhile, the related information such as edge and color distribution of transmission layer in the mixture image is used to assist the overall reflection layer recovery. In the large-scale experiments, our proposed model outperforms reflection removal-based SOTAs by more than 5.356 dB in PSNR, 0.116 in SSIM, and 0.057 in LPIPS.

**Keywords:** glass reflection; transmission removal; portrait identification; PatchGAN; perceptual loss



**Citation:** Li, Z.; Yao, H.; Shi, R.; Qiao, T.; Qin, C. Transmission Removal from a Single Glass Scene and Its Application in Photographer Identification. *Appl. Sci.* **2022**, *12*, 12484. <https://doi.org/10.3390/app122312484>

Academic Editor: Zhengjun Liu

Received: 28 October 2022

Accepted: 3 December 2022

Published: 6 December 2022

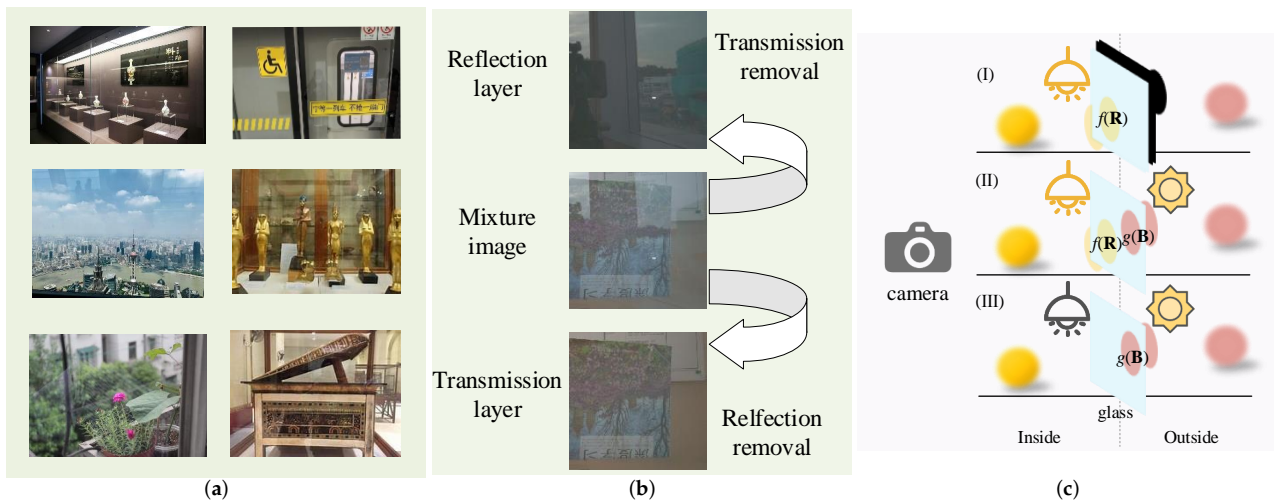
**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

An essential objective of computer vision is to extract information within an image to enhance the perception of the world while satisfying the visual effect [1–6]. According to the effect of the glass on the light, there is a significant difference in characteristic information in the obtained images that are captured through a glass [7]. In natural scenes, the mixture image with reflection usually contains two kinds of information: the dominant transmission layer and the non-dominant reflection layer. Most existing methods treat the reflection layer as noise, and their main goal is to recover the transmission layer from the mixture image. However, the reflection layer is as important as the transmission layer in many practical applications. The glass reflection information can be used to assist in scene and photographer identification. For instance, in some subway stations, museums, or jewelry stores with glass scenes, as shown in Figure 1a, the reflection images containing the photographer's facial features are more likely to be critical clues for surveillance investigation. This paper focuses on recovering a clear reflection image via transmission removal from the mixture image.



**Figure 1.** Image shooting principle under an ambient light environment. (a) Real-world glass scenes with reflections; (b) imaging principles; (c) reflection removal versus transmission removal.

Considering the influence of glass and combined with light analysis in the mixture scene, a mixture image can be regarded as a superposition of the transmission layer and reflection layer as

$$\mathbf{I} = g(\mathbf{B}) + f(\mathbf{R}) = g(\mathbf{B}_C, \mathbf{B}_E, \dots) + f(\mathbf{R}), \quad (1)$$

where  $\mathbf{I}$  represents a mixture image;  $g(\cdot)$  and  $f(\cdot)$  represent the changes of the object’s light before it reaches the sensor, and are caused by many complicated factors and difficult to be expressed in a specific mathematical formula;  $\mathbf{B}$  represents the light from the actual background image behind the glass, which contains the components such as the color distribution  $\mathbf{B}_C$  and the edge information  $\mathbf{B}_E$ , and  $\mathbf{R}$  represents the light from the reflection scene image. Figure 1b shows the imaging principle under an ambient light environment, where according to the object layer for reconstruction, i.e.,  $\mathbf{B}$  or  $\mathbf{R}$ , the enhancement problem of the mixture image with glass scene can be divided into two aspects: reflection removal and transmission removal, as shown in Figure 1c. On the one hand, due to the high transmittance rate of glass, in most occasions,  $g(\mathbf{B})$  is remarkably close to  $\mathbf{B}$ , which makes the transmission component dominate the mixture image  $\mathbf{I}$ . In our method, we no longer distinguish between  $g(\mathbf{B})$  and  $\mathbf{B}$ , i.e.,  $g(\mathbf{B}) = g(\mathbf{B}_C, \mathbf{B}_E, \dots) \approx \mathbf{B} = (\mathbf{B}_C, \mathbf{B}_E, \dots)$ . On the other hand, most of the reflection light  $\mathbf{R}$  from the reflection scenes will pass through the glass, and only a tiny amount will be reflected by the glass and captured by the camera, which leads to  $f(\mathbf{R}) < \mathbf{R}$ . In general,  $f(\mathbf{R})$  is much less than  $g(\mathbf{B})$ .

Since the extraction of  $f(\mathbf{R})$  is a regenerative process, we need to focus on generating results that express more detailed information, including content and style, for further analysis. In general, the content retains more distinct levels of the feature map than the style. At the end of the generator network, we add a super-resolution (SR) module to optimize and reconstruct high-quality details through a neural network via sampling from the traditional method.

Incorporating the above-mentioned motivations, in this paper, we propose a unified framework based on the generative adversarial network (GAN) [8] to remove the dominant transmission component from a mixture image via a comprehensive constraint. The transmission removal problem can be cast into the framework of mixture image minus reflection removal image; however, the dedicated network designed to recover the reflection image as much as possible is still worth further study. In order to have a better recovery effect than the results generated by the reflection removal-oriented approach, the analysis of different levels of feature maps and the selection of appropriate feature maps from different dimensions, including resolution, texture, and style, are involved in meeting this requirement. In addition, we also apply our framework to the application of photographer identification. The contributions of our work can be summarized as follows:

- We propose a transmission removal network to recover the reflection layer from the mixture image. In order to recover the weaker components from the mixture image, the edge and color distribution of transmission layer is introduced as a prior to guide the reflection layer separation. In a practical application scenario, we can replace the real transmission layer with a predicted one through reflection removal.
- By analyzing the performance exhibited at different levels of the feature maps, we improve the generated results in different dimensions such as resolution, texture, and style to enhance the visual perception of the estimated reflection layer. We also use PatchGAN as the discriminator to enhance the perceptual strength of the generated image and ground truth.
- We consider the portrait's appearance in the reflection layer as a separate problem and improve the style loss to solve the problem of the dim portrait in the reflection layer. Thus, the reflection layer image can be effectively applied to photographer identification application.

The remainder of this paper is organized as follows. Section 2 introduces the related works about reflection removal and transmission removal. Sections 3 and 4 describe the proposed method and analyze the experimental results, respectively, and Section 5 concludes this paper.

## 2. Related Works

As far as we know, the current research on images containing glass reflections mainly includes reflection removal, which is mainly used to recover the transmission layer image, and transmission removal, which is mainly used to separate the reflection layer image. Although the reflection removal problem has become a hot research topic in computer vision [9–19], the transmission removal problem is only just coming into focus [20]. In the following, the above two types of methods are reviewed separately.

### 2.1. Reflection Removal

Reflection removal treats reflection as a type of noise and aims at approximating the light emitted by the background scenes. The traditional methods [14,21–23] set the best segmentation boundary as the target to seek for the dominant transmission layer and the non-dominant reflection layer through the gradient, color, and other fields of pixel-level separation. Recent methods [10–13,16,17] mainly adopt the deep learning (DL) framework to solve this problem. Specifically, all the DL-based methods can be further divided into the following two categories:

(1) Multi-stage methods: obtaining the most relevant predicted information from an image, and then combining it with other leading information as the prior, thus forming a multi-stage network. For example, Fan et al. [19] proposed a two-stage cascaded network for edge prediction and image reconstruction. To be specific, they designed a network to estimate the transmission edge first and then used this predicted edge to estimate and reconstruct the reflection removal image. Similarly, Wan et al. [11] proposed a concurrent model to predict the edge details of the background. They further proposed a cooperative model [12] to utilize the edge information better. Lei et al. [13] proposed a simple yet effective reflection-free cue for robust reflection removal via subtracting the ambient image from the corresponding flash image in raw data space. Recently, Chang et al. [16] introduced three auxiliary techniques into their architecture. Unlike other reflection removal methods, they used the classifier network to determine whether the mixture image contained reflection.

(2) One-stage methods: reflection removal does not use the complicated network but focuses more on directly adding the prior input or constructing loss functions based on the relationships between layers. For example, Li and Brown [14] indicated that the transmission is smoother than the reflection; thus, they used the difference to achieve reflection removal with the distribution. Zhang et al. [10] proposed a network with the perceptual loss to improve the qualities of the generative image in content, then used an

exclusion loss to separate transmission and reflection from the gradient. Li et al. [17] proposed a cascaded network that iteratively refines the estimates of transmission and reflection layers to boost the prediction quality to each other. The above-proposed methods successfully remove reflection from both synthetic and real-world images.

## 2.2. Transmission Removal

Compared with the transmission restoration after removing the reflection, transmission removal pays more attention to exploring the undetectable features in the mixture image. In the traditional methods, exploring the differences between the two layers becomes a prerequisite for implementing the removal of one layer. For example, the information on the reflection layer is highly blurred compared to the transmission layer. Levin and Weiss [24] used the distribution to find pixel-level boundaries distinguishing the two layers. With the introduction of DL methods in recent years, relatively satisfactory results have been obtained in separating mixture images by learning features actively. However, this often requires complete prior knowledge as a guide. Based on the difference between the two layers in the gradient domain, most experiments pay more attention to using a gradient. Zhang et al. [10] first used the gradients of two layers simultaneously and used edge information as a prior. Similarly, Wan et al. [11,12] and Chang et al. [16] used a predicted network to take the edge information as a prior. Li et al. [17] proposed a cascaded network that the two layers can be mutually prior to improve the quality of prediction. Lei et al. [13] used the characteristic that the reflection layer only exists in the mixture image, then used the reflection-free flash-only cues as prior knowledge to guide the separation.

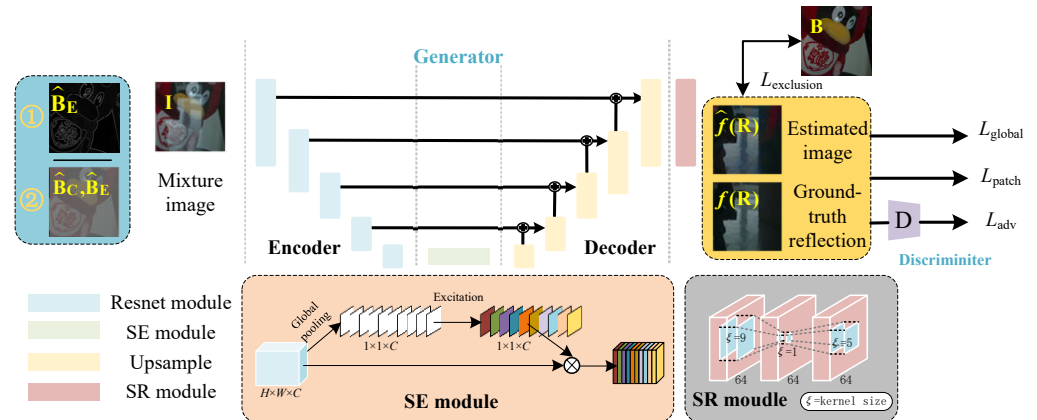
Compared with the estimation of the reflection layer as a byproduct in the above methods, Wan et al. [20] first raised the problem of reflection separation strictly. Specifically, their whole framework comprises two networks: the separation and enhancement networks. The separation network follows the framework of literature [25], and the main contribution of [20] focuses on improving the visual effect of the separated images by adding a shift-invariant loss and increasing the reflection image brightness via the corresponding ambient light environment image. However, the ground truth information should be retained as much as possible for the estimated reflection layer while meeting the visual requirements. In addition, we also need to consider the scenarios in which the corresponding ambient light environment image is difficult to obtain. In addition to the method [20], the most related one is proposed by Yano et al. [26]. However, the transmission in the mixture image explored by this method is not the dominant image, which is not in line with the actual situation and is not suitable for our discussion. Moreover, some methods [27,28] extracted facial features of bystanders from corneal reflection and put forward the matching theory, which makes it have application prospect in the criminal investigation. Wu et al. [29] also proposed another method to locate the objects from the reflections. Nishino et al. [30,31] estimated lighting from corneal reflections for the relighting applications.

In general, in addition to [20], transmission removal has not received widespread attention; there is still room for further research on obtaining more accurate reflection layer images. At the same time, targeted network design is needed for different application scenarios, such as photographer identification.

## 3. Proposed Method

This paper proposes a transmission removal method based on a GAN architecture with a PatchGAN discriminator and several modules for restoring the reflection layer. Figure 2 shows the flow diagram of the proposed method. In terms of network architecture design, the known transmission information about **B** and the mixture image **I** are taken as input. Although the actual transmission layer is not available in practice, we can substitute it by some approximate methods, such as the additional flash/ambient light photo approach used in [13], or even directly replace it with the estimated edge information of **B** through the existing reflection removal network. Here, we divide the information we may obtain about **B** into two cases as shown in Figure 2. Our experiment aims to remove **B** from **I** with

the loss functions that can constrain the generated results. First, we extract features from the inputs using Resnet, then analyze the features to reconstruct the predicted reflection layers via the SE module. The corresponding losses are used to constrain the content and style of the images, and finally the generated images are discriminated by a discriminator network. The remainder of this section presents the proposed overall network structure, loss function design, and application to photographer identification.



**Figure 2.** Flow diagram of the proposed method. The inputs contain two cases: (1) the estimated edge  $\hat{B}_E$  and the mixture image  $I$  and (2) the estimated  $\hat{B}_E$ ,  $\hat{B}_C$ , and the mixture image  $I$ .

### 3.1. Input Preprocessing

In a real scene, it is difficult to obtain the groundtruth transmission  $B$  directly. In our method, we need some information about  $B$  as the prior; one way is to use the edge predictor to extract the edge of  $B$  from  $I$ , which is denoted as  $\hat{B}_E$ . Another way is that when we can use some images that are consistent with  $B$  in terms of color distribution and edge information by means of prediction and re-photography. As mentioned in [13], the pure flash image without reflection can eliminate the influence of original data in the color domain after grayscale preprocessing. Therefore, we first perform grayscale preprocessing on the similar transmission layer. Compared to the ground truth of the transmission, it eliminates the interference of the color features. At the same time, because the reflection layer only exists in the mixture image, the grayscale transmission layer retains enough edge information and brightness to guide the separation and prediction of the reflection layer in the mixture image. The grayscaled transmission layer is denoted as  $(\hat{B}_C, \hat{B}_E)$ .

### 3.2. Network Architecture

Overall, we build the model based on the GAN architecture. The related information of transmission layer  $B$  and the mixture image  $I$  are used as input. By referring to the existing encoder–decoder architecture of general image-to-image translation methods [10,11,17,32], our generator also uses the U-Net as the backbone, combined with some image progressing methods, such as SR, image transfer, and image reconstruction. In addition, we use PatchGAN to replace the traditional discriminator network.

The reflection layer generated by the generator network can be formulated as follows:

$$\begin{aligned} \text{Case1 : } \hat{f}(R) &= G(I, \hat{B}_E), \\ \text{Case2 : } \hat{f}(R) &= G(I, \hat{B}_C, \hat{B}_E), \end{aligned} \tag{2}$$

where  $G(\cdot)$  represents the generator network. We intend to retain as much information as possible about reflection in the mixture image. The ordinary generator usually generates some edge information during image-to-image translation. Considering the practical applications, we lack information about the estimated results, so we divide the different priors into two cases, one with only the estimated edge  $\hat{B}_E$  of the transmission  $B$  as a prior



and one with the predicted transmission as a prior, which contains color distribution  $\hat{\mathbf{B}}_C$  and edge information  $\hat{\mathbf{B}}_E$ . Unlike traditional U-net, we use Resnet to replace the previous VGG network in obtaining the feature map. Its residual structure [33] will help us retain global information from higher levels. In addition, we re-select the feature map after feature extraction by adding the squeeze-and-excitation (SE) and SR modules to enhance the expression of the network in the learning progress.

In the encoder process, the convolutional layers in U-Net only compute the local image features, while the features like the average intensity affecting the overall visibility are not considered in [20]. This is very shallow for the information we can extract about the reflection from the mixture image. After extracting the features, we add the SE module to solve this problem. We add attention mechanisms to optimize what we learn and suppress unimportant modules. After the feature map is processed and all features are extracted, the image results are not predicted solely based on the local features.

Meanwhile, because the estimated image is usually blurred after the decoder process, the generated  $\hat{f}(\mathbf{R})$  is blurry when the edge of  $f(\mathbf{R})$  is close to that of  $\mathbf{I}$ . SR is a process of restoring a high-resolution image from a given low-resolution image [34,35], which is a classic application of computer vision. Specifically, the enhancement of the correlation in the image via adding bicubic interpolation, proposed by Dong et al. [36], can be used to improve the resolution of the restored image and improve the visual effect. Following [36], we use the idea of SR and add an SR module with three convolutional layers at the end of the generator. By convolution of different sizes, better compression effects are achieved within the region of the size range, resulting in higher resolution.

### 3.3. Loss Function

The loss function is mainly used for the network constraints the generated results by the generator. First, the adversarial loss is used as the main loss of GAN and assists in the overall control of the generated image  $\hat{f}(\mathbf{R})$ . Next, in our work, we consider reducing the correlation between the generated image of the output  $\hat{f}(\mathbf{R})$  and  $\mathbf{B}$ , and enhancing the correlations between  $\hat{f}(\mathbf{R})$  and  $\mathbf{I}$ , as well as  $\hat{f}(\mathbf{R})$  and  $\mathbf{R}$ . Therefore, the exclusion loss is used to reduce the connection on the gradient. Although the adversarial loss can realize the constraint on the generated results, it lacks the information more relevant to the ground truth. Therefore, for global loss, we use  $L_1$ ,  $L_2$ , and the structural similarity (SI) loss to constrain the consistency. In addition, we regard the estimated results as optimization of image transfer and further constrain the results from the two dimensions of content and style, thus introducing the style loss and content loss, respectively. The specific loss functions are described as follows.

#### 3.3.1. Adversarial Loss

As shown in Figure 3, different from the original GAN, the output of PatchGAN [8] as the discriminator is not a value but a matrix. Each element in the matrix represents the calculation of different receptive fields. Therefore, we can focus on more areas. Here, by using PatchGAN, the adversarial loss  $L_{adv}$  is composed of the loss of generator  $L_G$  and the loss of discriminator  $L_D$ . Note that the input of the generator is a concatenation of  $\mathbf{I}$  and  $\hat{\mathbf{B}}_E$  (or  $\mathbf{I}$ ,  $\hat{\mathbf{B}}_C$ ,  $\hat{\mathbf{B}}_E$ ) in this paper, and for the ease of description, denote this concatenation as  $\mathbf{I}_B$ . The ground truth for discriminator is  $f(\mathbf{R})$  in this paper. Losses  $L_G$  and  $L_D$  can be expressed as follows:

$$L_G(\mathbf{I}_B) = \mathbb{E}_{\mathbf{I}_B} \left[ \log \left( 1 - D(\mathbf{I}_B, \hat{f}(\mathbf{R}), N) \right) \right], \quad (3)$$

$$L_D(\mathbf{I}_B, f(\mathbf{R})) = \mathbb{E}_{\mathbf{I}_B, f(\mathbf{R})} [\log(1 - D(\mathbf{I}_B, f(\mathbf{R}), N))] + \mathbb{E}_{\mathbf{I}_B} \left[ \log \left( D(\mathbf{I}_B, \hat{f}(\mathbf{R}), N) \right) \right], \quad (4)$$

where  $D(x, y, N)$  represents the PatchGAN discriminator used to discriminate between  $x$  and  $y$  with an  $N \times N$  sized matrix output. The parameter  $N$  is set as 30 in our exper-

iments. In discriminator training, we hope  $D(\mathbf{I}_B, f(\mathbf{R}), N)$  is closer to the real label and  $D(\mathbf{I}_B, \hat{f}(\mathbf{R}), N)$  is closer to the fake label, where the real and fake labels are marked as 1 and 0, respectively. In general,  $L_D(\mathbf{I}_B, f(\mathbf{R}))$  as close to the real label as possible, that is,  $L_G(\mathbf{I}_B)$  should be small enough. In the end, the network parameters are continuously updated in the process of continuous optimization of adversarial loss  $L_{adv}$ .

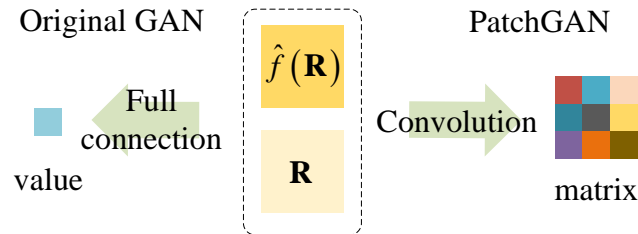


Figure 3. Comparison of original GAN and PatchGAN.

### 3.3.2. Global Loss

The SI Loss is used to measure the structural similarity between the estimated image and ground truth, we globally invoke the SI loss [11] as follows:

$$L_{SI}(f(R), \hat{f}(R)) = 1 - \frac{2\sigma_{f(R)\hat{f}(R)} + c_2}{\sigma_{f(R)}^2 + \sigma_{\hat{f}(R)}^2 + c_2}, \tag{5}$$

where  $c_2$  is a constant, here set to 0.0001. We combine both  $L_1$  and  $L_2$  losses as overall constraints, and the global loss function is combined as:

$$L_{global} = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_{SI}, \tag{6}$$

where  $\lambda_1$  to  $\lambda_3$  are the weights to balance different terms and set as 6, 4, and 3 in our experiments, respectively.

### 3.3.3. Patch-Level Loss

#### Exclusion Loss

The most significant characteristic of a mixture image is that the intensity of the two images on the gradient is not the same. The critical observation in [10] is that the edges of the transmission and the reflection layers are unlikely to overlap. The dominance of the transmission layer in the mixture image means that its edges will also dominate the gradient. Edge information is critical in guiding transmission removal. In the beginning, the matrix evaluates the gradient can be calculated as

$$\Psi_z(\mathbf{B}, \hat{f}(\mathbf{R})) = \tanh\left(\frac{\partial \mathbf{B}}{\partial z}\right) \otimes \tanh\left(\lambda_z \frac{\partial f(\mathbf{R})}{\partial z}\right), \tag{7}$$

$z = H \text{ or } V,$

where  $H$  and  $V$  stand for the horizontal and vertical directions, respectively,  $\lambda_z = \frac{\mathbb{E}|\partial \mathbf{B} / \partial z|}{\mathbb{E}|\partial f(\mathbf{R}) / \partial z|}$ ,  $z = H$  or  $V$ , are two normalization factors in the horizontal and vertical directions, respectively, which are designed to balance  $\mathbf{B}$  and  $\hat{f}(\mathbf{R})$ ,  $\otimes$  denotes the element-wise multiplication operation. Then, we calculate the exclusion loss to distinguish the transmission and the estimated reflection at the pixel level can be represented as

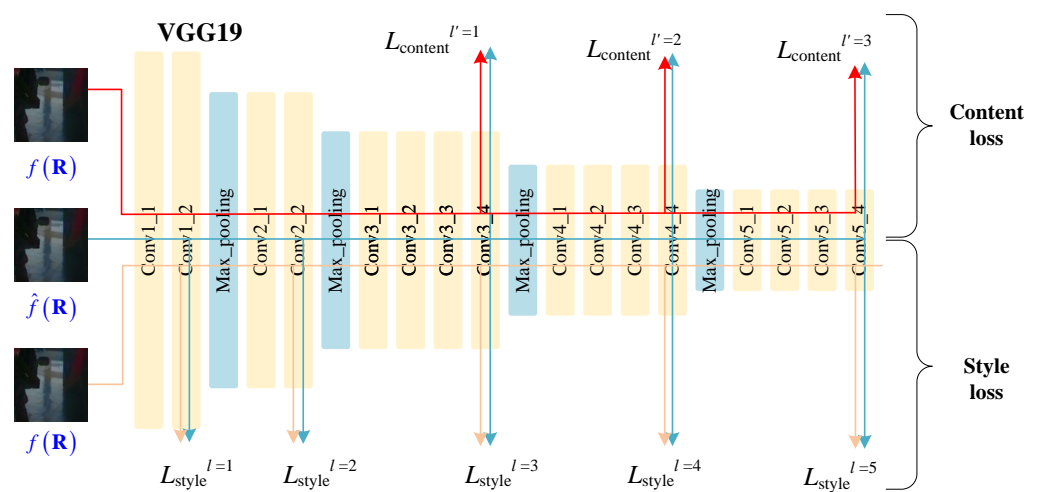
$$L_{excl}(\mathbf{B}, \hat{f}(\mathbf{R})) = \frac{1}{T+1} \sum_z \sum_t \left\| \Psi_z(\mathbf{B}_t, \hat{f}(\mathbf{R})_t) \right\|_F, \tag{8}$$

$t = 0, 1, \dots, T, z = H \text{ or } V,$

where  $\|\mathbf{A}\|_F$  represents the Frobenius norm of matrix  $\mathbf{A}$  and determined by the square root of the sum of the absolute squares of its elements, i.e.,  $\|\mathbf{A}\|_F = \sqrt{\sum a_{m,n}^2}$ , where  $a_{m,n}$  represents the element of  $\mathbf{A}$ ,  $T$  represents the times of image downsampling, which is set as 2 in our experiments, and subscript  $t$  represents the corresponding resized image.

### Style Loss

It is not enough to rely on the adversarial loss of the GAN network to constrain the style of the predicted image; this adversarial loss only grabs the style on the global image. We expect the estimated image to be visually close to the ground truth. In the previous methods [11,12], pixel-level  $L_1$  and  $L_2$  losses were used as the loss function to constrain the global information of the generated image and the ground truth. However, this will result in some local information loss, so the estimated images are often blurry and dim. The style feature maps can be extracted by constructing a Gram matrix [37], which takes advantage of the correlated feature map between the generated image and the target image to match in a manner. As mentioned in [10], there is color attenuation in the generated image. In contrast to restoring content using only pixel-level constraints, restoring style requires both global and local information and should be close to the ground truth. This makes it more challenging to preserve style than content. Some DL-based image style transfer methods can be used to obtain better visual quality in style on the generated results [37–39]. In our method, we use the style loss to consider the relevance of style; thus, the non-aligned color feature can help achieve the task of style transfer while preserving the ground truth. In detail, the style representation of images can be extracted by observing the activation of feature maps in the VGG19 network, as shown in Figure 4, and using the spatial correlation of their values.



**Figure 4.** The framework of using contextual loss. The VGG19 network is used to optimize the content and style of the resulting images.

First, the dimensional reduction is carried out for the features, and the feature map of  $\mathbf{X}$  at layer  $l$  is denoted as  $F^l(\mathbf{X})$  with the size of  $H_l \times W_l \times C_l$ , where  $C_l$  represents the number of channels in the feature map and the size of each channel is  $H_l \times W_l$ . The Gram matrix function  $\text{Gram}(\mathbf{X}, l)$  is defined as a  $C_l \times C_l$  matrix to characterize feature correlations of  $F^l(\mathbf{X})$ , and each element in this matrix, denoted as  $\text{Gram}(\mathbf{X}, l)_{i,j}$ , can be determined as

$$\text{Gram}(\mathbf{X}, l)_{i,j} = \sum_{k=1}^{H_l \times W_l} F_{ik}^l(\mathbf{X}) F_{jk}^l(\mathbf{X}), i, j = 1, 2, \dots, C_l, \tag{9}$$

where  $i$  and  $j$  represents the two channels of the feature map,  $k$  represents the element number in the feature matrix under each channel, and  $F_{ik}^l(\mathbf{X})$  denotes the  $k$ -th element of



the  $i$ -th channel of the feature map  $F^l(\mathbf{X})$ . Then calculate the relevance distance with MSE as follows:

$$E_l(f(\mathbf{R}), \hat{f}(\mathbf{R})) = \frac{\sum_B \|\text{Gram}(\hat{f}(\mathbf{R}), l) - \text{Gram}(f(\mathbf{R}), l)\|_2}{BH_l^2 W_l^2}, \tag{10}$$

where  $B$  represents the batch size reserved in the feature map.  $E_l(f(\mathbf{R}), \hat{f}(\mathbf{R}))$  is calculated by the average scoring distance of each element in the Gram matrix, which minimizes  $\text{Gram}(f(\mathbf{R}), l)$  and the estimated  $\text{Gram}(\hat{f}(\mathbf{R}), l)$ . Compared with the original method [39], we add the concept of batch calculation, so we only focus on the style transfer and speed up the generator’s learning of color domains. The specific style loss is defined as follows:

$$L_{\text{style}}(f(\mathbf{R}), \hat{f}(\mathbf{R})) = \sum_l w_l E_l(f(\mathbf{R}), \hat{f}(\mathbf{R})), l = 1, 2, \dots, 5, \tag{11}$$

where  $w_l$  represents the weight factor in each convolution layer, and we select “Conv1\_2”, “Conv2\_2”, “Conv3\_4”, “Conv4\_4”, and “Conv5\_4” as features in our network, as shown in Figure 4.

### Content Loss

After transmission, the estimated image that only satisfies the visual effect on color is not enough; it should obtain more contextual information about the actual reflection. In the previous processing of content information, perceptual loss [10] is used to constrain the high-level content information and low-level details of the generated image to satisfy the human perception. However, low-level features are enough to express the image’s texture, and too many features waste resources [25]. Therefore, instead of perceptual loss, we add content loss based on the pre-trained model VGG19, which is a complete feature extraction network trained under the ImageNet dataset. We use it to constrain the content, and the content loss function is defined as follows:

$$L_{\text{content}}(f(\mathbf{R}), \hat{f}(\mathbf{R})) = \sum_{l'} w_{l'} \left\| F^{l'}(f(\mathbf{R}), F^{l'}(\hat{f}(\mathbf{R})) \right\|_2, \tag{12}$$

$$l' = 1, 2, 3,$$

where  $w_{l'}$  denotes the weight factor, and we use the middle-frequency features with richer textural information “Conv3\_4”, “Conv4\_4”, and high-frequency feature “Conv5\_4” in our experiments and generate the output through  $L_2$  distance. The patch-level loss function is combined as:

$$L_{\text{patch}} = \lambda'_1 L_{\text{excl}} + \lambda'_2 L_{\text{style}} + \lambda'_3 L_{\text{content}}, \tag{13}$$

where  $\lambda'_1$  to  $\lambda'_3$  are the weights to balance different terms and set as 9, 3, and 2 in our experiments, respectively.

By considering the influence of image gradient, content, and style, the final loss function for the transmission removal network can be expressed as follow:

$$L_{\text{total}} = \lambda''_1 L_{\text{adv}} + \lambda''_2 L_{\text{global}} + \lambda''_3 L_{\text{patch}}, \tag{14}$$

where  $\lambda''_1$  to  $\lambda''_3$  are the weights to balance different terms and set as 1, 3, and 1 in our experiments, respectively.

### 3.4. Application to Photographer Identification

As one of the important applications of transmission layer removal, the photographer identification based on the recovered reflection layer image is of great importance in the investigation of shooting information. Figure 5 shows a schematic diagram of an application scenario for photographer identification, where unknown or suspicious photographers can be identified by the proposed transmission removal network as well as by several existing portrait recognition algorithms.

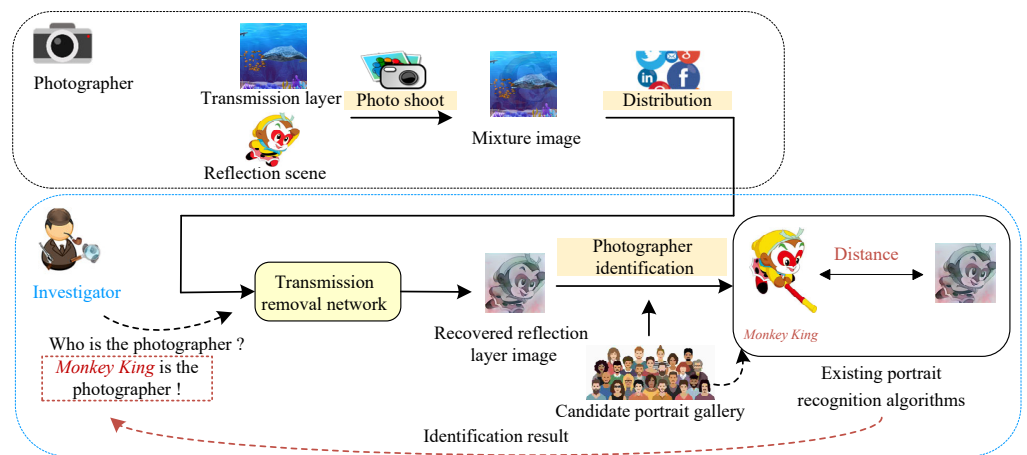


Figure 5. Schematic diagram of an application scenario for photographer identification.

As mentioned in Section 1, we have analyzed that when the face regions are present in a scene, they will provide more helpful information than in an ordinary environment scene. So we can only focus on the face region on the estimated images with a portrait. As mentioned in [20], the reflection component after reflection separation can be enhanced by the corresponding ground truth reflection scene image. Since the reflection layer is filled with brightness and color using the fixed reflection layer’s environment, the final effect was satisfied with the laboratory environment. However, we cannot always obtain the corresponding ground truth reflection scene image. Each image should reveal more facial information for easier analysis of images with portraits. Therefore, we can change the expression of the predicted image directly by changing the style image. Figure 6 shows an example of how to implement image transfer on the estimated image by using the style image. To distinguish transmission removal from the general scenarios, we reconsidered the differences in content and style. In detail, we use grayscale  $f(\mathbf{R})$  as the content image to retain content information and use color images with the corresponding semantic information as the style image to describe the distribution of colors. Note that when  $f(\mathbf{R})$  is originally a grayscale or dim image, we need to colorize it before setting it as a style image. These colorized images will better serve the application of photographer identification.

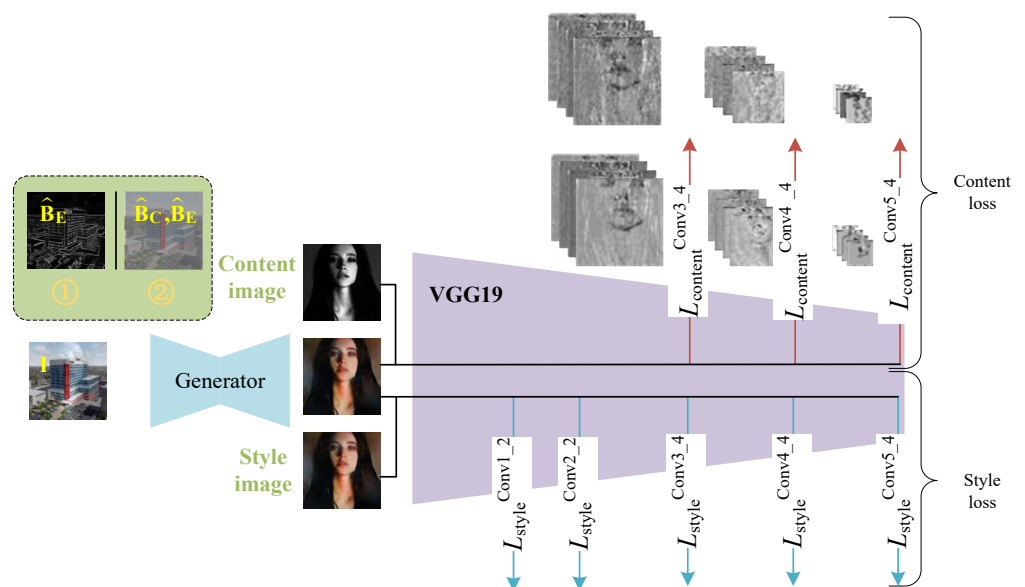


Figure 6. Content image and style image settings in a scene containing a portrait.

## 4. Results

The experiment preparation and results are presented in this section. We mainly demonstrated the superiority of the proposed network with the following two aspects: transmission removal in general scenes and some specific application scenes represented by photographer identification. In addition, the ablation study was also conducted to verify the efficacy of the proposed method further.

### 4.1. Experiment Preparation

#### 4.1.1. Comparative Methods

Since Wan et al. [20] first defined the reflection separation problem in 2020, few reflection-only separation networks have existed. However, many reflection-removal works also output the reflection layer as a byproduct, and we also compared with them. Table 1 lists the baseline comparative methods and their main technical routes.

**Table 1.** The baseline methods to be compared in this paper.

Method	Publication	Main Technical Route	TECR	TRCR
Wan et al. [11]	CVPR2019	Multi-scale edge guidance model	Yes	No
Zhang et al. [10]	CVPR2018	Dilated convolution model and exclusion loss	Yes	No
Yang et al. [15]	ECCV2019	Bi-directional estimation model	Yes	No
Li et al. [14]	CVPR2015	Distribution-based traditional model	Yes	N/A
Lei et al. [13]	CVPR2021	Free reflection cue guidance model	Yes	Yes
Chang et al. [16]	WACV2021	Edge guidance and recurrent decomposition model	Yes	Yes
Li et al. [17]	CVPR2020	Cascade network with LSTM module and reconstruction loss	Yes	Yes
Wan et al. [20]	CVPR2020	Dilated convolution model, U-Net, and shift-invariant loss	No	No

#### 4.1.2. Evaluation Metrics

We used three standard image metrics to analyze the generated images quantitatively. First, the recently proposed DL-based perceptual metric LPIPS [40] is adopted, which learns the reverse mapping of generated images to ground truth and forces the generator to reconstruct the reverse mapping of authentic images from fake images and prioritize the perceived similarity between them. The lower the LPIPS value is, the more similar the two images are, and vice versa. In addition, the SSIM and PSNR are also involved in our comparisons without loss of generality.

#### 4.1.3. Datasets

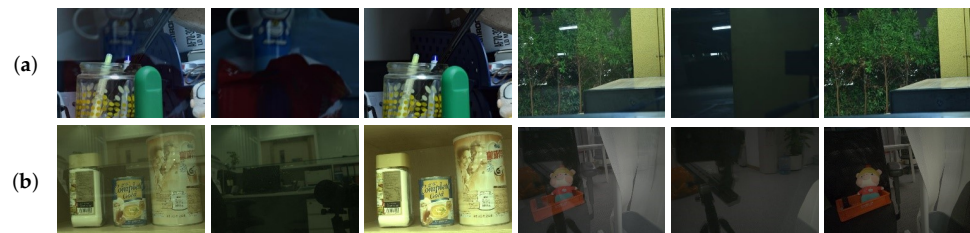
Since DL-based image processing is a data-driven problem, it is necessary to have a complete dataset. According to [41] a complete triplet data set is prepared, including the mixture image, transmission layer image, and reflection layer image. At present, the relevant methods are considering the scenes filmed in the real world and using different synthetic methods to augment the dataset. However, the existing dataset is not fully applicable to our experimental requirements. For example, when there are moving objects such as human faces in the reflection scene, it is difficult to obtain the corresponding pure reflection layers in the real-world scenes; thus, our results are difficult to evaluate intuitively. In addition to using existing data sets, we also supplemented our data set using the realistic synthetic methods as in the previous literature to overcome this difficulty.

#### Existing Dataset

We used two existing datasets in our evaluation.

- (1) The reflection removal method proposed in Wan et al. [41] provides a reasonable way to obtain the ground truth for the reflection component image. In detail, by putting a piece of black cloth behind the glass, only the camera can capture the reflection light reflected by the glass. The SIR<sup>2</sup> dataset (Downloaded from <https://sir2data.github.io>, accessed on 29 July 2022) in [41] has 500 real-world triplets.

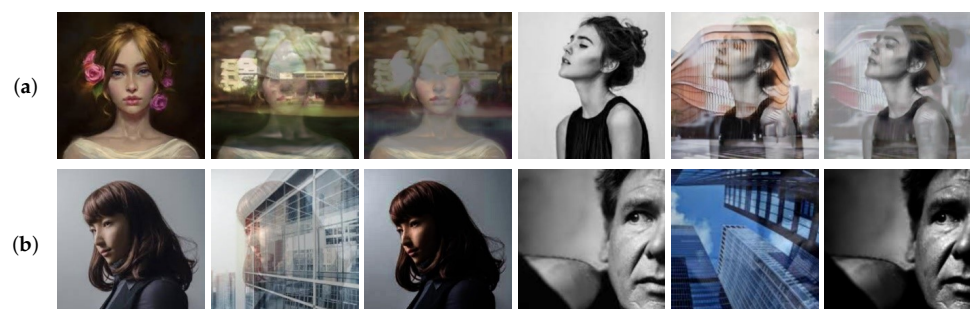
- (2) Recently, Lei et al. [13] proposed a method to obtain a pure flash transmission layer with no reflection. The article minimizes the effect of the reflection layer by combining the image before and after the flash. After preprocessing the image, the triplet image can be obtained as the dataset (Downloaded from [https://hkustconnect-my.sharepoint.com/personal/cleiaa\\_connect\\_ust\\_hk/Documents/Projects/cvpr2021-flash-rr/data.zip?ga=1](https://hkustconnect-my.sharepoint.com/personal/cleiaa_connect_ust_hk/Documents/Projects/cvpr2021-flash-rr/data.zip?ga=1), accessed on 29 July 2022). Their dataset consists of 4056 triplet images for training and 1012 triplet images for validation, both synthetic and real-world. Note that this dataset contains 35 triplet images from [41], so we eliminate them to obtain 4021 training triplet images. Some examples from the existing datasets are shown in Figure 7.



**Figure 7.** Some examples from the existing data set [13,41]. (a) from [41] and (b) from [13].

#### Synthetic Data Generation

Reflection layers in real-world datasets are usually relatively fixed scenes, while there is no relevant dataset for moving objects. The existing method of obtaining a pure reflection layer in a real-world scene is to use a black screen to block the light transmitted from the background image. The main difficulty is that we must simultaneously obtain the image pair of pure reflection images and mixture images with moving objects. The specific operation usually requires fixing the reflection image at the pixel level accuracy, but this is not easy to achieve ultimately. Therefore, we chose a synthetic method that can obtain the pure reflection layer and satisfy the realistic visual effect at the same time as the previous reflection removal experiments. Figure 8 shows the example results of two representative synthesis methods proposed in [15,32]. Although more mixture images closer to the real-world glass effect can be generated by [32], the coefficient matrix involved in the experiment confuses the relationship between the two layers, making it difficult to obtain pure reflection images. Considering the above issue, we adopted the method [15] to synthesize the data in the literature [32] and obtained 4000 sets of data. Note that for the first row of Figure 8, the synthetic mask was first trained with the network to interpret the relationship between the inputs and was then used to randomize the weights of the two input layers in the generated dataset. For the second row of Figure 8, the Gaussian convolution kernels were set as 5 and 7, the corresponding corrective parameters were set as 1 and 0, and the weights of the reflection layer were set as 0.2 and 0.3, respectively. After cropping the input transmission layer and the reflection layer to the same size, a one-step gamma correction was applied to the reflection layer image to modify the brightness and then proportional compositing.



**Figure 8.** Example results by using the two synthesis methods [15,32]. (a) for [32] and (b) for [15].

#### 4.1.4. Implementation and Training Details

For training of the transmission removal network in general scenarios, our training dataset contains 4521 image triplets from [13,41], the validation dataset contains 1012 image triples from [13], and we also added 50 real-world image triplets for testing from [13]. For training of network with portraits, the training dataset contains 3000 triplet images, and the validation dataset contains 1000 triplets.

We implemented our model using Pytorch. In our generator network, the input port was a combination of image **I** and one or two related information of the transmission layer. To distinguish between the scheme using edge information prediction (Case 1 in Figure 2) and the scheme using flash/ambient pair light (Case 2 in Figure 2), we denote the experimental results for both by Ours(a) and Ours(b), respectively. Note that for Ours(a), we chose the same input for transmission removal except for [13], and for Ours(b), we chose the same input as [13]. All source and target sizes were clipped randomly to 256. The feature extractor was based on the Resnet34, and its input channel number was set as 4. We used the extracted underlying information to represent the content so that the generated image was highlighted in the content, and the difference in the gradient was used to distinguish the content information. Gram matrix changes were made to the features to further constrain the style from source to target. We used a PatchGAN discriminator to constrain the style of the generated image and the size of the output matrix was set as 30. The sizes of the convolution kernels in the middle of the SR module were 9, 1, and 5.

We trained for 50 epochs with batch size 16. The entire model is learned from random initialization using the RMSProp optimizer [36] with a learning rate of  $5 \times 10^{-5}$  and clamp of the discriminator is set to the range of  $[-0.01, 0.01]$ . The whole training converged in approximately 7 h using a single GPU NVIDIA GeForce GTX 3090 Ti for the 4721 image pairs from the training data. We implemented random cropping for images with more than 640,000 pixels. A cyclic calculation with batch size 1 was used in the test network to calculate the mean value.

#### 4.2. Transmission Removal Results for General Scenarios

This sub-section evaluates our transmission removal results in general scenarios, both quantitatively and qualitatively. To be fair, we chose methods that provide testing codes for quantitative comparison and selected those methods that provide training codes for an intuitive comparison. Note that for the methods that have provided the training codes, we also retrained their network with the same dataset that we used to train our network. After training, we conducted tests using both real-world and synthetic images. First, Table 2 lists the quantitative comparison of the proposed method with other methods using three metrics of LPIPS, SSIM, and PSNR, where each metric is averaged on 50 real-world mixture images. For each metric of comparative methods, the number to the left of “/” represents the retraining results, and the number to the right represents the results with the originally trained models. Note that “-” stands for the methods that do not release the training code. Among all the methods, [10,11,14–17,20], and Ours(a) have the same input (i.e., a single image circumstance), while [13] and Ours(b) have another input case (i.e., a flash/ambient light image pair as input in addition to the mixture image.) As can be observed in the table, our results perform best among all the comparative methods, and the gains over the state-of-the-art methods are more than 5.356 dB in PSNR, 0.116 in SSIM, and 0.057 in LPIPS. Meanwhile, we found two noteworthy phenomena during our quantitative evaluations: (1) Except for [13], the methods that provide the training code have better metrics after retraining. This is likely due to the fact that different data sets still have a significant impact on the results of the existing reflection removal methods. Since most of the datasets we used are from [13], retraining has no significant effect on this method. (2) Among all methods, the method [15] performed the worst in all metrics, which may be because too much information about the corresponding transmission layer is obtained through the reflection estimation.

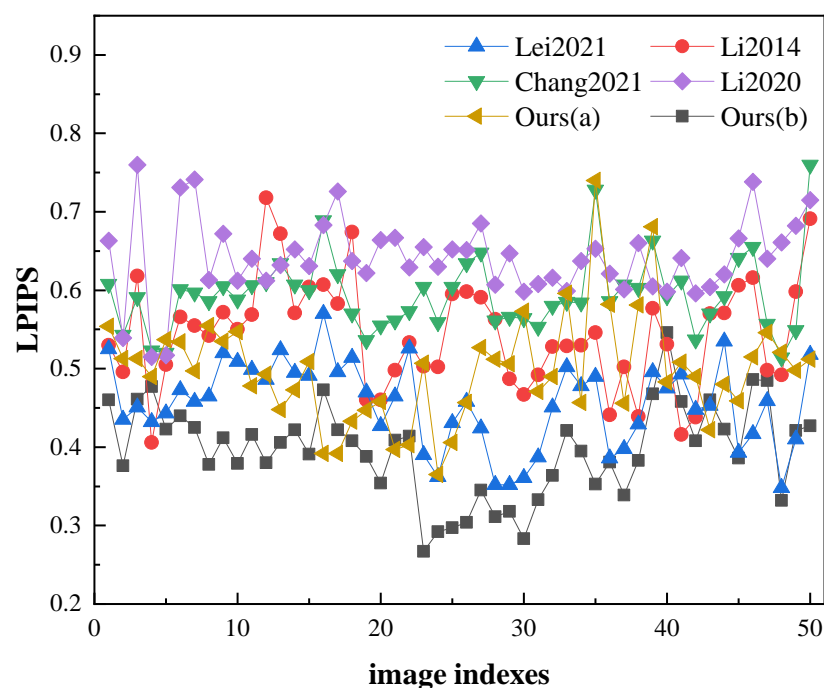


**Table 2.** Quantitative evaluations for transmission removal using three different error metrics of LPIPS, SSIM, and PSNR. (The best and second-best results are shown in bold and underlined, respectively.)

	LPIPS (↓)	SSIM (↑)	PSNR (dB) (↑)
Wan et al. [11]	-/0.512	-/0.716	-/21.934
Zhang et al. [10]/ Wan et al. [20] †	-/0.578	-/0.723	-/23.581
Yang et al. [15]	-/0.635	-/0.420	-/9.615
Li et al. [14] ‡	0.544	0.503	16.737
Chang et al. [16]	0.594/0.659	0.705/0.643	20.785/19.383
Li et al. [17]	0.641/0.673	0.734/0.625	23.657/20.264
Ours(a)	0.508	<u>0.800</u>	<u>24.883</u>
Lei et al. [13]	0.455/ <u>0.453</u>	0.729/0.735	23.583/23.703
Ours(b)	<b>0.396</b>	<b>0.851</b>	<b>29.059</b>

Note: “†” stands for that the separation network architecture applied in [20] is employed from [10]; “‡” represents the method [14] as a class of model-driven method that does not have a pre-training problem and therefore will only have one value in each metric; “-” stands for unavailable due to the absence of released training code.

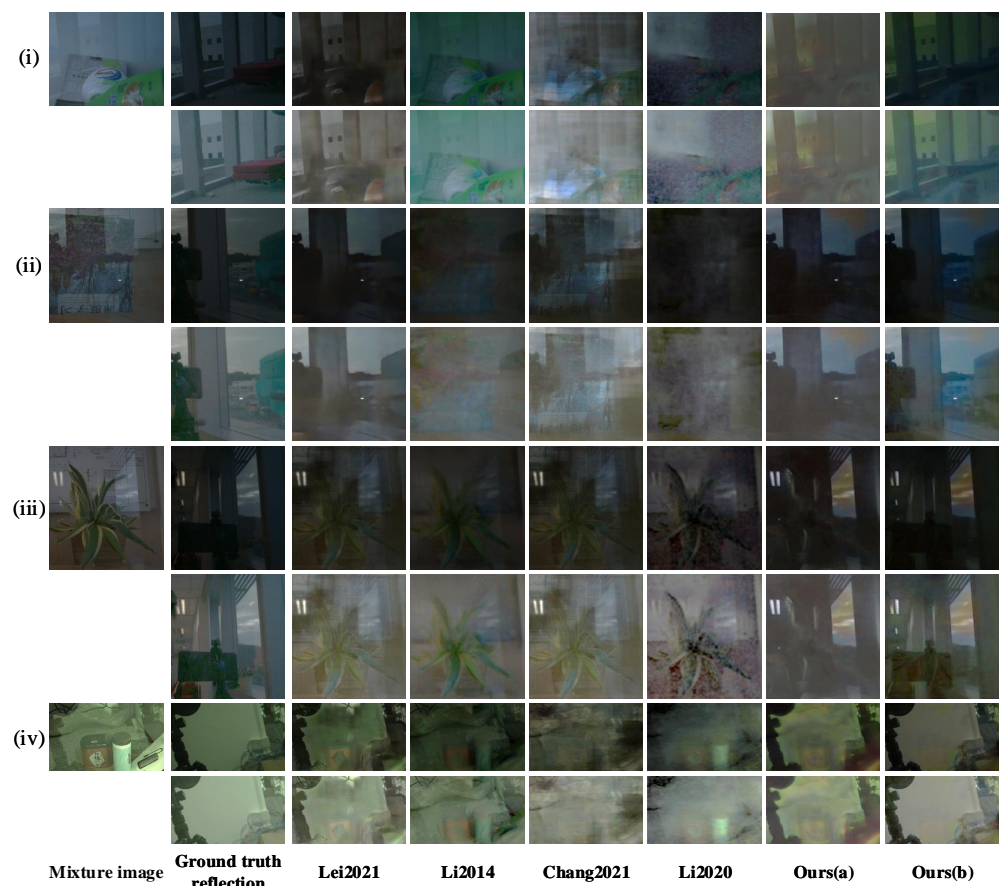
In order to show the superiority of the proposed method more intuitively, Figure 9 illustrates the LPIPS values on all 50 test images. Note that, in the figure, we mainly compare our method with [13,14,16,17], which have provided the training codes, and the data we use are the results obtained after retraining with these training codes. Our results are consistent with Table 2, showing better perceptual metrics than previous methods on almost every image.



**Figure 9.** LPIPS value comparisons of the transmission removal results on all 50 real-world mixture images. Lei2021: [13], Li2014: [14], Chang2021: [16], Li2020: [17].

Next, Figure 10 shows four examples of our results with the other four comparative methods of [13,14,16,17], where all the images were tested on the real-world dataset. In all experimental results, a simple histogram equalization operation was performed on

each transmission removal image for ease of observation, and the equalized version was placed below the corresponding original image of the recovery results. In fact, a visually better DL contrast enhancement method, such as Zero-DCE [42], LE-GAN [43], can be used here to enhance the visual effect further. It would even be possible to directly adopt the enhancement network from the literature [20], but from the concern of simply measuring the effect of transmission removal, this is not in our scope for this paper. As shown in the figure, the four reflection layer images generated by our method are visually closest to the ground truth reflections among all the methods. Specifically, the results of [14,17] are obscure and lack sufficient information, while the results of [13,16] contain relatively clearer reflection contextual information than [14,17], although they also contain some transmission information. Our results are clearer than [13,16] in general, especially in terms of retaining texture information, and our method also has a better separation effect in terms of transmission layer removal. It is worth pointing out that because the training code and dataset are not released in [10,20], we did not directly compare with them.



**Figure 10.** Qualitative comparison of the transmission removal results on real-world mixture images. (i)–(iv) correspond to four different test images. For each test image, the first row and second row are the direct outputs of all methods and their corresponding histogram equalized versions. Lei2021: [13], Li2014: [14], Chang2021: [16], Li2020: [17].

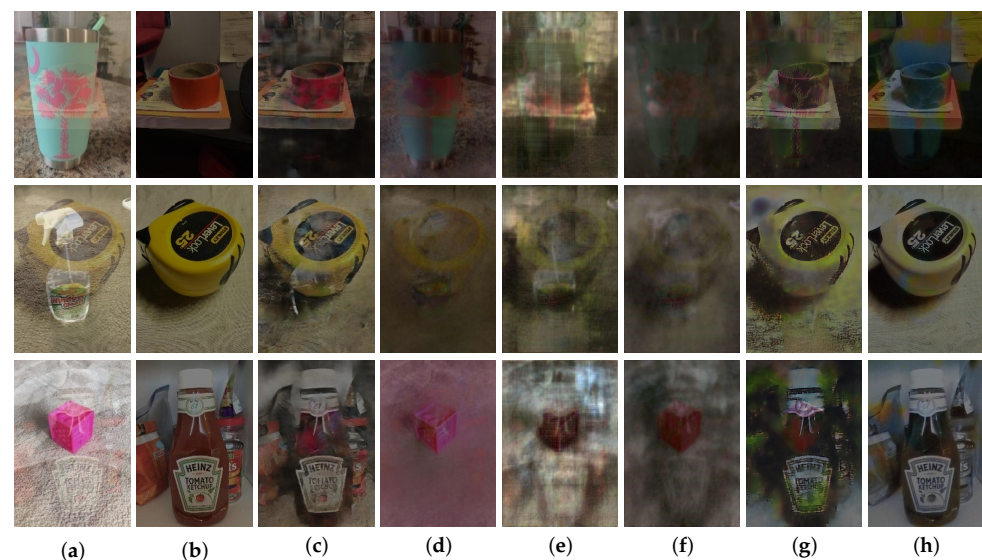
**Table 3.** Training time of the proposed network and other networks

	Training Time (hour)
Lei et al. [13]	17 <sup>†</sup>
Chang et al. [16]	21 <sup>†</sup>
Li et al. [17]	20 <sup>†</sup>
Ours(b)	7

Note: “†” stands for the retraining time for our training dataset.

Meanwhile, we also recorded the training time of [13,16,17], and our network in Table 3 when the epoch was set as 50. From the table, our network has the least training time in the same training set, which is more conducive to rapid comparison.

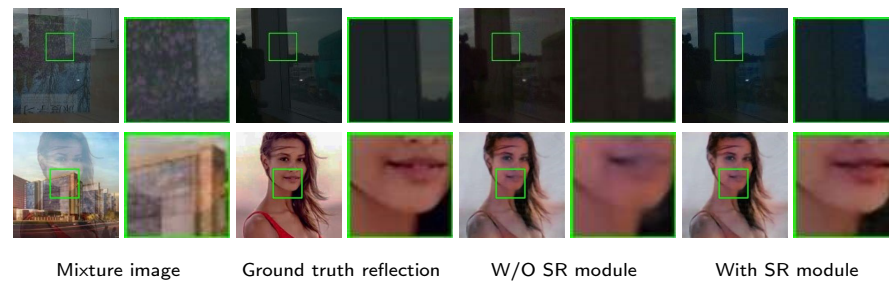
In the end, we also tested our method on some synthetic mixture images. Figure 11 shows three examples of transmission removal results of our method and comparison methods. From the results of the synthetic images, we can still conclude that our method has better visual effects and reflection layer details.



**Figure 11.** Qualitative comparison of the transmission removal results on synthetic mixture images. (a) Mixture image, (b) Ground truth reflection, (c) [13], (d) [14], (e) [16], (f) [17], (g) Ours(a), and (h) Ours(b).

#### 4.3. Ablation Study

This subsection performed some ablation experiments to verify the necessity of adding each module to the network. Figure 12 shows the comparison of the results with or without SR module for Ours(b). By adding the SR module, the generated image can achieve pixel-level detail representation in the corresponding region, and the overall image is clearer than without adding the SR module. Figure 13 compares the effect of with or without content constraint for Ours(b). Because the underlying information with richer texture information is used instead of global features, the details of the corresponding regions are more elaborated (see the edge information in the top row and the outline of eyes in the bottom row). Tables 4 and 5 present the quantitative evaluation results of the model without SR module and content loss for Ours(a) and Ours(b), respectively, again using the average results of 50 test images. Both tables compare the ablation experiments in terms of three metrics, with the difference that the inputs in Table 5 carry more information about the transmission layer. In general, the values of the metrics in Table 5 are better than those in Table 4. In addition, the results for the model with only the SR module, with only content constraints and completeness, also achieve an improvement in metrics compared to the model without the module and a decrease in SSIM of Table 4. On the one hand, this metric focuses more on the structural similarity between the two images, while the input of the overall network contains only a small amount of transmission layer information, which is incomplete for the guidance of the generated images. On the other hand, with the addition of the SR module and content loss, the overall model is more concerned with the consistency of the generated images with the pixel-level ground truth. In contrast, the input of Table 5 has more complete input information and learns the gradient as well as region block information during the bootstrapping process, so there will be further improvement in the metric. The best results can be achieved by combining the SR module and content loss as can be seen in the table.



**Figure 12.** Comparisons of visual effects of the proposed network with or without SR module for Ours(b). The green boxes are partially enlarged for a better view.



**Figure 13.** Comparisons of visual effects of the proposed network with or without content loss for Ours(b). The green boxes are partially enlarged for a better view.

**Table 4.** Quantitative ablation study of Ours(a) network for the SR module and content loss.

	LPIPS (↓)	SSIM (↑)	PSNR (dB) (↑)
W/O SR module and content loss	0.647	0.588	18.611
W/O SR module	0.568	0.844	23.313
W/O content loss	0.556	0.813	21.219
Complete	0.508	0.800	24.883

**Table 5.** Quantitative ablation study of Ours(b) network for the SR module and content loss.

	LPIPS (↓)	SSIM (↑)	PSNR (dB) (↑)
W/O SR module and content loss	0.521	0.813	23.769
W/O SR module	0.413	0.844	28.780
W/O content loss	0.437	0.811	25.559
Complete	0.396	0.851	29.059

#### 4.4. Transmission Removal for Photographer Identification

For the application of photographer identification, Figure 14 shows the transmission removal effects of our method. The recovered facial information can help us identify personal information. In detail, the Internet visual search tool (e.g., Bing Visual Search by Microsoft (<https://www.bing.com/visualsearch>, accessed on 29 July 2022)) and DeepFace [44] are used to evaluate the efficacy of photographer identification of the mixture images and estimated images. We selected the related Hollywood celebrity portraits from CACD2000 [45], as shown in Figure 15 as the target images for verification with our test images on DeepFace. Table 6 lists the verification results of two leading recognition platforms, where for the visual search tool, the output is the name of the image subject information, and for DeepFace, the output includes the verification results and matching distance with the target images. As can be seen from the table, most of the images recovered by other methods cannot be matched with the target image or even detect the face. In comparison, our results in the visual search tool can recognize faces well and identify the corresponding identity information. Moreover, the face matching by DeepFace is achieved with numerical output,



which means that we can still recognize face information even if there is an incomplete match.

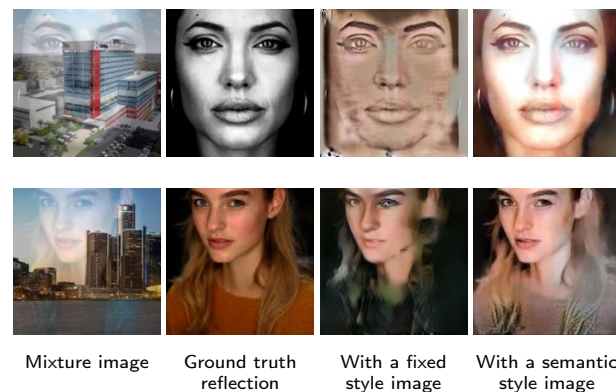


**Figure 14.** Four recovery examples of the reflection layers containing portrait scenes. (a) Mixture image, (b) [13], (c) [14], (d) [16], (e) [17], (f) Ours(a), (g) Ours(b).



**Figure 15.** Selected target images for verification from CACD2000 dataset.

Figure 16 shows the effects of different style image selection in the photographer identification application. The results in the third column are obtained by using a fixed face image as the style image. In contrast, the fourth column is obtained by using the corresponding image with semantic information as the style image. It can be seen that the results with semantic style images perform better than those with a fixed style image in terms of skin tone, eyebrows, and hair color. When the reflection in the mixture image is dim, the color images with the corresponding semantic information can achieve image colorization, as in the top row of Figure 16.



**Figure 16.** Comparisons of visual effects of the proposed network with a fixed style image or with a semantic style image. The top and bottom rows correspond to dim and vivid reflection images with portraits.



**Table 6.** Face identification results on two mainstream platforms. (Our results are highlighted in bold.)

Test Image	Recognition by Bing Visual Search	Verification by DeepFace	
		Verified	Distance
Figure 14(1-a)	An architecture	No face	N/A
Figure 14(1-b)	J. Lawrence	No face	N/A
Figure 14(1-c)	No result	No face	N/A
Figure 14(1-d)	No result	No face	N/A
Figure 14(1-e)	J. Lawrence	No face	N/A
Figure 14(1-f)	<b>J. Lawrence</b>	<b>False</b>	<b>0.6255</b>
Figure 14(1-g)	<b>J. Lawrence</b>	<b>True</b>	<b>0.2928</b>
Figure 14(2-a)	An architecture	No face	N/A
Figure 14(2-b)	No result	No face	N/A
Figure 14(2-c)	No result	No face	N/A
Figure 14(2-d)	No result	No face	N/A
Figure 14(2-e)	No result	No face	N/A
Figure 14(2-f)	<b>L. Dicaprio</b>	<b>True</b>	<b>0.2827</b>
Figure 14(2-g)	<b>L. Dicaprio</b>	<b>True</b>	<b>0.2387</b>
Figure 14(3-a)	An architecture	No face	N/A
Figure 14(3-b)	J. Carrey	No face	N/A
Figure 14(3-c)	No result	No face	N/A
Figure 14(3-d)	No result	No face	N/A
Figure 14(3-e)	No result	No face	N/A
Figure 14(3-f)	<b>J. Carrey</b>	<b>True</b>	<b>0.3157</b>
Figure 14(3-g)	<b>J. Carrey</b>	<b>True</b>	<b>0.2467</b>
Figure 14(4-a)	An architecture	No face	N/A
Figure 14(4-b)	No result	No face	N/A
Figure 14(4-c)	No result	No face	N/A
Figure 14(4-d)	No result	No face	N/A
Figure 14(4-e)	No result	No face	N/A
Figure 14(4-f)	<b>B. Cumberbatch</b>	<b>False</b>	<b>0.5538</b>
Figure 14(4-g)	<b>B. Cumberbatch</b>	<b>True</b>	<b>0.2938</b>

## 5. Conclusions

This paper proposes a reflection scene separation method by transmission removal to recover the highly reproduced reflection layer image from the mixed image with glass. The main framework of the proposed transmission removal network is based on GAN while taking into account the global information, where constraints of different dimensions are placed on the patch-level features of the generated images. Starting from the idea of separability of content and style in the generated images, we use different levels of contextual information to retain more details in the generated images and achieve clarity of details through interpolation and re-convolution operations of the SR module. Furthermore, we also extend our network to the application of photographer identification, which can effectively assist in exploring the secrets behind the photo. Our experiments show that both the transmission removal of the general scene and the scene with the portrait have satisfactory results. Our method no longer outputs the reflection layer as a byproduct compared to previous reflection removal methods. In addition, since we complement the perception of the reflection image on the glass image, some weak textural information in the mixture image can be perceived. Compared with the baseline method for reflection separation, whose contribution is to design a generic enhancement network to enhance the recovered reflection image, our proposed method focuses more on reducing the gap between the recovered reflection layer image and the ground truth reflection image. In fact, after our proposed network, it can also be migrated to the enhancement network in previous work for further data enhancement. However, we hold the following two perspectives: (1) the effect of recovering the reflection layer closer to the ground truth can provide more significant help for the subsequent processing; and (2) the subsequent enhancement of

the reflection layer should be differentiated for different application scenarios, such as the applications of photographer identification of reflection scenes proposed in this paper.

There are certainly still some improvements to our approach; for example, when there is a blur in the image due to shooting shake or rapid movement of the subject, our method is not suitable for solving the problem of reflection with ghosting. These are our future improvement directions.

**Author Contributions:** Conceptualization, Z.L. and H.Y.; methodology, Z.L. and R.S.; software, Z.L.; validation, Z.L., H.Y. and R.S.; formal analysis, T.Q.; investigation, H.Y. and C.Q.; resources, Z.L.; data curation, Z.L.; writing—original draft preparation, Z.L.; writing—review and editing, H.Y., R.S. and T.Q.; visualization, Z.L.; supervision, C.Q.; project administration, H.Y. and C.Q.; funding acquisition, H.Y. and C.Q. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Natural Science Foundation of China grant number 62172281, 62172280, U20B2051, 61702332, the Fundamental Research Funds for the Provincial Universities of Zhejiang grant number GK219909299001-007, the Natural Science Foundation of Shanghai grant number 21ZR1444600, and the STCSM Capability Construction Project for Shanghai Municipal Universities grant number 20060502300. The APC was funded by the STCSM Capability Construction Project for Shanghai Municipal Universities grant number 20060502300.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank the anonymous reviewers for their valuable suggestions, which helped to improve this paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Medhi, J.P.; Nirmala, S.; Choudhury, S.; Dandapat, S. Improved detection and analysis of Macular Edema using modified guided image filtering with modified level set spatial fuzzy clustering on Optical Coherence Tomography images. *Biomed. Signal Process. Control* **2023**, *79*, 104149. [\[CrossRef\]](#)
2. Hilal, A.; Alabdulkreem, E.; Alzahrani, J.; Eltahir, M.; Eldesouki, M.; Yaseen, I.; Motwakel, A.; Marzouk, R.; Mustafa, A. Political optimizer with deep learning-enabled tongue color image analysis model. *Comput. Syst. Sci. Eng.* **2023**, *45*, 1129–1143. [\[CrossRef\]](#)
3. Ibrahim, H.; Fahmy, O.M.; Elattar, M.A. License plate Image analysis empowered by Generative Adversarial Neural Networks (GANs). *IEEE Access* **2022**, *10*, 30846–30857.
4. Kubicek, J.; Penhaker, M.; Krejcar, O.; Selamat, A. Modern trends and Applications of Intelligent methods in Biomedical signal and Image processing. *Sensors* **2021**, *21*, 847. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Mambou, S.; Krejcar, O.; Selamat, A.; Dobrovolny, M.; Maresova, P.; Kuca, K. Novel thermal image classification based on techniques derived from mathematical morphology: Case of breast cancer. In Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering, Granada, Spain, 6–8 May 2020; pp. 683–694.
6. Rezaei, Z.; Selamat, A.; Taki, A.; Rahim, M.S.M.; Kadir, M.R.A. Automatic plaque segmentation based on hybrid fuzzy clustering and k nearest neighborhood using virtual histology intravascular ultrasound images. *Appl. Soft Comput.* **2017**, *53*, 380–395. [\[CrossRef\]](#)
7. Bach, H.; Neuroth, N. *The Properties of Optical Glass*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1998.
8. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
9. Li, Y.; Yan, Q.; Zhang, K.; Xu, H. Image Reflection Removal via Contextual Feature Fusion Pyramid and Task-Driven Regularization. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 553–565. [\[CrossRef\]](#)
10. Zhang, X.; Ng, R.; Chen, Q. Single image reflection separation with perceptual losses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4786–4794.
11. Wan, R.; Shi, B.; Duan, L.Y.; Tan, A.H.; Kot, A.C. CRRN: Multi-scale guided concurrent reflection removal network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4777–4785.
12. Wan, R.; Shi, B.; Li, H.; Duan, L.Y.; Tan, A.H.; Kot, A.C. CoRRN: Cooperative reflection removal network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2969–2982. [\[CrossRef\]](#)
13. Lei, C.; Chen, Q. Robust reflection removal with reflection-free flash-only cues. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14811–14820.

14. Li, Y.; Brown, M.S. Single image layer separation using relative smoothness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2752–2759.
15. Yang, J.; Gong, D.; Liu, L.; Shi, Q. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 654–669.
16. Chang, Y.C.; Lu, C.N.; Cheng, C.C.; Chiu, W.C. Single image reflection removal with edge guidance, reflection classifier, and recurrent decomposition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual Conference, 5–9 January 2021; pp. 2033–2042.
17. Li, C.; Yang, Y.; He, K.; Lin, S.; Hopcroft, J.E. Single image reflection removal through cascaded refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3565–3574.
18. Song, B.; Zhou, J.; Wu, H. Multi-stage Curvature-guided Network for Progressive Single Image Reflection Removal. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6515–6529. [[CrossRef](#)]
19. Fan, Q.; Yang, J.; Hua, G.; Chen, B.; Wipf, D. A generic deep architecture for single image reflection removal and image smoothing. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3238–3247.
20. Wan, R.; Shi, B.; Li, H.; Duan, L.Y.; Kot, A.C. Reflection scene separation from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2398–2406.
21. Wan, R.; Shi, B.; Hwee, T.A.; Kot, A.C. Depth of field guided reflection removal. In Proceedings of the IEEE International Conference on Image Processing, Phoenix, AZ, USA, 25–28 September 2016; pp. 21–25.
22. Springer, O.; Weiss, Y. Reflection separation using guided annotation. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 1192–1196.
23. Arvanitopoulos, N.; Achanta, R.; Susstrunk, S. Single image reflection suppression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4498–4506.
24. Levin, A.; Weiss, Y. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1647–1654. [[CrossRef](#)] [[PubMed](#)]
25. Mechrez, R.; Talmi, I.; Shama, F.; Zelnik-Manor, L. Maintaining natural image statistics with the contextual loss. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 427–443.
26. Yano, T.; Shimizu, M.; Okutomi, M. Image restoration and disparity estimation from an uncalibrated multi-layered image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 247–254.
27. Rafiq, M.; Bajwa, U.I.; Gilanie, G.; Anwar, W. Reconstruction of scene using corneal reflection. *Multimed. Tools Appl.* **2021**, *80*, 21363–21379. [[CrossRef](#)]
28. Jenkins, R.; Kerr, C. Identifiable images of bystanders extracted from corneal reflections. *PLoS ONE* **2013**, *8*, e83325. [[CrossRef](#)] [[PubMed](#)]
29. Wu, J.; Ji, Z. Seeing the unseen: Locating objects from reflections. In Proceedings of the Annual conference towards autonomous robotic systems, Bristol, UK, 25–27 July 2018; pp. 221–233.
30. Nishino, K.; Nayar, S.K. Eyes for relighting. *ACM Trans. Graph.* **2004**, *23*, 704–711. [[CrossRef](#)]
31. Nishino, K.; Belhumeur, P.N.; Nayar, S.K. Using eye reflections for face recognition under varying illumination. In Proceedings of the IEEE International Conference on Computer Vision, Beijing, China, 17–20 October 2005; pp. 519–526.
32. Wen, Q.; Tan, Y.; Qin, J.; Liu, W.; Han, G.; He, S. Single image reflection removal beyond linearity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3771–3779.
33. Zhang, K.; Sun, M.; Han, T.X.; Yuan, X.; Guo, L.; Liu, T. Residual Networks of Residual Networks: Multilevel Residual Networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1303–1314. [[CrossRef](#)]
34. Wang, Y.; Wang, L.; Wang, H.; Li, P. Resolution-Aware Network for Image Super-Resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 1259–1269. [[CrossRef](#)]
35. Xiang, X.; Zhu, L.; Li, J.; Wang, Y.; Huang, T.; Tian, Y. Learning Super-Resolution Reconstruction for High Temporal Resolution Spike Stream. *IEEE Trans. Circuits Syst. Video Technol.* **2021**. [[CrossRef](#)]
36. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 184–199.
37. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
38. Li, C.; Wand, M. Combining markov random fields and convolutional neural networks for image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2479–2486.
39. Johnson, J.; Alahi, A.; Li, F.F. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
40. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595.
41. Wan, R.; Shi, B.; Duan, L.Y.; Tan, A.H.; Kot, A.C. Benchmarking single-image reflection removal algorithms. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3922–3930.

42. Guo, C.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.; Cong, R. Zero-reference deep curve estimation for low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1780–1789.
43. Fu, Y.; Hong, Y.; Chen, L.; You, S. LE-GAN: Unsupervised low-light image enhancement network using attention module and identity invariant loss. *Knowl.-Based Syst.* **2022**, *240*, 108010. [[CrossRef](#)]
44. Serengil, S.I.; Ozpinar, A. Lightface: A hybrid deep face recognition framework. In Proceedings of the Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 15–17 October 2020; pp. 1–5.
45. Chen, B.C.; Chen, C.S.; Hsu, W.H. Cross-age reference coding for age-invariant face recognition and retrieval. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 768–783.