

Face Image Reflection Removal

Renjie Wan[†]Boxin Shi[★]Haoliang Li[†]Ling-Yu Duan[★]Alex C. Kot[†][†]School of Computer Science and Engineering, Nanyang Technological University, Singapore[★]National Engineering Laboratory for Video Technology, School of EECS, Peking University, China

Abstract

Face images captured through the glass are usually contaminated by reflections. The non-transmitted reflections make the reflection removal more challenging than for general scenes, because important facial features are completely occluded. In this paper, we propose and solve the face image reflection removal problem. We remove non-transmitted reflections by incorporating inpainting ideas into a guided reflection removal framework and recover facial features by considering various face-specific priors. We use a newly collected face reflection image dataset to train our model and compare with state-of-the-art methods. The proposed method shows advantages in estimating reflection-free face images for improving face recognition.

1. Introduction

As one of the most commonly observed subjects in computer vision, face images are often captured by various types of imaging sensors under unconstrained wild scenarios, which bring different types of distortions to the clear face images. When face images are captured behind a piece of glass, the reflection-contaminated face images not only unpleasantly affect the human perception, but also degrade the performance of visual computing algorithms focusing on face. Therefore, it is of great interest to remove the reflections and enhance the visibility of the human faces behind the glass.

Different from general objects or scenes, faces have its specific priors awarded by humans, even if a slight reflection (transmitted) distortion may significantly annoy human perception [21]. When reflection become stronger (non-transmitted), machine vision algorithms may fail due to the lost or distortion of important facial features. How to *remove non-transmitted reflections* and *recover important features for machine vision methods* pose unique challenges for face reflection removal.

Existing reflection removal methods [17, 32, 30, 6] can be directly applied to face images with reflections. How-

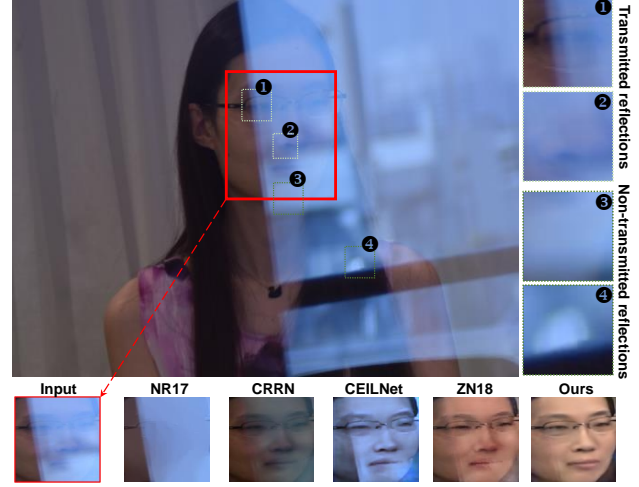


Figure 1. Examples of transmitted reflections with high transmittance, non-transmitted reflections with low transmittance, and the reflection removal results obtained by using NR17 [2], CRRN [30], CEILNet [6], ZN18 [39], and our method.

ever, due to ignorance of the specific facial priors and the non-transmitted reflections, artifacts on face largely remain on the recovered ‘reflection-free’ face image (e.g., the result obtained by CEILNet [6] in Figure 1). Thus, methods designed for generic reflections in arbitrary scenarios are not capable to deal with these challenges. To recover the facial information largely occluded by non-transmitted reflections, it is also straight-forward to integrate specific facial priors into the single image inpainting methods (e.g., [24, 38]). However, solely relying on learned representations from the training data to inpaint the reflection-contaminated region may not faithfully retain the lost face identity feature.

To conquer the above challenges, we first explore the complementary advantages from image inpainting and reflection removal to recover the facial information occluded by non-transmitted reflections. Then, to recover important facial features, we employ the guided removal framework with particular considerations on the feature level similarity and specific facial priors. Instead of only predicting the

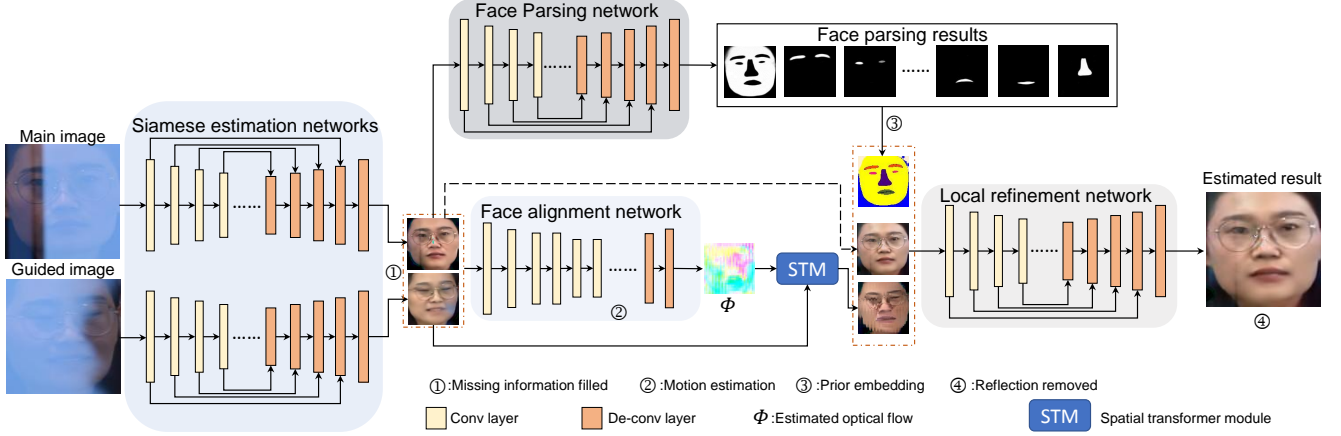


Figure 2. The framework of our proposed network. It consists of four networks with distinctive functions: the siamese estimation networks to roughly estimate the missing facial information (①), the face alignment network to compensate the different motion direction between two face images (②), the face parsing network to estimate the face parsing maps and do the prior embedding (③), and the local refinement networks to refine the local details (④).

missing information from the learned representations, the guided framework [31, 16] can provide more accurate identity details of the human faces due to the additionally guided information, from an additional face image of an image sequence with continuous face movement, which provide different facial pose or reflection properties. On the other hand, since the face similarity is compared in a compact feature space and the pixel level similarity adopted by previous reflection removal methods can hardly guarantee the identity consistency [38], we also embed the feature level similarity and other specific facial priors into the estimation process.

Our complete framework is shown in Figure 2, which includes four components: the siamese estimation network to roughly estimate the missing facial information, the face alignment network to compensate the different motions between two face images, the prior estimation network to embed the facial priors into the whole estimation process, and the local refinement network to refine the local details. Our major contributions are summarized as follows:

- We propose the first reflection removal framework that targets at the face images for improving human perception and facilitating machine vision algorithms.
- We propose an effective approach to remove non-transmitted reflections by mixing the merits of image inpainting and reflection removal.
- We recover important facial features by employing a guided removal framework with particular considerations on the feature level similarity and specific facial priors.
- We build the first face reflection image dataset to facilitate the research of reflection removal in a specific domain and accordingly perform quantitative and qualitative evaluation.

2. Related work

Reflection removal. Previous works on reflection removal can be roughly classified into two categories. The first category solves by using the non-learning based methods. For example, Li *et al.* [17] and Nikolas *et al.* [2] made use of the different blur levels of the background and reflection layers. Shih *et al.* [26] used the GMM patch prior to remove reflections with the visible ghosting effects. The hand-crafted priors adopted by these methods are based on the observations of some special properties between the background and reflection (*e.g.*, different blur levels [32, 17]) which is often violated in the general scenes especially when these properties are weakly observed.

The deep learning framework also benefits reflection removal problems. For example, Fan *et al.* [6] proposed a two-stage deep learning approach to learn the mapping between the mixture images and the estimated clean images. Recently, Wan *et al.* [30] also proposed a concurrent model to better preserve the background details. The method proposed by Zhang *et al.* [39] first utilized the generative model to better learn the mappings from the mixture image to the clean images. However, existing methods are all designed for general scenes, which have difficulty in preserving facial details in face image reflection removal problem.

Face image enhancement. Numerous methods have been proposed during the past decades to solve different face image enhancement problem including face hallucination [21], face deblurring [23], and face inpainting [19]. Recently, the end-to-end deep learning framework are introduced to solve this problem in a data-driven manner. For example, Li *et al.* [18] proposed a method based on the generative model to solve the face inpainting problem. Chen *et al.* [3] made full use of the geometry prior to solve the face super-resolution

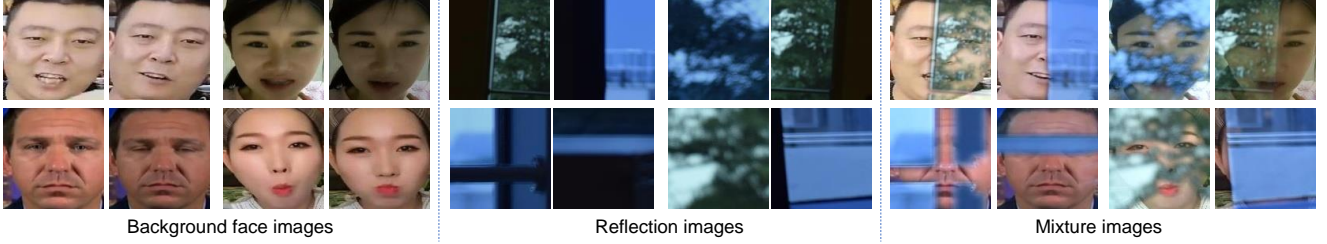


Figure 3. The background face images, reflection images, and the mixture images in our training dataset.

problem. Shen *et al.* [25] also proposed a method to solve the face deblurring problem by using the face semantic priors. However, the face image reflection removal problem has never been explicitly modeled and solved.

3. Proposed method

In this section, we describe the dataset, the design methodology of the proposed reflection removal network, the optimization process, and the training details.

3.1. Dataset preparation

The data-driven approaches need a large-scale dataset to learn the inherent reflection properties [30]. Previous methods [30, 6, 39] obtain the training dataset by using the following image formation model:

$$\mathbf{I} = \alpha \mathbf{B} + \beta \mathbf{R}, \quad (1)$$

where α and β are the mixing coefficients and \mathbf{I} , \mathbf{B} , and \mathbf{R} are the mixture image, the background image, and reflection image, respectively. The background images \mathbf{B} can be obtained from generic image datasets (*e.g.*, PASCAL [5] or COCO [20]) only when targeting the reflections at arbitrary scenes [30, 6, 39]. Accordingly, existing benchmark reflection removal datasets (*e.g.*, SIR² [29]) are not suitable for our task due to the scenery diversity. Although many face image datasets have been proposed (*e.g.*, CELEBA [22] and CASIA Webface dataset [36]), they are also not applicable for our problem since they mostly consider a fixed facial pose. To facilitate the training and evaluation of our approach, we build a large-scale face image training dataset collected from online resources and its corresponding evaluation dataset taken in the real world.

Training dataset. Our background face images in the training dataset are collected from Youtube by cropping face images from several consecutive video frames. The reflection images are taken by ourselves based on the method proposed in [30]. Then, we generate the mixture images by using Equation (1). To focus on the vital facial components, we further adopt the MTCNN [37] to crop the face portion. We show samples from our training dataset in Figure 3.

Our training dataset has two major characteristics: 1) **Diversity.** The face images in the training dataset are with

different races, expressions, and poses; 2) **Scale:** The training dataset have 15950 face images from approximately 450 people to meet the request of data-driven methods and each person is labeled by their corresponding person IDs.

Evaluation dataset. The images in the evaluation dataset are all taken in the real world by using different capturing devices (*e.g.*, DSLR cameras and mobile phones) with diverse scene settings. Except for the face images occluded by reflections, we also take its corresponding ‘groundtruth’ images with same identity information for further evaluation. The evaluation dataset has 450 images from 25 different person.

3.2. Network architecture

As shown in Figure 2, our network includes four parts to do the missing facial information estimation, motion compensation, prior estimation, and local refinement, respectively. Except that the face alignment network is largely based on an existing optical flow estimation network [7], other three networks all have a similar mirror-like framework with the encoder to capture the context information by contracting the feature channels step by step and decoder part to obtain the final results.

3.2.1 Siamese estimation networks

Due to the occlusions caused by the reflections, it is non-trivial to estimate facial priors (*e.g.*, facial landmark positions and parsing) from the reflection-contaminated input images directly. We first use the siamese estimation networks with shared weights to roughly estimate the coarse face images from the input image pair as:

$$\{\mathbf{B}_s^m, \mathbf{B}_s^g\} = \{\mathcal{G}_s(\mathbf{I}_m), \mathcal{G}_s(\mathbf{I}_g)\}, \quad (2)$$

where \mathcal{G}_s denotes one branch of the siamese estimation networks, \mathbf{I}_m and \mathbf{I}_g are the main image and guided image with different reflections or varying facial properties (*e.g.*, pose and illuminations), and \mathbf{B}_s^m and \mathbf{B}_s^g are the roughly recovered images corresponding to \mathbf{I}_m and \mathbf{I}_g , respectively.

Local context loss. Due to the regional property of reflections [28], the missing information occluded by reflec-

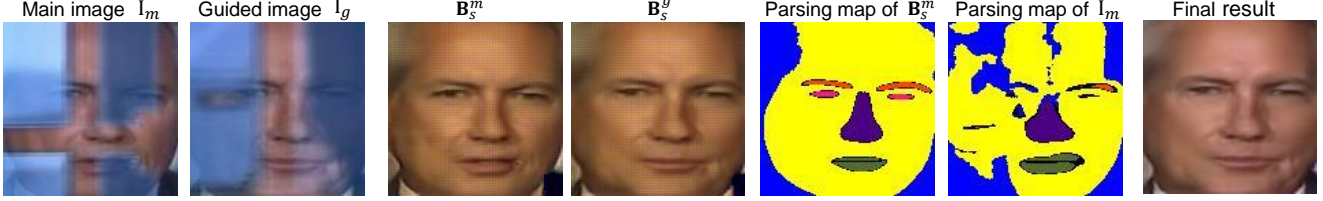


Figure 4. The intermediate results \mathbf{B}_s^m and \mathbf{B}_s^g obtained by the siamese estimation network. The parsing maps of \mathbf{B}_s^m and \mathbf{I}_m obtained by the face parsing network and the final result of the local refinement network rightmost.

tions cannot be well reconstructed by solely minimizing the global loss on the whole image. To solve this problem, we adopt the local context loss widely used by the image inpainting methods [24] as:

$$\mathcal{L}_c = \|\mathbf{W} \odot (\mathcal{G}_s(z) - z^*)\|_1, \quad (3)$$

where z is the input to the network, z^* is its corresponding ground truth, \odot is the element-wise product operation, and \mathbf{W} denotes the binary mask corresponding to the non-transmitted reflections. When the reflections occupy the whole image plane, Equation (3) degrades to the common global loss.

Adversarial loss. To roughly estimate the missing information occluded by the non-transmitted reflections, we employ the Conditional Wasserstein GAN as follows:

$$\begin{aligned} \mathcal{L}_{\text{adv}} = \min_{\mathcal{G}_s} \max_{D_s \in \mathcal{D}} & E_{z, z^* \sim \mathbb{P}_r} [D_s(z, z^*)] \\ & - E_{z \sim \mathbb{P}_r} [D_s(z, \mathcal{G}_s(z))], \end{aligned} \quad (4)$$

where D_s is the discriminator network, \mathcal{D} is the set of 1-Lipschitz functions and \mathbb{P}_r is the real data distributions. Our discriminator takes an input image with a size of 128×128 and has 6 strided convolutional layers followed by the ReLU activation function. In the last layer, we use the sigmoid function to generate the final result.

Combining the above terms, the loss function for the siamese estimation networks become:

$$\begin{aligned} \mathcal{L}_{\text{SEN}} = \alpha_e (\mathcal{L}_1(\mathbf{B}_s^m, \mathbf{B}^m) + \mathcal{L}_c(\mathbf{B}_s^m, \mathbf{B}^m) + \mathcal{L}_1(\mathbf{B}_s^g, \mathbf{B}^g) \\ + \mathcal{L}_c(\mathbf{B}_s^g, \mathbf{B}^g)) + \lambda_e \mathcal{L}_{\text{adv}}^m + \beta_e \mathcal{L}_{\text{adv}}^g, \end{aligned} \quad (5)$$

where \mathcal{L}_1 is the classical pixel-wise loss and $\alpha_e = 0.5$, $\lambda_e = 10^{-4}$, and $\beta_e = 10^{-4}$ are weights to balance different terms.

The siamese estimation networks can be regarded as a *single-image approach* to solve this problem. From the results shown in Figure 4, using such networks alone is not sufficient to efficiently remove the reflections on face. However, from the face parsing results shown in Figure 4, it helps to alleviate the difficulties for estimating the facial priors in the next stage by estimating some key facial components.

3.2.2 Face parsing network

Previous methods [18, 3] have proved the effectiveness of the face specific prior knowledge in preserving the essential appearance information and rough locations of the facial components.

To better recover the important facial features, we embed the facial prior into our estimation process by employing the face parsing network. We use the U-Net as the backbone but just keep the vital layers for the efficiency of the whole network to estimate the parsing maps of facial components as follows:

$$\mathbf{M} = \mathcal{G}_p(z), \quad (6)$$

where \mathcal{G}_p is the face parsing network, z denotes the input images of this network, and \mathbf{M} denotes the 11-channel semantic parsing map of the important facial components as shown in Figure 2.

3.2.3 Face alignment network

Due to the misalignment between the input image pair, their motion or pose inconsistency may increase the difficulty to recover image details [34]. To compensate the motion inconsistency, we adopt FlowNetS model [7] as the backbone to build the face alignment network. It takes the two roughly recovered images from the siamese estimation networks as the input and aims at estimating the optical flow from \mathbf{B}_s^m to \mathbf{B}_s^g as shown in Figure 2 as follows:

$$\Phi = \mathcal{G}_f(\mathbf{B}_s^m, \mathbf{B}_s^g), \quad (7)$$

where \mathcal{G}_f denotes the face alignment network and Φ denotes the estimated optical flow field. With the flow field Φ , the guided image can be warped to the main image by using the spatial transformer network [10] as:

$$\begin{aligned} \mathbf{B}_s^w(i, j) = \\ \sum_{h, w \in \mathcal{N}} \mathbf{B}_s^g(h, w) \mathbf{M}(0, 1 - |\Phi_{ij}^y - h|) \mathbf{M}(0, 1 - |\Phi_{ij}^x - w|), \end{aligned} \quad (8)$$

where $\mathbf{M} = \max(\cdot, 0)$, Φ_{ij}^x and Φ_{ij}^y denote the predicated x and y coordinates for the pixel $\mathbf{B}_s^w(i, j)$ and \mathcal{N} represents the four-pixel neighbors of $(\Phi_{ij}^x, \Phi_{ij}^y)$.

Landmark loss and face parsing loss. The classical FlowNetS model [7] is trained by using the supervised training strategy. However, due to the lack of corresponding ground truth optical flow, the supervised training strategy is not applicable in our settings. To solve this problem, an unsupervised training strategy is proposed in [11] by minimizing the MSE loss between a warped image and another non-warped image. However, due to the roughly estimate results $\mathbf{B}_s^m, \mathbf{B}_s^g$ may have different colors, the MSE loss cannot serve as a valid way to measure the difference. Thus, we propose to use the facial landmarks and the face parsing results to facilitate the training process.

In order to align \mathbf{B}_s^m and \mathbf{B}_s^g , the landmarks of \mathbf{B}_s^m and those of the warped image \mathbf{B}_s^w should be close to each other. We first get the facial landmarks corresponding to \mathbf{B}_s^m and \mathbf{B}_s^w and then define the landmark loss as follows:

$$\mathcal{L}_{lam} = \|\{\theta^m\}_{i,j=1}^{68} - \{\theta^s\}_{i,j=1}^{68}\|_2^2, \quad (9)$$

where $\{\theta^m\}_{i,j=1}^{68}$ and $\{\theta^s\}_{i,j=1}^{68}$ are the facial landmarks corresponding to \mathbf{B}_s^m and \mathbf{B}_s^w , respectively.

For the face parsing loss, we first feed \mathbf{B}_s^m and \mathbf{B}_s^w to the face parsing network to get corresponding parsing maps and then the face parsing loss is defined as:

$$\mathcal{L}_{fap} = \|\mathcal{G}_p(\mathbf{B}_s^m) - \mathcal{G}_p(\mathbf{B}_s^w)\|_2^2. \quad (10)$$

By combining the above the two terms, the loss function for the face alignment network becomes:

$$\mathcal{L}_{FAN} = \mathcal{L}_{lam} + \lambda_a \mathcal{L}_{fap}, \quad (11)$$

where $\lambda_a = 0.8$.

3.2.4 Local refinement network

To suppress the artifacts from the output of the siamese estimation networks (the second and third images shown in Figure 4) and better preserve the facial components, the main image after the siamese estimation networks \mathbf{B}_s^m (3 channels), the warped guided image (3 channels), and the probability maps of facial label (11 channels) are concatenated into a 17 channel tensors as the input to the local refinement network as:

$$\mathbf{B}^m = \mathcal{G}_o([\mathbf{B}_s^m, \mathbf{B}_s^w, \mathcal{G}_p(\mathbf{B}_s^m)]), \quad (12)$$

where \mathcal{G}_o is the local refinement network, \mathbf{B}^m denotes the final estimated results, and $\mathcal{G}_p(\mathbf{B}_s^m)$ is the probability maps of facial labels corresponding to \mathbf{B}_s^m .

Statistic identity loss. Existing reflection removal methods always aim at estimating the recovered images with higher PSNR [17] and/or SSIM values [30] by using different pixel-wise loss functions. Though these pixel-wise loss

functions are simple to calculate, the recovered face images estimated by them may have a larger difference from the ground truth since the face similarity is more properly defined in a compact feature space rather than the image pixel space [38]. Two perceptually indistinguishable face images with high SSIM values still have quite obvious feature-level differences [38].

To solve this problem, existing face restoration methods [25, 4] adopt the perceptual loss to measure the high-level feature similarity as:

$$\mathcal{L}_i = \sum_l \|\mathcal{F}_l(z^*) - \mathcal{F}_l(z)\|_2^2, \quad (13)$$

where \mathcal{F}_l denotes the l -th layer features from a pre-trained loss network (*e.g.*, VGG16 [27]) and z^* and z denote the estimate images and targets, respectively.

However, Equation (13) can only calculate the first-order statistics of the feature level differences. Previous methods [15, 12] have shown the important roles of the higher-order statistics in different tasks. Based on the perceptual loss used by previous methods [25, 4] in Equation (13), we propose a statistic identity loss to measure the feature level similarity on the basis of maximum mean discrepancy (MMD) in the local refinement network. As a kind of distribution divergence measurement derived from kernel embedding, MMD can measure the similarity of two distributions based on all-order moments as used in the two-sample testing problem [14]. Given two images z and z^* , MMD is defined as:

$$\text{MMD}(z, z^*) = \|\mu_{\mathbb{P}}(z) - \mu_{\mathbb{P}}(z^*)\|_{\mathcal{H}}, \quad (14)$$

where \mathcal{H} denotes the Hilbert space and $\mu_{\mathbb{P}}$ is defines as:

$$\mu_{\mathbb{P}} := \mu(\mathbb{P}) = \mathbb{E}_{z \sim \mathbb{P}}[\phi(\cdot)] = \mathbb{E}_{z \sim \mathbb{P}}[k(z, \cdot)]. \quad (15)$$

Here, $\phi : \mathbf{R}^d \rightarrow \mathcal{H}$ is a feature map, and $k(\cdot, \cdot)$ is the kernel function induced by $\phi(\cdot)$. Combining these, our statistic identity loss becomes:

$$\mathcal{L}_{sti} = \left\| \frac{1}{N_{z^*}} \phi(\mathcal{F}_l(z^*))^\top \mathbf{1}_{z^*} - \frac{1}{N_z} \phi(\mathcal{F}_l(z))^\top \mathbf{1}_z \right\|_F^2, \quad (16)$$

where $\mathbf{1}_{z^*}$ and $\mathbf{1}_z$ are all-one vectors with the size N_{z^*} and N_z , respectively. $\frac{1}{N_{z^*}} \phi(\mathcal{F}_l(z^*))^\top \mathbf{1}_{z^*}$ and $\frac{1}{N_z} \phi(\mathcal{F}_l(z))^\top \mathbf{1}_z$ are the empirical measure [8] of $\mu_{\mathbb{P}}(z^*)$ and $\mu_{\mathbb{P}}(z)$, respectively.

Local structural facial loss. Since human vision is more sensitive to the key components (*e.g.*, eyes, lips, and mouths) [25], instead of solely minimizing the global loss on the whole face image, we use the local structural facial loss similar to [25] to better preserve the facial information as follows:

$$\mathcal{L}_s(\mathbf{B}^m, \mathbf{B}^*) = \sum_{k=1}^K \|M_k(\mathcal{G}_p(\mathbf{B}_s^m))(\mathbf{B}^m - \mathbf{B}^*)\|_1, \quad (17)$$

where $M_k(\cdot)$ denotes the binary operation. We impose the local structural facial loss on eyebrows, eyes, noses, lips, and teeth.

Then the loss functions for the local refinement becomes:

$$\mathcal{L}_{\text{LRN}} = \lambda_o \mathcal{L}_1(\mathbf{B}^m, \mathbf{B}^*) + \alpha_o \mathcal{L}_s(\mathbf{B}^m, \mathbf{B}^*) + \beta_o \mathcal{L}_{sti}(\mathbf{B}^m, \mathbf{B}^*), \quad (18)$$

where $\lambda_o = 1.5$, $\alpha_o = 0.5$, and $\beta_o = 5$.

3.2.5 Overall loss function

Combining \mathcal{L}_{SEN} in Equation (5), \mathcal{L}_{FAN} in Equation (11), and \mathcal{L}_{LRN} in Equation (18), our overall loss function for training is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{SEN}} + \mathcal{L}_{\text{FAN}} + \mathcal{L}_{\text{LRN}}. \quad (19)$$

3.3. Implementation and training details

We have implemented our model using PyTorch. The complete training processes of our network can be divided into two stages: 1) We train the siamese estimation networks, the face parsing network, and the face alignment network separately to convergence. 2) We fix the face parsing network and then combine them with the local refinement network, and the entire network is fine-tuned again, which grant more opportunities to cooperate accordingly. The landmarks in Section 3.2.3 are obtained by using a pretrained landmark estimation network based on the MobileNet [9]. We adopt the pretrained LightCNN [33] as the face recognition model used in the statistics identity loss of Section 3.2.4. The learning rate for whole network training is set to 5×10^{-5} for the first 30 epochs and then decreases to 5×10^{-6} .

4. Experiments

We first compare the visual quality and quantitative errors of our method and state-of-the-art reflection removal approaches. We also conduct a user study to investigate how each method improves human perception. Then, another experiments on face identity recognition are conducted to investigate whether our proposed method can contribute to the high-level face recognition algorithms. At last, we conduct an ablation study to verify the effectiveness of each component in our network.

4.1. Comparison with the state-of-the-arts

We compare our method with state-of-the-art reflection removal methods, including ZN18 [39], CRRN [30], CEILNet [6], NR17 [2], and WS16 [32]. For fair comparison, we use the released codes of the above methods and train all models with the same training dataset for the data-driven methods (CRRN [30], CEILNet [6], and ZN18 [39]).

Visual quality comparison. We first show examples of recovered reflection-free face images by our method and other five methods in Figure 5 to check their visual quality. In these examples, our method removes reflections more effectively and recovers the details of the face images more clearly. All the non-learning based methods (NR17 [2] and WS16 [32]) cannot remove the non-transmitted reflections effectively and also downgrade the visual quality of the regions not covered by reflections. Though the data-driven based methods performs much better than the non-learning based methods, the final estimated results still remain visible artifacts and some key face components are also wrongly estimated (*e.g.*, ZN18 [39] in the second examples). CRRN [30] and CEILNet [6] cause serious color degradation in the estimated results. It is mainly due to linear dependency between the mixture image and background image of their image formation models.

Quantitative comparison. Since we consider the image capturing in a dynamic scenario, it is difficult to obtain the well-aligned ground truth like previous methods [29, 35]. Thus, the widely used error metrics (*e.g.*, PSNR and SSIM) are not suitable for our evaluations due to the lack of well-aligned ground truth. Instead, we evaluate the performances from the feature domains by using the high-level facial information. We use the OpenFace toolbox [1] to compute the identity distance between the ‘ground truth’ face images and different results obtained by using identity error defined as: $\mathbf{E}_{\text{Id}} = \|\mathcal{F}_E(\mathbf{B}) - \mathcal{F}_E(\mathbf{B}^*)\|_2^2$, where \mathcal{F}_E denotes the face recognition model used in the evaluations.

From the results shown in Figure 6, our method achieves the best identity scores, which demonstrates that the proposed method preserves the face identity well. The two non-learning methods WS16 [32] and NR17 [2] achieves even worse results than the baseline, which are consistent with the observations in the visual quality comparisons. The results of deep learning based methods are much better than that of the non-learning based methods. However, since CEILNet cannot well recover the color information, its performance also cannot beat CRRN [30] and ZN18 [39]. CRRN [30] and ZN18 [39] achieves similar performances. However, due to the wrongly recovered face components, its average scores is also worser than our proposed methods.

Human perception evaluations. To investigate how each method improves human perception on reflection-removed results, we conduct another experiments based on the user study scores. We invite 30 participants to judge all images in our evaluation dataset. The participants are required to give three rankings for different results. From the results shown in Figure 6, nearly 80% percent of our images are given the first rank, which are best among all methods. The two non-learning based methods generally fails on almost all images. The other three learning based methods



Figure 5. Examples of reflection removal results on the evaluation dataset, compared with CEILNet [6], CRRN [30], ZN18 [39], NR17 [2], and WS16 [32]. More results can be found in the supplementary materials.

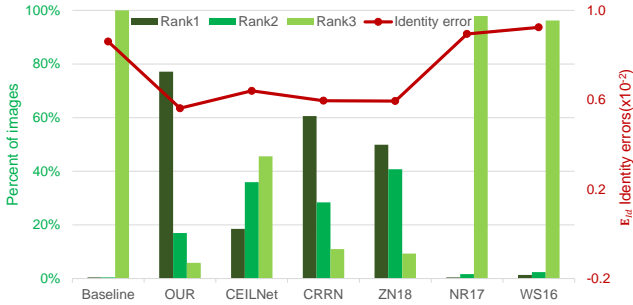


Figure 6. The human perception study and quantitative comparisons in terms of the Identity errors on the baseline, our method, CRRN [30], ZN18 [39], CEILNet [6], NR17 [2], and WS16 [32]. For the human perception study, we use the input mixture image as the baseline. For the quantitative comparison, lower identity error values are better and we use the errors between the ‘ground truth’ and the input mixture image as the baseline.

(CEILNet [6], CRRN [30], ZN18 [39]) perform much better than the non-learning based methods. The ranking of how each method performs in human perception evaluation (the higher rank-1 score the better) is generally consistent with quantitative comparison (the lower error the better).

Face recognition evaluations. The identity errors and human perception study in Figure 6 partly reveals the network ability of preserving the face identity information. In order to fully investigate whether our method can improve the accuracy of machine vision algorithms, we evaluate our estimated results in the task of face recognition. Given a probe face example, the goal of recognition is to find an

Table 1. Quantitative evaluation results using four different error metrics, and compared with FY17[6], NR17 [2], WS18 [28], and LB14 [17].

	Top-1	Top-3	Top-5	Top-10
Baseline	7.95%	2.5%	12.5%	15.91%
Ours	53.41%	69.32%	76.14%	85.23%
CRRN [6]	52.27%	60.23%	67.05%	72.73%
ZN18 [2]	50.00%	65.91%	72.73%	80.68%
CEILNet [28]	40.91%	59.09%	68.18%	80.68%
NR17 [17]	1.14%	9.09%	3.41%	3.41%
WS16 [32]	6.82%	9.09%	1.25%	1.82%

example from the gallery set that belongs to the same identity [18]. We randomly select 575 identities from the LFW dataset [13] and then merge it with the identities in our evaluation dataset to form an evaluation dataset with roughly 600 identities. Each identity has roughly the same amount of images in each set.

We use the Top-1, Top-3, Top-5, and Top-10 recognition accuracy to evaluate the performances. From the results shown in Table 1, our method achieves the highest recognition accuracy than all other methods. The non-learning based methods can not effectively increase the recognition rate and their results are even lower than the baseline. The learning based methods achieve much better results. However, the artifacts observed in Figure 5 downgrade their performances. ZN18 [39] achieves similar performances in the Top-1 part. However, the higher scores among other parts prove the effectiveness of our proposed method.



Figure 7. Examples of our complete model against our model with only the siamese estimation network (SEN), our model without the adversarial loss (AL) and local context loss (CL), our model without prior embedding (PE), and our model without the identity loss (IL).



Figure 8. Extreme examples with diverse reflections and degraded face color, compared with CRRN [30], CEILNet [6], ZN18 [39], and NR17 [2].

Table 2. Identity errors of our complete model against our model with only the siamese estimation network (SEN only), the model without the adversarial loss and local context loss (W/o AL and CL), the model without the identity loss (W/o IL), and the model without the prior embedding (W/o PE). Lower value is better.

Ours	SEN only	W/o AL and CL	W/o IL	W/o PE
0.562	0.592	0.611	0.598	0.606

4.2. Network analysis

Our framework consists of four parts, *i.e.*, the siamese estimation network (SEN), the face alignment network (FAN), the face parsing network (FPN), and the local refinement network (LRN). In this section, we have conducted several experiments to further analyze the contributions of the guided reflection removal framework and the influence of different loss functions.

The first one is to show the effectiveness of the guided reflection removal framework. We conduct this experiment by only keeping one branch of the SEN. As discussed in Section 3.2.1, this setting can be regraded as the straightforward single-image approach to solve this problem, which still contains obvious artifacts in the final estimated results as shown in Figure 7. The identity loss in Table 2 also proves this phenomenon, where it has relatively poor performance when compared with the complete model. Then, another experiment is conducted to verify the concepts leveraged from the image inpainting technique by removing the local context loss and adversarial loss. The two loss functions aim at estimating the missing information occluded by the non-transmitted reflections. From the results shown in Figure 7, without the two loss functions, the network fails to estimate the facial components occluded by the reflections.

Another experiment is to evaluate the contributions from the identity loss in the local refinement network. We train a network by removing the two loss functions. From the results shown in Table 2 and Figure 7, though the key components are successfully recovered, the color inconsistency between the regions with and without reflections are also very obvious. The lower face identity distance also prove its weakness. Then, we remove the prior embedding mechanism in the LRN, where the input to LRN reduces to a 6-channel tensors. From the results shown in Figure 7 and Table 2, the performances are similar to the results obtained by using the model without the identity loss.

5. Conclusion

We propose and solve the face image reflection removal problem in this paper. Different from the general scenes, the special properties of face images pose challenges on non-transmitted reflection and face identity feature recover for face reflection removal. To address these issues, we first leverage ideas from image inpainting to recover the key facial components occluded by reflections, and we then utilize the guided removal framework, prior embedding and statistics identity loss to better recover the important facial features. Based on the newly collected training dataset, our framework achieves better performances than existing methods on the proposed evaluation dataset.

Limitations. The performances of our method may drop when the reflections on faces become non-uniform as shown in Figure 8. However, even in this situation, our method still outperforms other methods. Another limitation of our work is from the evaluation dataset, as it is difficult to obtain the

face images with different identities. Though we have tried our best to cover more realistic scenarios, the diversity of our evaluation dataset is still limited. Our current dataset is suitable for a proof-of-concept purpose, and we are working on increasing the diversity of our evaluation dataset by including more identities and more challenging scenes (*e.g.*, the images capture by surveillance cameras).

References

- [1] B. Amos, B. Ludwiczuk, M. Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 2016.
- [2] N. Arvanitopoulos, R. Achanta, and S. Susstrunk. Single image reflection suppression. In *Proc. CVPR*, 2017.
- [3] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proc. CVPR*, 2018.
- [4] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. *arXiv preprint arXiv:1712.04695*, 2017.
- [5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Springer IJCV*, 88:303–338, 2010.
- [6] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf. A generic deep architecture for single image reflection removal and image smoothing. *Proc. ICCV*, 2017.
- [7] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. Van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. *arXiv preprint arXiv:1504.06852*, 2015.
- [8] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample problem. In *Proc. NIPS*, 2007.
- [9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [10] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Proc. NIPS*, 2015.
- [11] J. Y. Jason, A. W. Harley, and K. G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Proc. ECCV*, 2016.
- [12] P. Koniusz, F. Yan, P.-H. Gosselin, and K. Mikolajczyk. Higher-order occurrence pooling for bags-of-words: Visual concept detection. *IEEE TPAMI*, 39(2):313–326, 2017.
- [13] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua. Labeled faces in the wild: A survey. In *Proc. NIPS*. 2016.
- [14] H. Li, S. J. Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *Proc. CVPR*, 2018.
- [15] P. Li, J. Xie, Q. Wang, and W. Zuo. Is second-order information helpful for large-scale visual recognition. In *Proc. ICCV*, 2017.
- [16] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang. Learning warped guidance for blind face restoration. *Proc. ECCV*, 2018.
- [17] Y. Li and M. S. Brown. Single image layer separation using relative smoothness. In *Proc. CVPR*, 2014.
- [18] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *Proc. CVPR*, 2017.
- [19] D. Lin and X. Tang. Quality-driven face occlusion detection and recovery. In *Proc. CVPR*, 2007.
- [20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, 2014.
- [21] C. Liu, H.-Y. Shum, and W. T. Freeman. Face hallucination: Theory and practice. *Springer IJCV*, 75(1):115–134, 2007.
- [22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. ICCV*, 2015.
- [23] J. Pan, Z. Hu, Z. Su, and M.-H. Yang. Deblurring face images with exemplars. In *Proc. ECCV*, 2014.
- [24] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proc. CVPR*, 2016.
- [25] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang. Deep semantic face deblurring. *arXiv preprint arXiv:1803.03345*, 2018.
- [26] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman. Reflection removal using ghosting cues. In *Proc. CVPR*, 2015.
- [27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [28] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, W. Gao, and A. C. Kot. Region-aware reflection removal with unified content and gradient priors. *IEEE TIP*, 2018.

- [29] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot. Benchmarking single-image reflection removal algorithms. In *Proc. ICCV*, 2017.
- [30] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot. CRRN: Multi-scale guided concurrent reflection removal network. In *Proc. CVPR*, 2018.
- [31] R. Wan, B. Shi, A. Tan, and A. C. Kot. Sparsity based reflection removal using external patch search. In *Proc. ICME*.
- [32] R. Wan, B. Shi, A. H. Tan, and A. C. Kot. Depth of field guided reflection removal. In *Proc. ICIP*, 2016.
- [33] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE TIFS*, 13(11):2884–2896, 2018.
- [34] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman. Video enhancement with task-oriented flow. *arXiv preprint arXiv:1711.09078*, 2017.
- [35] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman. A computational approach for obstruction-free photography. *ACM TOG*, 34(4):79, 2015.
- [36] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.
- [37] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE SPL*, 23(10):1499–1503, 2016.
- [38] S. Zhang, R. He, Z. Sun, and T. Tan. Demeshnet: Blind face inpainting for deep meshface verification. *IEEE TIFS*, 13(3):637–647, 2018.
- [39] X. Zhang, N. Ren, and Q. Chen. Single image reflection separation with perceptual losses. In *Proc. CVPR*, 2018.