Unsupervised Single-Image Reflection Removal

Hamed RahmaniKhezri, Suhong Kim, and Mohamed Hefeeda, *Senior Member, IEEE*

***Abstract*— Reflections often degrade the quality of images by obstructing the background scenes. This is not desirable for everyday users, and it negatively impacts the performance of multimedia applications that process images with reflections. Most current methods for removing reflections utilize supervised learning models. These models require an extensive number of image pairs of the same scenes with and without reflections to perform well. However, collecting such image pairs is challenging and costly. Thus, most current supervised models are trained on small datasets that cannot cover the numerous possibilities of real-life images with reflections. In this paper, we propose an unsupervised method for single-image reflection removal. Instead of learning from a large dataset, we optimize the parameters of two cross-coupled deep convolutional neural networks on a target image to generate two exclusive background and reflection layers. In particular, we design a network model that embeds semantic features extracted from the input image and utilizes these features in the separation of the background layer from the reflection layer. We show through objective and subjective studies on benchmark datasets that the proposed method substantially outperforms current methods in the literature. The proposed method does not require large datasets for training, removes reflections from single images, and does not impose impractical constraints on the input images.**

***Index Terms*—Image reflection, unsupervised learning.**

I. Introduction

We frequently encounter unpleasant reflections when taking photos through transparent surfaces such as glass windows. These reflections reduce the visual quality and utility of the captured photos. Reflections may also significantly degrade the performance of multimedia applications such as object detection and face identification. Thus, removing reflection from images is an important problem for users and applications. Removing reflection is, however, a challenging research problem. Specifically, an image I containing reflection can be defined as a linear superposition of two image layers, background layer B and reflection layer R as:

$$I = B + R. \tag{1}$$

Equation (1) implies that the reflection removal problem is inherently *ill-posed*, since there are infinite valid decomposition pairs of B and R.

To address the difficulty of the reflection removal problem, some prior approaches utilize additional information such as motion cues from a *sequence* of images captured for the same scene [1]–[4]. In many practical scenarios, however, a sequence of images of the same scene may not be available, and thus these methods would fail. Other prior approaches make assumptions on the background and reflection layers, such as sparse gradient prior [5], blurriness of the reflection layer [6], and ghosting cues [7]. These approaches also fail when the assumptions do not hold, which regularly occurs because of the vast diversity of real-world images. Moreover, most prior works, especially recent ones that utilize deep learning models, require a large amount of training data. That is, most of them are supervised learning methods, which produce acceptable results on images somewhat similar to the ones seen in the training datasets. Collecting large training datasets for image reflection removal is challenging in practice, as it requires capturing each scene with and without reflection *at the same time*. Thus, most datasets in the literature tend to be small and do not cover a wide variety of reflection scenarios. Therefore, supervised learning methods may not produce good results because of the limited size of the datasets, especially on images that have different characteristics than those in the training datasets.

In this paper, we propose an *unsupervised* method for the *single-image* reflection removal problem, which, to the best of our knowledge, is the first unsupervised solution for such complex problem. The proposed method does not require large datasets for training, removes reflections from individual images, and does not make unrealistic assumptions on the input images. Despite the difficulty of designing unsupervised learning models, we believe they have the potential to address the complexity of the single-image reflection removal problem for wide diversity of images.

Our method builds on recent works which show that not all image priors must be learned from data. Rather, some of the image characteristics can be captured by the network structure itself. This is referred to as Deep Image Prior (DIP) [8], and it is used for some image restoration problems by optimizing the parameters of the untrained neural network to restore the target image from random noise. Gandelsman *et al.* [9] extended this idea by utilizing multiple DIPs to decompose images into their basic components, which can be useful for applications such as image dehazing, segmentation, watermark removal, and transparent layer separation. The generic image decomposition method in [9], however, requires multiple

Manuscript received 17 August 2021; revised 8 February 2022 and 11 June 2022; accepted 18 June 2022. Date of publication 23 June 2022; date of current version 30 October 2023. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Juyong Zhang. *(Corresponding author: Mohamed Hefeeda.)*

The authors are with the School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (e-mail: hamed_rahmani@sfu.ca; suhongk@sfu.ca; mhefeeda@sfu.ca).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Simon Fraser University's Research Ethics Board, application number 20200057.

Digital Object Identifier 10.1109/TMM.2022.3185929

inputs to solve the reflection separation problem. Specifically, this method either requires a sequence of images or two different mixtures of the background and reflection layers to address the *ambiguity* in the reflection removal problem, as indicated by (1). As mentioned earlier, in many cases a sequence of images of the same scene may not be available. And requiring two different mixtures of the background and reflection layers as *input* is not practical, as these layers are actually the outputs we are trying to obtain in the first place.

We present a new model which addresses the limitations of the multiple DIPs method, especially for the *single-image* reflection removal problem. Specifically, we first propose embedding high-level semantic information into the DIP, and we refer to it as *Perceptual DIP*. Second, we propose a *cross-feedback* structure of two Perceptual DIPs, where the output of one Perceptual DIP is weighted and fed back into the other DIP. Each Perceptual DIP captures the self-similarity nature of areas within each layer. The two Perceptual DIPs each capture the context of one of the two layers in the input image, and the cross-feedback structure allows our method to effectively separate layers in single images without any additional inputs. Thus, the proposed Perceptual DIP and the cross-feedback structure can address the ambiguity and difficulty of the single-image reflection removal problem.

The contributions of this paper are as follows.

- We present the first unsupervised method for the challenging single-image reflection removal problem. Given only a single image observation, our method successfully generates background and reflection layers, without any training data or additional information. The proposed method is composed of three main components: Perceptual DIP, cross-feedback, and refinement.
- We present a new architecture for the generator network in the Perceptual DIP component, which allows it to utilize both low-level image statistics and high-level perceptual information during the optimization.
- We design a cross-feedback structure that encourages perceptually more meaningful separation by jointly optimizing the parameters of two Perceptual DIPs, without requiring additional inputs.
- We present a semantically-guided in-painting neural network to refine the quality of the produced images after removing the reflection.
- We conduct a subjective study to compare our unsupervised method versus four state-of-the-art supervised methods for removing reflection [10]–[13]. The subjective study was approved by our university's Research Ethics Board. Fifty subjects participated in this study and evaluated the quality of the reflection separation achieved by all considered methods on 16 images chosen from datasets commonly used in prior works. The results show that, on real-world images with complex reflections, our unsupervised method substantially outperforms all prior works and successfully removes most of the reflections, without any training datasets. For example, an improvement in the Mean Opinion Score (MOS) by up to 37% can be achieved by our method compared to prior works. We also show that our method outperforms the unsupervised image decomposition method in [9], without requiring any additional inputs.
- We analyze the various components of the proposed method to show the importance and contribution of each component to the end result. We also analyze the limitations of the proposed method and the cases where it may not produce good results.

The rest of this paper is organized as follows. Section II summarizes the related work in the literature. Section III presents the proposed method. Section IV compares the performance of the proposed method against the closest works in the literature, and Section V concludes the paper.

II. Related Work

As mentioned in Section I, the image reflection removal problem is ill-posed and complex to solve. To address this complexity, several prior works assumed the availability of multiple images from a slightly moving camera for the same scene, which results in motion differences between the background and reflection layers due to their different depths with respect to the camera (motion parallax). Examples of such multi-image approaches for reflection removal include [1]–[4]. However, multiple images for the same scene may not always be available. Therefore, it is important and more practical to develop solutions for removing reflections from single images, which is the objective of this paper.

Several traditional, i.e., not neural network-based, prior works addressed the single-image reflection removal problem by imposing priors or assumptions on the reflection to make the problem tractable. Examples of these assumptions include the sparse prior of gradients and local features [5], blurrier reflection prior [6], ghosting cues [7], and different depth fields between the two layers [14].

More recent approaches for single-image reflection removal employ deep learning models and have been shown to outperform traditional ones. Examples of the most recent works in this direction include [10]–[13], [15]–[22]. We provide brief descriptions of these works in the following.

Fan *et al.* [15] introduce a solution using weakly supervised learning for training a single reflection removal model. Ma *et al.* [17] use unpaired supervision to design a weakly-supervised framework by integrating reflection generation and separation into a single model. Zhang *et al.* [16] propose a two-stage pipeline that utilizes edge hints of the background and reflection layers given by users to recover the missing details in the background layer.

Zhang *et al.* [10] utilize perceptual losses to improve the separation of the background layer from the reflection layer. Yang *et al.* [11] propose a cascade deep neural network (referred to as BDN) to estimate the background and reflection layers bidirectionally. Abico *et al.* [13] utilize a gradient constraint loss with generative adversarial networks to produce high-quality background layers. This approach is referred to as GCNet. Wei *et al.* [12] propose a framework with a context encoding module (called ERRNet) to handle the misalignment that usually occurs when collecting real datasets with pairs of images showing the captured scenes with and without reflections.

Prasad *et al.* [19] propose a lightweight deep learning model to remove reflection in two stages: processing the image using a deep architecture in the lower scales of the image and a progressive inference stage for higher scales, which is guided by the up-sampled lower scale outputs. Their architecture utilizes weight sharing, which allows it to perform faster and have fewer parameters compared to other methods. Niklaus *et al.* [20] introduce a model that uses stereo images as input to address the difficulty of the image reflection problem.

Zheng *et al.* [21] consider the absorption effect (which is, approximately, the average of the refractive amplitude coefficient map) in their formulation of the single image reflection removal problem. They propose a two-step solution that first estimates the absorption effect from an image with reflection, and, then recovers the transmission layer by taking the original image and the estimated absorption effect as inputs.

Wan *et al.* [18] design a model to recover the reflection layer from a mixed image. In our work, we separate the background layer from the reflection layer, while focusing on improving the quality of the recovered background layer as it typically represents the actual scene that users are interested in, whereas the reflection layer is mostly seen as obstructing that background scene. Wan *et al.* [22] address the reflection removal from face images, by incorporating inpainting ideas into a guided reflection removal framework. Their work focuses on face images and may not generalize to images with general scenes.

All of the above methods employ supervised-learning models, which require training datasets. Wan *et al.* [23] collect a dataset of real images with and without reflection, which is referred to as the single-image reflection dataset (SIR^2) [24], and it is frequently used as a benchmark for evaluating image reflection removal methods. In addition, some prior works generate synthetic datasets for the image reflection problem through various methods, including polarization pipeline [25], non-linear blending formulation [26], and generative adversarial training [27].

In our evaluations, we compare the proposed (unsupervised) method against four supervised methods for image reflection removal, which are Zhang el al. [10], BDN [11] GCNet [13] and EERNet [12]. These four methods represent the state-of-the-art and they published their codes and datasets, which allows us to conduct fair comparisons using common benchmark image datasets such as [24]. We could not include methods such as [18]–[22] in our comparisons as they did not release their codes or datasets. Furthermore, the four methods we compare against, Zhang *et al.* [10], BDN [11] GCNet [13] and EERNet [12], produce results with better or similar visual quality compared to other works, as shown in the evaluation sections of these papers.

It is important to notice that our method is unsupervised, yet we compare it against supervised methods to demonstrate its strength. A fairer comparison would have been against other unsupervised methods. However, we are not aware of any unsupervised methods in the literature. We note that Chandramouli *et al.* [28] proposed an unsupervised model for removing reflection from single *face* images. They use a generative model pre-trained on facial images as a deep image prior to suppress unwanted reflections from a single face image. Unlike our work, however, this method can only handle face images and does not generalize to other types of images with reflection. Thus, we could not compare our work against it.

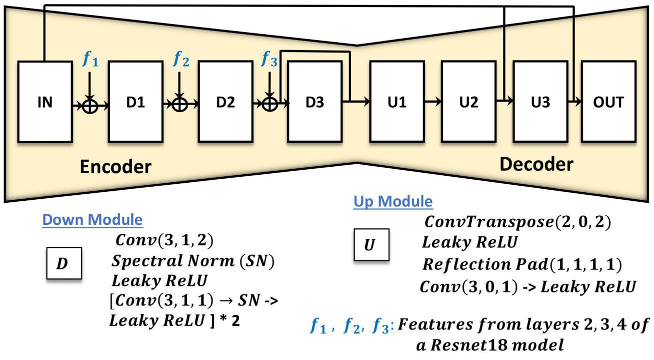
IN
D1
D2
D3
U1
U2
U3
OUT
f_1
f_2
f_3
Encoder
Decoder
Down Module
D
$Conv(3,1,2)$
$Spectral\ Norm\ (SN)$
$Leaky\ ReLU$
$[Conv(3,1,1) \rightarrow SN$ -> $Leaky\ ReLU$] * 2
Up Module
U
$ConvTranspose(2,0,2)$
$Leaky\ ReLU$
$Reflection\ Pad(1,1,1,1)$
$Conv(3,0,1)$ -> $Leaky\ ReLU$
f_1, f_2, f_3: $Features\ from\ layers\ 2,3,4\ of\ a\ Resnet18\ model$

Fig. 1. The structure of the proposed Perceptual DIP.

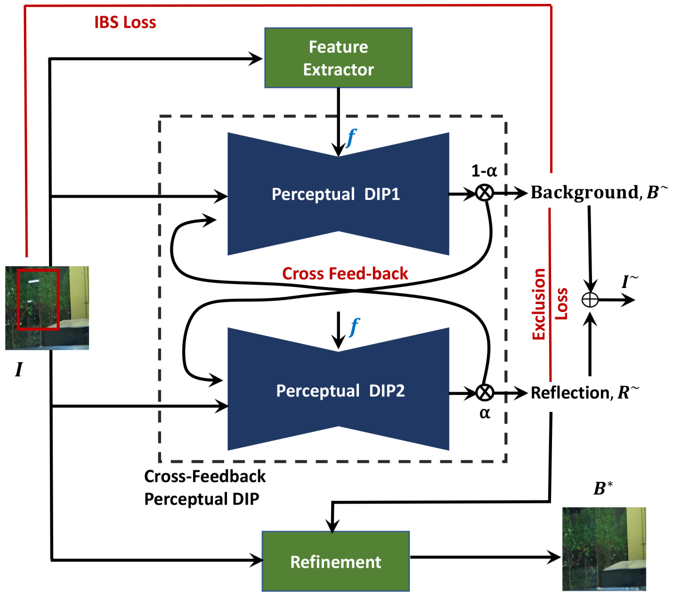

Fig. 2. Overview of the proposed method for image reflection removal. Two DIP networks with perceptual embedding are coupled with cross-feedback and loss functions, generating background and reflection layers from an input image. The (main) background layer goes through a final refinement stage.

Finally, we also compare against the unsupervised image decomposition method (Double-DIP) in [9], although, as mentioned in Section I, this method requires extra inputs that are typically not available in practice. We show that our proposed method outperforms Double-DIP, even when Double-DIP uses the extra inputs.

III. Proposed Method

A. Basic Elements

Prior works have shown that the entropy of small patches inside a natural image is smaller than the entropy across different images [29]. That is, patches of a natural image tend to have stronger internal self-similarity. For an image with reflection, this observation indicates that patches in the background

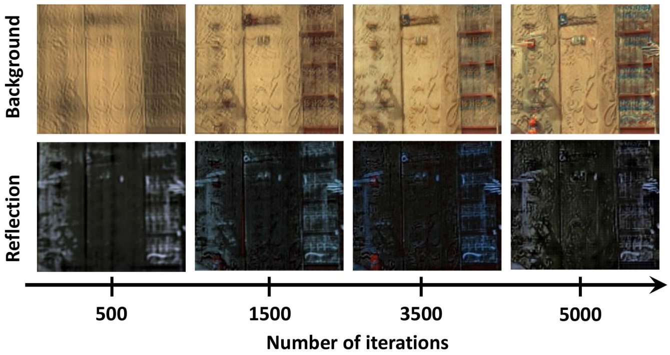

Fig. 3. The effect of cross-feedback. At early stages, up to 500 iterations, layers are separated mostly based on low-level features like colors and edges. However, at later stages, more semantic features get considered in the separation and the background and reflection layers start to exclude each other.

layer will likely have stronger self-similarity within this layer than across patches in the other reflection layer, and vice versa. To effectively utilize this observation in separating the reflection and background layers, we introduce two new structures: Perceptual DIP and Cross-Feedback Perceptual DIPs, which are explained in the following.

Perceptual DIP: Employing perceptual cues has shown remarkable advantages in capturing the semantic meanings in images, which improves the performance of various image processing tasks. Several recent deep-learning techniques improve the performance with the combination of two perceptual losses: a feature loss to measure some distance in the high-level feature space from a pre-trained perceptual network, and an adversarial loss to generate realistic images by training a separate discriminator network in parallel. However, computing the L1 or L2 distance between high-dimensional features is not sufficient to capture the real difference between them. In addition, an adversarial loss requires paired ground truth datasets of background and reflection layers to discriminate between real and fake data via supervised learning.

Reflection separation is a complex and ill-posed problem. To address this complexity and reduce ambiguity, we propose to utilize some high-level semantics. We propose perceptual embedding, which contains multi-level feature maps directly fed to the corresponding layers of an encoder, rather than leveraging perceptual losses.

Inspired by the perceptual discriminator [30], we design an encoder-decoder network with perceptual embedding, which is referred to as *Perceptual DIP*, as shown in Fig. 1. At the initialization step, the perceptual embedding module extracts multi-level features from a pre-trained image classifier. We chose ResNet18 [31] as our backbone structure of the perceptual module, which has four layers. We do not use the first layer output as the features from this layer are more sensitive to low-level information of the image, similar to those captured by DIP, while our goal is to incorporate high-level features. Then, the extracted feature maps are concatenated with the features of each layer in the encoder, which is constructed to fit well with the size of the perceptual embedding and the input image.

Cross-feedback Perceptual DIPs: We propose the coupling of two perceptual DIPs, where the output of one Perceptual DIP is fed back into the other DIP, as shown in Fig. 2. Each perceptual DIP iteratively captures similar small patches inside one of the two layers while excluding patches from the other layer. Once a perceptual DIP outputs its estimation, the corresponding cross-feedback estimation can be calculated from (1) at each iteration t as $\tilde{B}^c_t = I - \tilde{R}_t$ and $\tilde{R}^c_t = I - \tilde{B}_t$.

In Fig. 3, we show how the two Perceptual DIPs are excluding each other throughout the iterations, which enables our method to effectively separate the reflection layer from the background layer without additional inputs.

We note that we utilized dilated convolution in the last downsampler in the encoder of the Perceptual DIP. Dilated convolutions require far fewer parameters than conventional convolutions and they better capture local and global semantics within the image. We analyze the impact of the perceptual embedding on the reflection separation in Section IV-E.

B. Approach Overview

A high-level overview of the proposed method for single image reflection removal is depicted in Fig. 2. The figure shows two Perceptual DIPs with the cross-feedback idea discussed above. High-level features are first extracted from the input image using a simple image classifier. These features are fed to the two coupled Perceptual DIPs, which through iterations generate two different layers. Different types of loss functions are used to ensure good layer separation and minimize the distortion, as discussed in the following subsection. After convergence, the output of the cross-coupled Perceptual DIPs is given to a semantically-guided refinement step to produce images with high visual quality.

We define the structure of a Perceptual DIP as a parametric function $y = \mathcal{G}_\theta(x)$. Specifically, in our method, two Perceptual DIPs can be represented as $\hat{B}_t = \mathcal{G}_1(\tilde{B}^c_{t-1}, I)$ and $\hat{R}_t = \mathcal{G}_2(\tilde{R}^c_{t-1}, I)$ given an input image I and each cross-feedback, $\tilde{B}^c_{t-1} = I - \tilde{R}_{t-1}$ and $\tilde{R}^c_{t-1} = I - \tilde{B}_{t-1}$, at each iteration t. In addition, we add an external parameter α_t to control which Perceptual DIP network generates which image layer based on the following equation:

$$\begin{cases} \tilde{B}_t &= (1-\alpha_t)\cdot \hat{B}_t \\ \tilde{R}_t &= \alpha_t \cdot \hat{R}_t \end{cases} \tag{2}$$

where $\hat{B}_t$ and $\hat{R}_t$ are the direct outputs from the two Perceptual DIP networks. The range of α is between 0 and 0.5, as the range of (0.5, 1) would have the same effect. We set the initial value of α as 0.1, which implies that reflections are relatively weaker than the background scene in general cases. The impact of α in our model is analyzed in Section IV-E.

Algorithm 1 summarizes the proposed optimization method. The details of the loss functions are presented in the following.

C. Loss Functions

For a given input image I with reflection, our goal is to find a perceptually meaningful decomposition of I into $\tilde{B}$ and $\tilde{R}$ layers. We realize this goal by designing various loss functions and

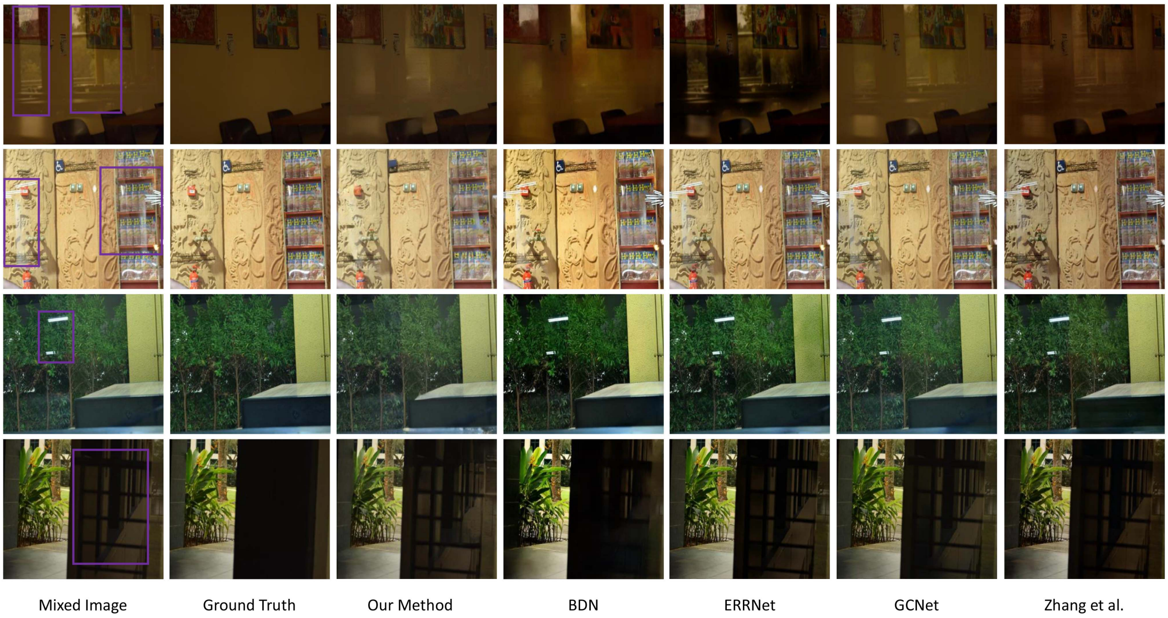

Fig. 4. Comparing our unsupervised method versus four supervised methods on dataset DS1.

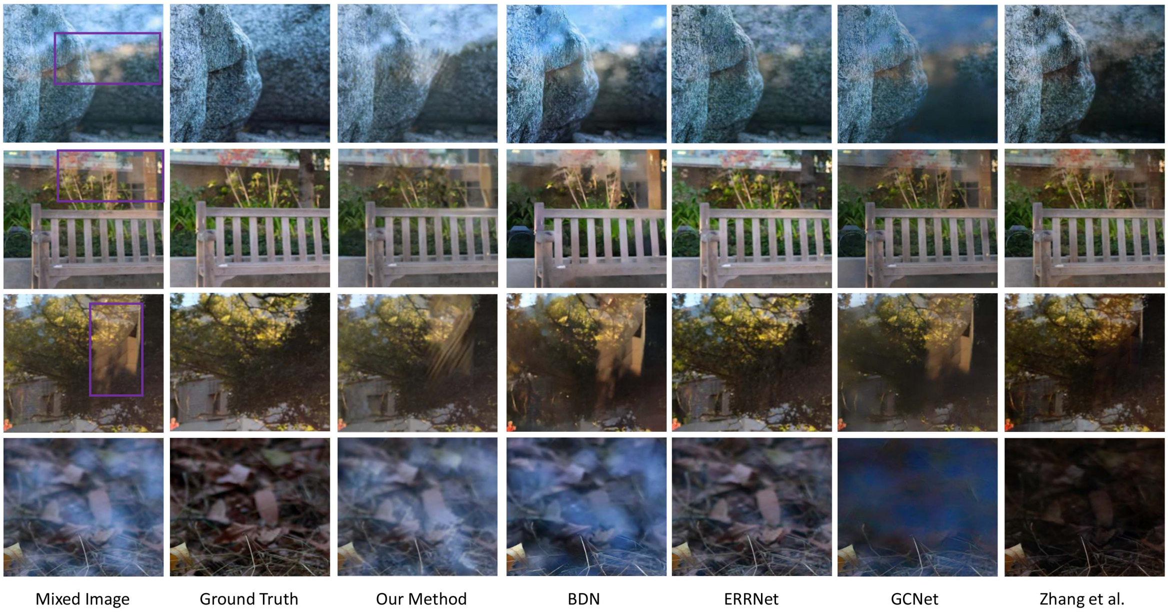

Fig. 5. Comparing our unsupervised method versus four supervised methods on dataset DS2.

integrating them into the model. These loss functions are: reconstruction loss, exclusive loss, similarity loss, and regularization loss. The total loss function can be written as:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{recon} + \lambda_2 \cdot \mathcal{L}_{excl} + \lambda_3 \cdot \mathcal{L}_{sim} + \lambda_4 \cdot \mathcal{L}_{reg}, \tag{3}$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are the corresponding weights for each loss function; we experimentally set the values of these weights. Once determined, we fixed all parameters throughout the entire evaluation. The details of each loss are explained below, while an ablation study to analyze the impact of each loss is presented in the Supplementary Materials.

Reconstruction Loss: We find that combining different types of reconstruction losses helps the network to converge faster. Thus, we define our reconstruction loss as:

$$\begin{aligned} \mathcal{L}_{recon} &= \mathcal{L}_{color} + \omega_1 \cdot \mathcal{L}_{gray} + \omega_2 \cdot \mathcal{L}_{grad}, \\ \mathcal{L}_{color} &= \|I - \tilde{I}\|_2, \\ \mathcal{L}_{gray} &= \|c(I) - c(\tilde{I})\|_2, \\ \mathcal{L}_{grad} &= \| \bigtriangledown I - \bigtriangledown \tilde{I}\|_1, \end{aligned} \tag{4}$$

where $c(\cdot)$ is the conversion function from RGB image to grayscale image, and $\bigtriangledown(\cdot)$ denotes the gradient of the input with

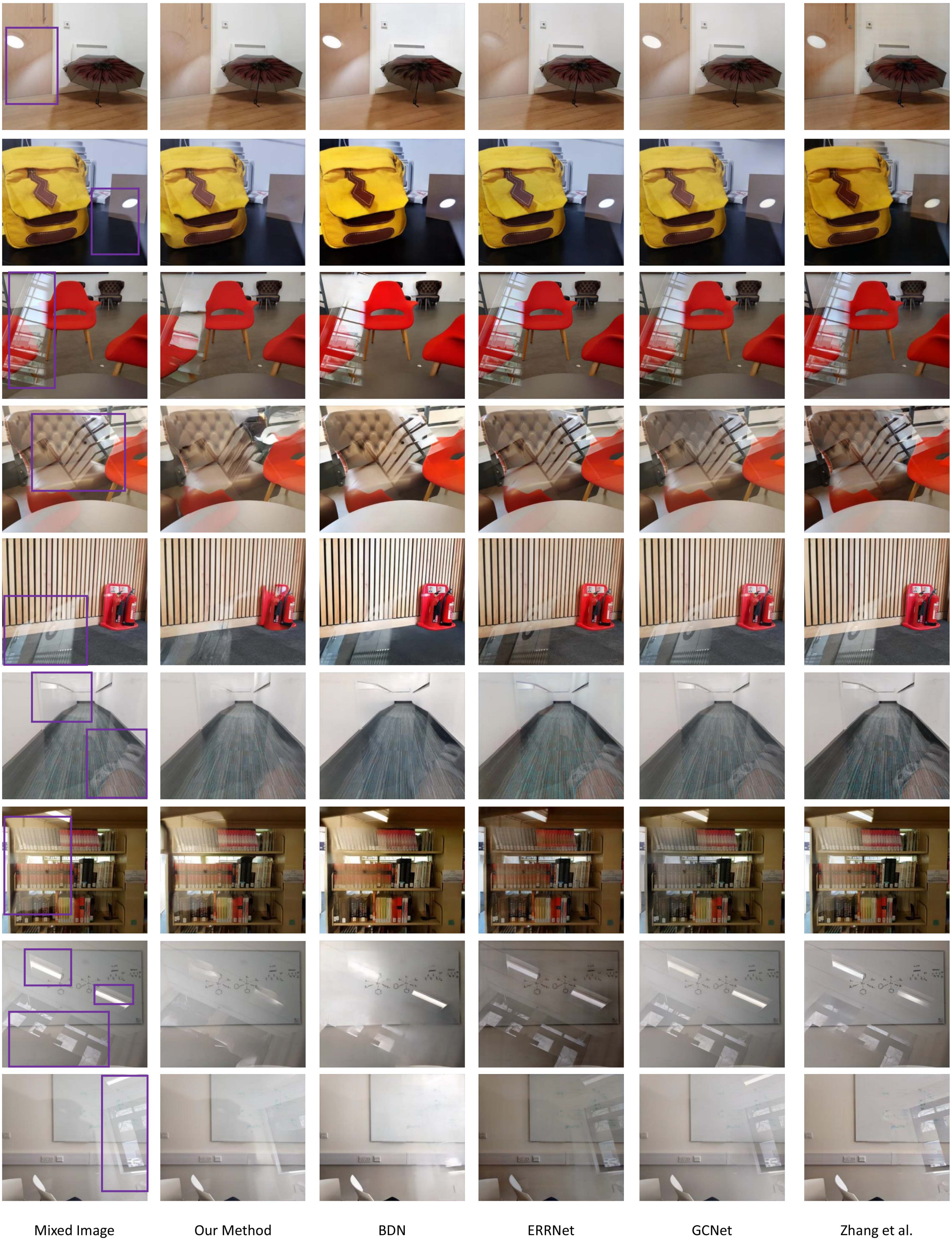

Fig. 6. Comparing our unsupervised method versus four supervised methods on dataset DS3.

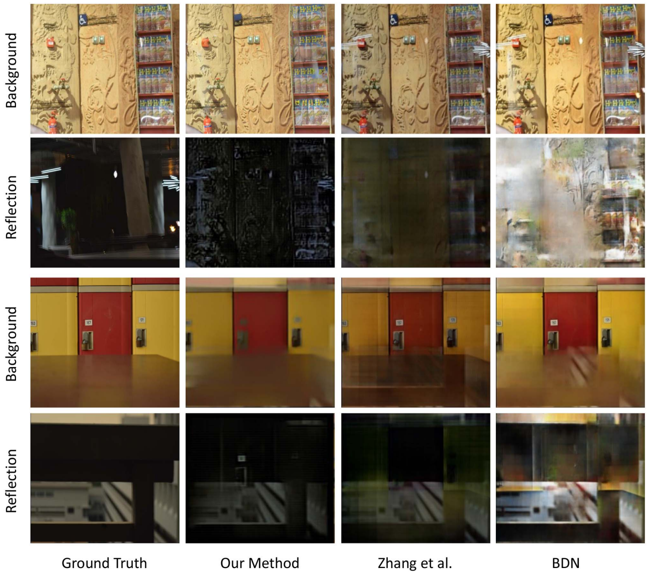

Fig. 7. Comparison of the separation quality produced by our method versus BDN [11] and Zhang *et al.* [10] methods.

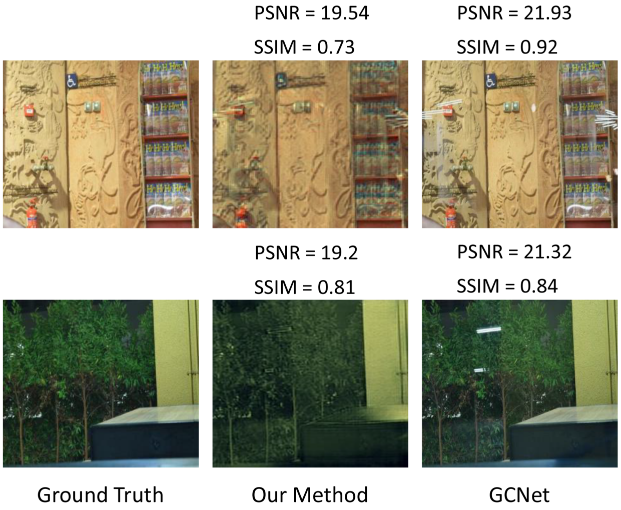

Fig. 8. Comparison between the output of our model and GCNet to show the importance of the visual quality over the objective PSNR and SSIM metrics. Although GCNet's output achieved better PSNR and SSIM, it did not remove much of the reflection, whereas our method removed most of the reflection.

Algorithm 1: Optimization Algorithm.

```
Input: The image $I$ with reflection
Output: Decomposed layers, $\tilde{B}$ and $\tilde{R}$
initialize $\tilde{B}_0 = \tilde{R}_0 = I, \alpha_0 = 0.1$
for $t = 0$ to $T$: //$T$ is set to 5,000 iterations
    $\tilde{B}_t = (1 - \alpha_t) \cdot \mathcal{G}_1(I - \tilde{R}_{t-1})$
    $\tilde{R}_t = \alpha_t \cdot \mathcal{G}_2(I - \tilde{B}_{t-1})$
    Compute the gradients of $\mathcal{L}_{total}$ w.r.t. $\tilde{B}_t, \tilde{R}_t, \alpha_t$
    Update $\tilde{B}_t, \tilde{R}_t, \alpha_t$ using the Adam optimizer [32]
    $\tilde{B}^c_t = I - \tilde{R}_t$
    $\tilde{R}^c_t = I - \tilde{B}_t$
end for
return $\tilde{B}_t, \tilde{R}_t$
```

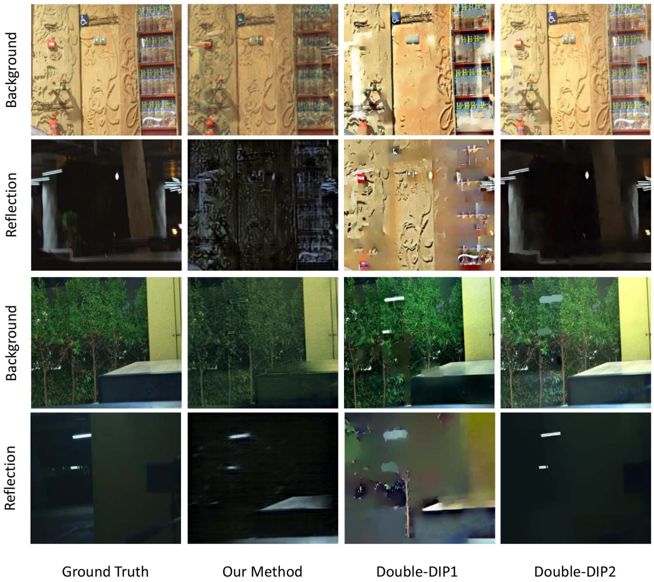

Fig. 9. Comparing our method against the unsupervised Double-DIP method [9].

the Sobel filter, which we use in the initial steps. The main reconstruction loss is a pixel-wise $\mathcal{L}2$ distance between the given image and the recombined image in the RGB color space. We also design the same $\mathcal{L}2$ losses both in the gray space ($\mathcal{L}_{gray}$) and in the gradient domain ($\mathcal{L}_{grad}$). We find that $\mathcal{L}_{gray}$ enhances the generated output and $\mathcal{L}_{grad}$ makes the network more robust.

Exclusion Loss: The exclusion loss aims to minimize the correlation between edges of the background layer and the reflection layer at multiple spatial resolutions. Thus, similar to [10], we define the exclusion loss as:

$$\mathcal{L}_{excl} = \sum_{n=1}^{N} \|norm(\bigtriangledown \tilde{B}_n) \odot norm(\bigtriangledown \tilde{R}_n)\|_F, \qquad (5)$$

where n is the image downsampling factor, as exclusion loss minimizes the correlation between edges of background and reflection at multiple spatial resolutions. For each n in (5), the image is downsampled by a factor of 2, and we chose N as 3 in our experiments. $norm(\cdot)$ is the normalization in gradient fields of the two layers, $\odot$ is the element-wise multiplication, and $\| \cdot \|_F$ denotes the Frobenius norm.

Similarity Loss: We design the similarity loss function with two components: Cross-Consistent loss $\mathcal{L}_{cc}$ and the Input-Background-Similarity (IBS) loss $\mathcal{L}_{IBS}$.

Our goal is to empower the model to make the layers exclude one another, such that each generated layer should be similar to its corresponding cross-feedback from the other network as well as its previous output. The Cross-Consistent loss contributes to this goal, and it is defined as:

$$\mathcal{L}_{cc} = \|\tilde{B}_t - (I - \tilde{R}_{t-1})\|_2 + \|\tilde{R}_t - (I - \tilde{B}_{t-1})\|_2, \qquad (6)$$

Our observation suggests that although the reflection could be evident in an image, the dominating part of the image is the background. Thus, we would like the produced background layer to resemble the input image. The IBS loss tries to make

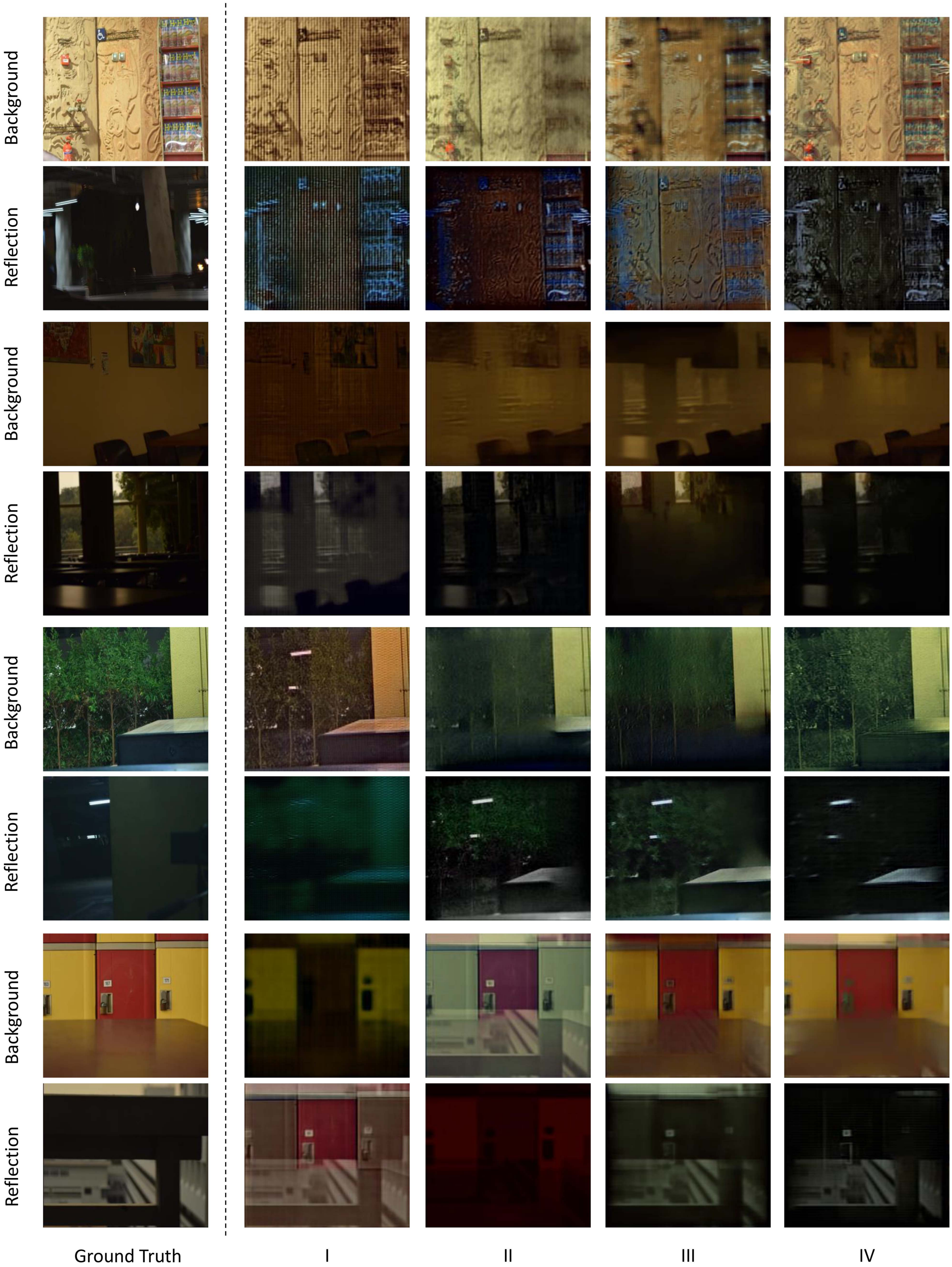

Fig. 10. Ablation study to analyze the impact of different losses in four different scenarios in two real images: “I”: Using only the Reconstruction Loss, “II”: Reconstruction + Exclusion, “III”: Reconstruction + Exclusion + Regularization Loss, and “IV”: All the losses.

the produced background layer similar to the input image, and it is defined as:

$$\mathcal{L}_{IBS} = \omega_3 \cdot (\|\tilde{B}_t - I\|_2 + \mathcal{L}_{percep}),$$
$$\mathcal{L}_{percep} = \lambda_m \cdot \sum_m \|f(\tilde{B}_t) - f(I)\|_1, \tag{7}$$

Since we do not want complete similarity between the input image and the output to avoid a case where the model keeps on generating the reflection in the background, we found through experiments that the L_2 loss with a small effect is more suitable in both preserving details and separating reflection. In addition, the IBS loss also includes a perceptual component, which as shown in prior works, e.g., [33], helps in producing visually pleasing images. The perceptual component is defined based on the activation of the 19-layer VGG network [34] trained on ImageNet. The $f(\cdot)$ operator is the activation of an image at a certain level, and the perceptual loss calculates the L_1 distance between the activation of two images at each level. λ_m is a balancing weight for each layer and we put the largest weight to emphasize low-level features and edges. We used convolution layers similar to [12].

Combining the two loss components, the similarity loss is given by:

$$\mathcal{L}_{sim} = \mathcal{L}_{cc} + \mathcal{L}_{IBS}. \tag{8}$$

Notice that the relative weights for the terms in (8) are controlled by the parameters in the (7).

Regularization Loss: We regulate the network under three priors: a total-variance loss $\mathcal{L}_{TV}$[35], a total-variance balance loss $\mathcal{L}_{TVB}$ that we developed, and a ceiling rejection loss $\mathcal{L}_{ceil}$ [4], which are defined as follows:

$$\mathcal{L}_{reg} = \gamma_1 \cdot (\mathcal{L}_{TV} + \mathcal{L}_{TVB}) + \mathcal{L}_{ceil},$$
$$\mathcal{L}_{TV} = \| \bigtriangledown \tilde{B}_t\|_1 + \| \bigtriangledown \tilde{R}_t\|_1,$$
$$\mathcal{L}_{TVB} = \| \bigtriangledown \tilde{B}_t\|_1 - \| \bigtriangledown \tilde{R}_t\|_1,$$
$$\mathcal{L}_{ceil} = \sum_m f(\tilde{B}_t, I, m) + f(\tilde{R}_t, I, m),$$
$$f(x,y,m) = \begin{cases} \|x_m - y_m\|_1 & if\ x_m > y_m \\ 0 & otherwise \end{cases}, \tag{9}$$

where m denotes each image pixel. While a total-variance loss boosts the spatial smoothness in both generated scenes, our total-variance balance loss penalizes the system when one of the networks is giving up on generating the output (degeneration problem) by balancing the total gradients of each output. Also, the ceiling rejection loss constrains each pixel whose intensity is larger than the input one, helping to resolve the color ambiguity.

D. Refinement

The cross-coupled Perceptual DIPs generate images for the background and reflection layers. In the generation process, there are multiple downsampling and upsampling operations. During these operations, some details of the input image can be lost, which may result in an output with poor visual quality even if the layers are perfectly separated. To address this issue, we add a final stage to the proposed model to refine the output.

The refinement model is inspired by recent works on image in-painting and restoration, e.g., the contextual in-painting method in [36]. The contextual in-painting method [36] requires user-specified masks for areas that have damage in the image. We adapt this contextual in-painting method to the reflection removal problem as follows. Reflections in images can be thought of as obstructions that cause damage to images. Thus, we consider the reflection layer extracted by our cross-coupled Perceptual DIPs as obstructions (damages) to the main background layer in the image. We then create a mask based on this reflection layer and use it to *fix* the damages (reflections in this case) in the full-resolution input image using the contextual in-painting method, without requiring any user-specified masks as in [36].

IV. Evaluation

We evaluate the performance of the proposed unsupervised method and compare it against the state-of-the-art supervised methods for image reflection removal in the literature using a subjective study as well as multiple objective metrics. In addition, we analyze the impact of various components of the proposed method. We also compare our method against the *unsupervised* image decomposition method in [9] and its *limited* application to the image reflection removal problem.

We note that the images presented in this paper contain subtle reflections and thus they are best viewed digitally and zoomed in to see these details and differences.

A. Experimental Setup

Datasets: We assess the performance of the proposed method using three datasets, referred to as DS1, DS2, and DS3. These datasets contain images with diverse reflection characteristics for indoor and outdoor scenes, and they have been used to evaluate prior methods for image reflection removal in the literature, including the ones compared against in this paper.

The first dataset, DS1, comes from [24]. There are hundreds of images in the dataset available from [24]. However, there are only 55 real-world images with reflections having corresponding ground truth background and reflection layers, which we use as our DS1. An image in this dataset is first captured through a glass barrier, which produces a mixed image with reflection and background layers. Then, the ground truth reflection layer is captured by putting a sheet of black paper behind the glass. The ground truth background is later captured by removing the glass.

The second dataset, DS2, contains 20 images [10]. This dataset has a ground truth for the background layer only. Images are captured by a camera on a tripod with a portable glass in front of the camera. The ground truth background is captured after removing the glass. The third dataset, DS3, is collected from the Kaggle website [37] and it includes 1,000 image pairs with and without reflections from 108 different scenes.

Methods Compared Against: We compare the proposed method against four state-of-the-art methods, which are BDN [11], GCNet [13], ERRNet [12], and Zhang *et al.* [10]. All of these methods use supervised deep learning models and have been shown to outperform prior works. We use the implementations released by the authors of these works in our evaluation, to ensure fair comparisons. Some works, e.g., [18]–[22], did not release their codes and datasets, and thus we could not include them in the evaluation.

Implementation Details: To address the complexity of the reflection removal problem, our model uses multiple losses. We set the relative weights of different losses experimentally. However, once the parameters are determined, they do not change for all experiments. We set λ_1, λ_2, λ_3 and λ_4 to 1.5, 0.13, 1.0, and 1.0, respectively. For the reconstruction loss, we set the value of ω_1 and ω_2 to 0.09 and 0.07. The value of γ_1, the regularization loss coefficients, is set to 0.003. As for ω_3 in the similarity loss, we set it to 0.1. The λ_m parameters are set according to their values in prior works that used the same structure as detailed in [12].

Since our method is based on optimizing the model parameters for each single image, the batch size is set to 1, and the parameters are updated with a learning rate of 0.0001 until the number of iterations (epochs) reaches 5500.

B. Comparison Using Subjective Study

Reflections in images come in different forms and vary substantially based on numerous factors such as the illumination of the scene, the nature of the surfaces causing the reflections, and the angles of capturing images. Reflections may also cover small parts of an image or the entire image. Reflections can be very subtle or they could dominate an image and obstruct objects in it. Thus, performance evaluation of reflection removal methods should account for these complex and inter-dependent issues. However, objective metrics, such as PSNR and SSIM, can only partially capture the performance of the reflection removal methods, as they typically focus on comparing pixels and they cannot accurately consider the above issues. Although demanding and time consuming, subjective studies provide more accurate assessment of the performance of reflection removal methods, where humans can consider various aspects in the evaluation of the quality of the produced results.

We conducted a subjective study to compare the quality of the produced images by our method against those produced by four supervised reflection removal methods. **The study was approved by the Research Ethics Board of our university**. A total of 50 subjects participated in this study, where 34% of the subjects were female. The subjects have various education and work backgrounds and are from different age groups: 72% are between 18–25 years old, 24% between 26–35, and 4% are older than 35.

The experiments were conducted through web forms, where a subject is shown an input image that contains reflection along with the outputs produced by five reflection removal methods: BDN [11], GCNet [13], ERRNet [12], Zhang *et al.* [10], and ours. The web form contains two rows of images, where the image in the leftmost column in the first row is the input image with reflection, with purple boxes indicating where reflections are located. The other images are the reflection-removed versions of the image produced using the considered methods. We ask subjects to give a score between 1 (Poor) and 5 (Excellent) for each generated image indicating the "quality of reflection removal". We ask subjects to consider whether the method has removed the reflection while preserved image visual quality. We explain and show examples to subjects before they start ranking. The names of the used reflection removal methods are not shown to subjects and the order of showing the results changes randomly for each input image.

TABLE I
SUMMARY STATISTICS OF THE SUBJECTIVE STUDY

	Average MOS	Median MOS
BDN [11]	2.68	2.75
GCNet [13]	2.49	2.5
ERRNet [12]	2.87	2.84
Zhang et al. [10]	2.74	2.75
Our	**3.82**	**3.94**

Each of the 50 subjects evaluated the quality of removing reflections from 16 representative and diverse images chosen from DS1, DS2, and DS3. Thus, in total, we collected $50 \times 16 = 800$ data points.

A summary of the results is given in Table I. The table compares the average and median of the Mean Opinion Score (MOS) computed across all users and images for the five considered methods. The results in Table I show that our method substantially outperforms all prior works, despite being unsupervised and not requiring any training data. For example, the median MOS resulted from our method is 3.94, which is 37% higher than the best median MOS resulted from prior works (2.87 produced by ERRNet [12]).

C. Visual and Objective Comparisons

Visual Comparisons: We present samples of our results to visually compare the proposed method versus the state-of-the-art methods in Figs. 4, 5, and 6, on datasets DS1, DS2, and DS3, respectively. In these figures, we draw rectangles showing some areas that have reflections. The input to all methods is shown on the left, which is an image with reflection. These figures show only the background layer of each image after removing the reflection layer. We analyze the reflection layer later.

The results in the Figs. 4, 5, and 6 show that our method produces better (or at least the same) reflection removal than the supervised methods that require a substantial amount of training data. For example, in the sample images of the second row and third row in Fig. 4, all methods except ours failed to detect and remove the reflection. Similarly, for the sample in the fourth row, our method generated an output close to the ground truth background, whereas the other methods failed to remove the reflection in the image. As for the first row, our model has managed to locate and remove the reflection better than the other methods. Similar observations can be made on the results in Figs. 5 and 6.

We further analyze the quality of the layer separation of different methods in Fig. 7. This figure shows both the background and reflection layers produced by various methods and compares

TABLE II
COMPARING OUR METHOD AGAINST SUPERVISED METHODS USING THE SSIM AND PSNR METRICS. B: BACKGROUND, R: REFLECTION

Dataset	DS1			
Metric	PSNR		SSIM	
	B	R	B	R
BDN [11]	22.01	9.01	0.86	0.31
GCNet [13]	24.53	—	0.92	—
Zhang et al. [10]	21.13	20.88	0.87	0.64
ERRNet [12]	23.86	—	0.88	—
Our Method	20.52	20.28	0.82	0.41

TABLE III
COMPARING OUR METHOD AGAINST DOUBLE-DIP METHOD USING THE SSIM AND PSNR METRICS. B: BACKGROUND, R: REFLECTION

Dataset	DS1			
Metric	PSNR		SSIM	
	B	R	B	R
Double-DIP1 [9]	16.61	10.02	0.73	0.39
Double-DIP2 [9]	16.53	20.35	0.65	0.66
Our Method	20.52	20.28	0.82	0.41

them against each other and the ground truth. We show the results for only our method as well as the BDN [11] and Zhang *et al.* [10] methods, as they were the ones that produced the best results from prior works, as indicated in Figs. 4, 5, and 6. As Fig. 7 shows, our method produces a cleaner separation of the background and reflection layers.

Objective Comparisons: As we mentioned above, objective image quality metrics, including PSNR and SSIM, do not accurately measure the quality of separating the reflection layer from the background layer, which is the main goal of our method. Instead, they measure the quality of the produced images, even if the separation of the layer was not done properly. Nonetheless, for completeness, we compare our method versus others using the PSNR and SSIM objective metrics. The results for dataset DS1 are presented in Table II, which shows that our method results in somewhat smaller SSIM and PSNR values than some of the other methods.

We illustrate the shortcomings of the PSNR and SSIM in assessing the performance of the reflection removal methods in Fig. 8, where we compare the produced background layer of our method versus the one produced by GCNet. As the figure shows, GCNet produced a background that is similar to the input image *without* removing much of the reflection. Thus, the computed PSNR and SSIM values are high, despite the poor performance in the main task at hand (removing reflection). On the other hand, our method removed most of the reflection from the image and produced images with acceptable PSNR and SSIM values.

We note that the PSNR metric is sensitive to the variations in pixel intensity between the produced image and the reference image. Effectively removing reflections from an image indicates that the pixel values could substantially change compared to the original image, leading to lower PSNR values, as is the case in our method. Similarly, the SSIM metric measures the structural similarity between two images. And since removing reflections from an image can change different parts of the image (e.g., by removing objects reflecting on the background scene), the structural similarity between the image before and after removing reflections is expected to decrease.

Remark: We note that the performance of prior supervised methods heavily depends on the used datasets in the training and their performance typically degrades on images that do not have similar ones in the training datasets, which is usual as real-life images have numerous varieties. In contrast, our method exploits both high-level and low-level statistics of an image to find two layers that are as close as possible to a natural image. It optimizes the parameters of the model on each input sample separately, which means that it learns the image statistics of the input and uses them to separate the input into two layers.

D. Comparison Against the Double-DIP Unsupervised Layer Separation Method

As mentioned in Section I, the unsupervised image decomposition method in [9] requires a sequence of images or two different mixtures of the background and reflection layers to address the ambiguity in the reflection removal problem. Although requiring two different mixtures of the background and reflection layers is not practical, since we do not know these layers beforehand, we compare the proposed method against the unsupervised method in [9], which we refer to as Double-DIP.

To be able to compare against Double-DIP, we use images in dataset DS1, because they have ground truth background and reflection layers. This enables us to create the mixtures of background and reflection layers needed by Double-DIP to function. As there was no specific method in [9] for mixing the two layers, we experimented with two different configurations, referred to as Double-DIP1 and Double-DIP2. For Double-DIP1, we mix the original (ground truth) background layer with the reflection layer that was modified by a Gaussian kernel. For Double-DIP2, we linearly add the background and reflection layers with a higher weight for the reflection layer. We expect Double-DIP2 to produce better results as it solves a simpler problem with linear combinations of the ground truth layers. We used the Double-DIP implementation released by the authors of [9]. We realize that Double-DIP1 and Double-DIP2 only represent two possible combinations. However, the main point here is that the Double-DIP method requires *unrealistic* inputs to solve the single-image reflection removal problem. Nonetheless, we compare our method against Double-DIP as it represents the closest work in the literature that considered unsupervised models for the complex single-image reflection removal problem.

Fig. 9 shows sample results comparing our method versus Double-DIP. The results in the figure show that our method produces better separation quality, despite not needing any extra inputs. For example, as shown in the first two rows, our method performed better and separated the reflection from the background, whereas Double-DIP1 and Double-DIP2 failed to remove the reflection.

Next, we compare our method versus Double-DIP using PSNR and SSIM in Table III. The table shows that our method achieves higher PSNR and SSIM values, especially for the background layer. As commented before, PSNR and SSIM indicate the quality of the produced images, but they may not consider the layer separation quality.

E. Analysis of Our Method and Ablation Study

We conduct a detailed analysis of various components of the proposed method.

Ablation Study–Impact of Different Losses: Our method utilizes four types of losses: reconstruction loss, exclusion loss, similarity loss, and regularization loss. Since the reconstruction loss performs the most important role in the problem definition, we adjusted the weights of other losses based on this loss to obtain better separation results. Thus, we evaluate the impact of the different losses by adding each loss sequentially to the reconstruction loss as shown in Fig. 10. Since we utilize high-level features of perceptual embeddings, the separation result in the second column from the left in Fig. 10, when using only reconstruction loss, looks reasonable but not sufficient due to the ambiguity between the two layers. We add the exclusion loss to make the model decompose the input sample into two layers having different contents based on edge information. The results in the third column in Fig. 10 show better separation but still have some small artifacts. While the results in the fourth column might be similar to the ones in the third, the regularization term brings improvement in the speed of convergence and robustness of the model. We enhance the model with a cross-feedback structure and its corresponding loss to perform well even when the gradient information of the reflection layer is not enough. By adding the similarity loss, we obtain our best output shown in the last column in Fig. 10, which shows more solid separation in colors and shapes, in addition to its help on convergence and robustness.

Impact of α: The parameter α gives different weights to the background and reflection layers that are generated during the iterations and fed back to the two perceptual DIPs. We conducted experiments to analyze the impact of α by varying the value of α within its rage, which is between 0.0 and 0.5. Two sample results for $\alpha = 0.1$ and 0.4 are shown in Fig. 11. Our experiments show that the impact of α diminishes as we get closer to 0.5, as its influence on the two Perceptual DIPs becomes equal. In addition, smaller values of α tend to yield better layer separation results, as these values assign lower weights to the reflection layer. This is in line with the observation that the reflection layer tends to have lower pixel intensity than the background layer in natural images. Through experimentation, we found that α values around 0.1 produced the best results.

Perceptual Embedding: We analyze the impact of the perceptual embedding on the reflection separation using multiple images with different degrees of reflection. Recall that we modify a ResNet18 model to extract these features. We trained this model using two common datasets of objects: ImageNet [38] and Places365 [39]. This training does not need any datasets for image reflection removal and is done once.

Fig. 12 shows the importance of the perceptual embedding in separating the background layer from the reflection layer for two sample images. The results in the figure also indicate that using the Places365 dataset yields better layer separations than using the ImageNet dataset. This is because the Places365 dataset has more images for indoor and outdoor scenes, which usually exist in many reflection removal problems.

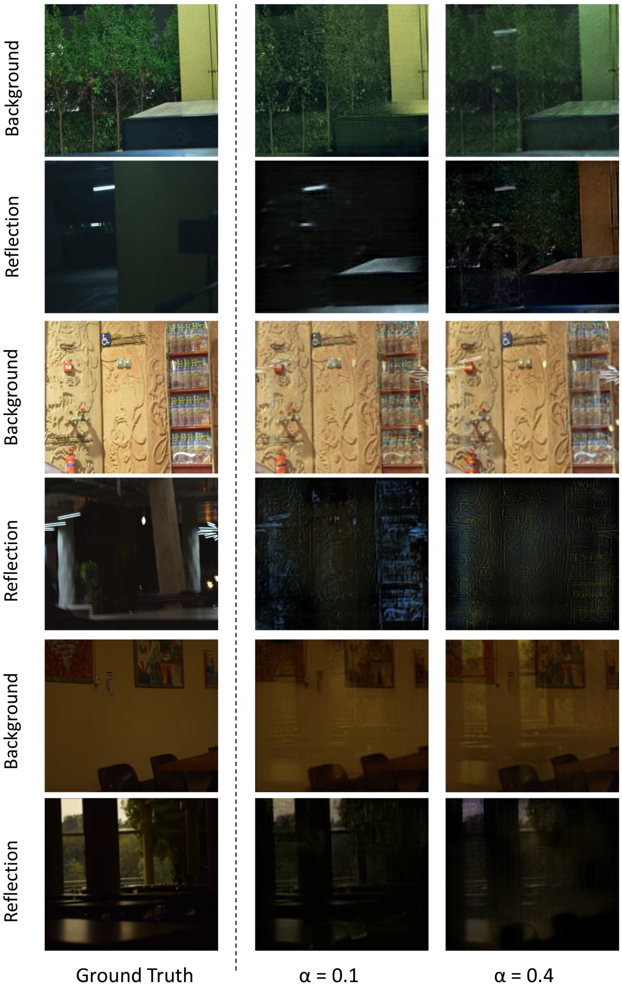

Fig. 11. Impact of α on the layer separation.

F. Limitations of the Proposed Method

As mentioned before, reflections in images can have many different forms, and removing these reflections is an ill-posed and complex problem. We analyze the cases in which our method fails to properly remove the reflections, and we contrast the performance of our method versus the performance of the two methods that produced the best results in our experiments, which are ERRNet [12] and Zhang *et al.* [10]. We note that ERRNet produces only the background layer, and thus in the figures the reflection images of ERRNet are not shown.

Through our experiments, we identified three challenging situations in which our method (and others) failed to remove the reflection: (i) dominant reflection, (ii) weak reflection, and (iii) dark images. An example of the first case is shown in Fig. 13, where the reflection is so strong that it dominates most of the objects in the background and makes it hard for our method

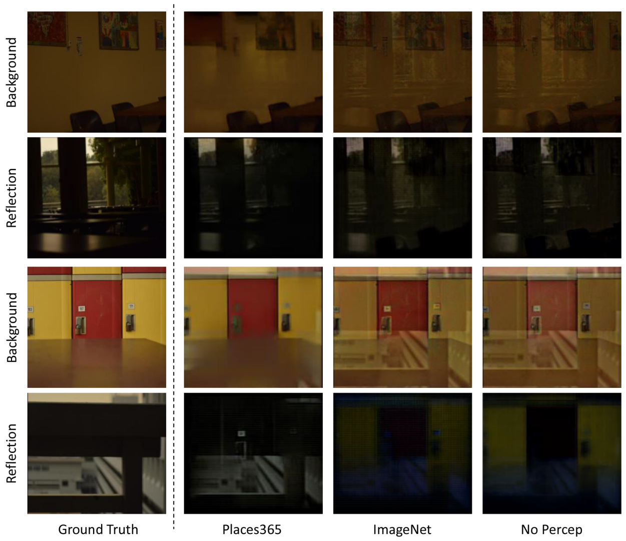

Fig. 12. The impact of Perceptual Embedding on layer separation.

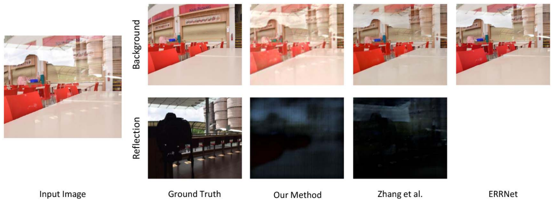

Fig. 13. Performance of our method and others in case of dominant reflections.

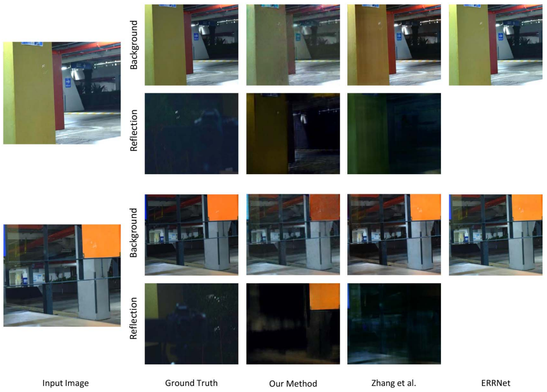

Fig. 14. Performance of our method versus others in the cases of weak reflection (top) and dark images (bottom).

to separate the background layer from the reflection layer. Examples of the other two cases are shown in Fig. 14, where in the second case, the reflection exists in the image, but it is very subtle, and in the third case, the image does not have enough illumination, and the pixels of the background and reflection layers cannot easily be distinguished.

V. Conclusion and Future Work

We have presented an unsupervised method for single-image reflection removal. To the best of our knowledge, this is the first unsupervised work for removing reflection from individual images of natural scenes. We have proposed a novel architecture of cross-coupled *Perceptual DIPs* that is capable of capturing not only the low-level statistics of a natural image but also the high-level semantic cues. We have also designed an optimization scheme using multiple loss functions without training on any dataset, which addresses the ambiguity of the single-image reflection removal problem, and leads to good separation results for natural images. Both qualitative and quantitative evaluations using real datasets show that our method outperforms the state-of-the-art supervised models. They also show that our method significantly outperforms the closest unsupervised approach in the literature, which, unlike our method, requires additional inputs to function.

The work in this paper can be extended in multiple directions. For example, the quality of the separated layers can further be improved by incorporating recent image restoration and inpainting techniques, which can potentially address some of the extreme reflection cases, e.g., when the reflection is strong and dominates the background scene.

References

[1] C. Sun et al., “Automatic reflection removal using gradient intensity and motion cues,” in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 466–470.

[2] A. Nandoriya, M. Elgharib, C. Kim, M. Hefeeda, and W. Matusik, “Video reflection removal through spatio-temporal optimization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2411–2419.

[3] J.-B. Alayrac, J. Carreira, and A. Zisserman, “The visual centrifuge: Model-free layered video representations,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2457–2466.

[4] Y. Yingda et al., “Deep reflection prior,” 2020, *arXiv:1912.03623*.

[5] 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.

[6] 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.

[7] 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.

[8] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9446–9454.

[9] Y. Gandelsman, A. Shocher, and M. Irani, “‘Double-DIP’: Unsupervised image decomposition via coupled deep-image-priors,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11026–11035.

[10] 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.

[11] 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.

[12] 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.

[13] R. Abiko and M. Ikehara, “Single image reflection removal based on GAN with gradient constraint,” *IEEE Access*, vol. 7, pp. 148790–148799, 2019.

[14] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot, “Depth of field guided reflection removal,” in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 21–25.

[15] 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 Int. Conf. Comput. Vis.*, 2017, pp. 3238–3247.

[16] H. Zhang et al., “Fast user-guided single image reflection removal via edge-aware cascaded networks,” *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2012–2023, Aug. 2020.

[17] D. Ma, R. Wan, B. Shi, A. Kot, and L. Duan, “Learning to jointly generate and separate reflections,” in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 2444–2452.

[18] R. Wan, B. Shi, H. Li, L.-Y. Duan, and A. C. Kot, "Reflection scene separation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2395–2403.

[19] B. H. P. Prasad, G. R. K. S., L. R. Boregowda, K. Mitra, and S. Chowdhury, "V-DESIRR: Very fast deep embedded single image reflection removal," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2390–2399.

[20] S. Niklaus et al., "Learned dual-view reflection removal," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3713–3722.

[21] Q. Zheng et al., "Single image reflection removal with absorption effect," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13395–13404.

[22] R. Wan, B. Shi, H. Li, L.-Y. Duan, and A. Kot, "Face image reflection removal," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 385–399, Feb. 2021.

[23] R. Wan et al., "CoRRN: Cooperative reflection removal network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 2969–2982, Dec. 2020.

[24] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3922–3930.

[25] Y. Pang, M. Yuan, Q. Fu, and D.-M. Yan, "Reflection removal via realistic training data generation," in *Proc. ACM SIGGRAPH Posters*, 2020, pp. 1–2, doi: 10.1145/3388770.3407419.

[26] Q. Wen et al., "Single image reflection removal beyond linearity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3771–3779.

[27] D. Lee, M.-H. Yang, and S. Oh, "Generative single image reflection separation," 2018, *arXiv:1801.04102*.

[28] P. Chandramouli and K. Vaishnavi Gandikota, "Blind single image reflection suppression for face images using deep generative priors," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 3315–3323.

[29] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 977–984.

[30] D. Sungatullina, E. Zakharov, D. Ulyanov, and V. Lempitsky, "Image manipulation with perceptual discriminators," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 579–595.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, *arXiv:1412.6980*.

[33] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 730–734.

[35] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5188–5196.

[36] Y. Zeng, Z. Lin, H. Lu, and V. M. Patel, "Image inpainting with contextual reconstruction loss," 2021, *arXiv:2011.12836*.

[37] Single-image-reflection-removal-dataset. [Online]. Available: https://www.kaggle.com/siboooo/singleimagereflectionremovaldataset

[38] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[39] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jan. 2018.

Hamed RahmaniKhezri received the B.Sc. degree in electrical engineering from Tehran University, Tehran, Iran, in 2019, and the M.Sc. degree in computing science from Simon Fraser University, Burnaby, BC, Canada, in 2021. His research interests include machine learning and multimedia systems.

Suhong Kim received the B.Sc. degree in mechanical and control system engineering from Handong Global University, Pohang, South Korea, in 2012, and the M.Sc. degree in computing science from Simon Fraser University, Burnaby, BC, Canada, in 2020. Her research interests include machine learning, computer vision, and multimedia systems.

Mohamed Hefeeda (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Mansoura University, Mansoura, Egypt, in 1994 and 1997, respectively, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2004. He is a Professor in the School of Computing Science at Simon Fraser University (SFU), Burnaby, BC, Canada, where he served as the Director of the School between 2018 and 2023. He founded and leads the Network and Multimedia Systems Laboratory (http://nmsl.cs.sfu.ca) at SFU. His research interests include multimedia systems, mobile and wireless video streaming, immersive video processing and delivery, and network systems and protocols. He has authored or co-authored more than 150 papers and multiple granted patents. Dr. Hefeeda was the recipient of the prestigious NSERC Discovery Accelerator Supplements (DAS) awards in 2011, which is granted to a select group of distinguished researchers from all Science and Engineering disciplines in Canada. His research on mobile multimedia systems has resulted in multiple patents and conference awards (e.g., ACM MMSys Best Paper, ACM Multimedia Best Demo, and IEEE Innovation Best Paper), and has been featured in several news venues, including ACM Tech News, World Journal News, and CTV British Columbia. He served on the editorial boards of premier journals, such as the *ACM Transactions on Multimedia Computing*, *Communications and Applications (TOMM)*, where he was named the Best Associate Editor in 2014. He served on the organization committees and/or Co/Chaired several international conferences, such as ACM MMSys, ACM MM, ICME, and NOSSDAV.