Unsupervised Single-Image Reflection Removal

by

SeyedHamed RahmaniKhezri

B.Sc., University of Tehran, 2019

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Fall 2021

Declaration of Committee

Name:	**SeyedHamed RahmaniKhezri**
Degree:	**Master of Science (Computing Science)**
Title:	**Unsupervised Single-Image Reflection Removal**
Examining Committee:	**Chair:** Khaled Diab University Research Associate, Computing Science
	Mohamed Hefeeda Supervisor Professor, Computing Science
	Jiangchuan Liu Committee Member Professor, Computing Science
	Ali Mahdavi-Amiri Examiner Assistant Professor, Computing Science

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

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 a vast 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 thesis, 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 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 individual images, and does not impose constraints or assumptions on the input images.

Keywords: Image Reflection; Unsupervised Learning; Deep Image Prior

Dedication

To mom and dad, who have always been and will be there for me, to my sister; and to those who are no longer with us but we carry their hopes with us.

Acknowledgements

I would like to state my sincere gratitude to my supervisor Dr. Hefeeda for his unremitting support and patience during this path. Also, I would like to thank Suhong Kim and Mohammad Nourbakhsh, who have been very helpful in the project. Finally, I am also thankful to the committee members, Dr. Diab as the chair, Dr. Liu as supervisor, and Dr. Mahdavi-Amiri as examiner.

Table of Contents

List of Tables

List of Figures

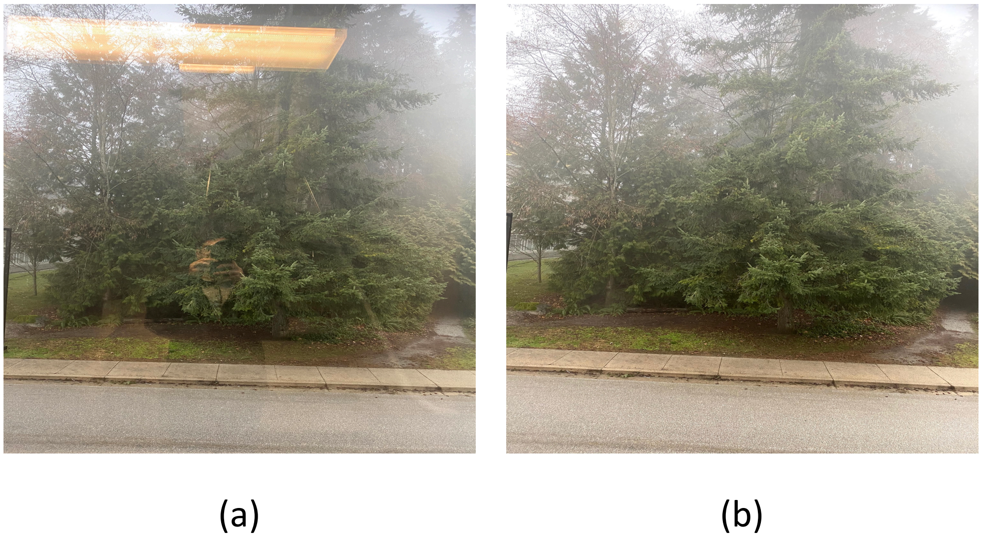


Figure 1.1: (a) A scenario in which reflections occur. The image shows a case where the photographer stands behind the glass (reflective surface) while taking the image. (b) A scenario in which the background scene is captured without any reflection appearing.

Chapter 1

Introduction

We frequently encounter unpleasant reflections when taking photos through transparent surfaces such as glass windows. Also, we might encounter reflections from non-transparent surfaces like water or car bodies. These reflections reduce the visual quality and utility of the captured photos as in Figure 1.1. We have addressed the reflection from transparent surfaces. 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, e.g., removing reflection caused by windshield images captured by surveillance cameras to see inside cars [5].

Removing reflection is, however, a challenging research problem. Reflection removal in a natural image can be interpreted as a layer decomposition 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:

$$I = B + R. \tag{1.1}$$

Eq. (1.1) 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 [22, 17, 3, 7]. 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 [13], blurriness of the reflection layer [14], and ghosting cues [20]. These approaches also fail when the assumptions do not hold, which regularly occurs because of the vast diversity of real-world images and thus, these low-level priors are not general enough in real cases. Moreover, most prior works, especially recent ones that utilize deep learning models, require a large amount of training data. 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 on images with different characteristics than those in the training datasets.

1.1 Contributions

In this thesis, 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 a complex problem. Our method builds on recent works which show that not all image priors must be learned from data. Instead, some of the image characteristics can be captured by the network structure itself. This is referred to as Deep Image Prior (DIP) [24], and it is suitable 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 essential components, which can be helpful in applications such as image dehazing, segmentation, watermark removal, and transparent layer separation. The generic image decomposition method in [9], however, requires multiple 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 Eq. (1.1). As mentioned earlier, a sequence of images of the same scene is not available in many cases. Moreover, requiring two different mixtures of the background and reflection layers as *input* is not practical, as these layers are 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, which can capture only low-level statistics of natural images like edges through the handcrafted structure of the network, 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 distribution of small patches within each separate layer (background and reflection) is simpler (more uniform) than in the image with reflection, resulting in strong internal self-similarity. 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 work can be summarized 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 or assumptions.

- The proposed method comprises three main parts: Perceptual DIP, cross-feedback, and refinement. The first one is a new architecture of the generator network by embedding semantic features, allowing the network to utilize both low-level image statistics and high-level perceptual information during the optimization. The cross-feedback structure encourages perceptually more meaningful separation by jointly optimizing two Perceptual DIPs' parameters without requiring additional inputs. The refinement part employs a semantically-guided in-painting neural network to improve 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 [33, 30, 28, 2]. Our university's Research Ethics Board approved the subjective study. 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 our unsupervised method substantially outperforms all prior works on real-world images with complex reflections 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 rigorously analyze the various components of the proposed method and conduct ablation studies to show the importance and contribution of each component to the end result.

1.2 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 summarizes the related work in the literature. Chapter 3 presents the proposed method. Chapter 4 compares the performance of the proposed method against the closest works in the literature, and Chapter 5 concludes the thesis. Details about the subjective evaluation form are given as an Appendix.

Chapter 2

Related Work

The problem of image reflection removal has been explored in two general directions: using traditional methods, and through deep-learning. In this chapter, we review the related works in these two directions.

2.1 Traditional Methods

As mentioned in Chapter 1, 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 multiple-image approaches for reflection removal include [22, 17, 3, 7]. 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 work.

Multiple traditional (i.e., not neural network-based) prior works addressed the single-image reflection removal problem by imposing priors or assumptions on reflection to make the problem tractable. Examples of these assumptions include the sparse prior of gradients and local features [13], blurrier reflection prior which penalizes large reflection gradients [14], ghosting cues [20], and different depth of fields between the two layers that is used for edge labelling and layer separation [27]. Still, just adding these priors would not be sufficient due to the variety of natural images.

2.2 Deep-learning based Methods

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 [6, 32, 15, 33, 30, 2, 28]. We provide brief descriptions of these works in the following.

Fan et al. [6] introduce a solution using weakly supervised learning for training a single reflection removal model. They synthesize a training samples database that captures the background and reflection statistics and replaces prior knowledge injected through explicit gradient penalization or energy minimization with a particular deep network to capitalize on this form of weak supervision. Ma et al. [15] use unpaired supervision to design a weakly-supervised framework by integrating reflection generation and separation into a single model. Zhang et al. [32] 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. [33] utilizes perceptual losses to improve the separation of the background layer from the reflection layer. Yang et al. [30] propose a cascade deep neural network (referred to as BDN) to estimate background and reflection layers bidirectionally. Abico et al. [2] introduce a gradient constraint loss along with the generative adversarial networks to produce high-quality background layers. This approach is referred to as GCNet. Wei et al. [28] propose an enhanced 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.

All of the above methods employ supervised-learning models, which require training datasets. Wan et al. [25] collect a dataset of real images with and without reflection, which is referred to as the single-image reflection dataset (SIR^2) [26] and is frequently used as a benchmark for evaluating image reflection removal algorithms. In addition, some prior works generate synthetic datasets for the image reflection problem through physically-based polarization pipeline [18], non-linear blending formulation [29], and generative adversarial training [12].

In our evaluations, we compare the proposed (unsupervised) method against four supervised methods for image reflection removal, which are Zhang el al. [33], BDN [30] GCNet [2] and EERNet [28]. These four methods represent the state-of-the-art, and they outperform prior ones. We utilize benchmark real image datasets, including [26]. In addition, we compare against the unsupervised image decomposition method (Double-DIP) in [9], although, as mentioned in Chapter 1, this method requires extra inputs that are typically not available in practice. We show that the proposed method outperforms Double-DIP, even when Double-DIP uses the extra inputs.

Finally, we note that Chandramouli et al. [4] 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, this method can only handle face images and does not generalize to other types of images with reflection. Thus, we could not compare against it.

Chapter 3

Proposed Method

This chapter describes the proposed solution for the single-image reflection removal problem.

3.1 Basic Elements and Approach Overview

At a high level, the proposed method works as follows. Given an Image I with reflection, the method produces a reflection-free background B^*. The method first decomposes the Image I into an estimated background layer $\tilde{B}$ and an estimated reflection layer $\tilde{R}$ through an unsupervised manner which involves iterative optimization steps. Then, the method uses an in-painting model as a refinement step, generating a refined background B^*.

Prior works have shown that the empirical entropy of small patches inside a natural image is much smaller than the entropy across different images [36]. 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 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 explained in the following.

Perceptual DIP: Employing perceptual cues has shown remarkable advantages in capturing semantic meanings for various image-related tasks. Several recent deep-learning techniques improve the performance by combining 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 L1 or L2 distance between high-dimensional features is insufficient to capture the real difference between them. In addition, an adversarial loss requires paired ground-truth datasets of background and reflection layers to discriminate real and fake data via supervised learning.

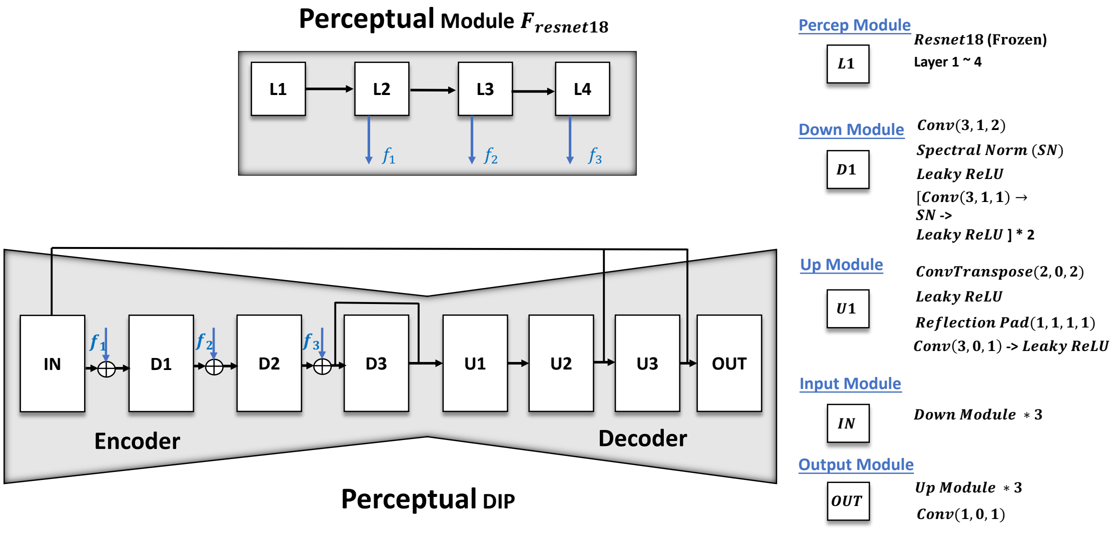


Figure 3.1: The structure of the proposed Perceptual DIP. High-level features of the input image are extracted from ResNet18, and are used in the Perceptual DIP as $f1$, $f2$, and $f3$.

Reflection separation is a low-level vision task, but it is a complex and ill-posed problem. To address this complexity and reduce ambiguity, we 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 [23], we design an encoder-decoder style network with perceptual embedding, which is referred to as *Perceptual DIP*, as shown in Figure 3.1. At the initialization step, the perceptual embedding module extracts multi-level features from a pre-trained image classifier. We chose ResNet18 [10] as our backbone structure of the perceptual module, which has four layers. We skip the first layer output because features from this layer are more sensitive to low-level information of the image, similar to those captured by DIP, while our expectation for this module is to incorporate high-level features. Then, the extracted feature maps are concatenated with the features of each layer in the encoder, constructed to fit well with the size of the perceptual embedding and the input image.

Cross-feedback Perceptual DIPs: We propose coupling of two perceptual DIPs, where the output of one Perceptual DIP is fed back into the other DIP, as shown in Figure 3.2. 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 Eq. (1.1) at each iteration t as $\tilde{B}^c_t = I - \tilde{R}_t$ and $\tilde{R}^c_t = I - \tilde{B}_t$.

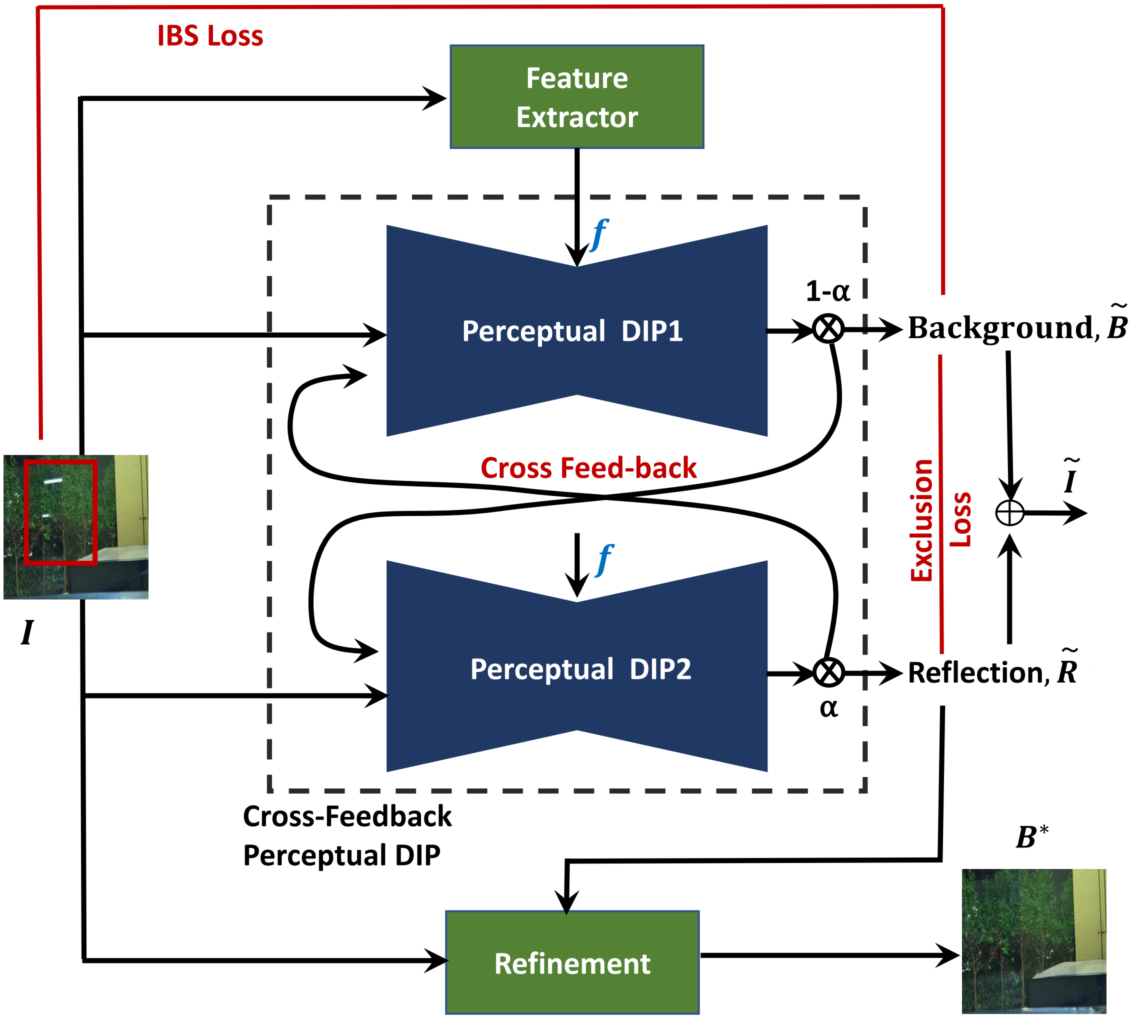


Figure 3.2: Overview of the proposed method for image reflection removal. Two DIP networks with perceptual embedding are coupled with cross-feedback and loss functions, generating background and reflection layers from an input image. The (main) background layer is achieved after going through a final refinement stage.

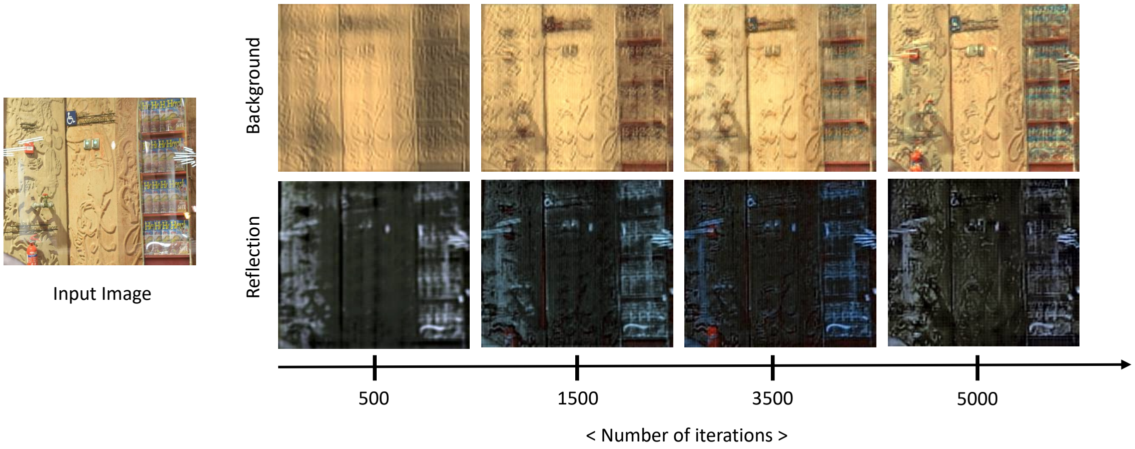


Figure 3.3: The effect of cross-feedback. At the early stage, up to 500 iterations, both layers are separated based on low-level features like color and edges. However, at iteration 500, the reflection layer restores objects while the background recovers other parts of the scene by excluding each other, similar to human perception such that at iteration 5000, our model manages to remove the reflection from the background, separating it into another layer.

In Figure 3.3, 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 study the impact of the perceptual embedding on the reflection separation in Chapter 4.5.

Approach Overview: A high-level overview of the proposed method for single image reflection removal is depicted in Figure 3.2. 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. 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. The details of the used loss functions and refinement step are presented in Chapters 3.2 and 3.3.

3.2 Optimization and Losses

Optimization Scheme and Perceptual DIPs: 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

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 [11]
    $\tilde{B}^c_t = I - \tilde{R}_t$
    $\tilde{R}^c_t = I - \tilde{B}_t$
**end for**
**return** $\tilde{B}_t, \tilde{R}_t$
```

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:

$$\begin{cases} \tilde{B}_t & = (1 - \alpha_t) \cdot \hat{B}_t \\ \tilde{R}_t & = \alpha_t \cdot \hat{R}_t \end{cases} \tag{3.1}$$

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, implying that reflections are relatively weaker than the background scene in general cases. The impact of α in our model is evaluated in Chapter 4.5.

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

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

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{recon} + \lambda_2 \cdot \mathcal{L}_{excl} + \lambda_3 \cdot \mathcal{L}_{sim} + \lambda_4 \cdot \mathcal{L}_{reg}, \tag{3.2}$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are the corresponding weights for each loss functions; 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: As we only have one single image without any pairs of ground-truth, we find that combining different types of reconstruction losses helps the network to converge faster. It should be noted that our model is based on the reconstruction of the input image by combining the extracted background and reflection layer from our model. It is intuitive that the reconstructed image $\tilde{I}$ should be similar to the original input of the optimization iterations. Thus, we define our reconstruction loss as:

$$\begin{aligned} \mathcal{L}_{recon} &= \mathcal{L}_{color} + \omega_1 \cdot \mathcal{L}_{gray} + \omega_2 \cdot \mathcal{L}_{grad}, \\ \mathcal{L}_{color} &= \|I - \tilde{I}\|_2, \\ \mathcal{L}_{gray} &= \|c(I) - c(\tilde{I})\|_2, \\ \mathcal{L}_{grad} &= \| \bigtriangledown_x I - \bigtriangledown_x \tilde{I}\|_1 + \| \bigtriangledown_y I - \bigtriangledown_y \tilde{I}\|_1, \end{aligned} \tag{3.3}$$

where $c(\cdot)$ is the conversion function from RGB image to gray-scale image, and $\bigtriangledown_{x,y}(\cdot)$ denotes the gradient of the input with the Sobel filter. The main reconstruction loss is a pixel-wise $\mathcal{L}2$ distance between the given image and the recombined image $\tilde{I}$ in the RGB color space. We also design the same $\mathcal{L}2$ losses both in the gray color 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 and helps to prevent the model from generating blurry images. As seen in figure 4.8, having only the reconstruction loss will result in a simple layer separation, in which the images are neither smooth nor meaningful. Still, the only purpose they serve is to have their linear combination as close as possible to the input image.

Exclusion Loss: The exclusion loss aims to minimize the correlation between two edges of the background layer and the reflection layer at multiple spatial resolutions. This will allow us to capture more contextual information from various scales, which can consider different scales of both low-level and high-level information. Thus, similar to [33], we define the exclusion loss as:

$$\mathcal{L}_{excl} = \sum_{n=1}^{N} \|norm(\bigtriangledown \tilde{B_n}) \odot norm(\bigtriangledown \tilde{R_n})\|_F, \tag{3.4}$$

where n is the image downsampling factor, as exclusion loss minimizes the correlation between background edges and reflection at multiple spatial resolutions, each time in Eq. (3.4) the image is downsampled by a factor 2, and we chose N as 3 in the experiment. $norm(\cdot)$ is the normalization in gradient fields of the two layers, $\odot$ is the element-wise multiplication, and $\| \cdot \|_F$ denotes the Frobenius norm.

The key observation is that the edges of the transmission and the reflection layers are unlikely to overlap through the samples, as An edge in I should be caused by either B or R, but not both. This loss is effective in separating the background and reflection layers at the pixel level. As shown in figure 4.8, disabling this loss would make the separation of two layers weaker, especially in sections like edges. In addition, more residual reflections may remain visible in the extracted background.

Similarity Loss: The proposed model exploits cross-feedback to empower the network to exclude one another under the assumption that each generated layer should be similar to its corresponding cross-feedback from the other network as well as its previous output. We call the first constraint as the **cross-consistent loss** $\mathcal{L}_{cc}$ and is defined as follows:

$$\mathcal{L}_{cc} = \|\tilde{B}_t - (I - \tilde{R}_{t-1})\|_2 + \|\tilde{R}_t - (I - \tilde{B}_{t-1})\|_2, \tag{3.5}$$

Our observation suggests that although reflection is evident in an image with reflection, the most dominating part of the image is the background. The work in [34] shows that for a deep network to produce visually pleasing images, the error function should be perceptually motivated. l_1 preserves colors and luminance. 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 l_2 loss with small effect is more suitable in both preserving details and separating reflection. We define the **Input-Background-Similarity (IBS) loss** as follows:

$$\begin{aligned}\mathcal{L}_{IBS} &= \omega_1 \cdot \|\tilde{B}_t - I\|_2 + \omega_2 \cdot \mathcal{L}_{percep}, \\ \mathcal{L}_{percep} &= \lambda_m \cdot \sum_m \|f(\tilde{B}_t) - f(I)\|_1,\end{aligned} \tag{3.6}$$

Semantic reasoning about the scene would benefit the task of reflection removal [33]. A feature loss combining low-level and high-level features from a perception network serves our purpose. Perceptual loss is defined based on the activation of the 19-layer VGG [21] trained on ImageNet. $f(\dot{)}$ operator is the activation of an image at a certain level, and the perceptual loss calculates l_1 distance between activation of two images at each level. λ_m is a balancing weight for each layer, and we put the most significant weight to emphasize low-level features and edges. We used convolution layers similar to [28]. As figure 4.8 illustrates, this loss would help to increase the details and color consistency of the extracted background by having a weak comparison against the original input, and through working with the other loss terms, it won't be similar in the reflection characteristics.

Combining the two losses mentioned above, we get:

$$\mathcal{L}_{sim} = \mathcal{L}_{cc} + \mathcal{L}_{IBS}. \tag{3.7}$$

Regularization Loss: We regulate the network under three priors: a total-variance loss $\mathcal{L}_{TV}$[16], a total-variance balance loss $\mathcal{L}_{TVB}$ that we applied on our own, and a ceiling rejection loss $\mathcal{L}_{ceil}$[7], which are defined as follows:

$$\begin{aligned}
\mathcal{L}_{reg} &= \gamma_1 \cdot \mathcal{L}_{TV} + \gamma_2 \cdot \mathcal{L}_{TVB} + \mathcal{L}_{ceil}, \\
\mathcal{L}_{TV} &= \| \bigtriangledown \tilde{B}_t \|_1 + \| \bigtriangledown \tilde{R}_t \|_1, \\
\mathcal{L}_{TVB} &= \| \bigtriangledown \tilde{B}_t \|_1 - \| \bigtriangledown \tilde{R}_t \|_1, \\
\mathcal{L}_{ceil} &= \sum_m f(\tilde{B}_t, I, m) + f(\tilde{R}_t, I, m), \\
f(x, y, m) &= \begin{cases} \|x_m - y_m\|_1 & if\ x_m > y_m \\ 0 & otherwise \end{cases},
\end{aligned} \tag{3.8}$$

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 gives 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.

All coefficients of our loss are fine-tuned through experiments.

3.3 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, resulting in 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 [31]. This contextual in-painting method requires user-specified masks for areas that have damages in the image. We adapt the contextual in-painting method to the reflection removal problem as follows. Reflections in images can be thought of as obstructions that cause damages in 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 from this reflection layer based on [8] and use it to *fix* the damages (reflections in this case) in the full-resolution input image using the contextual in-painting method.

Chapter 4

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 and 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.

4.1 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 commonly been used to evaluate prior methods for image reflection removal in the literature, including the ones compared against in this paper.

DS1 [26] consists of are hundreds of images. 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 the glass, 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. And the ground truth background later is captured by removing the glass.

The second dataset, DS2, contains 20 images [33]. This dataset has a ground truth for the background layer only. Images are captured through 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 [1] 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 [30], GCNet [2], ERRNet [28], and Zhang et al. [33]. 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 comparisons.

Implementation Details: We experimentally set the values of weights of different losses. We set λ_1, λ_2, λ_3 and λ_4 as 1.5, 0.13, 1.0, and 1.0, respectively. For the reconstruction loss, we set the value of ω_1 and ω_2 as 0.09. The values of γ_1 and γ_2, the regularization loss coefficients, are set to 0.003 and 0.003 in our experiments. As for ω_1 and ω_2 in the similarity loss, we set them 0.1. Note that decreasing the value of similarity loss will result in a better separation of reflection. Since our method is based on optimizing the model parameters on the single image input with the size of 224*224, the batch size is set as 1 and the parameters are updated with a learning rate of 0.0001 until the number of iteration (epochs) reaches 5500.

4.2 Comparison using Subjective Study

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 through. **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 participant were female. The participants 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 [30], GCNet [2], ERRNet [28], Zhang et al. [33], 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 columns 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 the subjects to consider whether the method has removed the reflection while preserved image visual quality. 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. More details about the subjective study are presented in the Appendix.

Each of the 50 participants 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 4.1. 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 4.1 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 [28]).

We also show more statistics of our studies in the Appendix.

Figure 4.1 separates the average scores over all the scenes into five bins, and indicates how many users' average score for our method is in each of these score ranges. Through this histogram, we see around 94% of users had average scores of three or higher over all the images generated by our method.

Table 4.1: Summary statistics of the subjective study.

	Average MOS	**Median MOS**
BDN [30]	2.68	2.75
GCNet [2]	2.49	2.5
ERRNet [28]	2.87	2.84
Zhang et al. [33]	2.74	2.75
Ours	**3.82**	**3.94**

4.3 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 Figure 4.2, Figure 4.3, and Figure 4.4, on datasets DS1, DS2, and DS3, respectively. Samples are chosen randomly from images that would have a visible reflection. 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 Figures 4.2, 4.3, and 4.4 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 Figure 4.2, 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 models 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 Figures 4.3 and 4.4.

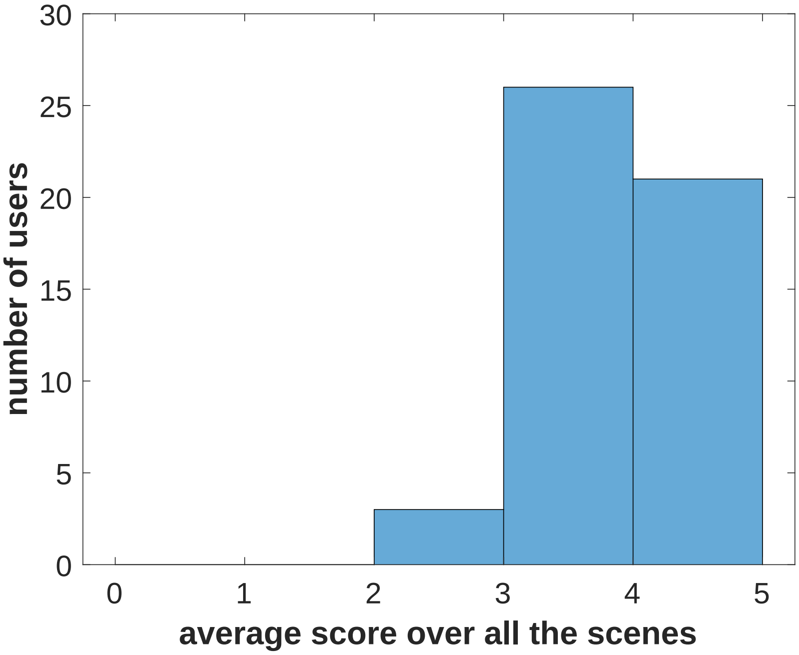


Figure 4.1: Number of users with average scores in each range.

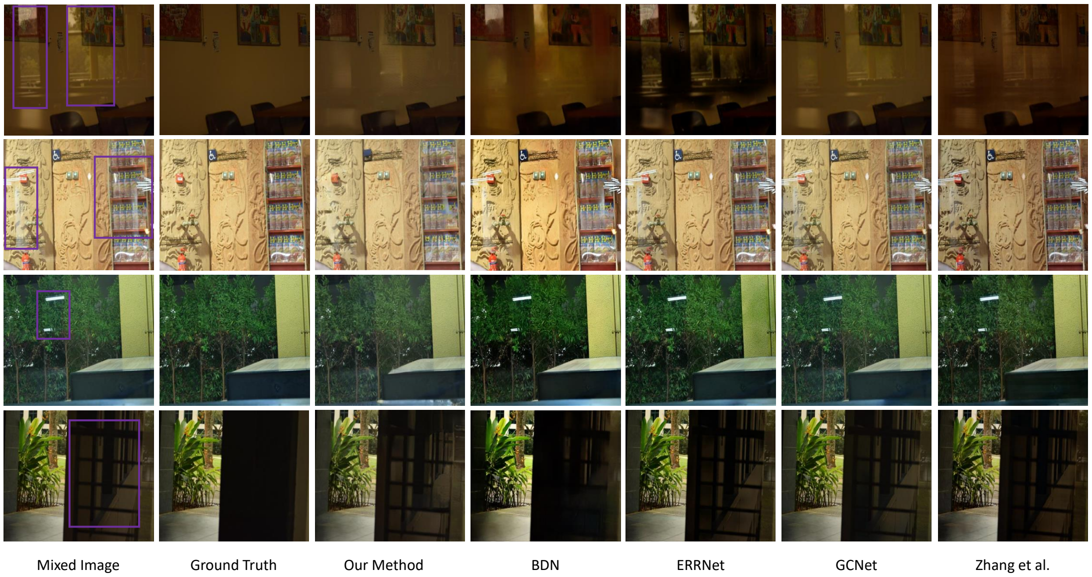


Figure 4.2: Comparing our unsupervised method versus four supervised methods on dataset DS1.

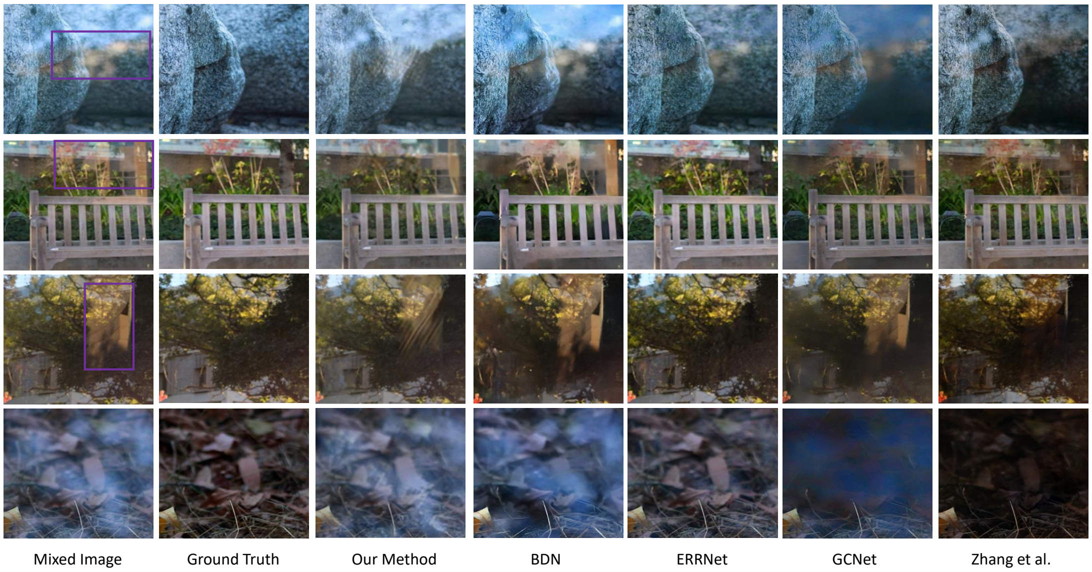


Figure 4.3: Comparing of our unsupervised method versus four supervised methods on dataset DS2.

We further analyze the quality of the layer separation of different methods in Figure 4.5. This figure shows both the background and reflection layers produced by various methods and compares them against each other and the ground truth. We show the results for only our method as well as the BDN [30] and Zhang et al. [33] methods, as they were the ones that produced the best results from prior works, as indicated in Figures 4.2, 4.3, and 4.4. As Figure 4.5 shows, our method produces a cleaner separation of the background and reflection layers.

Objective Comparisons: Next, we compare our method versus others using the PSNR and SSIM objective metrics. The results for dataset DS1 are presented in Table 4.2, which shows that our method results in somewhat smaller SSIM and PSNR values than some of the other methods. We note the SSIM and PSNR do not measure the quality of separation. Instead, they measure the quality of the produced images, even if the separation of the layer was not done properly. We illustrate this in Figure 4.6, 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 too much 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 removes most of the reflection from the image and produces images with acceptable PSNR and SSIM values.

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 has

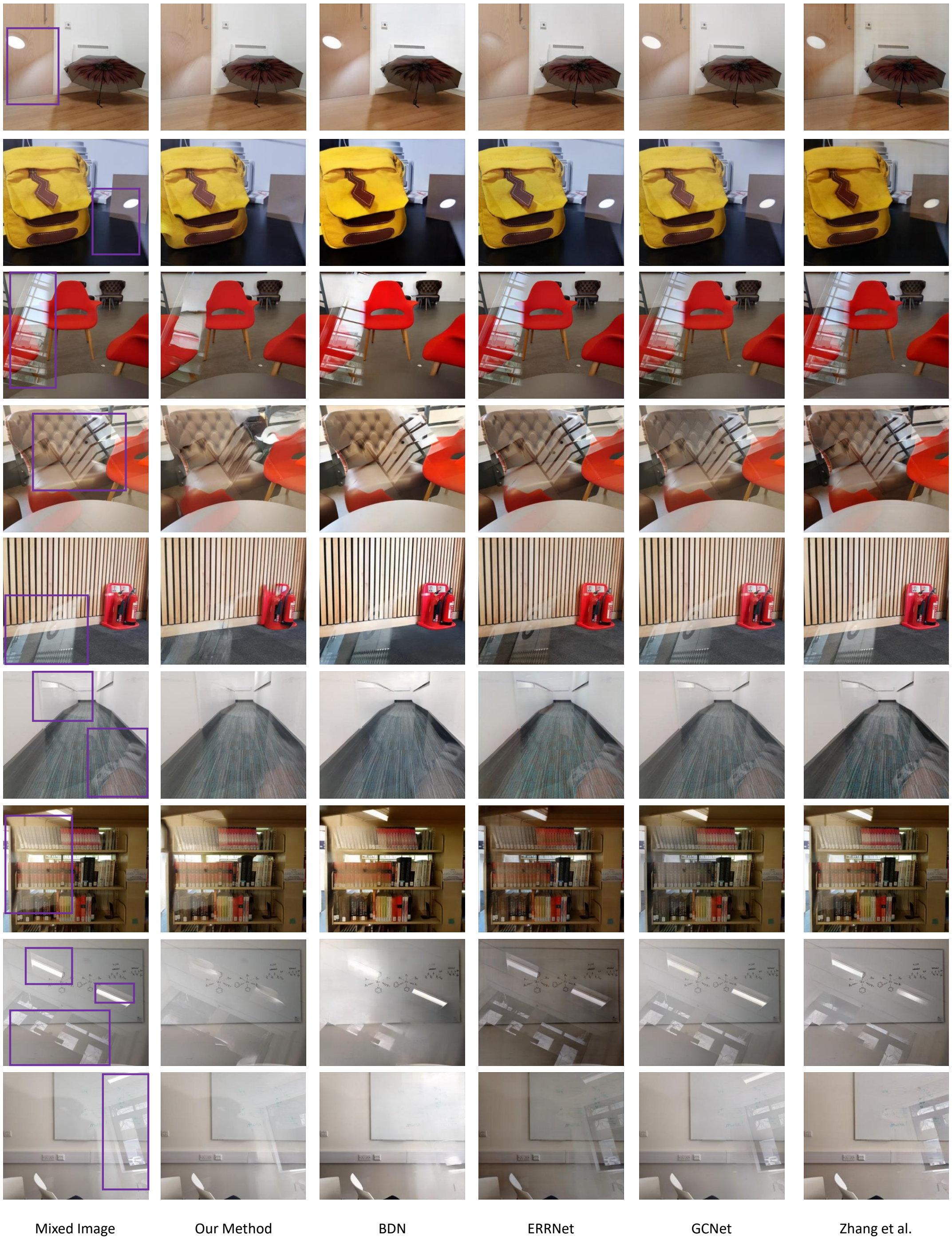


Figure 4.4: Comparing of our unsupervised method versus four supervised methods on dataset DS3. Eight of the shown nine images were used in the subjective study.

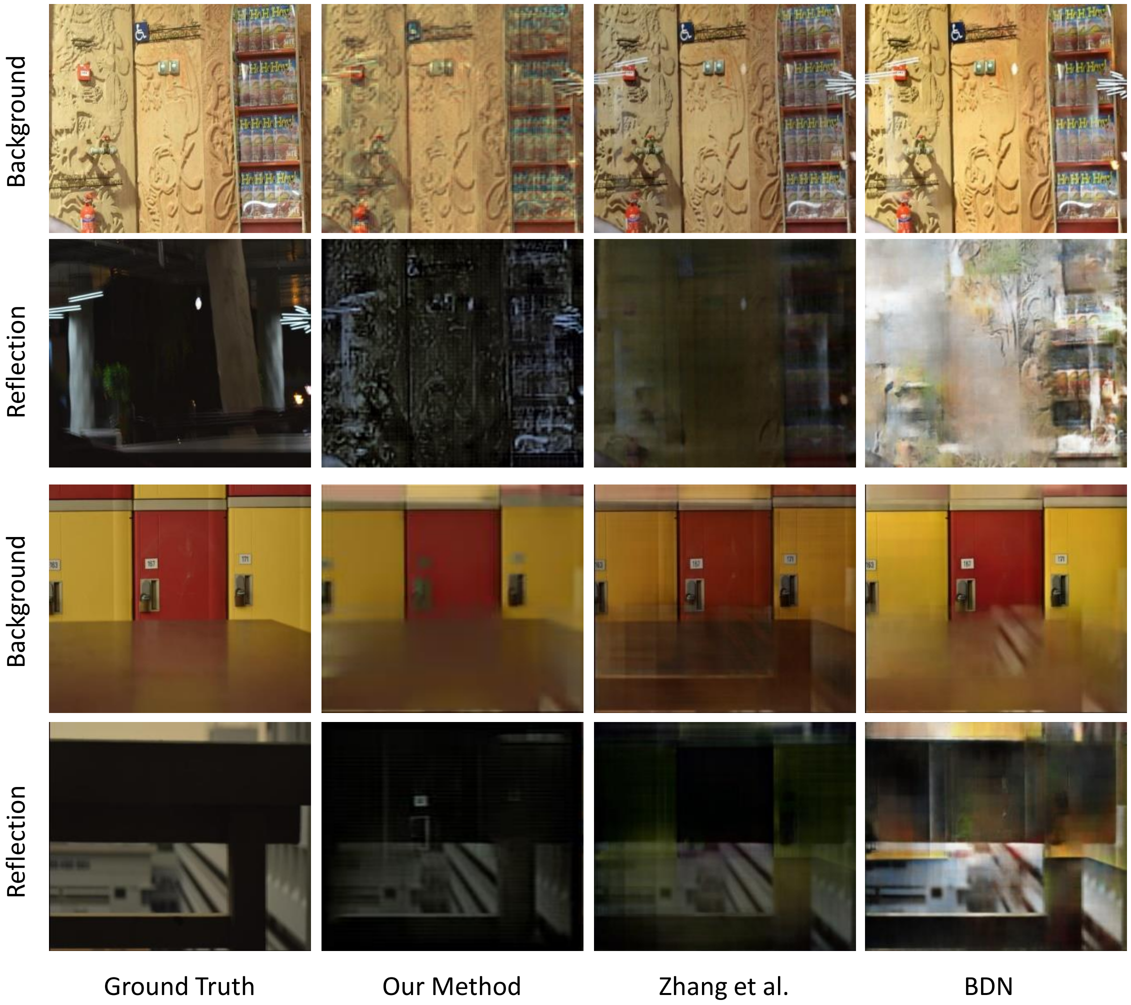


Figure 4.5: Comparison of the separation quality produced by our method versus BDN [30] and Zhang et al. [33] methods.

numerous varieties. In contrast, our model 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 sample and uses them to separate the input into two layers.

Table 4.2: 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 [30]	22.01	9.01	0.86	0.31
GCNet [2]	24.53	—	0.92	—
Zhang et al. [33]	21.13	20.88	0.87	0.64
ERRNet [28]	23.86	—	0.88	—
Ours	20.52	20.28	0.82	0.41

4.4 Comparison against the Double-DIP Unsupervised Layer Separation Method

As mentioned in chapter 1, 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 is referred 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

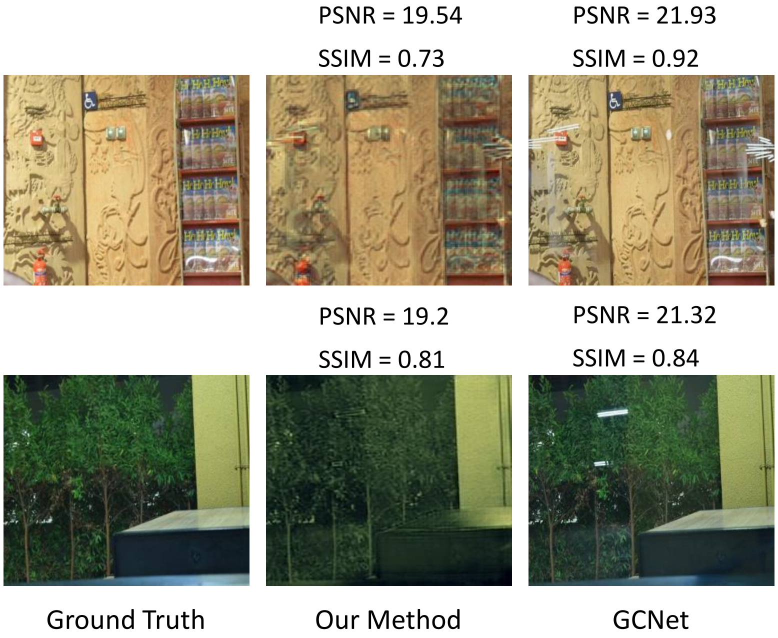


Figure 4.6: Comparison between the output of our model and GCNet to show the importance of the visual quality over the objective PSNR and SSIM metrics. Although GCNet's output achieved better PSNR and SSIM, it did not remove much of the reflection, whereas our method removed most of the reflection.

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.

Figure 4.7 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 4.3. 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.

Table 4.3: 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
Ours	20.52	20.28	0.82	0.41

4.5 Analysis of our Method and Ablation Study

In this chapter, 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 Figure 4.8. Since we utilize high-level features of Perceptual Embeddings, the separation result in the second column from the left in Figure 4.8, 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 Figure 4.8 show better separation but still has some small artifacts. While the results the fourth column are 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 joining the similarity loss, we can obtain our best output shown in the last column in Figure 4.8, which shows more solid separation in color 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 by varying the value of α within its rage, which is between 0.0

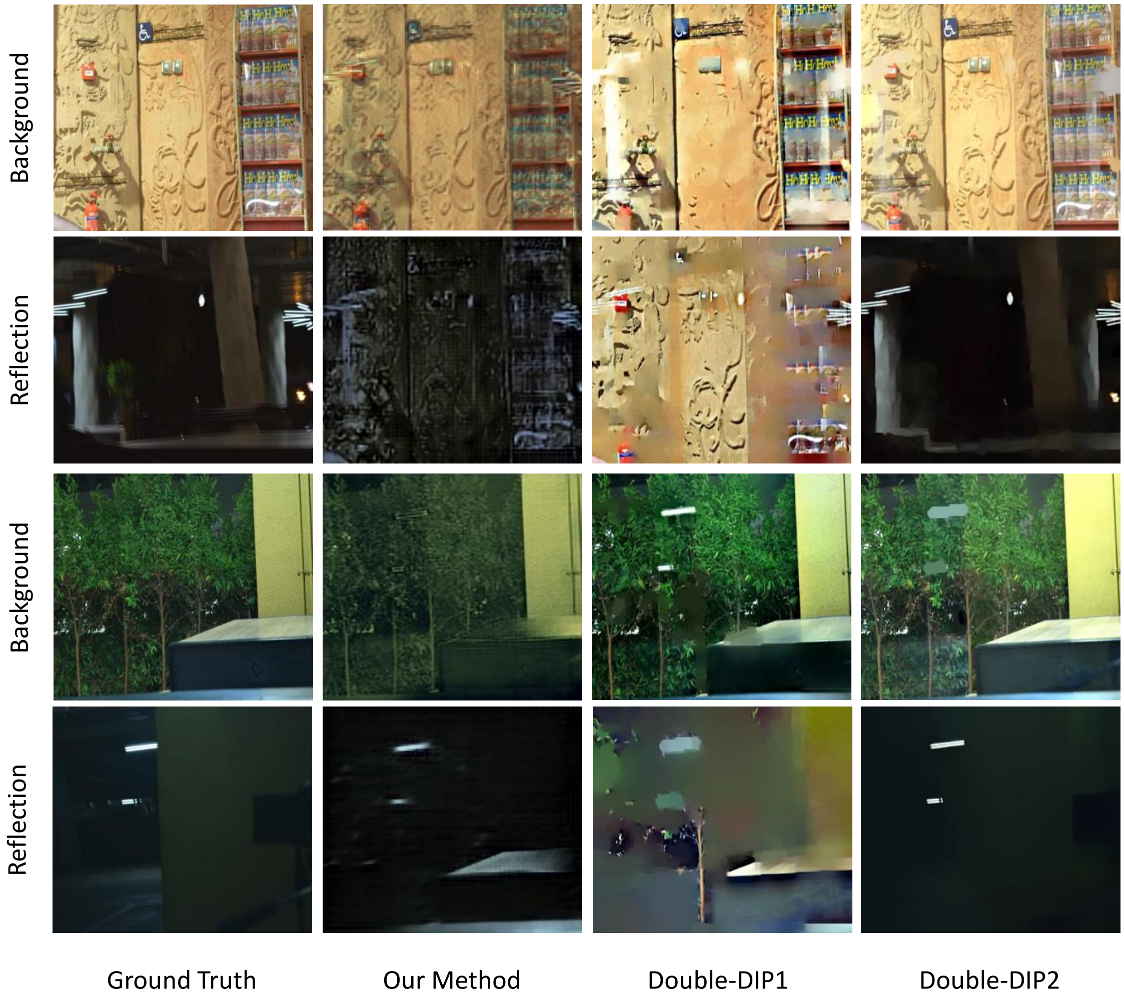


Figure 4.7: Comparing our method against the unsupervised Double-DIP [9].

and 0.5. Two sample results for $\alpha = 0.1$ and 0.4 are shown in Figure 4.9. 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 inline 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 small α values around 0.1 resulted in the best results.

Perceptual Embedding: We analyze the impact of the perceptual embedding on the reflection separation using multiple images with different degrees of reflections. Recall that we modify a ResNet18 model to extract these features. We trained this model using two common datasets of objects: ImageNet [19] and Places365 [35]. This training does not need any datasets for image reflection removal and is done once.

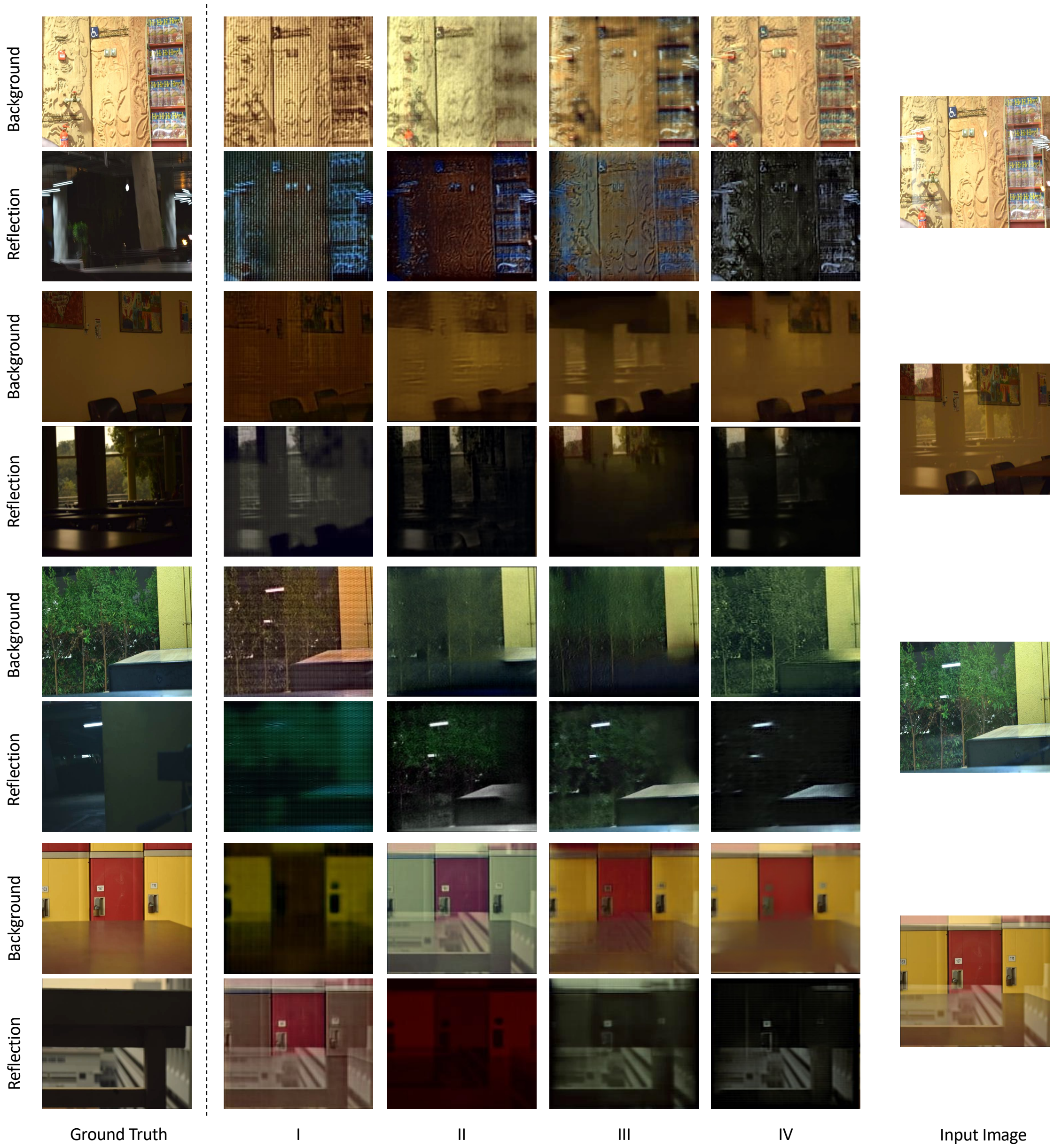


Figure 4.8: Ablation study to analyze the impact of different losses in four different scenarios in two real images: "I": Using only the Reconstruction Loss, "II": Reconstruction + Exclusion, "III": Reconstruction + Exclusion + Regularization Loss, and "IV": All the losses.

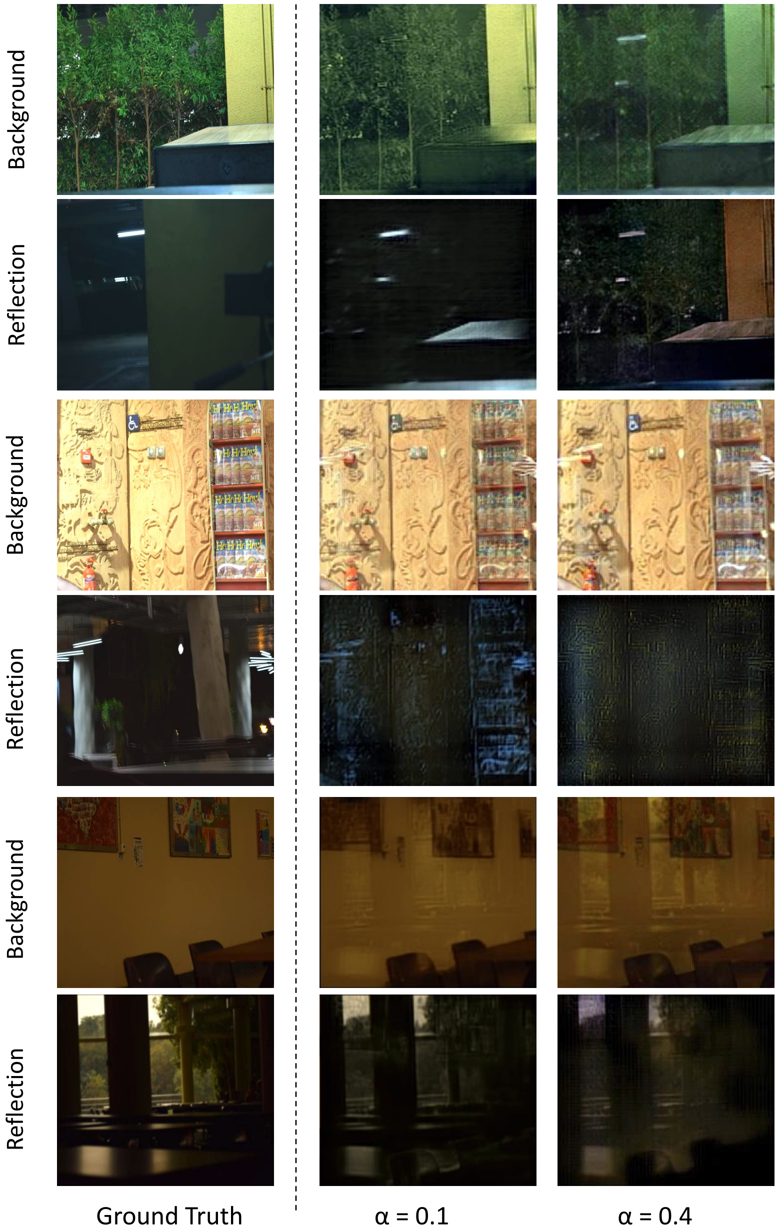


Figure 4.9: Impact of α on the layer separation.

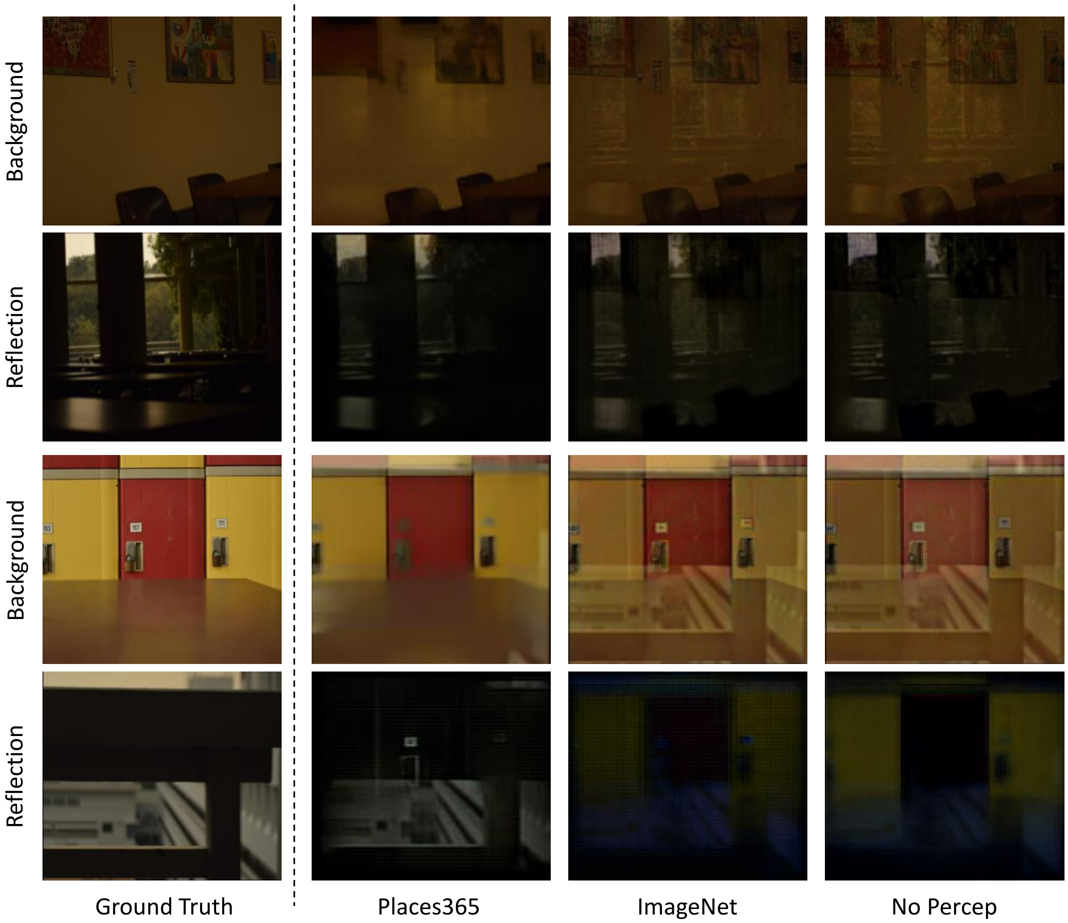


Figure 4.10: The impact of Perceptual Embedding on layer separation.

Figure 4.10 shows the importance of the perceptual embedding in separating the background layer from the reflection layer for two sample images. The results int 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.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this thesis, we considered the problem of single-image reflection removal. Prior computational methods approached this problem by making assumptions, which are impractical in real-world scenarios. Recent deep learning based methods, on the other hand, utilize supervised learning models, which are heavily reliant on datasets to be general and well-performed, which is scarce when it comes to non-synthetic datasets.

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 for natural scenes using only a single image. 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 using Deep Image Priors but also the high-level semantic cues through the perceptual module. We have also designed an optimization scheme using multiple loss functions without training on any dataset, which significantly resolves the ambiguity of single-image separation and leads to good separation results for natural images. Both qualitative and quantitative evaluations on real datasets show that our method outperforms the state-of-the-art supervised models. It also significantly outperforms the closest unsupervised method in the literature, which, unlike our method, requires additional inputs to function.

5.2 Future Work

The work in this paper can be extended in multiple directions. For example, the refinement stage of the proposed method can further be improved to remove any visual artifacts that may occur around areas with strong reflections.

Another extension is to dynamically adjust the value of α. In the current setting, we fix α in a way to indicate that the intensity of the reflection layer is less than the background layer, which captures the most realistic scenarios. However, this may fail when the reflection is too strong or more dominant. In order to solve this problem, future work can consider using a neural network to estimate α from labeled data.

Bibliography

[1] Single-image-reflection-removal-dataset. `https://www.kaggle.com/siboooo/singleimagereflectionremovaldataset`.

[2] Ryo Abiko and Masaaki Ikehara. Single image reflection removal based on gan with gradient constraint. *IEEE Access*, 7:148790–148799, 2019.

[3] Jean-Baptiste Alayrac, Joao Carreira, and Andrew Zisserman. The visual centrifuge: Model-free layered video representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2457–2466, 2019.

[4] Paramanand Chandramouli and Kanchana Vaishnavi Gandikota. Blind single image reflection suppression for face images using deep generative priors. In *roceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[5] Chi-Rung Chang, Kuan-Yu Lung, Yi-Chung Chen, Zhi-Kai Huang, Hong-Han Shuai, and Wen-Huang Cheng. Stop hiding behind windshield: A windshield image enhancer based on a two-way generative adversarial network. In *Proceedings of the ACM Multimedia Asia*, MMAsia '19, New York, NY, USA, 2019. Association for Computing Machinery.

[6] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3238–3247, 2017.

[7] Qingnan Fan, Yingda Yin, Dongdong Chen, Yujie Wang, Angelica Aviles-Rivero, Ruoteng Li, Carola-Bibiane Schnlieb, Dani Lischinski, and Baoquan Chen. Deep reflection prior. *arXiv preprint arXiv:1912.03623*, 2019.

[8] R. FU, P. KUANG, Y. ZHOU, H. YAN, and T. ZHENG. Area-aware reflection detection and removal for single image. In *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*, pages 307–310, 2019.

[9] Yosef Gandelsman, Assaf Shocher, and Michal Irani. "double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. Generative single image reflection separation. *arXiv preprint arXiv:1801.04102*, 2018.

[13] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1647–1654, 2007.

[14] Yu Li and Michael S Brown. Single image layer separation using relative smoothness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2759, 2014.

[15] D. Ma, R. Wan, B. Shi, A. Kot, and L. Duan. Learning to jointly generate and separate reflections. In *2019 IEEE/CVF International Conference on Computer Vision*, pages 2444–2452, 2019.

[16] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[17] Ajay Nandoriya, Mohamed Elgharib, Changil Kim, Mohamed Hefeeda, and Wojciech Matusik. Video reflection removal through spatio-temporal optimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2411–2419, 2017.

[18] Youxin Pang, Mengke Yuan, Qiang Fu, and Dong-Ming Yan. Reflection removal via realistic training data generation. In *ACM SIGGRAPH 2020 Posters*, SIGGRAPH '20, New York, NY, USA, 2020. Association for Computing Machinery.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[20] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman. Reflection removal using ghosting cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3193–3201, 2015.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, San Diego, CA, USA, May 7-9, Conference Track Proceedings*, 2015.

[22] Chao Sun, Shuaicheng Liu, Taotao Yang, Bing Zeng, Zhengning Wang, and Guanghui Liu. Automatic reflection removal using gradient intensity and motion cues. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 466–470, 2016.

[23] Diana Sungatullina, Egor Zakharov, Dmitry Ulyanov, and Victor Lempitsky. Image manipulation with perceptual discriminators. In *Proceedings of the European Conference on Computer Vision*, pages 579–595, 2018.

[24] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[25] R. Wan, B. Shi, H. Li, L. Y. Duan, A. H. Tan, and A. C. Kot. Corrn: Cooperative reflection removal network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(12):2969–2982, 2020.

[26] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3922–3930, 2017.

[27] Renjie Wan, Boxin Shi, Tan Ah Hwee, and Alex C Kot. Depth of field guided reflection removal. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 21–25. IEEE, 2016.

[28] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019.

[29] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3771–3779, 2019.

[30] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of the European Conference on Computer Vision*, pages 654–669, 2018.

[31] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M. Patel. Cr-fill: Generative image inpainting with auxiliary contexutal reconstruction, 2021.

[32] 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*, 22(8):2012–2023, 2020.

[33] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4786–4794, 2018.

[34] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017.

[35] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[36] M. Zontak and M. Irani. Internal statistics of a single natural image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 977–984, 2011.

Appendix A

Subjective Study

Our user study is approved by the Research Ethics Board of SFU. The objective of this user study is to rank the quality of some single-image reflection removal methods. The reflection images used in this survey are from public datasets. Reflection are seen in diverse, different scenes, with different intensity and shape.

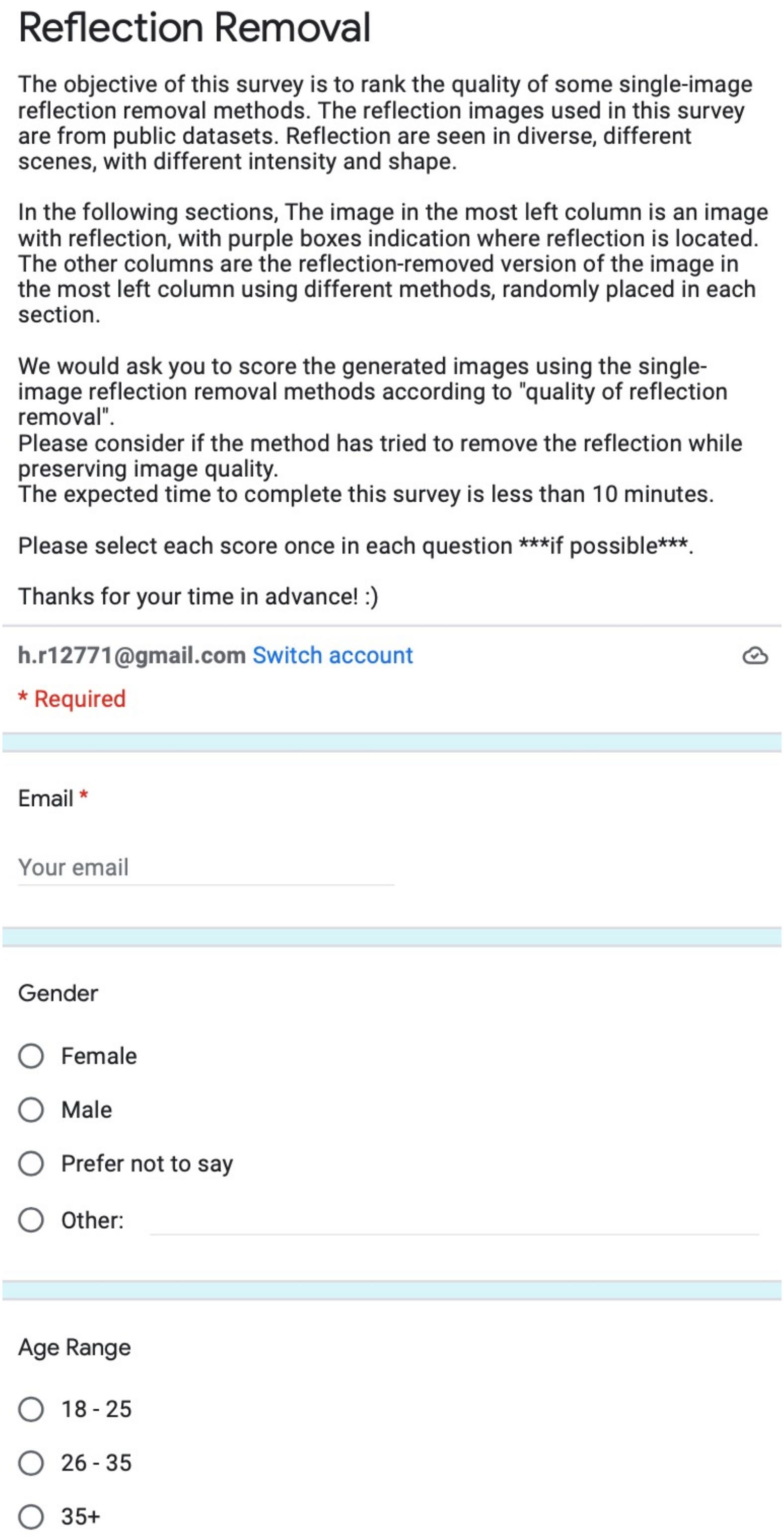
The image in the most left column is an image with reflection, with purple boxes indicating where reflection is located. The other columns are the reflection-removed version of the image in the most left column using different methods, randomly placed in each section.

In the subjective study and as shown in Fig A.1, the subject is first presented with an overview of the study and its objectives. Then, the subject completes brief basic information about themselves, including their age and gender. Furthermore, the subject is shown an example explaining the reflection removal task as in Figure A.2. Then, as shown in Figure A.3, successive scenes are shown to the subject to rank.

Users are asked to score the generated images using the single-image reflection removal methods according to "quality of reflection removal", considering if the method has tried to remove the reflection while preserving image quality.

Tables A.1, A.2 and A.3 show the average, median and standard deviation of all the users' scores over each scene respectively.

These tables compare the average and median of the Mean Opinion Score (MOS) computed across all users for each image for the five considered methods. The results in Table A.1 and A.2 show that our method substantially outperforms all prior works over all images, and also over all users which are demonstrated at the bottom row. As seen in Table A.2 the median MOS resulting from our method is 3.94, which is 37% higher than the best median MOS resulting from prior works (2.87 produced by ERRNet [28]).
This is also observed in Table A.1, in which the average MOS resulting from our method is higher than other methods, which are all supervised methods unlike ours.



Reflection Removal

The objective of this survey is to rank the quality of some single-image reflection removal methods. The reflection images used in this survey are from public datasets. Reflection are seen in diverse, different scenes, with different intensity and shape.

In the following sections, The image in the most left column is an image with reflection, with purple boxes indication where reflection is located. The other columns are the reflection-removed version of the image in the most left column using different methods, randomly placed in each section.

We would ask you to score the generated images using the single-image reflection removal methods according to "quality of reflection removal".
Please consider if the method has tried to remove the reflection while preserving image quality.
The expected time to complete this survey is less than 10 minutes.

Please select each score once in each question ***if possible***.

Thanks for your time in advance! :)

h.r12771@gmail.com Switch account

* Required

Email *

Your email

Gender

○ Female

○ Male

○ Prefer not to say

○ Other:

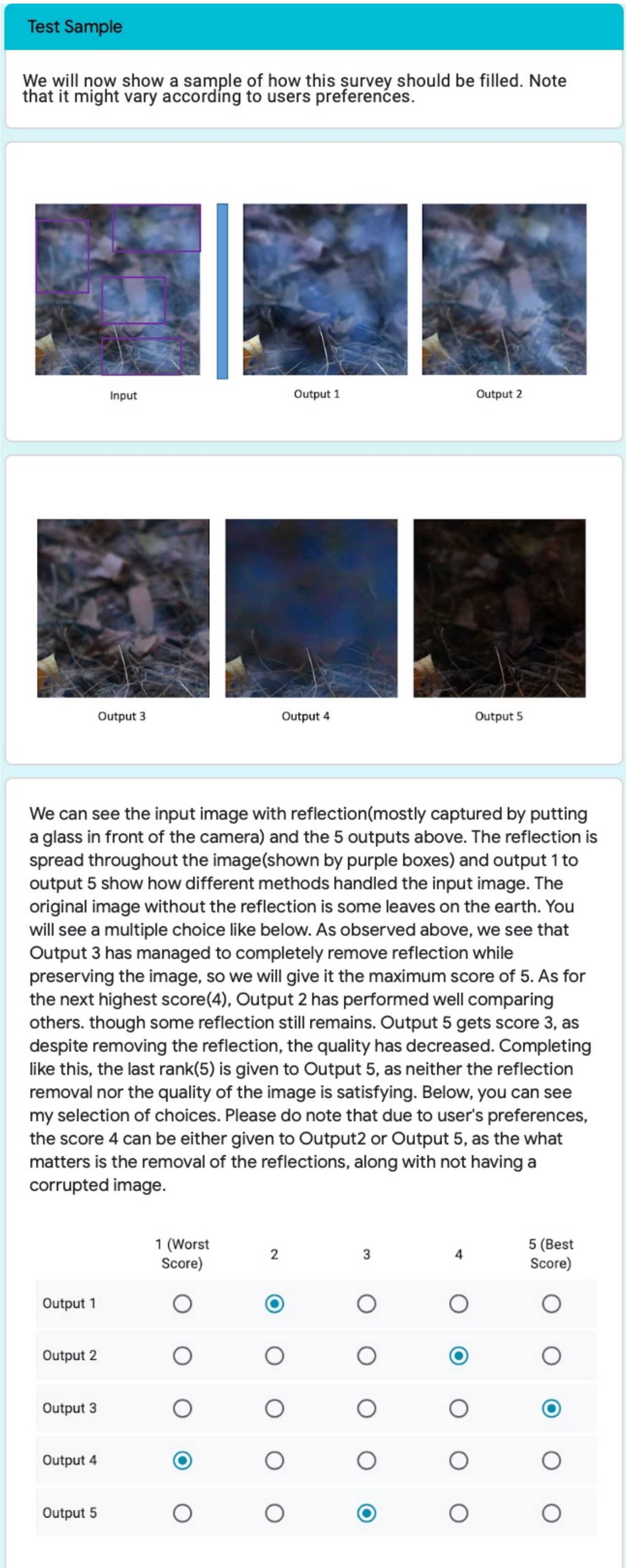
Age Range

○ 18 - 25

○ 26 - 35

○ 35+

Figure A.1: Subjective evaluation form. Details about each user is gathered at the beginning of the survey.



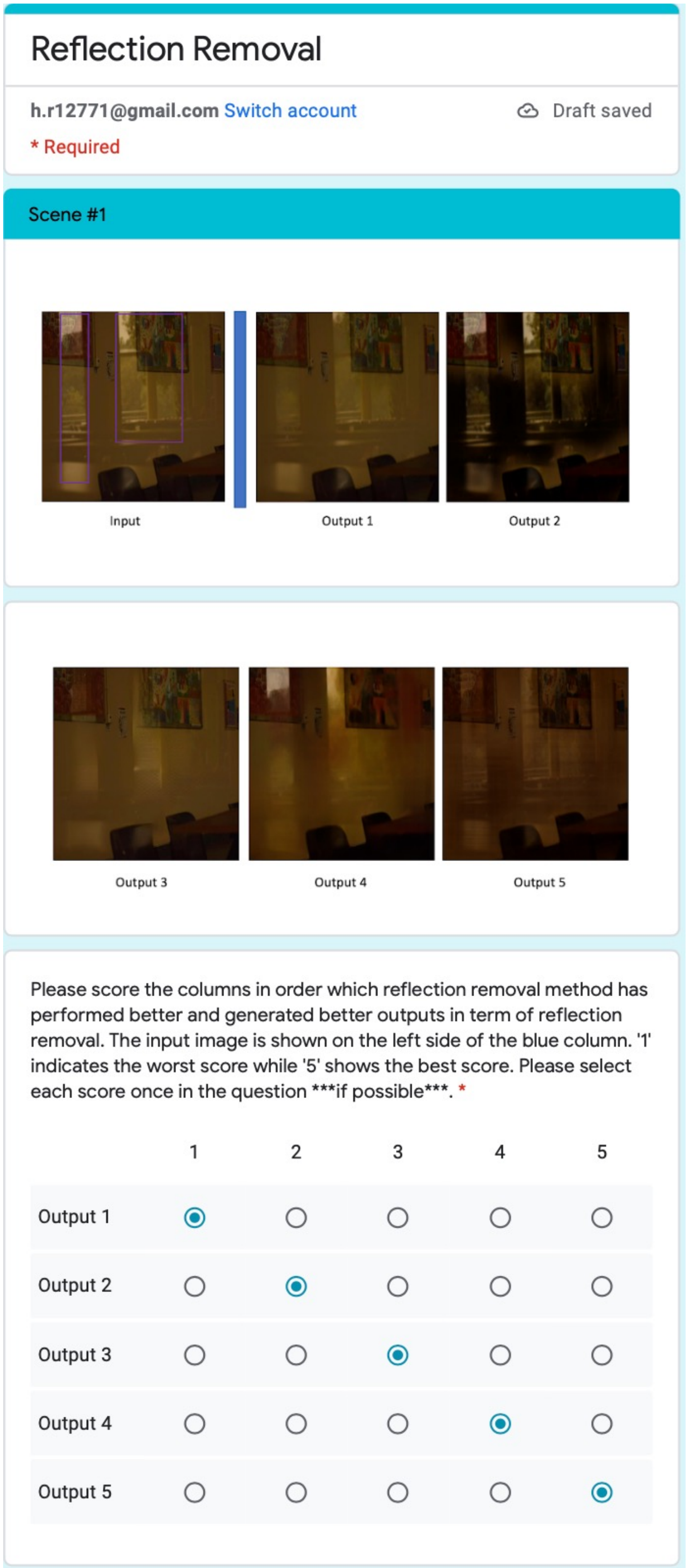
Test Sample

We will now show a sample of how this survey should be filled. Note that it might vary according to users preferences.

We can see the input image with reflection(mostly captured by putting a glass in front of the camera) and the 5 outputs above. The reflection is spread throughout the image(shown by purple boxes) and output 1 to output 5 show how different methods handled the input image. The original image without the reflection is some leaves on the earth. You will see a multiple choice like below. As observed above, we see that Output 3 has managed to completely remove reflection while preserving the image, so we will give it the maximum score of 5. As for the next highest score(4), Output 2 has performed well comparing others. though some reflection still remains. Output 5 gets score 3, as despite removing the reflection, the quality has decreased. Completing like this, the last rank(5) is given to Output 5, as neither the reflection removal nor the quality of the image is satisfying. Below, you can see my selection of choices. Please do note that due to user's preferences, the score 4 can be either given to Output2 or Output 5, as the what matters is the removal of the reflections, along with not having a corrupted image.

	1 (Worst Score)	2	3	4	5 (Best Score)
Output 1	○	●	○	○	○
Output 2	○	○	○	●	○
Output 3	○	○	○	○	●
Output 4	●	○	○	○	○
Output 5	○	○	●	○	○

Figure A.2: A test case to introduce a user to the demanded tasks and how to score the images subjectively.



Reflection Removal

h.r12771@gmail.com Switch account

Draft saved

* Required

Scene #1

Please score the columns in order which reflection removal method has performed better and generated better outputs in term of reflection removal. The input image is shown on the left side of the blue column. '1' indicates the worst score while '5' shows the best score. Please select each score once in the question ***if possible***. *

	1	2	3	4	5
Output 1	●	○	○	○	○
Output 2	○	●	○	○	○
Output 3	○	○	●	○	○
Output 4	○	○	○	●	○
Output 5	○	○	○	○	●

Figure A.3: Scene 1.

Table A.1: Average scores of the subjective study for each scene over each method.

	Average over all users				
Evaluated Scenes	BDN	GCNet	ERRNet	Zhang	OurMethod
Scene I	3.38	2.88	1.3	2.56	4.68
Scene II	3.24	2.3	2.64	2.82	3.68
Scene III	2.14	2.58	2.9	3.44	4.14
Scene IV	1.98	2.88	4.44	3.38	2.78
Scene V	2.64	2.16	2.58	2.24	4.78
Scene VI	3.68	2.18	1.72	2.44	4.56
Scene VII	1.98	3.04	3.72	4.10	2.24
Scene VIII	1.46	2.30	3.72	2.64	4.66
Scene IX	1.58	2.06	4.70	4.04	2.84
Scene X	3.42	2.46	2.36	2.68	2.88
Scene XI	1.86	3.70	4.28	2.22	2.58
Scene XII	2.12	2.34	2.12	2.84	4.82
Scene XIII	2.32	2.30	2.14	2.58	4.14
Scene XIV	4.62	2.20	2.00	2.12	3.66
Scene XV	2.92	2.14	3.30	1.78	4.32
Scene XVI	3.58	2.46	2.02	2.02	4.42
Average MOS	2.68	2.49	2.87	2.74	**3.82**

Table A.2: Median scores of the subjective study for each scene over each method

	Median over all users				
Evaluated Scenes	BDN	GCNet	ERRNet	Zhang	OurMethod
Scene I	4	3	1	2	5
Scene II	4	2	3	3	4
Scene III	2	3	3	4	5
Scene IV	2	3	5	4	3
Scene V	3	2	3	2	5
Scene VI	4	2	1	2	5
Scene VII	2	3	4	4	2
Scene VIII	1	2	4	3	5
Scene IX	1	2	5	4	3
Scene X	4	3	2	3	3
Scene XI	2	4	5	2	3
Scene XII	2	2	2	3	5
Scene XIII	2	2	2	3	5
Scene XIV	5	2	2	2	4
Scene XV	3	2	3	1	5
Scene XVI	4	3	2	2	5
Median MOS	2.75	2.50	2.84	2.75	**3.94**

Table A.3: Standard deviation of the subjective study's score for each scene over each method

	Standard deviation over all users				
Evaluated Scenes	BDN	GCNet	Zhang	OurMethod	ERRNet
Scene I	1.04	0.86	0.90	0.64	0.59
Scene II	1.50	1.18	0.96	1.20	1.51
Scene III	1.04	1.13	1.01	1.43	1.21
Scene IV	1.05	1.12	1.06	1.03	1.13
Scene V	1.34	1.20	1.11	1.05	0.65
Scene VI	0.99	0.84	1.13	0.89	0.79
Scene VII	1.10	1.21	0.99	1.10	1.25
Scene VIII	0.79	0.90	0.72	0.80	0.87
Scene IX	0.73	0.64	0.71	0.61	0.85
Scene X	1.68	1.08	1.07	1.18	1.43
Scene XI	0.96	0.98	1.13	0.83	1.27
Scene XII	1.02	1.01	1.18	1.17	0.66
Scene XIII	1.23	1.02	1.17	1.18	1.35
Scene XIV	1.01	1.02	1.00	1.07	0.66
Scene XV	0.95	1.00	1.02	0.95	1.09
Scene XVI	1.04	0.87	0.99	1.13	1.12