Reflection Removal with NIR and RGB Image Feature Fusion

Yuchen Hong, Youwei Lyu, Si Li, Gang Cao, Boxin Shi, *Senior Member, IEEE*

***Abstract*—Removing undesirable reflections in photographs benefits both human perceptions and downstream computer vision tasks, but it is a highly ill-posed problem based on a single RGB image. Different from RGB images, near-infrared (NIR) images captured by an active NIR camera are less likely to be affected by reflections when glass and camera planes form certain angles, while textures on objects could "vanish" in some situations. Based on this observation, we propose a cascaded reflection removal network with an image feature fusion strategy to utilize auxiliary information in active NIR images. To tackle the insufficiency of training data, we propose a data generation pipeline to approximate perceptual properties and the reflection-suppressing nature of active NIR images. We further build a dataset with synthetic and real images to facilitate the research. Experimental results show that the proposed method outperforms state-of-the-art reflection removal methods in both quantitative metrics and visual quality.**

***Index Terms*—Reflection removal, deep learning, feature fusion, near-infrared image.**

I. Introduction

REFLECTION contamination is commonly confronted when photographing in front of windows or glass, which significantly degrades the quality of captured images, as users prevailingly attempt to obtain reflection-free background images. In consequence, the reflection removal problem, which is targeted at removing reflections and recovering the background, has become an active research area in the computer vision and computational photography community [1]–[5].

The reflection removal problem is challenging due to its ill-posed nature (*i.e.*, unknown variables are twice as many as equation numbers). Before the prevalence of deep learning, non-learning methods are widely used for reflection removal, which requires sophisticatedly handcrafted priors observed from specific scenes such as the gradient sparsity prior [6]–[8], the relative smoothness [9], [10] and ghosting cues [11]. However, such methods are ineffective occasionally as the desired low-level priors merely reveal local characteristics of reflections, which are weak in generalization and easy to fail in certain scenes, *e.g.*, reflections with similar content of background scenes are hard to be separated.

Y. Hong and B. Shi are with National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, China (e-mail: yuchenhong.cn@gmail.com, shiboxin@pku.edu.cn).

B. Shi is also affiliated with the Peng Cheng Laboratory, Shenzhen, China and Beijing Academy of Artificial Intelligence, Beijing, China.

Y. Lyu and S. Li are with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China (e-mail: youweilv@bupt.edu.cn, lisi@bupt.edu.cn).

G. Cao is with Beijing Academy of Artificial Intelligence, Beijing, China (e-mail: caogang@baai.ac.cn).

Corresponding author is S. Li.

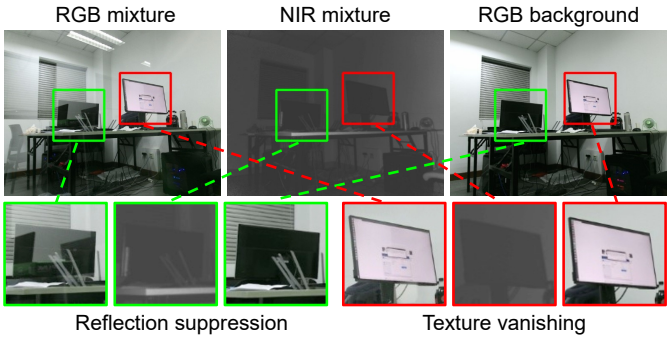


Fig. 1: An example of the reflection suppression and texture vanishing phenomenon, with close-up views displayed at the bottom. RGB and NIR mixture images are captured in front of a piece of glass simultaneously and the RGB background image is captured without the glass. Compared with the RGB mixture image (left), reflections in green boxes are significantly suppressed in the NIR mixture image (middle), bringing about the auxiliary contextual information which is consistent with the RGB background image (right). The texture vanishing phenomenon is illustrated in red boxes, where textures on the LED monitor are invisible in the NIR mixture image (middle) compared with RGB images (left and right), since the monitor does not emit light containing NIR spectral components.

In recent years, methods based on deep learning develop rapidly, which have been demonstrated to be effective in reflection removal with a single image as the input [1]–[4], [12], [13]. Due to the reliance on features learned solely from input mixture images, the performance of these methods is highly relevant to the similarity between training and testing scenes. User guidance [5] or auxiliary information independent from image contents [14]–[18] can help to relieve such a restriction.

Following the popularity of Kinect V2, active near-infrared (NIR) cameras have become easily available for non-professional users (*e.g.*, for smartphone users, Huawei P40 Pro and Samsung Galaxy Note10 have such cameras). NIR images captured by such cameras are physically less sensitive to reflections when glass and camera planes form certain angles, which contain crucial clues for reflection removal. However, the texture vanishing phenomenon may appear in NIR images due to the different physical properties between NIR and visible light. This phenomenon describes a situation where textures on objects could "vanish" if emitted lights from light sources do not contain NIR components or reflected NIR intensities are consistent across a single material. An example is illustrated

in Fig. 1, which presents the reflection suppression and the texture vanishing phenomenon in NIR images.

To address the above issues, our preliminary work, near-infrared image guided reflection removal network (NIR^{2}Net) [17], for the first time introduces active NIR images into reflection removal pipeline and proposes a two-stream framework with multi-stage feature guidance strategy, which shows more promising reflection removal performance compared with RGB image based methods [2]–[4]. However, NIR^{2}Net [17] utilizes the auxiliary information in NIR images at decoders of its two sub-networks, indicating that the guidance for reflection separation and background recovery is only conducted on the latter part of the network, which neglects global influence. Besides, the data generation pipeline of NIR^{2}Net [17] ignores certain perceptual disparity between NIR and RGB images, which degrades the generalization capacity of the network on real data.

In this work, we analyze the differences of light transmission characteristics between passive RGB imaging and active NIR imaging to further demonstrate the reflection-suppressing property of the active NIR imaging. In contrast to the two-stream framework with multi-stage feature guidance in NIR^{2}Net [17], we propose the **NIR** and RGB feature fused **R**eflection **R**emoval **Net**work (NIR^{3}Net). The network architecture is illustrated in Fig. 2. NIR^{3}Net is composed of three modules: the feature fusion module (FFM) for the fusion and enhancement of multi-scale contextual features from NIR and RGB images, the feature refinement module (FRM) for the exploration of auxiliary information and the removal of reflections in feature space, and the background recovery module (BRM) for the estimation of RGB background images. Compared with NIR^{2}Net [17], the feature fusion strategy and the cascaded network architecture render the exploitation of intrinsic correlations between NIR mixture images and background scenes to be more sufficient and effective. Besides, we replace the difference loss in NIR^{2}Net [17] with a simple but effective gradient loss to diminish the influence of the texture vanishing phenomenon. Furthermore, we improve the data generation pipeline by considering more appearance differences between NIR and RGB images, which generates data more conforming to real distributions. Our major contributions are summarized as follows:

- We propose a cascaded reflection removal network via NIR and RGB image feature fusion, which can deal with the impact of the texture vanishing phenomenon.
- We propose a data generation pipeline to approximate physical and perceptual properties of active NIR images.
- We build a reflection removal dataset containing synthetic and real data of NIR and RGB images, which promotes the generality of network models and facilitates future research in this area.

II. Related work

A. Reflection removal

Reflection removal has become an active research area in computer vision community for more than decades. For non-learning methods, handcrafted priors observed from reflection-contaminated images are widely adopted to facilitate solving the ill-posed problem. Based on the gradient sparsity prior derived from statistics of natural scenes [6], [7], Levin and Weiss utilize the iterative reweighted least squares optimization approach to separate reflections and background layers with assistance of users [8]. Li and Brown [9] exploit the relative smoothness to solve the layer separation problem. Wan *et al.* [10] utilize the smoothness prior and Depth of Field (DoF) confidence maps to distinguish edges of reflection and background layers for the subsequent separation. Shih *et al.* [11] take the ghosting cues into consideration and use a GMM model to remove reflections. Wan *et al.* [19] integrate gradient and content priors jointly to achieve background and reflection separation. Though above priors exploit differences of visual properties between background and reflection layers in certain scenarios, they are likely to fail in more complicated real-world situations.

Thanks to the comprehensive modeling capacity of deep learning, learning-based methods become prevalent in reflection removal. CEILNet [1] adopts the traditional two-stage framework which predicts edge maps and background layers successively. Zhang *et al.* [2] propose a neural network with perceptual loss to emphasize the independence of background and reflection layers in the gradient domain. CRRN [20] and CoRRN [3] combine the gradient inference and the image inference in one unified mechanism to remove reflections concurrently. ERRNet [4] embeds context modules in the network and exploits the unaligned data to enhance the generality of the model. Wen *et al.* [21] synthesize reflection images with learned non-linear blending mask and accomplish reflection removal based on such non-linearity. LBCLN [12] proposes a cascaded refinement approach with convolutional LSTM network structure to refine estimation of background and reflection layers iteratively. Kim *et al.* [13] generate data with physically-based rendering and restore the background layer considering the various impacts of glass and lens.

Restricted by limited clues (*e.g.*, the defocus of the reflection layer) for separating the reflection and background layer from a reflection-contaminated image, single-image methods perform poorly in images with strong reflections or where the content of reflections is similar to background layers. Therefore, auxiliary information is introduced to facilitate reflection removal. Zhang *et al.* [5] involve user interaction to indicate background and reflection layers and propose a two-stage pipeline for reflection removal. A series of work exploits characteristics of the polarization to accomplish reflection removal using polarized images with different polarization angles [15], [16], [22]. Sun *et al.* [14] use the shape and edge information in depth maps to guide reflection removal, which has limited capability in recovering details in background layers due to the texture-less appearance of depth maps. Besides, as flash images can provide auxiliary information of background layers and less interfered by reflections due to the active imaging, Chang *et al.* [18] utilize a pair of no-flash and flash image to remove reflections via a siamese dense network. To avoid the drawback that the active visible light imaging is prone to be affected by other visible light sources, our method is based on the more reliable active NIR imaging since common indoor lighting rarely covers the NIR spectrum [23].

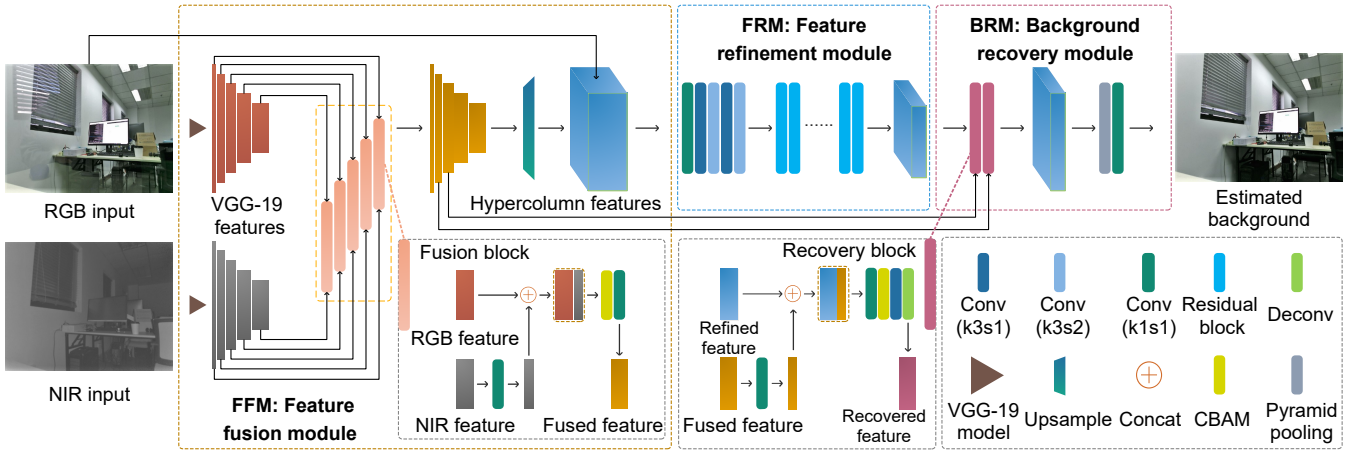


Fig. 2: The network architecture of NIR^{3}Net. Three modules: the feature fusion module (FFM), the feature refinement module (FRM), and the background recovery module (BRM) are cascaded to accomplish reflection removal successively.

B. Near-infrared imaging

NIR imaging can be divided into two categories: the passive imaging and the active imaging. The passive NIR imaging is often implemented by attaching NIR pass filters in front of lenses [24], [25], with intensities of captured images to be determined by the NIR component of ambient light. Owing to the unique physical and perceptual properties compared to RGB images (*e.g.*, higher contrast of natural and artificial objects [26] and less atmospheric scattering [27]), passive NIR images have been utilized for various computer vision tasks such as dehazing [27], shadow detection [28], semantic segmentation [29], and intrinsic image decomposition [25]. Utilizing active NIR projectors, the active NIR imaging has been widely applied to 3D sensing devices (*e.g.*, Kinect V1 and V2), which are leveraged for computer vision tasks like geometry refinement [23] and robot navigation [30]. Exploiting the reflection-suppressing property of the active NIR imaging, this work utilizes the detailed content information about background layers in captured active NIR images to achieve reflection removal with a feature fusion strategy.

III. Proposed Method

In this section, we describe the reflection-suppressing property of the active NIR imaging, the design methodology of the proposed network architecture with elaborate loss functions, and implementation details of network training.

A. Reflection-suppressing active NIR imaging

The key constraint of the proposed method is the observation that the majority of images taken through the glass by active NIR cameras are hardly affected by reflections, except that large angles are formed by the imaging plane and the glass plane (say $> 80°$, indicating a large incidence angle of lights from reflection scenes). In contrast, corresponding RGB mixture images are always blended with undesirable artifacts compared with NIR mixture images. Fig. 3 shows how incidence angles of lights from reflection scenes influences the reflection suppression phenomenon in active NIR imaging.

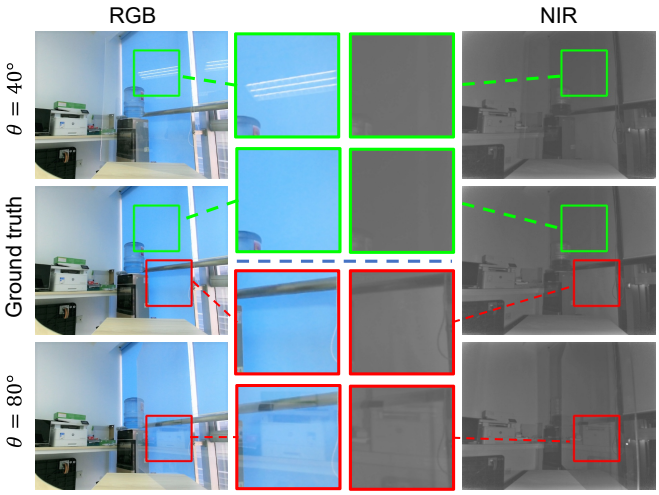


Fig. 3: Examples of how incidence angles of lights from reflection scenes θ influence glass reflections in NIR mixture images. When the angle is relatively small as $\theta = 40°$ (the first row), compared with the ground truth (the second row), reflections of the lamp (green boxes) are invisible in the NIR mixture image. When θ increases to $80°$ (the third row), reflections of the printer (red boxes) can be observed in both RGB and NIR mixture images, indicating that a large incidence angle (hardly encounters in practice) diminishes the reflection-suppression property of the active NIR imaging.

We model the light propagation process based on the Fresnel equation [31], and simplified light paths of the passive RGB imaging and the active NIR imaging are illustrated in Fig. 4. For the passive RGB imaging in Fig. 4(a), suppose I_b, I_r to be the intensity of the background and reflection scene respectively, we define the intensity of visible lights received by the camera as I_{rgb}, which can be expressed as:

$$I_{rgb} = I_r\mathcal{R}(\theta) + I_b[1 - \mathcal{R}(\theta)], \quad (1)$$

where θ represents the incidence angle and $\mathcal{R}(\theta)$ is the corresponding relative strength of the reflective component [32].

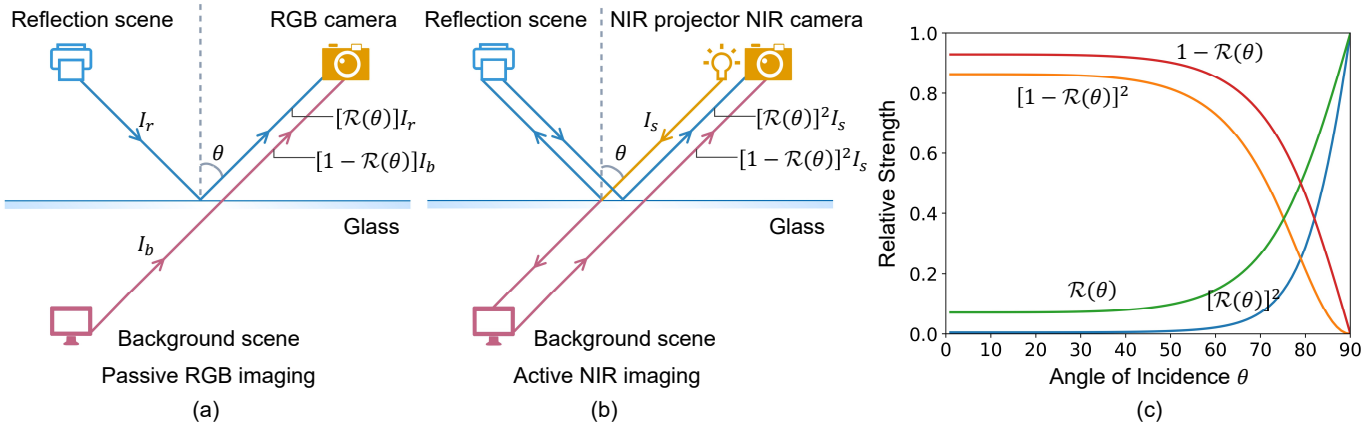


Fig. 4: Transmission characteristics of lights in different imaging approaches. Simplified light paths of (a) the passive RGB imaging and (b) the active NIR imaging. (c) Curves of relative strengths.

Compared with the passive RGB imaging, the active NIR imaging utilizes active light sources which contributes to its unique light path as shown in Fig. 4(b). Suppose I_s is a beam of NIR light emitted from the projector, and it is split into reflected and transmitted components when reaching the glass plane. The NIR image records the intensity of the received NIR lights as I_{nir}, thus this simplified propagation process can be expressed as:

$$I_{nir} = I_s[\mathcal{R}(\theta)]^2 + I_s[1-\mathcal{R}(\theta)]^2, \quad (2)$$

where $\mathcal{R}(\theta)$ is the same meaning as in Equation 1.

Under the assumption that the emitted NIR lights and ambient visible lights are unpolarized, the curves of relative strengths in Equation 1 and 2 are plotted in Fig. 4(c). When θ is small, the relative strength of $\mathcal{R}(\theta)$ keeps stable at a non-zero value while $[\mathcal{R}(\theta)]^2$ is approximate to zero, which result in the reflection contamination in the passive RGB imaging and the reflection suppression in the active NIR imaging, respectively. The reflection components can finally dominate the intensity of the NIR images until θ increases over a large angle (say $> 80°$), while it rarely happens in practice. On the basis of the above analysis, we design a cascaded framework with a fusion strategy for the exploitation of reflection-suppressed active NIR images to achieve high fidelity reflection removal in RGB images.

B. Network architecture

Given an RGB mixture image $\mathbf{M}$ contaminated by reflections and an NIR mixture image $\mathbf{I}$ with auxiliary information of the background scene, our task is to recover the background layer $\mathbf{B}$. To explore the inherent correlation between the reflection-free $\mathbf{B}$ and reflection-suppressed $\mathbf{I}$, we develop a cascaded convolutional neural network as shown in Fig. 2, which fuses multi-stage features from RGB and NIR mixture images as enhanced inputs with channel and spatial attention mechanism embedded into the network design. Given an RGB-NIR mixture image pair $\{\mathbf{M}, \mathbf{I}\}$, we denote the whole estimation process as:

$$\mathbf{B}^\star = \mathcal{F}(\mathbf{M}, \mathbf{I}; \xi), \quad (3)$$

where $\mathcal{F}$ presents the network to be trained with parameters ξ, and $\mathbf{B}^\star$ is the estimated RGB background.

As shown in Fig. 2, the proposed network takes RGB and NIR mixture images as inputs, which consists of three cascaded components: the feature fusion module (FFM), the feature refinement module (FRM), and the background recovery module (BRM). FFM extracts features of RGB and NIR inputs with a VGG-19 network [33] and conducts the fusion on RGB and NIR features with the same spatial resolution, obtaining a fused feature pyramid with multi-scale semantic information. As the input augmentation strategy of utilizing hypercolumn features has been proved to be effective in reflection removal [2], [4], we transform the fused feature pyramid into hypercolumn features to help the network to learn about semantic cues. FRM condenses hypercolumn features and explores auxiliary information about background scenes provided with NIR features to render the reflection removal problem less ill-posed. Finally, BRM accomplishes the background recovery process to output the estimated reflection-free RGB background image. The detailed structure of FFM, FRM, and BRM will be introduced as follows.

Feature fusion module (FFM). FFM is composed of a pretrained VGG-19 model [33] and five fusion blocks which share similar structure except for input and output feature channels. As the VGG-19 network [33] is designed for image recognition problem, we remove the last four layers (i.e. three fully connected layers and a softmax layer) of it to maintain its capacity of feature extraction and make it adapt to our image-to-image translation problem. A fusion block consists of two convolutional blocks and a convolutional block attention module (CBAM) [34] to calculate channel-wise and spatial attention maps for adaptive feature refinement. Each convolutional block contains a convolutional layer with kernel size equal to 1×1 for channel adjustment and feature enhancement, followed by an activation layer with ReLU function.

FFM takes RGB and NIR mixture images as inputs and extracts their features of each image by the VGG-19 model [33]. Then RGB and NIR features with the same spatial resolution are fed into the corresponding fusion blocks to obtain fused

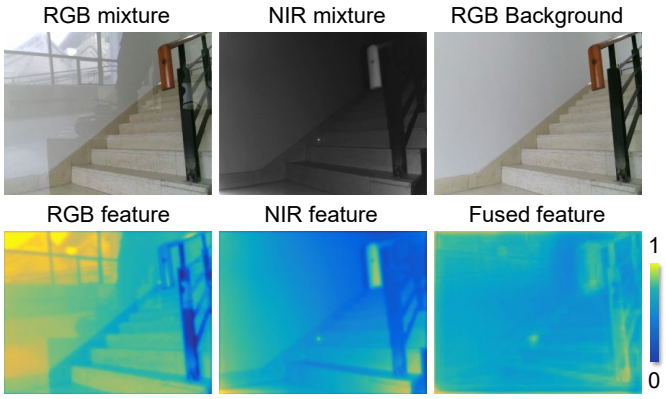


Fig. 5: An example of the inner feature presentation of FFM, where reflections are significantly suppressed in the fused feature compared with the RGB feature.

features. As previous methods [5], [8] demonstrated, solving the reflection removal problem relies more on the auxiliary edge information of background scenes rather than the absolute intensity information. Therefore, we compress the channels of input NIR features by the anterior convolutional block in each fusion block, since the reflection-suppressed active NIR images are utilized for indicating contextual information (especially edges) of background scenes. After channel compression, NIR features are concatenated with RGB features and then pass through a CBAM [34] for global and local feature enhancement. Finally, fused features are generated by the posterior convolutional block.

FFM conducts the above feature fusion on RGB and NIR features extracted by the VGG-19 model [33] to obtain a feature pyramid, and acquires the target hypercolumn features with abundant contextual information using interpolation and concatenation operations, generating the augmented input for the following network components. An example which indicates the inner presentation of features in FFM is illustrated in Fig. 5. As can be observed, the fused feature benefits from the reflection-suppressing property of the NIR feature, which facilitates the following feature refinement.

Feature refinement module (FRM). FRM comprises five convolutional blocks and thirteen residual blocks which are connected successively. The convolutional blocks share the similar structure, each of which is composed of a convolutional layer followed by an activation layer with the ReLU function while different in kernel sizes and strides. The first convolutional block with a 1×1 kernel condenses channels of hypercolumn features and reduces network parameters, and the rest convolutional blocks consist of 3×3 kernels. The strides of the first, second, and fourth convolutional block are set to 1, and set to 2 in the rest two blocks as down-sampling operation to decrease the computational cost of subsequent network blocks. Similar to [4], we employ residual blocks with channel attention mechanism for better capacity of feature refinement and faster convergence. With the cascaded structure, FRM refines and condenses hypercolumn features from FFM to exploit the intrinsic correlations between RGB background and NIR mixture images in feature space, and decreases the undesirable reflection context information progressively.

Background recovery module (BRM). BRM is designed for restoring the reflection-free RGB background images, which consists of two recovery blocks for feature up-sampling and elaboration, a pyramid pooling module as implemented in [4] to obtain a global-scene representation considering multi-scale spatial context, and a convolutional block for final background image estimation. To conserve distinct details and avoid gradient vanishing, we conduct skip connection to feed the preceding features in the fused feature pyramid and the corresponding features in deeper layers into recovery blocks together. Each recovery block contains three convolutional blocks, a CBAM [34] and a transposed convolutional block. Input features are enhanced and fused by convolutional blocks and CBAM [34], then elaborated by the transposed convolutional block. The processed features with the same spatial resolution of the original input image are converted into the final representation through the pyramid pooling module, and ultimately utilized by the last convolutional block to estimate the RGB background image.

C. Loss function

Pixel-wise loss. Considering the simplicity and the cost of computation, we penalize the pixel-wise discrepancy between the ground truth $\mathbf{B}$ and the estimated background $\mathbf{B}^\star$ with mean squared errors (MSE). The loss is defined as:

$$\mathcal{L}_{\text{pixel}}(\mathbf{B}, \mathbf{B}^\star) \quad = \quad \|\mathbf{B} - \mathbf{B}^\star\|_2^2. \tag{4}$$

Structural similarity loss. Simply utilizing pixel-wise loss generates results with blurry regions which degrades the visual quality. To tackle this problem, the structural similarity index (SSIM) [35] which conforms to human perception closely and measures the similarity of the luminance, contrast, and structure between two images in an image pair $\{\mathbf{z}, \mathbf{z}^\star\}$, is introduced to form a loss function. The SSIM index is defined as followed:

$$\text{SSIM}(\mathbf{z}, \mathbf{z}^\star) = \frac{(2\mu_{\mathbf{z}}\mu_{\mathbf{z}^\star} + c_1)(2\sigma_{\mathbf{z}\mathbf{z}^\star} + c_2)}{(\mu_{\mathbf{z}}^2 + \mu_{\mathbf{z}^\star}^2 + c_1)(\sigma_{\mathbf{z}}^2 + \sigma_{\mathbf{z}^\star}^2 + c_2)}, \tag{5}$$

where c_1 and c_2 are regularization constants, $\mu_{\mathbf{z}}$ and $\mu_{\mathbf{z}^\star}$ are the means of $\mathbf{z}$ and $\mathbf{z}^\star$, $\sigma_{\mathbf{z}}$ and $\sigma_{\mathbf{z}^\star}$ are the variances of $\mathbf{z}$ and $\mathbf{z}^\star$, and $\sigma_{\mathbf{z}\mathbf{z}^\star}$ represents their covariance. Considering the common setting of loss functions in deep learning, we define our structural similarity loss as:

$$\mathcal{L}_{\text{ssim}}(\mathbf{B}, \mathbf{B}^\star) = 1 - \text{SSIM}(\mathbf{B}, \mathbf{B}^\star). \tag{6}$$

Feature loss. Feature loss is designed for measuring the discrepancy between $\mathbf{B}$ and $\mathbf{B}^\star$ in feature space. We combine features with both low-level and high-level contextual information from the VGG-19 model [33] to form the feature loss, which is denoted as follows:

$$\mathcal{L}_{\text{feat}}(\mathbf{B}, \mathbf{B}^\star) = \sum\nolimits_i \lambda_i \left\|\Phi_i(\mathbf{B}) - \Phi_i(\mathbf{B}^\star)\right\|_1, \tag{7}$$

where $\{\lambda_i\}$ are the weights for equilibrium of multi-stage feature differences, and Φ_i presents the i-th convolutional layer in the VGG-19 model. Similar to [2], the layers as 'conv1_2',

Fig. 6: Examples of our dataset containing both real (SCENE, OBJECT, and DISPLAY) and synthetic data.

'conv2_2', 'conv3_2', 'conv4_2', and 'conv5_2', are utilized in our experiments.

Adversarial loss. The gradient statistics of reflection-contaminated mixture images are proved to be different from clear background images, which indicates that images with and without reflections follow different statistical distributions [3]. Similar to previous reflection removal methods [2], [4], [12], [13], we utilize the adversarial loss for our network optimization, which helps to diminish unrealistic color attenuation and reflection residuals in the estimated background images. The loss is defined as:

$$\mathcal{L}_{\text{adv}}(\mathbf{B},\mathbf{B}^{\star}) = -\log(\mathcal{D}(\mathbf{B},\mathbf{B}^{\star})) - \log(1-\mathcal{D}(\mathbf{B}^{\star},\mathbf{B})), \quad (8)$$

where $\mathcal{D}$ presents a relativistic discriminator network with details in [36], and for the real-fake image pair $\{\mathbf{B},\mathbf{B}^{\star}\}$, $\mathcal{D}(\mathbf{B},\mathbf{B}^{\star})$ measures the probability that $\mathbf{B}$ is more realistic than $\mathbf{B}^{\star}$.

Gradient loss. In NIR spectra, numbers of colorants and dyes are transparent and image intensity values remain consistent across a single material [29]. Therefore, objects in NIR images tend to miss important textures such as figures or icons visible in RGB images, causing context clues from NIR images to confuse the background recovery process and generate results with undesirable blurry regions. To cope with this problem, we design a gradient loss to measure the gradient differences between $\mathbf{B}$ and $\mathbf{B}^{\star}$ and constrain the network to remain texture details, which is denoted as follows:

$$\mathcal{L}_{\text{grad}}(\mathbf{B},\mathbf{B}^{\star}) = \alpha \left\|\nabla\mathbf{B} - \nabla\mathbf{B}^{\star}\right\|_1 + \beta \text{SI}(\nabla\mathbf{B},\nabla\mathbf{B}^{\star}), \quad (9)$$

where α and β are balancing weights set as 0.5 and 1, respectively. SI is a factor of SSIM that focuses on the structural similarity [37], [38], which is defined as:

$$\text{SI}(\mathbf{z},\mathbf{z}^{\star}) = \frac{2\sigma_{\mathbf{z}\mathbf{z}^{\star}}+c}{\sigma_{\mathbf{z}}^2+\sigma_{\mathbf{z}^{\star}}^2+c}, \quad (10)$$

where factors are the same definitions as in Equation 5.

Above all, our loss function is summarized to be:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \omega_1\mathcal{L}_{\text{pixel}} + \omega_2\mathcal{L}_{\text{ssim}} + \omega_3\mathcal{L}_{\text{feat}} \\ + \omega_4\mathcal{L}_{\text{adv}} + \omega_5\mathcal{L}_{\text{grad}},\end{aligned} \quad (11)$$

where $\mathcal{L}_{\text{pixel}}$ measures the discrepancy in color space, $\mathcal{L}_{\text{ssim}}$, $\mathcal{L}_{\text{feat}}$ and $\mathcal{L}_{\text{adv}}$ punish differences based on contextual information and human perception, with $\mathcal{L}_{\text{grad}}$ to reduce the impact of texture vanishing phenomenon in NIR images. Empirically, weights are set as $\omega_1 = 1$, $\omega_2 = 0.5$, $\omega_3 = 0.1$, $\omega_4 = 0.01$, and $\omega_5 = 1$ throughout our experiments.

D. Implementation details

Our model is implemented using PyTorch [39]. We train the model end-to-end for 100 epochs with Adam [40] optimizer. The learning rate is set to 10^{-4} initially and decreases to 10^{-5} at epoch 50. The weights are initialized as in [41].

IV. DATA PREPARATION

Due to the insufficiency of available datasets for reflection removal with RGB and NIR images, we build a dataset containing both real (Section IV.A) and synthetic (Section IV.B) data to facilitate network training and evaluation. The training dataset contains 5000 sets of synthetic images and 400 sets of patches extracted from real images, and the data for evaluation come from three real datasets captured by Kinect V2. An image set is composed of an RGB mixture image, an NIR mixture image, and an RGB background image (if available). All images are resized to the resolution of 224×288 pixels to reduce the computational cost. Examples of the dataset are illustrated in Fig. 6.

A. Real data

Though previous works collected several datasets with mixture and background images in visible spectra [2], [12], [38], they are not suitable for the proposed method due to the lack of mixture images captured in NIR spectra. To facilitate the training and evaluation of the proposed method, we use a Kinect V2 which is equipped with an RGB camera and an active NIR camera to capture real data. RGB and NIR mixture images are captured in front of a piece of glass, and RGB background images are captured without the glass. The resolution and field of view (FoV) of original captured NIR and RGB images are different [30], thus data pre-processing is required. Inspired by previous methods [38], [42], RGB images are coarsely aligned to NIR images through key point matching, homographic transformation, and image cropping. However, misalignment caused by scene depth variation still exists and usually causes 5-20 pixel shifts, and we consider such misalignment in synthetic data generation to diminish its influence (see details in Section IV.B).

Fig. 7: An example of the result generated from SimNet, which is more similar to the real NIR image in appearance than the grayscale image.

We capture three real datasets in total, which are denoted as SCENE, OBJECT, and DISPLAY, respectively. Since the operating range of Kinect V2 is from 0.5 to 4.5 meters [30], the SCENE dataset mainly includes various indoor scenes and contains 110 image triplets. The OBJECT dataset is obtained by capturing several solid objects which are commonly seen in daily life (paper cups, plush toys, *etc.*), and it contains 10 image triplets. The DISPLAY dataset is mainly captured in museums and contains 10 RGB-NIR mixture image pairs (as ground truth RGB background images are infeasible). We use 80 image triplets from the SCENE dataset to generate a part of training data, and the rest of the above three datasets are utilized for comprehensive evaluation of the proposed method.

B. Synthetic data

To satisfy the data-driven needs of our learning-based method, we propose a data generation approach to approximate physical and perceptual properties of active NIR images. We further build a synthetic dataset using SUN RGB-D database [43], which contains 5000 data triplets for network training.

RGB image. Based on the assumption adopted by previous works [1], [3], [4] that background layers are sharper than reflections as people tend to focus on background scenes when photographing, we generate RGB mixture images with the method proposed in [1]. Background and reflection images are randomly selected from the SUN RGB-D database [43], and mixture images are synthesized with them.

NIR image. Previous research shows that NIR lights reveal fairly different scattering patterns and physical properties from those of visible lights [29], which results in the salient appearance disparity of RGB and NIR images, including brightness of materials and the texture vanishing phenomenon. Besides, an active NIR imaging device receives the reflected NIR light emitted from its source to form an image, in which the relation between the intensity and depth from the light source follows the inverse square law [23]. Taking the above premises into consideration, we generate images from RGB background images to approximate the reflection-suppressing property of the active NIR imaging.

Instead of simply converting RGB background images to grayscale images with empirical formulation as the first step in our preliminary work [17], we transform RGB images using a convolutional neural network with the structure similar to [4], which is denoted as SimNet. The network is trained on the dataset proposed in [29] which consists of passive RGB and NIR image pairs, so that the network is able to model the appearance differences of RGB and NIR images and to approximate NIR images perceptually. As displayed in Fig. 7, the first two columns show a pair of real RGB and NIR image, and the last two columns contain the grayscale image converted from the RGB background and the result generated by SimNet. Obviously, the result from SimNet is more similar to the NIR image in perceptual properties than the grayscale image, especially the consistent intensity of the machine surface due to the uniform reflectance of a single material in the NIR band.

To approximate the texture vanishing phenomenon, we use semantic segmentation labels in the SUN RGB-D database [43] and select five classes of objects empirically to remove textures in the corresponding regions of results from SimNet. Then we employ depth maps as weighting masks for the processed images to simulate the inverse square law of the intensity variation with depth [23]. Finally, since RGB images and active NIR images in the real datasets exist spatial misalignment as mentioned in Section IV.A, we randomly apply image translation on the simulated NIR images, which introduces priors of misalignment between RGB and NIR images and ensures the proposed method to be "invariant" for slight spatial shifts through network training.

V. EXPERIMENTS

Following the comparison principle of previous auxiliary-information-guided reflection removal methods [5], [14]–[16], [18], [22], which mainly focus on comparing with methods based on the single image and methods with exactly the same inputs (in addition to RGB images), we conduct experiments with several state-of-the-art single-image methods [3], [4], [12], [13] (Section V.A) and our preliminary work NIR^{2}Net [17] (Section V.B) that also adopts the setup of utilizing active NIR images to guide reflection removal. Furthermore, we conduct ablation studies and a sensitivity analysis on the proposed method to validate the network design and the loss function combination strategy (Section V.C).

A. Comparison with single-image methods

We conduct quantitative and visual quality comparisons with several state-of-the-art single-image-based reflection removal methods, including Dong *et al.* (denoted as DX21) [44], IBCLN [12], Kim *et al.* (denoted as KH20) [13], and CoRRN [3]. For the sake of fairness, we use codes provided by their authors and finetune the above state-of-the-art methods using our training dataset with parameters suggested in their papers.

Quantitative comparison. The quantitative experiments are conducted on the SCENE and OBJECT dataset. For error metrics, we utilize the peak signal-to-noise ratio (PSNR) [45], the structural similarity index (SSIM) [35], the normalized cross correlation (NCC) [46], and the least mean square error (LMSE) [47]. The higher SSIM, PSNR, NCC, and lower LMSE indicate better restoration qualities of background images. As results shown in Table I, our method accomplishes state-of-the-art performance and consistently exceeds other methods in all the metrics, which indicates less image distortion and quality degradation of our estimated results. Besides, the

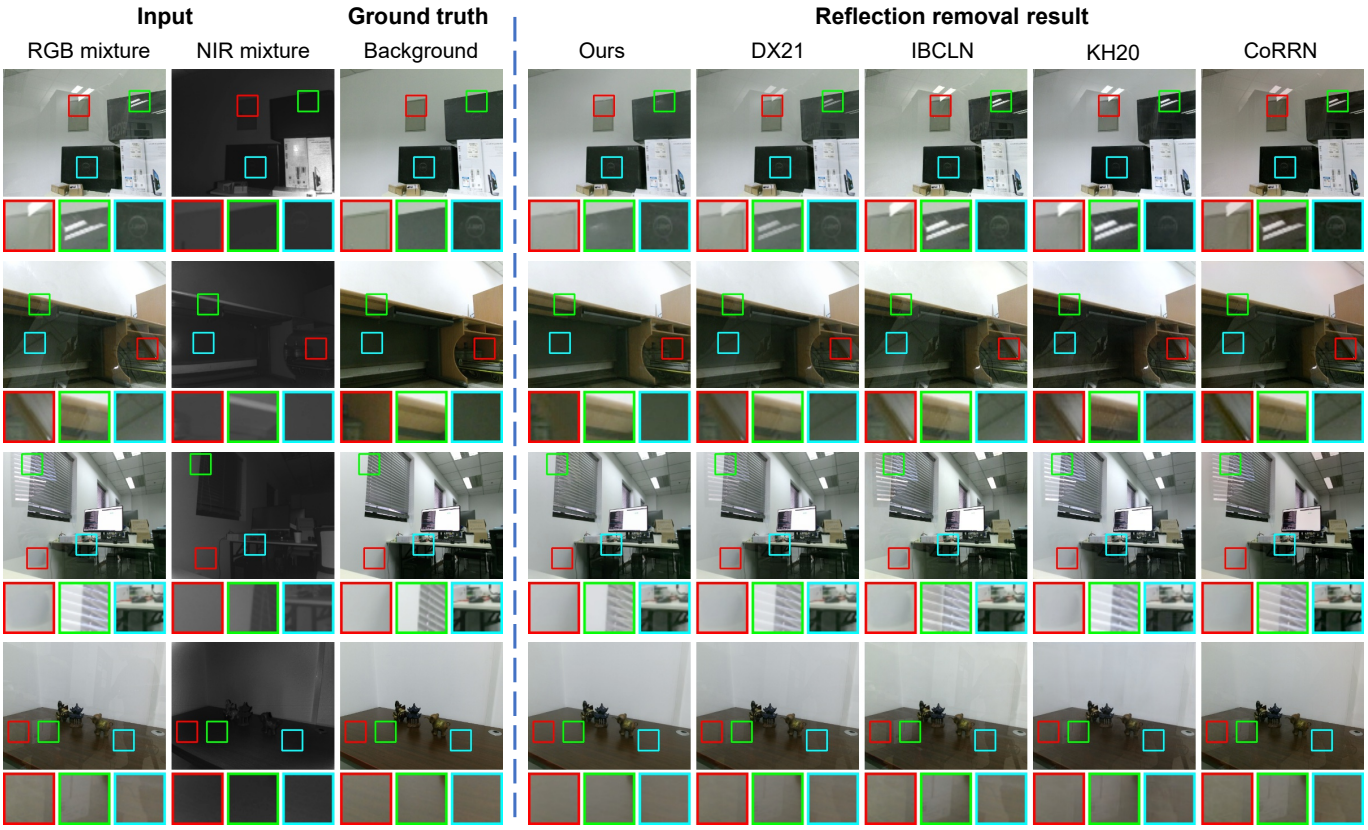


Fig. 8: Visual quality comparison results on the SCENE (the first three rows) and OBJECT (the last row) dataset, compared with several single-image methods (DX21 [44], IBCLN [12], KH20 [13], and CoRRN [3]). Close-up views are displayed at the bottom of each image (with patch brightness $\times 2$ for better visualization). Zoom in for details.

TABLE I: Quantitative results of our method on the SCENE and OBJECT dataset, compared with several single-image methods (DX21 [44], IBCLN [12], KH20 [13], and CoRRN [3]). ↑ (↓) indicates larger (smaller) values are better. Bold numbers indicate the best performing results.

Dataset (size)	Index	Methods				
		DX21	IBCLN	KH20	CoRRN	Ours
Scene (30)	PSNR↑	22.687	22.267	22.959	22.649	**24.866**
	SSIM↑	0.831	0.816	0.828	0.839	**0.860**
	NCC↑	0.944	0.939	0.944	0.942	**0.956**
	LMSE↓	0.013	0.014	0.013	0.012	**0.011**
Object (10)	PSNR↑	28.189	28.007	28.145	27.864	**30.694**
	SSIM↑	0.921	0.910	0.909	0.916	**0.951**
	NCC↑	0.990	0.989	0.989	0.991	**0.995**
	LMSE↓	0.002	0.002	0.002	0.002	**0.001**
Average(40)	PSNR↑	24.063	23.702	24.256	23.953	**26.323**
	SSIM↑	0.854	0.840	0.848	0.858	**0.883**
	NCC↑	0.956	0.952	0.955	0.954	**0.966**
	LMSE↓	0.010	0.011	0.010	0.010	**0.009**

outperformance on both SCENE and OBJECT dataset also demonstrates better generalization capacity of our method than other state-of-the-art methods.

Visual quality comparison. For visual quality comparison, examples of the estimated RGB background images of the state-of-the-art methods and our method are displayed in Fig. 8 (the SCENE and OBJECT dataset) and Fig. 9 (the DISPLAY dataset). Due to the lack of auxiliary priors, single-image methods tend to fail in regions with sharp and strong reflections (the first row in Fig. 8 and the second row in Fig. 9). Besides, if the reflections are similar to background scenes (the third row in Fig. 8), single-image methods can hardly distinguish reflections from background images and retain the majority of mixture images. Therefore, the estimated background images are still suffering from the reflections.

For the proposed method, active NIR images provide auxiliary contextual information about background scenes, which helps to locate reflection regions and guide background recovery, and contributes to our remarkable performance in the majority of real scenarios. Admittedly, for situations where NIR images can only provide limited guidance, the proposed method may generate locally imperfect results. To be specific, in the green box of the third row in Fig. 8, unclear edges in NIR images (caused by the limited resolution of the NIR camera) bring about the regional blur in results. In the red box of the second row in Fig. 8, insufficient auxiliary information for image color and illumination lead to intensity degradation since strong reflections have been suppressed while background recovery encounters difficulties. However, despite the above limited cases, the proposed method still outperforms single-

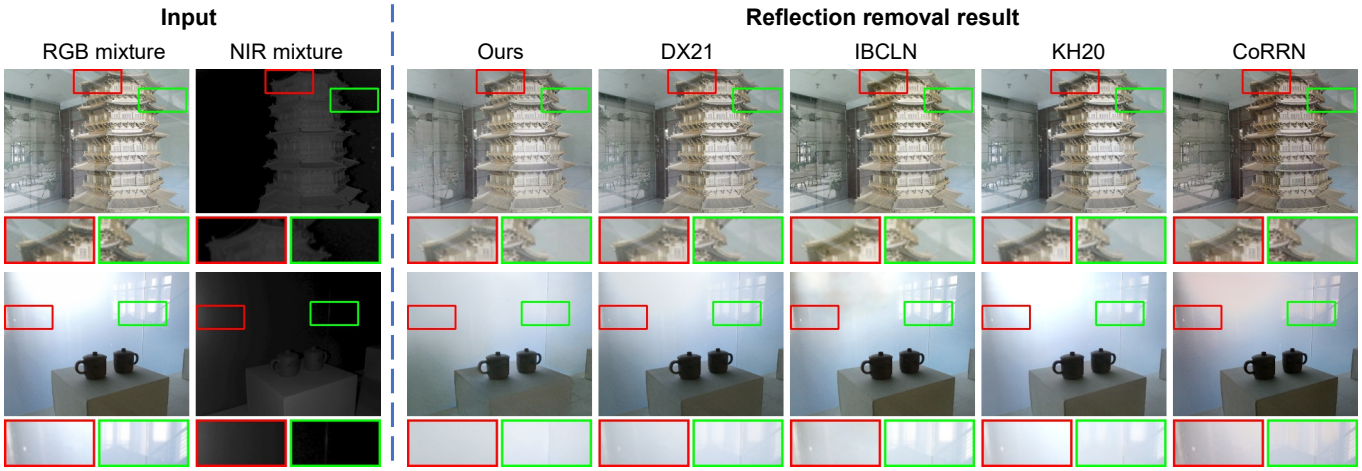


Fig. 9: Visual quality comparison results on the DISPLAY dataset, compared with several single-image methods (DX21 [44], IBCLN [12], KH20 [13], and CoRRN [3]). Close-up views are displayed at the bottom of each image (with patch brightness $\times 2$ for better visualization). Zoom in for details.

TABLE II: Comparisons on the computational complexity (input image size equal to 224×288), compared with several single-image methods (DX21 [44], IBCLN [12], KH20 [13], and CoRRN [3]).

Metric	Method				
	DX21	IBCLN	KH20	CoRRN	Ours
Params	32.79M	21.61M	27.54M	59.51M	37.21M
FLOPs	329.28G	386.16G	322.230G	75.53G	345.43G

image methods in both reflection suppression and global recovery fidelity of background images, which demonstrates the efficacy of introducing active NIR mixture images in the reflection removal pipeline. Besides, the visually pleasant results on the DISPLAY dataset shown in Fig. 9 further validate our generalization capacity on real-world glass reflections.

Computational complexity. To evaluate the computational complexity, we compare the model size and computational cost of the proposed method with state-of-the-art single-image methods (DX21 [44], IBCLN [12], KH20 [13], and CoRRN [3]). As shown in Table II, when the input image size is 224×288, the number of parameters and FLOPs of the proposed method are comparable to single-image methods, indicating that the proposed method achieves the trade-off between the computational complexity and the model performance.

B. Comparison with NIR^{2}Net

We further conduct quantitative and qualitative comparisons with NIR^{2}Net [17], the preliminary version of the proposed method, which also accomplishes reflection removal with RGB and NIR inputs. NIR^{2}Net [17] applies a simple concatenation operation to leverage multi-stage guidance of NIR features in the decoder of its network. While the proposed method adopts the feature fusion strategy at the beginning of the network to incorporate channel-wise and spatial context, which strengthens the ability of the network to learn more intrinsic correlations of NIR and RGB images, generating results with better image quality. Quantitative results in Table III and qualitative results in Fig. 10 show the effectiveness of our feature fusion strategy.

TABLE III: Quantitative comparisons of our model against NIR^{2}Net [17] and different network settings. Bold numbers indicate the best performing results.

Method	Error metric			
	PSNR↑	SSIM↑	NCC↑	LMSE↓
NIR^{3}Net (Ours)	**24.866**	**0.860**	**0.956**	**0.011**
NIR^{2}Net	24.393	0.851	0.952	0.012
W/o fusion	22.966	0.839	0.950	0.015
W/o backbone	23.667	0.845	0.952	0.013
W/o FRM	23.021	0.841	0.951	0.014
W/o $\mathcal{L}_{\text{grad}}$	24.102	0.848	0.951	0.013
$\mathcal{L}_{\text{pixel}}$ only	23.988	0.847	0.950	0.012

C. Ablation study

In this section, we conduct several ablation studies to investigate the effectiveness of individual network modules and loss functions. Experiments are conducted on the SCENE dataset. As quantitative results listed in Table III and qualitative results shown in Fig. 10, our complete model obtains the best performance among all network settings.

For network modules, we validate the effectiveness of the feature fusion strategy, the backbone network, and the feature refinement module by comparing to following variants: 'W/o fusion' that only inputs with RGB mixture images while lacking auxiliary NIR information, 'W/o backbone' that replaces the VGG-19 model [48] with a simple convolutional layer, 'W/o FRM' that removes the feature refinement module. As Table III and Fig. 10 shown, the variant 'W/o fusion' performs poorly due to the lack of instructive clues about background images contained in NIR images, which validates the effectiveness of

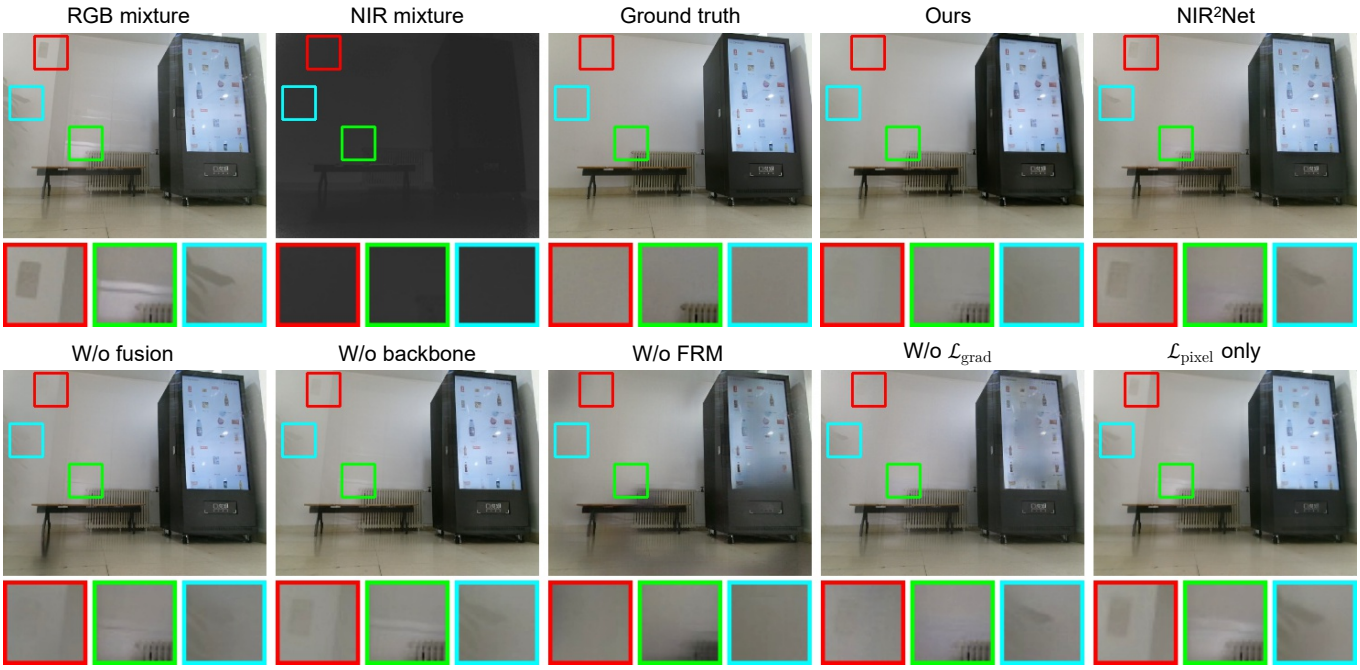


Fig. 10: Visual quality comparison of our model against NIR^{2}Net [17] and other different network settings. Close-up views are displayed at the bottom of each image (with patch brightness $\times 2$ for better visualization). Zoom in for details.

TABLE IV: Sensitivity analysis for weights of loss functions on the SCENE dataset. Bold numbers indicate the best performing results.

Weight (loss)	Value	Error metric			
		PSNR↑	SSIM↑	NCC↑	LMSE↓
ω_1 ($\mathcal{L}_{\text{pixel}}$)	1	**24.866**	**0.860**	**0.956**	**0.011**
	0.1	24.549	0.855	0.953	0.012
	0.01	24.212	0.852	0.951	0.014
ω_2 ($\mathcal{L}_{\text{ssim}}$)	1	24.843	**0.862**	0.955	**0.011**
	0.5	**24.866**	0.860	**0.956**	**0.011**
	0.1	24.574	0.856	0.954	0.012
	0.01	24.295	0.848	0.952	0.012
ω_3 ($\mathcal{L}_{\text{feat}}$)	1	22.770	0.839	0.940	0.014
	0.1	**24.866**	**0.860**	**0.956**	**0.011**
	0.01	24.355	0.856	0.953	0.012
ω_4 ($\mathcal{L}_{\text{adv}}$)	0.1	24.128	0.849	0.949	0.013
	0.01	**24.866**	**0.860**	**0.956**	**0.011**
	0.001	24.367	0.854	0.954	**0.011**
ω_5 ($\mathcal{L}_{\text{grad}}$)	1	**24.866**	**0.860**	**0.956**	**0.011**
	0.1	24.479	0.855	0.954	0.012
	0.01	24.256	0.852	0.953	0.012

our feature fusion strategy and the positive influence of our data generation pipeline. Due to the absence of the backbone network which can extract and aggregate multi-level contextual information, the variant 'W/o backbone' suffers from the degradation of performance, but it still outperforms 'W/o fusion' as it retains the utilization of NIR images. The performance of 'W/o FRM' decreases significantly, indicating the necessity of the feature refinement module to exploit the intrinsic correlations between RGB background and NIR mixture images in the feature space and decrease the undesirable reflection context information progressively.

Since the majority of loss functions in this paper are inherited from our preliminary work [17] (except for gradient loss $\mathcal{L}_{\text{grad}}$), which also have been validated in previous reflection removal methods [2]–[4], we mainly conduct ablation studies by disabling the gradient loss $\mathcal{L}_{\text{grad}}$ (denoted as 'W/o $\mathcal{L}_{\text{grad}}$') and only using pixel-wise loss $\mathcal{L}_{\text{pixel}}$ (denoted as '$\mathcal{L}_{\text{pixel}}$ only') for network training. As illustrated in Fig. 10, the variant 'W/o $\mathcal{L}_{\text{grad}}$' generates results with over smoothing and retains more reflection residuals compared with the complete model, especially in regions where the texture vanishing phenomenon occurs. For the variant '$\mathcal{L}_{\text{pixel}}$ only', the recovery quality of background images degrades significantly, which demonstrates the effectivity of the combination of loss functions. We further conduct sensitivity analysis for weights of loss functions, which is shown in Table IV. $\mathcal{L}_{\text{pixel}}$, $\mathcal{L}_{\text{ssim}}$, and $\mathcal{L}_{\text{grad}}$ are relatively more stable with the variation of their corresponding weights, while $\mathcal{L}_{\text{feat}}$ and $\mathcal{L}_{\text{adv}}$ would degrade the performance of the proposed method if their weights are not properly set (*e.g.*, large weights of these two parameters make the model hard to get converged).

To further validate the effect of active NIR images in our method, we conduct experiments by inputting NIR images with RGB mixture images containing weak reflections (from the DISPLAY dataset) or clear RGB background images containing no reflection (from the OBJECT dataset). As examples shown in Fig. 11, our method removes weak reflections and retains the image contents of background scenes, demonstrating that our method leverages the auxiliary contextual information in active NIR images and accomplishes robust background recovery.

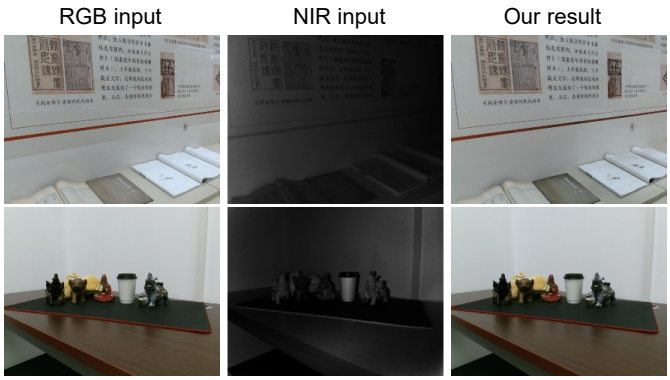


Fig. 11: Examples of our method tested on images with weak reflections (the first row) or without reflection (the second row).

VI. Conclusion

We propose a cascaded learning-based framework for reflection removal with a feature fusion strategy of RGB mixture images and reflection-suppressed active NIR images. To cope with the insufficiency of data, we propose a data generation pipeline considering perceptual properties and the reflection-suppressing nature of active NIR images to simulate data which conform to real distributions. We further create a dataset composed of both synthetic and real data for network training and evaluation. The quantitative and visual quality comparisons conducted on our dataset demonstrate the effectiveness of the proposed method.

Limitations. In our present synthetic data generation pipeline, though physical and perceptual properties of NIR images are considered, the imitation of texture vanishing is empirical, which still shows discrepancies with real images. For data capture, time-of-flight (ToF) devices containing active NIR cameras (*e.g.*, Kinect V2 and Intel RealSense L515) are required, while their limited operating ranges restrict the practicability of the proposed method in outdoor scenarios. Besides, as active NIR cameras use block filters to filter out visible light, the proposed method is currently not suitable for off-the-shelf surveillance devices which do not have such block filters due to the size limitation.

In the future, we will make efforts to improve the imitation method for texture vanishing and enhance the generalization capacity of our method to deal with more challenging scenes. Moreover, since ToF cameras have become prevalent in the design of the smartphone lens system, the proposed method owns the potential to be deployed on smartphones.

Acknowledgment

This work is supported by National Key R&D Program of China (2020AAA0105200) and National Natural Science Foundation of China under Grant No. 62136001.

References

[1] Q. Fan, J. Yang, G. Hua, B. Chen, and D. P. Wipf, "A generic deep architecture for single image reflection removal and image smoothing." in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

[2] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] R. Wan, B. Shi, H. Li, L.-Y. Duan, A.-H. Tan, and A. K. Chichung, "CoRRN: Cooperative reflection removal network," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[4] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang, "Single image reflection removal exploiting misaligned training data and network enhancements," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] H. Zhang, X. Xu, H. He, S. He, G. Han, J. Qin, and D. Wu, "Fast user-guided single image reflection removal via edge-aware cascaded networks," *IEEE Transactions on Multimedia*, 2020.

[6] A. Levin, A. Zomet, and Y. Weiss, "Learning to perceive transparency from the statistics of natural scenes," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2002.

[7] A. Levin, A. Zomet, and Y. Weiss, "Separating reflections from a single image using local features," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2004.

[8] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2007.

[9] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2014.

[10] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot, "Depth of field guided reflection removal," in *Proceedings of International Conference on Image Processing (ICIP)*, 2016.

[11] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3193–3201.

[12] C. Li, Y. Yang, K. He, S. Lin, and J. E. Hopcroft, "Single image reflection removal through cascaded refinement," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] S. Kim, Y. Huo, and S.-E. Yoon, "Single image reflection removal with physically-based training images," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] J. Sun, Y. Chang, C. Jung, and J. Feng, "Multi-modal reflection removal using convolutional neural networks," *IEEE Signal Processing Letters (SPL)*, 2019.

[15] Y. Lyu, Z. Cui, S. Li, M. Pollefeys, and B. Shi, "Reflection separation using a pair of unpolarized and polarized images," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2019.

[16] C. Lei, X. Huang, M. Zhang, Q. Yan, W. Sun, and Q. Chen, "Polarized reflection removal with perfect alignment in the wild," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] Y. Hong, Y. Lyu, S. Li, and B. Shi, "Near-infrared image guided reflection removal," in *Proceedings of International Conference on Multimedia and Expo (ICME)*, 2020.

[18] Y. Chang, C. Jung, J. Sun, and F. Wang, "Siamese dense network for reflection removal with flash and no-flash image pairs," *International Journal of Computer Vision (IJCV)*, 2020.

[19] 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 Transactions on Image Processing (TIP)*, 2018.

[20] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "CRRN: Multi-scale guided concurrent reflection removal network," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[21] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han, and S. He, "Single image reflection removal beyond linearity," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] P. Wieschollek, O. Gallo, J. Gu, and J. Kautz, "Separating reflection and transmission images in the wild," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[23] G. Choe, J. Park, Y.-W. Tai, and I. So Kweon, "Exploiting shading cues in kinect IR images for geometry refinement," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2014.

[24] M. Brown and S. Süsstrunk, "Multi-spectral SIFT for scene category recognition," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2011.

[25] Z. Cheng, Y. Zheng, S. You, and I. Sato, "Non-local intrinsic decomposition with near-infrared priors," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.

[26] C. Gyeongmin, K. Seong-Heum, I. Sunghoon, L. Joon-Young, N. Srinivasa G., and K. In So, "RANUS: RGB and NIR urban scene dataset for deep scene parsing," *IEEE Robotics and Automation Letters*, 2018.

[27] C. Feng, S. Zhuo, X. Zhang, L. Shen, and S. Süsstrunk, "Near-infrared guided color image dehazing," in *Proceedings of International Conference on Image Processing (ICIP)*, 2013.

[28] D. Rüfenacht, C. Fredembach, and S. Süsstrunk, "Automatic and accurate shadow detection using near-infrared information," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014.

[29] N. Salamati, D. Larlus, G. Csurka, and S. Süsstrunk, "Incorporating near-infrared information into semantic image segmentation," *arXiv preprint arXiv:1406.6147*, 2014.

[30] P. Fankhauser, M. Bloesch, D. Rodriguez, R. Kaestner, M. Hutter, and R. Siegwart, "Kinect V2 for mobile robot navigation: Evaluation and modeling," in *Proceedings of International Conference on Advanced Robotics (ICAR)*, 2015.

[31] E. Hecht, *Optics*. Pearson, 2015.

[32] N. Kong, Y.-W. Tai, and J. S. Shin, "A physically-based approach to reflection separation: from physical modeling to constrained optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[34] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[35] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conference on Signals, Systems & Computers*. IEEE, 2003.

[36] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," in *ICLR*, 2018.

[37] S.-H. Sun, S.-P. Fan, and Y.-C. F. Wang, "Exploiting image structural similarity for single image rain removal," in *Proceedings of International Conference on Image Processing (ICIP)*, 2014.

[38] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2019.

[40] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[42] C. Kim, S. Yun, S.-W. Jung, and C. S. Won, "Color and depth image correspondence for Kinect v2," in *Advanced Multimedia and Ubiquitous Engineering*, 2015.

[43] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun RGB-D: A RGB-D scene understanding benchmark suite," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2015.

[44] Z. Dong, K. Xu, Y. Yang, H. Bao, W. Xu, and R. W. Lau, "Location-aware single image reflection removal," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2021.

[45] Q. Hu-ynh Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, 2008.

[46] J.-C. Yoo and T. H. Han, "Fast normalized cross-correlation," *Circuits, systems and signal processing*, 2009.

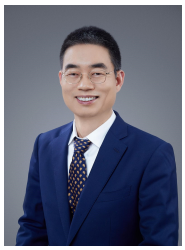
[47] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, "Ground truth dataset and baseline evaluations for intrinsic image algorithms," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2009.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

Yuchen Hong received the B.E. degree from Beijing University of Posts and Telecommunications in 2020. He is currently studying at School of Computer Science, Peking University. His research interests include computational photography and computer vision.

Youwei Lyu received his B.S. degree from Beijing University of Posts and Telecommunications in 2019. He is currently studying at Beijing University of Posts and Telecommunications. His current research interests are centered around computational photography and physics-based vision.

Si Li received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2012. She is currently an Associate Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. Her current research interests include computer vision and machine learning.



Gang Cao received the BE degree from Zhengzhou University, the MS degree from Xi'an Jiaotong University, and the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 1996, 2000, and 2004. He is currently a Research Professor at Beijing Academy of Artificial Intelligence.

Boxin Shi received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Assistant Professor and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper Runner-Up at ICCP 2015 and selected as Best Papers from ICCV 2015 for IJCV Special Issue. He has served as an editorial board member of IJCV and an area chair of CVPR/ICCV. He is a senior member of IEEE.