IEEE Open Journal of Signal Processing

Received 22 August 2023; revised 28 November 2023; accepted 29 November 2023. Date of publication 9 January 2024; date of current version 30 January 2024. The review of this article was arranged by Associate Editor E. Vera.

Digital Object Identifier 10.1109/OJSP.2024.3351472

Face Reflection Removal Network Using Multispectral Fusion of RGB and NIR Images

HUI LAN, ENQUAN ZHANG, AND CHEOLKON JUNG (Member, IEEE)

School of Electronic Engineering, Xidian University, Xi'an 710071, China

CORRESPONDING AUTHOR: CHEOLKON JUNG (email: zhengzk@xidian.edu.cn).

This work was supported by the National Natural Science Foundation of China under Grant 62111540272.

ABSTRACT Images captured through glass are usually contaminated by reflections, and the removal of them from images is a challenging task. Since the primary concern on photos is face, the face images with reflections annoy viewers severely. In this article, we propose a face reflection removal network using multispectral fusion of color (RGB) and near infrared (NIR) images, called FRRN. Due to the different spectral wavelengths of visible light [380 nm, 780 nm] and near infrared [780 nm, 2526 nm], NIR cameras are not sensitive to the visible light and thus NIR images are less corrupted by reflections. NIR images preserve structure information well and can guide the restoration process from reflections on the RGB images. Thus, we adopt multispectual fusion of RGB and NIR images for reflection removal from a face image. FRRN consists of one encoder model (contextual encoder model (CEM)) and two decoder models (NIR inference decoder model (NIDM) and image inference decoder model (IIDM)). CEM captures features from shallow to deep layers on the scene information while suppressing the sparse reflection component. NIDM infers NIR image to facilitate multi-scale guidance for reflection removal, while IIDM estimates the transmission layer with the guidance of NIDM. Besides, we present the reflection confidence generation module (RCGM) based on Laplacian convolution and channel attention-based residual block (CARB) to represent the reflection confidence in a region for reflection removal. To train FRRN, we construct a large-scale training dataset with face image and reflection layer (RGB and NIR images) and its corresponding test dataset using JAI AD-130 GE camera. Various experiments demonstrate that FRRN outperforms state-of-the-art methods for reflection removal in terms of visual quality and quantitative measurements.

INDEX TERMS Convolutional neural networks, deep learning, reflection removal, multispectral fusion, near infrared.

I. INTRODUCTION

Reflections from glass significantly degrade the image quality by obstructing, deforming and blurring the scene [1]. Since face is the primary concern in images, face images are frequently captured by various imaging devices in our daily life with a high quality requirement. When face images are captured through glass, they are inevitably contaminated by reflections. Different from general objects or scenes, faces contain the specific priors awarded by humans. A slight reflection distortion may significantly annoy human visual perception [2]. Moreover, the reflections affect the performance of many computer vision tasks such as face recognition and visual surveillance by the lost or distorted facial features. The degradation caused by reflection is different from the degradation caused by blur due to the combination of two different scenes. Thus, it is difficult to restore image quality by sharpening based method [3], [4], [5]. Therefore, it is required to remove reflections and enhance the quality of face images. In general, reflection in an image can be formulated as follows [6], [7], [8]:

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

where α and β are the mixing coefficients, and $\mathbf{I}$, $\mathbf{B}$, and $\mathbf{R}$ are the reflection-contaminated RGB image, transmission layer, and reflection layer, respectively. Since $\mathbf{B}$ and $\mathbf{R}$ are estimated from $\mathbf{I}$, the reflection removal is an ill-posed problem that has many or even infinite solutions.

FIGURE 1. Face reflection removal results by Face Reflection Removal Network (FRRN). Left to right: Input RGB images contaminated by reflection, input NIR images, the reflection removal results by FRRN, and the ground truth. Obvious reflections in the input RGB images are less visible in the input NIR images. Thus, NIR images contain more structure information, which can be used to guide the RGB image restoration process.

Up to the present, many deep learning approaches to reflection removal have been proposed by researchers and have achieved outstanding performance [6], [7], [9], [10]. However, most of them are designed to solve reflections from general scenes which are relatively weak. Thus, for face images with strong reflections, they remain artifacts in face images after reflection removal. With the popularity of near-infrared (NIR) cameras, non-professional users (e.g. for smartphone users, Huawei P30 Pro and Honor Magic2 deploy such lenses) have more access to take NIR images. Due to the different spectral wavelengths of visible light and NIR, i.e. [380 nm, 780 nm] and [780 nm, 2526 nm], NIR images are more robust to the reflection than RGB images in the scene [11]. That is, the reflection of NIR light is weaker than that of visible light on most object surfaces, and thus the image degradation caused by the reflection can be reduced in NIR images. Moreover, in low light condition, the NIR camera has a high sensitivity to the NIR spectral band, which effectively captures the NIR radiance of the target object and generates better textures than the visible light camera. An example is illustrated in Fig. 1, which presents the reflection-suppression property of NIR imaging without color.

In this article, we propose a face reflection removal network using multi-spectral fusion of RGB and NIR images, called FRRN. To deal with the face reflection removal problem, we exploit the feature discrepancy between NIR and RGB images caused by the different sensitivity to reflection. Fig. 1 shows two face reflection removal results by FRRN. The input RGB images are contaminated by reflection, while the input NIR images contain little reflection because the PC monitor only emits visible light. NIR images contain accurate structure information of the scene, thus they can guide the RGB image restoration process. Thus, the obvious reflections in the input RGB images are successfully removed by FRRN. Moreover, we generate a reflection confidence map based on Laplacian convolution and channel attention-based residual block (CARB) to represent the reflection degree in a region. FRRN consists of four main components: 1) Context encoder module (CEM) to extract features from shallow to deep layers suppressing sparse reflection residuals, 2) NIR inference decoder module (NIDM) to exploit reflection-suppressed information from NIR images, 3) Image inference decoder module (IIDM) to distinguish the transmission layer for the reflection layer with the guidance of NID, and 4) Reflection confidence generation module (RCGM) to obtain the reflection confidence map $\mathbf{C}$, which indicates the intensity of reflection in $\mathbf{I}$. Experimental results show that FRRN successfully removes reflections in face images and outperforms state-of-the-art methods in both visual quality and quantitative measurements. Fig. 2 illustrates the whole architecture of the proposed FRRN.

Our major contributions are summarized as follows:

- We propose a face reflection removal network using multi-spectral fusion of RGB and NIR images, called FRRN. Since NIR images are more robust to the reflection with clearer textures than RGB images in the scene, we adopt the multi-spectral fusion of RGB and NIR images for reflection removal in reflection-contaminated image. To our knowledge, this is the first work of applying multi-spectral fusion to the face reflection removal.
- We generate a reflection confidence map based on Laplacian convolution and channel attention-based residual block (CARB), named the reflection confidence generation module (RCGM). Laplacian convolution extracts the edges caused by strong reflection and represents the reflection confidence in a region. CARB retains feature channels of importance. The reflection confidence map helps FRRN to extract features in the reflection layer while recovering the high-quality transmission layer.
- We build a network architecture for FRRN that consists of one context encoder module (CEM) branch and two decoder branches (NIR inference decoder module (NIDM) branch and image inference decoder module (IIDM) branch). CEM extracts multi-scale and multi-spectral features from the input RGB image, NIR image and reflection confidence map to suppress sparse reflection residuals. NIDM captures reflection-suppressed information through NIR inference, while IIDM branch distinguishes the transmission layer from the reflection layer with the guidance of NIDM.
- We generate a real dataset for training and testing that contains both transmission and reflection layers of RGB and NIR images. We use JAI AD-130 GE camera to simultaneously capture RGB and NIR image pairs consisting of transmission and reflection layers. We synthesize 10,000 images for training and 300 images for testing. Moreover, we generate 40 real images for testing.

II. RELATED WORK

A. REFLECTION REMOVAL

Traditional Methods: Due to the ill-posed nature of the reflection removal problem, traditional methods have employed image priors to contain the special property of the transmission and reflection layers. Li et al. [12] and Nikolas et al. [13]

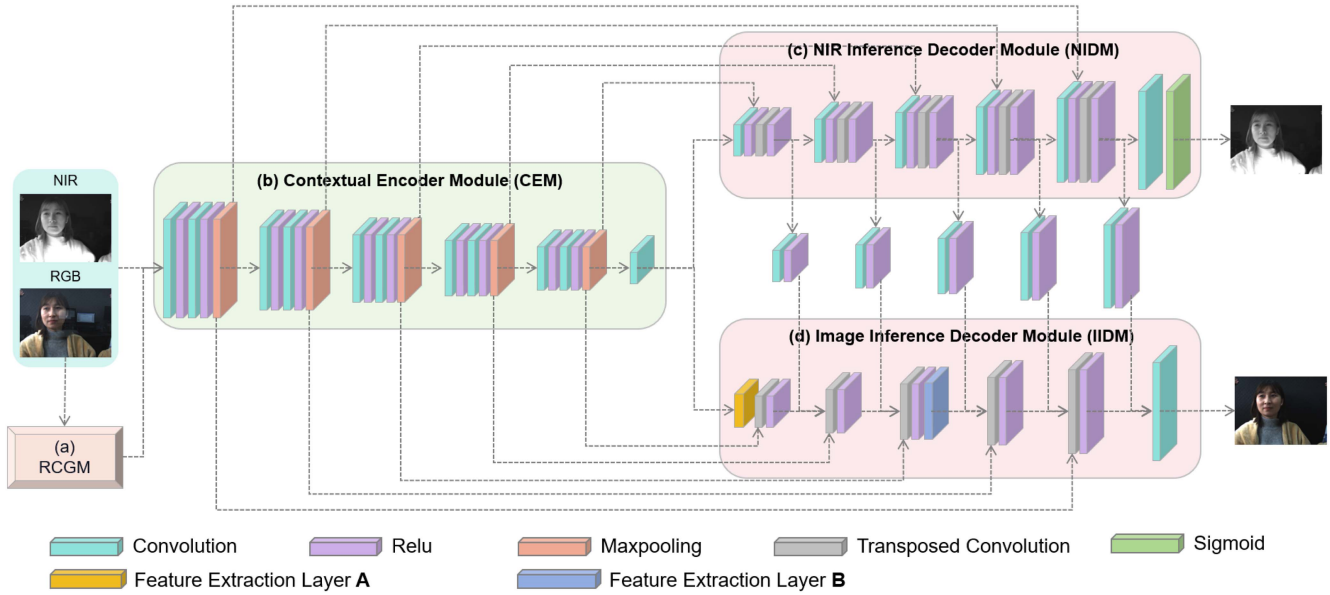


FIGURE 2. **Whole architecture of the proposed FRRN. FRRN contains four main components: 1) Context encoder module (CEM) to extract features from shallow to deep layers suppressing sparse reflection residuals, 2) NIR inference decoder module (NIDM) to capture reflection-suppressed information in the NIR inference process, 3) Image inference decoder module (IIDM) to distinguish the transmission layer from the reflection layer with the guidance of NIDM, and 4) Reflection confidence generation module (RCGM) to obtain the reflection confidence map.**

made use of the different blur levels of the transmission and reflection layers. Levin et al. [14] leveraged the sparsity prior of gradients when decomposing an image into reflection and transmission layer. However, they relied on manual labels for the transmission and reflection edges, which were quite labor-intensive and might fail in textured regions. Shih et al. [15] used Gaussian mixture model (GMM) patch prior to remove reflections causing ghosting effects. The handcrafted priors adopted by them were based on the relationship between the transmission and reflection layers. However, different blur levels [12], [16] were not suitable for the general scenes, especially when they were weakly observed. Some methods used a set of images taken from different viewpoints to remove reflections [17], [18]. Xue et al. [18] exploited motion cues between the transmission and reflection layers from multiple viewpoints and assumed the glass was closer to the camera. The projected motion of the two layers were different due to the parallax. The motion of each layer could be represented by parametric models, such as translative motion [19], affine transformation and homography [18]. It was also beneficial to separate the gradients of the original image into reflection and transmission gradients. The separation provided gradient-domain constraints to facilitate the layer separation in single image reflection removal (SIRR) [20]. Detecting reflection dominated regions in an image was achieved by depth-of-field analysis [16], [21].

Through the combination of the motion and traditional cues, the non-learning based methods using the multiple images as the input can show more reliable results when the input data are appropriately prepared. However, the requirement for special facilities of capturing limits such methods for practical use, especially for mobile devices or images downloaded from the Internet.

Deep Learning Based Methods: Recently, deep learning has achieved outstanding performance in both high-level and low-level vision tasks. Its outstanding ability of feature extraction provides promising solutions to reflection removal. Deep learning based methods are data-driven and try to learn task-specific features to solve the SIRR problem in the feature space [1]. Fan et al. [6] proposed a deep neural network, named CEIL-Net, to first regress an edge map and then reconstruct the transmission layer based on it. Paramanand et al. [22] proposed a two-stage deep learning approach to learn edge features for reflections using light field data. Zhang et al. [8] first utilized a generative model to better learn a mapping from $\mathbf{I}$ to the clean images. Yang et al. [23] proposed a bidirectional network (BDN), also a two-stage network, that the reflection layer in the first stage was used as auxiliary information to guide reconstruction of the transmission layer in the second stage. Li et al. [9] proposed a recurrent network based on LSTM [24] units (IBCLN) to refine the predicted reflection and transmission layers iteratively. Chang et al. [25] used pairs of flash and no-flash images to remove reflection. Hong et al. [11] proposed a two-stream NIR image guided reflection removal network to introduce NIR images into the reflection removal pipeline. Face reflection removal has also been studied in [26]. In the data-driven approach, the dataset construction is critical to the success of deep learning-based reflection removal methods. To this end, Jin et al. [27] proposed multiple data generation models. Wen et al. [28] proposed SynNet to generate images with reflections beyond linearity. Wei et al. [29] introduced an alignment-invariant

loss to utilize misaligned images as the real-world training dataset. In recent years, Kim et al. [30] proposed a physics-based rendering method to render images with reflections, and considered the reflection and refraction of light in glasses to obtain realistic rendering results. Existing methods are mostly designed for general scenes, but face image reflection removal needs to recover face details more precisely. When strong reflection contains in face images, they cannot effectively remove them and might cause blur of the face details.

B. NEAR-INFRARED IMAGING

Due to the low power consumption and low interference properties compared to RGB images, near-infrared (NIR) images are increasingly used in machine vision tasks. NIR imaging methods can be divided into two categories: active NIR imaging and passive NIR imaging. The active NIR imaging systems use an actively emitted light source such as a laser or infrared LED to illuminate the target and obtain an image. They can control the intensity and direction of the light, thereby reducing or suppressing the reflection of the target itself. Active NIR imaging has been widely applied to 3D sensing devices such as Kinect V1 and V2 for geometry refinement [31] and robot navigation [32]. Sun et al. [33] used shape and edge information contained in depth images from Kinect V2 to guide reflection removal which has limited capability of recovering transmission details due to the texture-less appearance of depth images. However, the passive NIR imaging systems rely on its own radiation of the target that does not require additional light sources and cannot control the reflection degree of the target [34], [35], [36]. In this work, we capture the paired RGB and NIR images simultaneously by JAI AD-130 GE camera (passive RGB and NIR imaging) to contain rich textures due to the reflection-suppression property as shown in Fig. 1.

III. PROPOSED METHOD

A. DATA PREPARATION

Real-world image datasets play an important role in studying physics-based computer vision tasks [37], [38]. Although the reflection removal problem has been studied for more than decades, publicly available datasets are rather limited. However, the data-driven methods need a large-scale dataset to learn the reflection property in images. Since the previous work has mainly focused on arbitrary scenes, the transmission image **B** can be obtained from generic image datasets (e.g., **PASCAL** [39] or **COCO** [40]). Existing benchmark reflection removal datasets (e.g., **SIR^2**[41]) were constructed in this way, thus they are not suitable for our face reflection removal task due to their scenery diversity. Although there are many face image datasets (e.g., **CELEBA** [42] and **CASIA** webface dataset [43]), they are also not suitable for our task since they mostly consider a fixed facial pose and contain only RGB images. Thus, we construct a large-scale face image dataset for training and its corresponding evaluation dataset using JAI AD-130 GE camera. As shown in Fig. 3, JAI AD-130 GE camera is a 2-CCD multispectral prism camera that simultaneously captures two images with different spectra in a single camera: one visible color image from 400–700 nm and one near infrared (NIR) image from 750–900 nm. It adopts 2 pieces of Sony 1/3 in ICX447 CCD sensor: Bayer color CCD and NIR black/white CCD. Through the prism spectroscopic technology, it can project the coaxial light incident from the same lens to two CCD sensors, thus RGB and NIR images from two CCDs have a completely consistent viewing angle. Thus, it does not need calibration. However, if RGB and NIR images are not registered, [44], [45], [46] can be used to identify the edges of reflection regions. Under the full resolution of 1296 × 966, the frame rate of each CCD can reach 31fps. Thus, we use this camera to capture the face reflection-contaminated RGB image **I**, reflection layer **R** and their corresponding NIR images **N**. The reflection images are taken by putting a black piece of article behind the glass while moving the camera and the glass around, which is similar to [18], [41] as shown in Fig. 3. The camera is configured with varying exposure parameters and aperture sizes under a fully manual mode to capture images in different scenes. We provide some samples in Fig. 4, and our dataset has two major characteristics:

FIGURE 3. Top: JAI AD-130 GE camera and its graphical user inferface (GUI). Bottom: Image capturing process using JAI AD-130 GE camera.

Diversity: For face transmission layers, we collect 800 RGB and NIR image pairs of about 80 people in different illumination conditions and scenes (both indoor and outdoor). For reflection layers, we take them at different illumination conditions to include both strong and weak reflections, and adjust the focal lengths randomly to create different blur levels of reflection. Moreover, the reflection layers are taken from a great diversity of both indoor and outdoor scenes, e.g., streets, parks, and inside office buildings.

Scale: The whole dataset contains 800 image pairs (about 80 people) for face transmission layer and 1000 image pairs for reflection layers. Moreover, we collect 40 pairs of real reflection images with their corresponding ground truth for visual quality comparison. The resolution of the real reflection

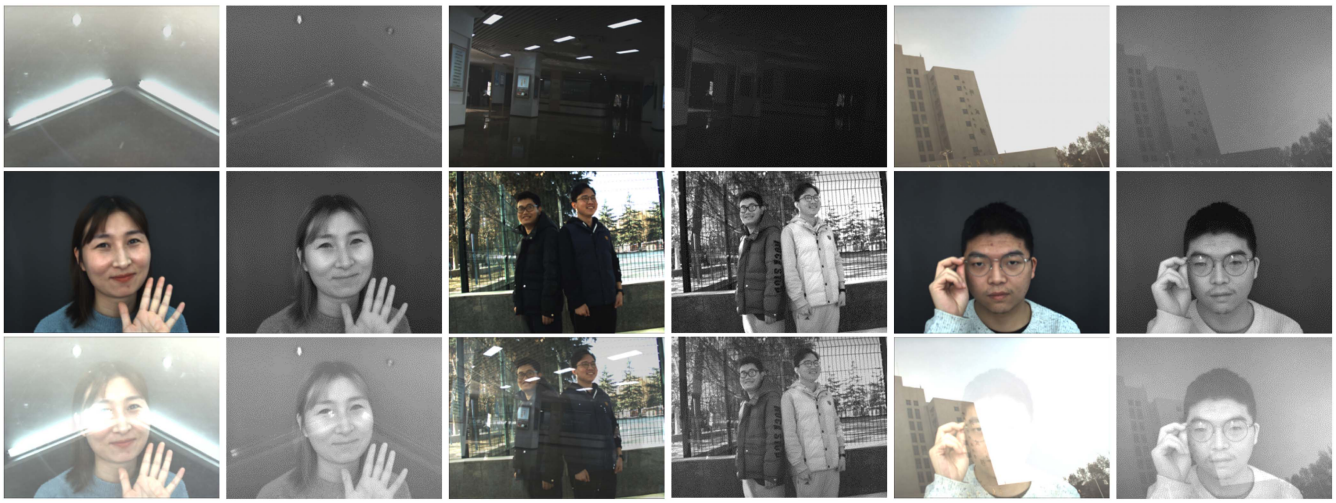


FIGURE 4. Samples of the reflection layer (first row), face transmission layer (second row) in the dataset, and their synthetic images (third row). We provide pairs of RGB and NIR images with diverse illumination conditions, focal lengths, and scenes.

images is 1296 × 966. We put a glass in front of the person to capture reflection-contaminated images. Then, we removed the glass to acquire the ground truth of real images.

It is difficult to obtain large quantities of real reflection images with the corresponding ground truth. Thus, we use (1) to generate dataset for training and testing. Our synthetic image **I** is generated by adding the reflection layer to the face transmission layer with different weighting factors. To ensure a sufficient amount of data, α and β are randomly sampled from 0.6 to 1 and 0.3 to 0.6, respectively. We construct 10,000 (10 K) image pairs for training and 300 image pairs for testing. We further augment the synthetic images with rotation and flipping as dataset pre-processing. The synthetic images cover more reflections in the real world, resulting in enhancing reflection removal and robustness of FRRN.

B. NETWORK ARCHITECTURE

Given a reflection-contaminated RGB image **I**, we aim to recover the face transmission layer **B** under the guidance of NIR image **N**. To accomplish this, we develop a multi-scale learning convolutional neural network which consists of four modules to process **I** and **N** simultaneously and to recover the **B** progressively as shown in Fig. 2. We concatnate two source images of **I** and **N** into FRRN to simultaneously produce restored images $\mathbf{B}^*$ and $\mathbf{N}^*$. We formulate the whole estimation process as follows:

$$(\mathbf{B}^*, \mathbf{N}^*) = \mathcal{F}(\mathbf{I}, \mathbf{N}, \theta) \tag{2}$$

where $\mathcal{F}$ presents the network to be trained with parameters θ, and $\mathbf{B}^*$, $\mathbf{N}^*$ are the estimated clear RGB transmission layer and clear NIR layer, respectively.

Reflection Confidence Generation Module: Compared with RGB images, NIR images are less corrupted by reflections. The information covered by reflection in RGB images can be provided by the corresponding NIR images, but for some special reflections the NIR images are even more affected

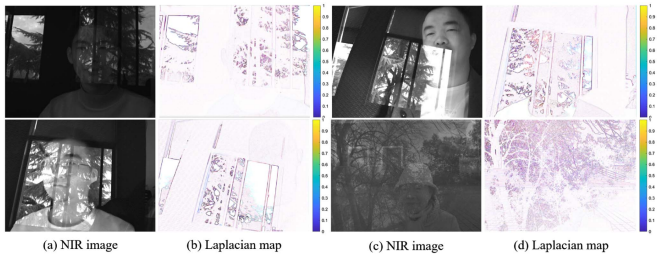


FIGURE 5. NIR images with the corresponding reflection confidence maps obtained by RCGM. Compared with NIR images, strong reflections in the reflection confidence maps are preserved well while the face edges are almost disappeared.

than RGB images. Thus, we introduce a reflection confidence generation module (RCGM) based on Laplacian operator to estimate the single channel reflection confidence map **C** from the input reflection-contaminated RGB image **I**. As shown in Fig. 5, NIR images are also severely corrupted by reflections in some situations. However, in reflection confidence maps, the reflection area is almost completely preserved, but the face edges are almost disappeared. Based on this observation, we use reflection confidence map to facilitate the reflection detection.

Contextual Encoder Module: We concatenate the reflection-contaminated RGB image **I**, NIR image **N** and a reflection confidence map **C** as input of the contextual encoder module (CEM). The three input images have the same spatial resolution. Here, **I** is 3-channel image, while **N** and **C** are the 1-channel image. Thus, the input of the CEM encoder has total 5 channels. Inspired by the perceptual loss [47], we adopt the pretrained VGG-16 network [48] for feature extraction as shown in Fig. 2. Since the input of CEM has total 5 channels, we first use two convolutional layers and one Maxpooling layer to extract initial features from the input images. Then, the other layers of CEM are completed

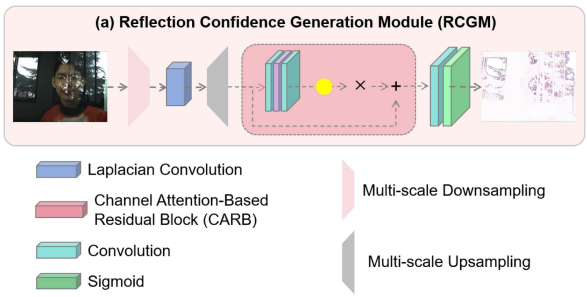
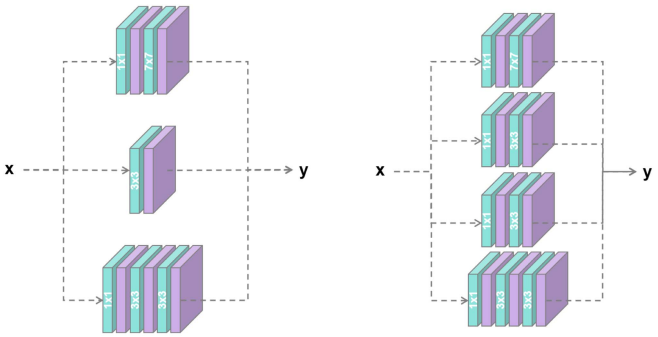


FIGURE 6. Network architecture of the reflection confidence generation module. We use convolution with Laplacian kernel to calculate Laplacian edges and channel attention based residual block (CARB) to get efficient multi-channel Laplacian features.



(a) Feature extraction layer **A** (b) Feature extraction layer **B**

FIGURE 7. Network structures of the feature extraction layers A and B. We use several parallel convolutional layers with different kernel size to extract rich features.

by the pretrained VGG-16 network. VGG16 adopts several consecutive 3×3 convolution kernels instead of larger ones. For a given receptive field, multi-layer nonlinear layers can increase the network depth to ensure that more complex features are learned, and the cost is relatively small (less parameters), and the 3×3 convolution kernel is conducive to better maintaining the image features. VGG-16 has 3 fully connected layers, which consumes a lot of computing resources. After compromise, we replace the last fully connected layer with a 3×3 convolutional layer to reduce the calculation and make it adapt to the reflection removal task. After layer-by-layer learning in CEM, the weak reflections in the input images are initially suppressed, which is helpful for the subsequent network.

NIR Inference Decoder Module: The NIR inference decoder module (NIDM) takes the LR features extracted by CEM to exploit the reflection-suppressed context information in the transmission layer and provide multi-scale guidance for the entire recovery process. NIDM is composed of convolutional layer and transposed convolutional layers with stride 2, and the final layer is an activation layer with sigmoid function. To make full use of the image details and avoid the gradient vanishing problem, the features from CEM are connected to its corresponding layers in NIDM with the same spatial resolution.

Image Inference Decoder Module: The image inference decoder module (IIDM) is utilized to extract high-level semantic transmission features, and facilitate the face transmission recovery under the guidence of NIDM. The input of IIDM is the same as the input of NIDM, which is the features extracted by CEM. Inspired by CRRN [7], we use feature extraction layer-A/B layers from Inception-ResNet-v2 [49] as shown in Fig. 7, which consists of several parallel convolutional layers with different kernel size to extract rich features. Convolutional layers between NIDM and IIDM to map the features with reflections into the feature space with relatively fewer reflections. Similar to NIDM, the feature maps from CEM are concatenated to its corresponding layers in IIDM with the same spatial resolution to conserve the distinct details and avoid the gradient vanishing problem.

TABLE 1. Quantitative Measurements Among FRRN and Three State-of-The-Art Methods (RAGN [50], IBCLN [9], CORRN [51]) in Terms of PSNR, SSIM, Model Size (MS) and Runtime (RT)

Methods	$PSNR \pm SD$	$SSIM \pm SD$	MS(MB)	RT(s)
RAGN	29.2657 ± 3.9565	0.9388 ± 0.0410	60.73	**14.67**
CORRN	21.8687 ± 3.8816	0.9242 ± 0.0422	59.51	19.56
IBCLN	19.4000 ± 3.9928	0.7963 ± 0.1431	**58.75**	27.07
FRRN	$\mathbf{30.3186 \pm 3.3544}$	$\mathbf{0.9403 \pm 0.0304}$	59.76	20.51

SD represents standard deviation. We provide average PSNR, SSIM and runtime for 300 test images in the synthetic dataset. Bold numbers represent the best performance.

As shown in Fig. 6, RCGM uses convolution with Laplacian kernel to calculate Laplacian edges. We downsample the input reflection-contaminated RGB image $\mathbf{I}$ to $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$ of the original size for multi-scale Laplacian feature learning. Then, we utilize a kernel with weights initialized to be a 3×3 Laplacian kernel, denoted by $\mathbf{k}_L = [0, -1, 0; -1, 4, -1; 0, -1, 0]$, to extract Laplacian edge information from multi-scale images. After multi-scale upsampling, the Laplacian edge features are recovered to the original size. CARB focuses on more important features, and the final convolutional layer is used to obtain $\mathbf{C}$ for the reflection removal of FRRN.

C. LOSS FUNCTION

Pixel-level Loss: For the image reflection removal, the pixel-level loss is essential to recover the transmission layer $\mathbf{B}^*$. In this work, we use a $\mathcal{L}_1$ loss to measure the pixel-level similarity between the predicted $\mathbf{B}^*$ and its ground truth $\mathbf{B}$ as follows:

$$\mathcal{L}_1 = |\mathbf{B}, \mathbf{B}^*|_1 \tag{3}$$

Composition Loss: RCGM generates a reflection confidence map $\mathbf{C}$ to facilitate the reflection removal, and the quality of the confidence map largely affects the recovery results of the subsequent network. Thus, we use the composition loss for RCGM to guide the training of confidence prediction. Since $\mathbf{C}$ contains lots of information about the edges caused by strong reflection, $(1 - \mathbf{C})$ can be denoted as the weight of transmission (the value in $\mathbf{C}$ has been normalized). In (1) and

FIGURE 8. Visual quality comparison among FRRN and three state-of-the-art methods (RAGN [50], IBCLN [9], CORRN [51]) on the synthetic dataset. FRRN successfully recovers the facial details in the regions occluded by reflections that are closest to the ground truth even in extremely harsh conditions.

composition loss can be formulated as follows:

$$\begin{aligned} \mathbf{I}^* &= (1-\mathbf{C}) \circ \mathbf{B}^* + \mathbf{R}^* \\ \mathcal{L}_{\mathbf{C}} &= \mathcal{L}_{MSE}\left(\mathbf{I}, \mathbf{I}^*\right) \end{aligned} \tag{4}$$

where $\circ$ is an element-wise production operation; $\mathbf{I}^*$ is the synthesized reflection-contaminated RGB image from the estimated transmission layer $\mathbf{B}^*$ and the reflection layer $\mathbf{R}^*$, respectively. $\mathcal{L}_{MSE}$ indicates the mean squared error, and $\mathbf{R}^*$ is obtained by $(\mathbf{I}-\mathbf{B}^*)$

Structural Similarity Loss: To generate the results satisfying with the human perception, we adopt *SSIM* loss [52] to measure the similarity between predicted $\mathbf{B}^*$ and $\mathbf{N}^*$ with their corresponding ground truth. *SSIM* measures the similarity of structure, contrast and luminance between two images. We use $\mathbf{B}^*$ and $\mathbf{B}$ as an example for formulated as follows:

$$SSIM\left(B, B^*\right) = \frac{\left(2\mu_B\mu_{B^*} + C_1\right)\left(2\sigma_{BB^*} + C_2\right)}{\left(\mu_B^2 + \mu_{B^*}^2 + C_1\right)\left(\sigma_B^2 + \sigma_{B^*}^2 + C_2\right)} \tag{5}$$

where μ_B and μ_{B^*} are the means of $\mathbf{B}$ (transmission ground truth) and $\mathbf{B}^*$ (estimated transmission layer), σ_B and σ_{B^*} are the variance of $\mathbf{B}$ and $\mathbf{B}^*$, σ_{BB^*} is the corresponding covariance. Since a higher *SSIM* represents better performance, we use $\mathcal{L}_{\mathbf{SSIM}}$ to minimize the prediction cost as follows:

$$\mathcal{L}_{\mathbf{SSIM}} = 1 - \mathrm{SSIM} \tag{6}$$

Gradient Aware Loss: The main difference between the images with and without reflections can be found from their gradient level statistics, especially in the case of strong reflections. To improve the removal quality, we introduce a gradient aware loss function $\mathcal{L}_{\mathbf{Q}}$ to preserve gradient information in the predicted $\mathbf{B}^*$ and $\mathbf{N}^*$. Based on the output transmission layer $\mathbf{B}^*$ and the transmission ground truth $\mathbf{B}$, we use Sobel edge operator to yield the edge strength g_i and the orientation δ_i at each pixel i. The relative strength $G_i^{BB^*}$ and orientation value $\Delta_i^{BB^*}$ between $\mathbf{B}^*$ and $\mathbf{B}$ are defined as follows:

$$\begin{aligned} G_i^{BB^*} &= \begin{cases} \frac{g_i^{B^*}}{g_i^B}, & \text{if } g_i^B > g_i^{B^*} \\ \frac{g_i^B}{g_i^{B^*}}, & \text{if } g_i^B \le g_i^{B^*} \end{cases} \\ \Delta_i^{BB^*} &= 1 - \frac{\left|\delta_i^B - \delta_i^{B^*}\right|}{\pi/2} \end{aligned} \tag{7}$$

In the training stage, we use Sigmoid function $f(x, y)$ as smooth approximation, and rewrite (7) as follows:

$$\begin{aligned} G_i^{BB^*} &\approx f\left(g_i^{B^*}, g_i^B\right) \times \frac{g_i^B}{g_i^{B^*}} + \left(1 - f\left(g_i^{B^*}, g_i^B\right)\right) \times \frac{g_i^{B^*}}{g_i^B} \\ \Delta_i^{BB^*} &= 1 - \frac{\left|\delta_i^B - \delta_i^{B^*}\right|}{\pi/2} \end{aligned} \tag{8}$$

FIGURE 9. Visual quality comparison among FRRN and three state-of-the-art methods (RAGN [50], IBCLN [9], CORRN [51]) on the real dataset. FRRN successfully recovers the facial details in the regions occluded by reflections that are closest to the ground truth even in extremely harsh conditions.

The edge information preservation is then defined as:

$$Q_i^{BB^*} = G_i^{BB^*} \times \Delta_i^{BB^*} \tag{9}$$

Since higher $Q_i^{BB^*}$ represents better performance, we use $\mathcal{L}_{\mathbf{Q}}$ to minimize the prediction cost as follows:

$$\mathcal{L}_{\mathbf{Q}}\left(B, B^*\right) = 1 - Q_i^{BB^*} \tag{10}$$

The loss function for the transmission layer is:

$$\mathcal{L}_{\mathbf{B}} = \mathcal{L}_1(\mathbf{B}, \mathbf{B}^*) + \mathcal{L}_{\mathbf{SSIM}}(\mathbf{B}, \mathbf{B}^*) + \mathcal{L}_{\mathbf{Q}}(\mathbf{B}, \mathbf{B}^*) \tag{11}$$

the loss function for NIR layer is :

$$\mathcal{L}_{\mathbf{N}} = \mathcal{L}_{\mathbf{SSIM}}(\mathbf{N}, \mathbf{N}^*) + \mathcal{L}_{\mathbf{Q}}(\mathbf{N}, \mathbf{N}^*) \tag{12}$$

Thus, the total loss is as follows:

$$\mathcal{L}_{\mathbf{total}} = \mathcal{L}_{\mathbf{C}}(\mathbf{I}, \mathbf{I}^*) + \mathcal{L}_{\mathbf{B}} + \mathcal{L}_{\mathbf{N}} \tag{13}$$

IV. EXPERIMENTAL RESULTS

We implement FRRN using PyTorch framework. For training, CEM is based on a pretrained VGG16 model [48], which is connected with RCGM, NIDM and IIDM. The entire network is fine-tuned end-to-end, which grants the four sub-networks more opportunities to cooperate accordingly. We synthesize 10,000 image quads of the input reflection RGB/NIR and the ground truth RGB/NIR for training and 300 quads for test, and 40 quads of real reflection images for visual comparison. During the training process, all images are resized to 512×384 (height H and width W are divisible by 2^5) and randomly flipped. The initial learning rate is 1×10^{-4}, which is decayed by factor 0.8 at every two epoch. We optimize the objective function by Adam optimizer. The batch size and the number of epochs are 2 and 50, respectively. The training and test of FRRN are performed on a PC with a NVIDIA 1080ti GPU with 11 GB memory. To evaluate the performance of FRRN, we perform comparison of FRRN with state-of-the-art reflection removal methods on the dataset captured by us. We first use all image quads to evaluate both quantitative measurements and visual quality. We then conduct experiments to compare the influence of the different loss functions to the final performances with the generalization ability. Finally, we perform ablation experiments on loss function and NID to verify their effectiveness of loss functions on the performance.

A. QUANTITATIVE MEASUREMENTS

We quantitatively compare FRRN with RAGN [50], CORRN [51] and IBCLN [9] on datasets in terms of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM), and provide the results in Table 1. The numbers represent average performance for all 300 image quads. It can be observed that FRRN achieves the best performance in both PSNR and SSIM among them. The SSIM results indicate that FRRN preserves better structural information in faces than the others, while the PSNR results indicate that FRRN achieves

TABLE 2. Ablation Study on the Loss Function

Model	$PSNR \pm SD$	$SSIM \pm SD$
W/O $\mathcal{L}_{\mathbf{C}}$	26.8301 ± 3.6870	0.8527 ± 0.0283
W/O $\mathcal{L}_{\mathbf{SSIM}}$	27.2972 ± 3.0414	0.8225 ± 0.0256
W/O $\mathcal{L}_{\mathbf{Q}}$	27.7371 ± 3.3546	0.8393 ± 0.0264
W/O NIDM	27.4822 ± 3.2939	0.9017 ± 0.0259
W/O NIR and NIDM	28.0907 ± 3.5265	0.9145 ± 0.0555
FRRN	$\mathbf{30.3186} \pm 3.3544$	$\mathbf{0.9403} \pm 0.0304$

We evaluate the performance of FRRN without composition loss ($\mathcal{L}_{\mathbf{C}}$), without SSIM loss ($\mathcal{L}_{\mathbf{SSIM}}$), without gradient aware loss ($\mathcal{L}_{\mathbf{Q}}$), without NIDM and without NIR image. Bold numbers represent the best performance.

the best reconstruction in faces whose appearance is closest to the ground truth. The smallest standard deviation (SD) indicates that FRRN achieves the most stable performance in the whole test dataset. We also measure average runtime of them in Table 1, which shows that FRRN balances model size and runtime in the face reflection removal. The results verify that FRRN achieves outstanding performance in various real-world scenes.

B. VISUAL COMPARISON

Fig. 8 shows the reflection removal results for different methods on highly reflection-contaminated images. These images are from the datasets that we have synthesized. They contain large area of reflections with strong highlights so that the reflections can not be removed well by other methods. In contrast, FRRN removes most annoying reflections while keeping high-frequency details in the transmission layer. Thus, FRRN generates nearly the same results as the ground truth even in highly reflection-contaminated images. Moreover, we evaluate their performance on real test images with reflections in Fig. 9. The real dataset contains 40 reflection-contaminated RGB images with its corresponding ground truth and NIR images. We put a glass in front of the person to capture reflection-contaminated images and then remove the glass to acquire the ground truth of real images. As shown in the figure, FRRN successfully removes most reflections in images, and produces nearly the same results as the ground truth. Since the reflection confidence map provides accurate edges in the reflection layer, it helps FRRN distinguish between face transmission layer and reflection layer. Moreover, the loss function accelerates the network convergence thanks to the help of NIR image during training.

C. ABLATION STUDY

To verify the effectiveness of the loss function and the contribution of NIDM and NIR image to IIDM, we conduct several ablation studies by re-training FRRN and testing on the synthetic test dataset. Table 2 shows the ablation study results. FRRN without $\mathcal{L}_{\mathbf{C}}$ performs the worse among them, which means the reflection confidence map can effectively provide reflection information to help the network distinguish between the reflection and the face, especially in the regions with strong reflections. $\mathcal{L}_{\mathbf{SSIM}}$ and $\mathcal{L}_{\mathbf{Q}}$ are helpful for recovering the details of $\mathbf{B}^*$ and $\mathbf{N}^*$. Without the NIDM branch, the IIDM branch cannot effectively recover details in the transmission layer. If we remove the input NIR image and the NIDM branch, then FRRN is similar to the single image reflection removal network. It is difficult to estimate the accurate background layer and reflection layer from a reflection-contaminated RGB image by FRRN. Through the ablation studies, it can be seen that FRRN effectively uses the information of NIR image to remove the reflection and restore an accurate transmission layer. At the same time, the loss function based on reflection confidence generation helps FRRN to preserve more details of the face image.

V. CONCLUSION

We have proposed FRRN for face reflection removal based on the fusion of RGB and NIR images. Based on the observations that NIR images captured by NIR sensors are robust to reflections, we have built a multi-spectral CNN that consists of CEM, NIDM, IIDM and RCGM. CEM extracts multi-scale features from shallow to deep layers gradually in RGB nd NIR images while suppressing the sparse reflection component. NIDM exploits reflection-suppressed information from NIR images, while IIDM estimates face transmission layer with the guidance of NIDM features. RCGM generates a reflection confidence map to measure the reflection-dominance degree in a region. Moreover, we have used composition loss and gradient aware loss to facilitate reflection removal restoring details in the transmission layer. Various experiments on both synthetic and real images verify that NIR images are successfully used for reflection removal and FRRN outperforms state-of-the-art methods in terms of visual quality and quantitative measurements.

In the future work, we will consider more practical cases of reflections for FRRN contained in natural scenes to generate natural-looking results.

REFERENCES

[1] A. Amanlou, A. A. Suratgar, J. Tavoosi, A. Mohammadzadeh, and A. Mosavi, "Single-image reflection removal using deep learning: A systematic review," *IEEE Access*, vol. 10, pp. 29937–29953, 2022.

[2] C. Liu, H. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *Int. J. Comput. Vis.*, vol. 75, pp. 115–134, 2007.

[3] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10265–10274.

[4] L. Yu, D. Liu, H. Mansour, and P. T. Boufounos, "Fast and high-quality blindmulti-spectral image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5403417.

[5] L. Yu, D. Liu, H. Mansour, P. T. Boufounos, and Y. Ma, "Blindmulti-spectral image pan-sharpening," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 1429–1433.

[6] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proc. IEEE Conf. Comput. Vis.*, 2017, pp. 3238–3247.

[7] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "CRRN: Multi-scale guided concurrent reflection removal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4777–4785.

[8] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4786–4794.

[9] C. Li, Y. Yang, K. He, S. Lin, and J. E. Hopcroft, "Single image reflection removal through cascaded refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3565–3574.

[10] Z. Chi, X. Wu, X. Shu, and J. Gu, "Single image reflection removal using deep encoder-decoder network," 2018, *arXiv:1802.00094*.

[11] Y. Hong, Y. Lyu, S. Li, and B. Shi, "Near-infrared image guided reflection removal," in *Proc. IEEE Conf. Multimedia Expo.*, 2020, pp. 1–6.
[12] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2752–2759.
[13] N. Arvanitopoulos, R. Achanta, and S. Susstrunk, "Single image reflection suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4498–4506.
[14] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1647–1654, Sep. 2007.
[15] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3193–3201.
[16] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot, "Depth of field guided reflection removal," in *Proc. IEEE Conf. Image Process.*, 2016, pp. 21–25.
[17] K. Gai, Z. Shi, and C. Zhang, "Blind separation of superimposed moving images using image statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 19–32, Jan. 2012.
[18] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, "A computational approach for obstruction-free photography," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–11, 2015.
[19] E. Be'Ery and A. Yeredor, "Blind separation of superimposed shifted images using parameterized joint diagonalization," *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 340–353, Mar. 2008.
[20] Z. Dong, K. Xu, Y. Yang, H. Bao, W. Xu, and R. W. Lau, "Location-aware single image reflection removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2020, pp. 5017–5026.
[21] 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 Trans. Image Process.*, vol. 27, no. 6, pp. 2927–2941, Jun. 2018.
[22] P. Chandramouli, M. Noroozi, and P. Favaro, "ConvNet-based depth estimation, reflection separation and deblurring of plenoptic images," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 129–144.
[23] J. Yang, D. Gong, L. Liu, and Q. Shi, "Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 654–669.
[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[25] Y. Chang, C. Jung, J. Sun, and F. Wang, "Siamese dense network for reflection removal with flash and no-flash image pairs," *Int. J. Comput. Vis.*, vol. 128, pp. 1673–1698, 2020.
[26] R. Wan, B. Shi, H. Li, L.-Y. Duan, and A. C. Kot, "Face image reflection removal," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 385–399, 2021.
[27] M. Jin, S. Süsstrunk, and P. Favaro, "Learning to see through reflections," in *Proc. IEEE Conf. Comput. Photogr.*, 2018, pp. 1–12.
[28] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han, and S. He, "Single image reflection removal beyond linearity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3771–3779.
[29] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang, "Single image reflection removal exploiting misaligned training data and network enhancements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8178–8187.
[30] S. Kim, Y. Huo, and S.-E. Yoon, "Single image reflection removal with physically-based training images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5164–5173.
[31] G. Choe, J. Park, Y.-W. Tai, and I. S. Kweon, "Exploiting shading cues in kinect IR images for geometry refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3922–3929.
[32] P. Fankhauser, M. Bloesch, D. Rodriguez, R. Kaestner, M. Hutter, and R. Siegwart, "Kinect v2 for mobile robot navigation: Evaluation and modeling," in *Proc. IEEE Int. Conf. Adv. Robot.*, 2015, pp. 388–394.
[33] J. Sun, Y. Chang, C. Jung, and J. Feng, "Multi-modal reflection removal using convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 26, no. 7, pp. 1011–1015, Jul. 2019.
[34] M. Brown and S. Süsstrunk, "Multi-spectral SIFT for scene category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 177–184.
[35] Z. Cheng, Y. Zheng, S. You, and I. Sato, "Non-local intrinsic decomposition with near-infrared priors," in *Proc. IEEE Conf. Comput. Vis.*, 2019, pp. 2521–2530.
[36] N. Salamati, D. Larlus, G. Csurka, and S. Süsstrunk, "Incorporating near-infrared information into semantic image segmentation," 2014, *arXiv:1406.6147*.
[37] B. Shi, Z. Wu, Z. Mo, D. Duan, S.-K. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3707–3716.
[38] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 7, pp. 1794–1809, Jul. 2018.
[39] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[40] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[41] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *Proc. IEEE Conf. Comput. Vis.*, 2017, pp. 3922–3930.
[42] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Conf. Comput. Vis.*, 2015, pp. 3730–3738.
[43] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.
[44] L. Yu and M. T. Orchard, "Accurate edge location identification based on location-directed image modeling," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 2971–2975.
[45] L. Yu and M. T. Orchard, "Location-directed image modeling and its application to image interpolation," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 2192–2196.
[46] L. Yu, "Determining accurate locations of edges in natural images: A phase-based, nonparametric framework," Ph.D. dissertation, Rice Univ., Houston, TX, USA, 2016.
[47] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[49] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-V4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[50] Y. Li, M. Liu, Y. Yi, Q. Li, D. Ren, and W. Zuo, "Two-stage single image reflection removal with reflection-aware guidance," *Appl. Intell.*, vol. 53, pp. 19433–19448, 2023.
[51] R. Wan, B. Shi, H. Li, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "CoRRN: Cooperative reflection removal network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 2969–2982, Dec. 2020.
[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.