Reflection Segmentation for Single Image Reflection Removal

A Thesis

submitted by

ROHAN SINGH JAIN

in partial fulfilment of the requirements
for the award of the degree of

BACHELOR AND MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS

August 2020

THESIS CERTIFICATE

This is to certify that the thesis titled **Reflection Segmentation for Single Image Reflection Removal**, submitted by **Rohan Singh Jain**, to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelor and Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Kaushik Mitra
Research Guide
Assistant Professor
Dept. of Electrical Engineering
IIT-Madras, 600036
Place: Chennai

ACKNOWLEDGEMENTS

First and foremost, I express my sincere gratitude to my research guide, Prof. Kaushik Mitra, for his consistent guidance, motivation and support for my research work. Our regular meetings facilitated me in learning how to solve research problems. I want to thank him for granting me the freedom and opportunity to work on the topic of my interest and work at my pace.

I would also like to thank Pawan Prasad (ee19d005), a PhD scholar at the Computational Imaging Lab with whom I have worked on this project jointly. He was the one who came up with the problem statement and I acknowledge his important contributions to this project. The real dataset was entirely collected by him.

Last but not the least, I express my gratitude to my beloved parents Mrs. Nitisha Jain, Mr. Sudeep Singh Jain and my twin brother Rohit Singh Jain for their constant support throughout my stay at IIT Madras.

ABSTRACT

KEYWORDS: Reflection removal, image restoration, reflection detection, semantic segmentation, deep learning

An image captured through a glass plane usually contains both of a target transmitted scene behind the glass plane and a reflected scene in front of the glass plane. Removing undesirable reflections from a photo taken in front of a glass is of great importance for enhancing the efficiency of visual computing systems. In this thesis, we use a semantic segmentation network to extract regions of real and synthetic mixed images affected by reflection. We propose two approaches for estimation of the ground truth reflection binary masks of synthetic images to avoid cumbersome manual pixel-level annotation. The estimated reflection binary mask, along with the original input mixed image, can be used subsequently by an encoder-decoder architecture to guide reflection removal from images. We propose a synthetic dataset consisting of $50,000$ mixed images synthesized from real pairs of the targeted transmitted scene and the reflected scene to simulate real-world reflections. A real dataset of 622 image pairs of the mixed image and the transmitted scene is also captured.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS i

ABSTRACT ii

LIST OF FIGURES viii

LIST OF TABLES ix

ABBREVIATIONS x

NOTATION xi

1 INTRODUCTION 1

1.1 Thesis outline 1

1.2 Problem description : single image reflection removal 1

1.3 Focus of this thesis 3

2 LITERATURE REVIEW 4

2.1 Categorization of reflection removal methods 4

2.2 Traditional methods for single image reflection removal 5

2.3 Deep learning based methods for single image reflection removal . . 6

3 DATASETS 9

3.1 Ideal dataset for single image reflection segmentation and removal . 9

3.2 Synthetic datasets 10

3.2.1 Linear mix dataset 11

3.2.2 CEILNet synthetic dataset 13

3.2.3 PLNet synthetic dataset 14

3.2.4 Convex blurring dataset 17

3.2.5 Focused reflection dataset 18

3.2.6 Ghosting dataset 18

3.3 Real datasets 21

4 ESTIMATION OF GROUND TRUTH FOR REFLECTION SEGMENTATION 24

4.1 Estimation of ground truth reflection strength maps and reflection binary masks for synthetic data 24

4.1.1 Proposed approach 1 : absolute differencing of the mixed image and transmission layer intensities 25

4.1.2 Proposed approach 2 : proportion of reflection intensity in the mixed image intensity 26

4.1.3 Qualitative estimation results for synthetic data 26

4.1.4 Analysis of estimation results on synthetic data 27

4.2 Experiments for estimation of ground truth reflection strength maps and reflection binary masks for real data 27

5 REFLECTION SEGMENTATION 36

5.1 Network and training details 36

5.2 Quantitative evaluation on synthetic data 39

5.3 Qualitative evaluation on synthetic and real data 40

6 KEY RESULTS and SUMMARY 48

7 SCOPE FOR FUTURE WORK 49

LIST OF FIGURES

1.1 Some examples of images captured through glass : the first row contains the mixed images (I) and the second row contains the corresponding target transmitted images without undesirable reflections (T). 2

2.1 A categorization of methods of reflection removal in images 4

3.1 This figure shows the effect of increasing α in Eqn. 3.1 for the same T and R. From left to right : transmission layer T, reflection layer R, synthesized mixed image I with $\alpha = 0.5$, synthesized mixed image I with $\alpha = 0.7$ and synthesized mixed image I with $\alpha = 0.9$ 11

3.2 Examples of some synthesized mixed images via linear mixing of T and R. From left to right in each row : input mixed image I, transmission layer T, transmission layer T scaled by α, reflection layer R and reflection layer R scaled by $(1 - \alpha)$ 12

3.3 This figure shows the effect of the increasing the blurring of the reflection layer while synthesizing the mixed image I as per the CEILNet data synthesis procedure [1] in Eqns. 3.2 - 3.7. From left to right : transmission layer T, reflection layer R, synthesized mixed image I with $\sigma = 2$, synthesized mixed image I with $\sigma = 4$ and synthesized mixed image I with $\sigma = 5$ 14

3.4 Examples of some synthesized mixed images via the CEILNet data synthesis procedure [1]. From left to right in each row : mixed image I, transmission layer T, reflection layer R and reflection layer R blurred by a gaussian blurring kernel of size 11×11 and σ drawn randomly from $U(2, 5)$. 15

3.5 Examples of some synthesized mixed images via the PLNet data synthesis procedure [2]. From left to right in each row : input mixed image I, transmission layer T, the scaled transmission layer T, reflection layer R and reflection layer R after gaussian blurring and application of vignette mask. 16

3.6 Examples of some synthesized mixed images via convex addition of T and blurred R, as mentioned in Subsection 3.2.4. From left to right in each row : input mixed image I, transmission layer T, the scaled transmission layer T, reflection layer R and reflection layer R after gaussian blurring and scaling. 17

3.7 Examples of some synthesized mixed images with focused reflections. From left to right in each row : input mixed image I, transmission layer T, reflection layer R and reflection layer R blurred by a gaussian blurring kernel of size 11×11 and σ drawn randomly from $U(0.5, 1.5)$ 19

3.8 Examples of some synthesized mixed images with ghosting of reflections. From left to right in each row : input mixed image I, transmission layer T, transmission layer T scaled by α, reflection layer R and ghosted (or double shifted) reflection layer R. 20

3.9 The physical and mathematical image formation models of the three major types of reflection : focused, defocused and ghosting. 21

3.10 Examples of some pairs of (I, T) in the proposed real dataset. In each row, the image at the left is the mixed image I and the image at the right is the transmission layer T. 23

4.1 Some examples of reflection strength maps and reflection binary masks estimated for mixed images synthesized by linear mixing of T and R. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2. 28

4.2 Some examples of reflection strength maps and reflection binary masks estimated for mixed images synthesized as per the CEILNet data synthesis procedure. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2. 29

4.3 Some examples of reflection strength maps and reflection binary masks estimated for mixed images synthesized as per PLNet data synthesis procedure. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2. 30

4.4 Some examples of reflection strength maps and reflection binary masks estimated for mixed images synthesized as per Subsection 3.2.4. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2. 31

4.5 Some examples of reflection strength maps and reflection binary masks estimated for mixed images with focused reflections. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2. 32

4.6 Some examples of reflection strength maps and reflection binary masks estimated for mixed images with ghosting effect. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2. 33

4.7 Some examples of reflection strength maps and reflection binary masks estimated for real mixed images from the SIR^2 dataset. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1 with $\alpha = 0.6$, rbm_1 obtained by setting a threshold of 0.2 on rsm_1, reflection strength map rsm_2 estimated via approach 1 with $\alpha = 0.8$ and rbm_2 obtained by setting a threshold of 0.2 on rsm_2. 35

5.1 Plot showing the evolution of train MSE loss, test MSE loss, F1 score on the training dataset, train AUROC value, F1 score on the test dataset and test AUROC value for CNet. 37

5.2 Plot showing the evolution of train MSE loss, test MSE loss, F1 score on the training dataset, train AUROC value, F1 score on the test dataset and test AUROC value for PNet. 38

5.3 Plot showing the evolution of train MSE loss, test MSE loss, F1 score on the training dataset, train AUROC value, F1 score on the test dataset and test AUROC value for FNet. 38

5.4 Plot showing the evolution of train MSE loss, test MSE loss, F1 score on the training dataset, train AUROC value, F1 score on the test dataset and test AUROC value for GNet. 39

5.5 Some reflection segmentation results of CNet on the CEILNet synthetic dataset. In each row, from left to right : the mixed image I, transmission layer T, ground truth reflection binary mask as estimated in Subsection 4.1.1, output reflection strength map and output reflection binary mask 41

5.6 Some reflection segmentation results of PNet on the PLNet synthetic dataset. In each row, from left to right : the mixed image I, transmission layer T, ground truth reflection binary mask as estimated in Subsection 4.1.1, output reflection strength map and output reflection binary mask 42

5.7 Some reflection segmentation results of FNet on the focused reflection dataset. In each row, from left to right : the mixed image I, transmission layer T, ground truth reflection binary mask as estimated in Subsection 4.1.1, output reflection strength map and output reflection binary mask 43

5.8 Some reflection segmentation results of GNet on the ghosting dataset. In each row, from left to right : the mixed image I, transmission layer T, ground truth reflection binary mask as estimated in Subsection 4.1.1, output reflection strength map and output reflection binary mask . . 44

5.9 Some reflection segmentation results on real images from CEILNet real dataset [1]. In each column, from top to bottom : mixed image I, rsm output by CNet, rsm output by PNet, rsm output by FNet, rsm output by GNet, rbm output by CNet and rbm output by PNet. 45

5.10 Some more reflection segmentation results on real images from CEILNet real dataset [1]. In each column, from top to bottom : mixed image I, rsm output by CNet, rsm output by PNet, rsm output by FNet, rsm output by GNet, rbm output by CNet and rbm output by PNet. . . . 46

5.11 Some reflection segmentation results on real images from CEILNet real dataset [1] where none of the networks perform well. In each column, from top to bottom : mixed image I, rsm output by CNet, rsm output by PNet, rsm output by FNet, rsm output by GNet, rbm output by CNet and rbm output by PNet. 47

LIST OF TABLES

3.1 Table describing the six datasets synthesized by us. Here $I, \tilde{T}$ and $\tilde{R}$ denote the mixed image, effective transmission layer and effective reflection layer respectively. 10

3.2 A comparison of our real dataset and existing real datasets on the basis of dataset size, diversity and spatial alignment procedures implemented. 22

5.1 The train AUROC and test AUROC values of the four different reflection segmentation networks trained on four synthetic datasets : **(i)** CEILNet synthetic dataset, **(ii)** PLNet synthetic dataset, **(iii)** focused reflection dataset and **(iv)** ghosting dataset 40

ABBREVIATIONS

DL	Deep Learning
SIRR	Single Image Reflection Removal
MIRR	Multiple Image Reflection Removal
CNN	Convolutional Neural Network
DoF	Depth of Field
CEILNet	Cascaded Edge and Image Learning Network [1]
PLNet	Perceptual Loss Network [2]
SIR^2	Single Image Reflection Removal real dataset [3]
MSE	Mean Squared Error
NCC	Normalized Cross Correlation
SSIM	Structural Similarity
SI	Structured Index
RID	Reflection Image Dataset [4]
cGAN	Conditional Generative Adversarial Network
ERRNet	Enhanced Reflection Removal Network [5]
HVS	Human Visual System
FCN	Fully Convolutional Network
RANSAC	Random Sample Consensus
RGB	Red Green Blue color model for images
YUV	$Y'UV$ colour space
ORB	Oriented FAST and Rotated Brief features
BCE	Binary Cross Entropy loss
CNet	proposed network trained on the CEILNet synthetic dataset
PNet	proposed network trained on the PLNet synthetic dataset
FNet	proposed network trained on the focused reflection dataset
GNet	proposed network trained on the ghosting dataset
ROC	Receiver Operating Characteristic curve
AUROC	Area Under Receiver Operating Characteristic curve

NOTATION

I	the mixed image containing undesirable reflection scene components
T	the transmitted scene devoid of any undesirable reflection scene components
R	the undesirable reflected scene
α	glass transmittance
β	glass reflectance
$L1$	$L1$ loss function (Least Absolute Deviations)
$L2$	$L2$ loss function (Least Square Errors)
σ	standard deviation of the gaussian blurring kernel
$\circledast$	convolution operator
G	gaussian blurring kernel
K	ghosting kernel, unless otherwise mentioned
$U\ (a, b)$	uniform probability distribution between a and b
$\tilde{T}$	effective transmission layer
$\tilde{R}$	effective reflection layer
y	intensity channel in the $Y^{'}UV$ colour space of image
rsm	reflection strength map
rbm	reflection binary mask
t	threshold for reflection binary mask

CHAPTER 1

INTRODUCTION

1.1 Thesis outline

The main contributions of this thesis are organized as follows :

- Chapter 1 : provides **(i)** a detailed description and motivation of the problem of reflection removal from images and the major challenges associated with it and **(ii)** the primary focus of this thesis.
- Chapter 2 : contains **(i)** a brief overview of the traditional methods for reflection removal using single and multiple images and **(ii)** a brief overview of the deep learning based approaches used for single image reflection removal.
- Chapter 3 : describes **(i)** the characteristics of an ideal dataset for single image reflection segmentation and reflection removal, **(ii)** the various procedures used for synthesizing 50000 mixed images I from real pairs of the targeted transmitted scene T and the reflection scene R and **(iii)** the proposed and existing real datasets.
- Chapter 4: describes **(i)** the proposed approaches used to estimate the ground truth reflection strength maps and reflection binary masks for the synthetic datasets and **(ii)** the experiments performed to estimate the ground truth reflection strength maps and reflection binary masks for real data.
- Chapter 5 : contains **(i)** the network, training and implementation details for the task of reflection segmentation, **(ii)** quantitative evaluation of the proposed reflection segmentation network on the synthesized datasets and **(iii)** qualitative evaluation of the proposed reflection segmentation networks on the synthesized and real datasets.
- Chapter 6 : contains a summarization of the major takeaways from this thesis.
- Chapter 7 : describes the scope of future work for the task of accurate reflection segmentation and reflection removal in single images.

1.2 Problem description : single image reflection removal

Capturing images through a transparent glass is unavoidable in many daily scenarios such as looking through a window or in front of a glass show case at the museum.

However, when taking pictures through glass, light reflection occurs on glass planes, which reduces the visibility of target transmitted scenes behind the glass planes and thus degrades the performance of computer vision techniques such as text recognition, object detection and object classification. Images taken under such circumstances usually have the objects of interest overlaid by the undesirable reflections of the scene behind the camera. The mixture image is composed of two components, the background target objects behind the glass and the reflected objects in front of the glass, in a weighted additive manner.

Figure 1.1: Some examples of images captured through glass : the first row contains the mixed images (I) and the second row contains the corresponding target transmitted images without undesirable reflections (T).

Reflection removal aims at removing the reflection (while obtaining the clear background or transmitted scene) from the mixture image using one or more shots, where the former is a highly ill-posed problem. For CNN-based single image reflection removal, our focus herein, the challenge originates from at least two sources: **(i)** the extraction of a background image layer devoid of reflection artifacts is fundamentally ill-posed because of twice the number of unknowns as equations, and **(ii)** training data from real-world scenes, is exceedingly scarce because of the difficulty in obtaining ground truth labels of the background or the transmitted scene. Mathematically speaking, it is typically assumed that a captured image I is formed by a linear combination of a background or transmitted layer T and a reflection layer R. Obviously, when given access only to I, there exist an infinite number of feasible decompositions.

Further compounding the problem is the fact that unlike in the case of rain and shadow removal, the structures and properties of reflections can be similar to that of the background. This makes it difficult to simultaneously remove the reflections and restore the contents in the background. This can make them difficult to distinguish even

for human observers in some cases, and simple priors that might mitigate this ambiguity are not available except under specialized conditions.

Although CNNs can perform a wide variety of visual tasks, at times exceeding human capabilities, they generally require a large volume of labeled training data. Unfortunately, real reflection images accompanied with densely-labeled, ground-truth transmitted layer intensities are scarce. Consequently, previous learning-based approaches have resorted to training with synthesized images and / or small real-world data captured from specialized devices. However, existing image synthesis procedures are heuristic and the domain gap may jeopardize accuracy on real images. On the other hand, collecting sufficient additional real data with precise ground-truth labels is tremendously labor-intensive.

1.3 Focus of this thesis

Most of the existing learning-based single image reflection removal methods first estimate a prior image (like semantic map of the ground truth transmission layer T, or the edge map of the ground truth transmission layer T) and then use the prior image to guide the process of reflection removal from the input mixed image I. This thesis focuses on solving the problem of single image reflection removal guided by the reflection segmentation map of the input mixed image I. We provide results of the reflection segmentation map, which is essentially a reflection binary mask whose pixel values are set to 1 if reflection is present at the pixel location, and 0 otherwise. The primary focus of this thesis lies in accurately estimating a reflection binary mask of the input mixed image I. Although we don't provide results of reflection removal, we believe that accurately estimating the reflection binary masks will help us eventually achieve state-of-the-art results in reflection removal from single images.

CHAPTER 2

LITERATURE REVIEW

In this chapter, in Section 2.1, we first provide a categorization of the reflection removal methods in images. In Section 2.2, we briefly discuss the various non-learning based approaches in literature to remove reflection from single and multiple input images. This is followed by Section 2.3, which provides an overview of the existing deep learning based approaches in literature to remove reflections from single input images.

2.1 Categorization of reflection removal methods

The existing methods for reflection removal in images can be classified on the basis of **(i)** number of input images used (single or multiple) and **(ii)** the approach used (learning or non-learning). Single image reflection removal (SIRR) methods use only one input mixed image whereas multiple image reflection removal (MIRR) methods utilize slightly differing multiple shots of the same scene through glass. Traditional methods apply additional priors to make the problem of reflection removal less ill-posed whereas the deep learning based methods employ CNNs so that the handcrafted priors can be replaced by data-driven learning.

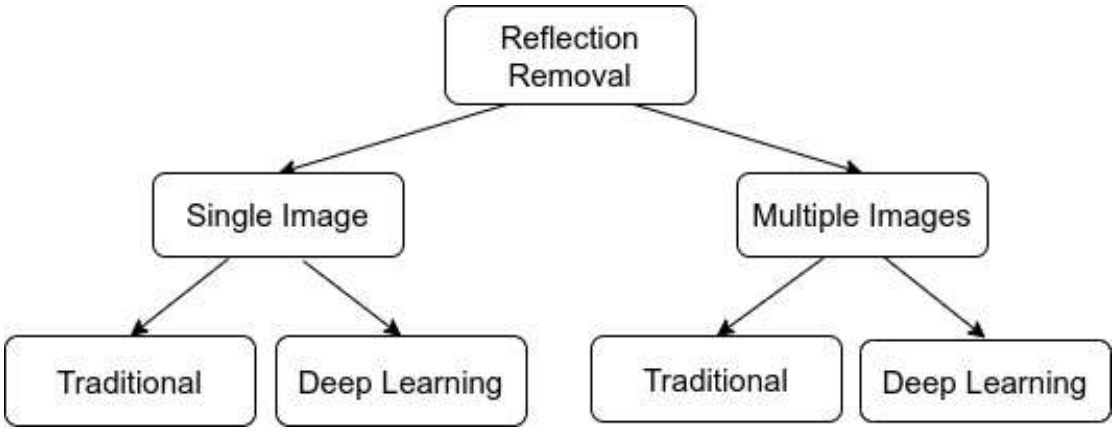

Figure 2.1: A categorization of methods of reflection removal in images

In this thesis, our focus lies on SIRR using DL based methods. We will discuss

the traditional methods for SIRR in brief and move on to describing existing DL based methods for SIRR in slightly more detail.

2.2 Traditional methods for single image reflection removal

The focus of this thesis is on reflection removal from single images. Existing works utilizing multiple input images of e.g., flash / non-flash pairs [6], different polarization [7], multi-view or video sequences [8, 9, 10, 11, 12, 13, 14, 15, 16] will not be considered here.

Reflection removal from a single image is a massively ill-posed problem. Additional priors are needed to solve the otherwise prohibitively-difficult problem in traditional optimization-based methods. In [17], user annotations are used to guide layer separation jointly with a gradient sparsity prior [18]. [19] introduces a relative smoothness prior where the reflections are assumed to be blurry and thus their large gradients are penalized. [20] explores a variant of the smoothness prior where a multi-scale Depth of Field (DoF) confidence map is utilized to perform edge classification. [21] exploits the ghost cues for layer separation. [22] proposes a simple optimization formulation with an l_0 gradient penalty on the transmitted layer inspired by image smoothing algorithms.

Traditional methods often impose certain priors or assumptions to target particular type(s) of reflection such as shifted double reflection, and thus have difficulty to generalize to other types of reflection. Despite the fact that decent results can be obtained by these methods when their assumptions hold, the vastly-different imaging conditions and complex scene content in the real world render their generalization problematic. When the structures and patterns of the background are similar to those of the reflections, the non-learning based methods have difficulty in simultaneously removing reflections and recovering the background. All these approaches rely on low-level info and are limited in cases where a high level understanding of image is needed.

2.3 Deep learning based methods for single image reflection removal

Due to the advantages in robustness and performance, there is an emerging interest in applying neural networks to SIRR. Most methods estimate a prior image from the mixed image I first. This prior image acts as auxiliary information along with the input mixed image I to recover the transmission layer T. Fan et al. [1] provide the first neural network model (CEILNet) to solve this ill-posed problem. They propose a linear method for synthesizing images with reflection for training, and use an edge map as auxiliary information to guide the reflection removal. Wan et al. [4] employ two cooperative sub-networks that predict the transmission layer intensity and gradients concurrently. Both of these works [1, 4] utilize edge or gradient information of the captured layer I, motivated by the idea that the reflection layers are usually not in focus and thus blurry as compared to the transmission layers. From the edge information of the captured image I, the edge map of the transmission image T is predicted and used in estimating the transmission result. Other works that use edge or gradient information of the transmission layer T include [23] and [24]. In [24], the authors improve upon [4] by using shared encoders for the image and gradient decoder as opposed to the independent encoders utilized in [4]. This allows for more cooperation between the image and gradient decoders. [25] predicts reflection layers which are then used as auxiliary information in a subsequent network to estimate the target transmission layer T. [26] tries to imitate the HVS by using a semantic segmentation map of the transmission layer T as a prior image for reflection removal.

Typical training losses include a combination of **(i)** L1 loss, **(ii)** L2 loss and **(iii)** gradient loss. In several recent methods, improved formulations of the objective function are presented. These include the adoption of perceptual losses [27] to account for both low-level and high-level image information [2, 28, 29, 23, 30]. In these works, images are fed to VGG-19 [31] pre-trained on ImageNet, and comparisons are made based on extracted multi-stage features. Adversarial losses have also been applied, specifically to improve the realism of predicted transmission layers [2, 32, 25, 5, 33, 34]. [35] proposes a cGAN based framework conditioned on the input mixed image I with a U-Net [36] based generator and FCN based classifier as discriminator.

Another direction of study focuses on datasets for training. Moving beyond improvements for the linear synthesis method in [1] and [2], Wen et al. [33] synthesize training data with learned non-linear alpha blending masks that better model the real-world imaging conditions. These masks are also used in forming a reconstruction loss that guides the prediction of transmission layers. To deal with the insufficiency of densely-labeled and properly aligned training data, Wei et al. [5] present a technique for utilizing misaligned real-world images as the training data, since they are less burdensome to acquire than aligned images and are more realistic than synthetic images. [34] proposes a framework where unlike most SIRR methods, image generation and separation are not treated as two separate stages. For image generation, a more general non-linear mapping from (T, R) to I is learnt.

Yet another approach [37, 38] involves utilizing multiple mixed images with the same transmission layer T and different reflection layers R for training and using single image during inference stage.

Many networks [30, 28, 37] adopt a residual learning framework, i.e. estimate the residual $I - T$ instead of the transmission layer T. This is based on the fact that residual learning is known to have faster convergence and the reflection layer is relatively consistent in terms of luminance and color, thus being more tractable to learn for the generator as compared to the transmission layer T.

For quantitative evaluation on real data, the authors in [3] conclude that MSE (Mean squared error) is a bad error metric and instead introduce four other metrics : **(i)** Local MSE, **(ii)** NCC (Normalized Cross Correlation), **(iii)** Structural Similarity (SSIM) and **(iv)** SI (Structured Index). Local MSE evaluates the local structure similarity by calculating the similarity of each local patch. NCC is used because the ground truth T and estimated T can have different intensities, which can be compensated for by subtracting their mean values. SSIM is a perceptually motivated error metric, which evaluates the similarity of two images on the basis of luminance, contrast, and structure as the human eyes do. SI evaluates only structural similarities. [4] introduces regional SI and regional SSIM (SSIM_r and SI_r respectively) as additional metrics for quantitative comparison of SIRR methods. This is deemed necessary by the authors because of their observation that due to the regional properties of reflections, many existing reflection removal methods (non DL based) may downgrade the quality of whole images,

although they can remove the local reflections cleanly. The reflection dominant regions are manually labelled and the SSIM and SI values at these regions are evaluated. The error metric SI usually shows best consistency with visual quality.

No method is able to completely remove reflections, and various artifacts are visible in most of the results. The performances of the methods on bright backgrounds are much better than those on dark backgrounds, which indicates that removing strong reflection components in dark backgrounds is still challenging for all methods. Most networks are also limited in input images that have focused reflections, or when the reflection layer has very similar structure as the transmission layer.

CHAPTER 3

DATASETS

In this chapter, we first briefly discuss the characteristics of the ideal dataset for single image reflection segmentation and reflection removal in Section 3.1. Subsequently, in Section 3.2, we describe the various procedures implemented by us for synthesizing the input mixed images I using real pairs of (T, R). This is followed by Section 3.3, where the proposed and existing real-world reflection image datasets are described.

3.1 Ideal dataset for single image reflection segmentation and removal

In order to model the diversity in real-world reflections well, the ideal dataset for single image reflection segmentation and removal should have the following characteristics :

- T and R should have a sufficient amount of both indoor and outdoor scenes.
- T should have varied illuminations, like direct sunlight, cloudy sky light, twilight, low light conditions etc. This is because many existing state-of-the-art methods don't perform well in mixed input images having weak transmitted light.
- I should contain varied blur levels of the reflection scene, i.e. it should contain both focused and defocused reflections. This is to ensure that the method also learns to remove reflections as sharp as T. Most of the current methods don't perform well on mixed input images having sharp reflections.
- To account for ghosting effects, (arising from shifted reflections on the two surfaces of glass) thickness of the glass should be varied while capturing I. The phenomena of shifted reflections increases with increase in glass thickness.
- I should have both localized reflection and reflection spread almost throughout I.
- I should have some regions where the background is completely removed. This is to account for saturated reflection intensities, for example, when R has strong light sources. The network needs to learn to perform inpainting in such cases.

3.2 Synthetic datasets

We will now describe the procedures implemented by us for synthesizing input mixed images I using real T and R. We synthesize 6 different synthetic datasets : **(i)** linear mix dataset (Subsection 3.2.1), **(ii)** CEILNet synthetic dataset (Subsection 3.2.2), **(iii)** PLNet synthetic dataset (Subsection 3.2.3), **(iv)** convex blurring dataset (Subsection 3.2.4), **(v)** focused reflection dataset (Subsection 3.2.5) and **(vi)** ghosting dataset (Subsection 3.2.6). Table 3.1 presents the salient features of the datasets synthesized by us. The mixed image I is synthesized via $I = \tilde{T} + \tilde{R}$, where $\tilde{T}$ and $\tilde{R}$ denote the effective transmission layer and effective reflection layer respectively.

Synthetic dataset	Number of I synthesized	Source of real T and R	$\tilde{T}$ and $\tilde{R}$	target reflections
Linear mix dataset	7642	PASCAL VOC 2012 [39]	$\tilde{T} = \alpha T,\ \tilde{R} = (1-\alpha)R$	no particular reflection
CEILNet synthetic dataset	7642	PASCAL VOC 2012 [39]	$\tilde{T} = T, \tilde{R} =$ gaussian blurred and gamma corrected R	localized, defocused and saturated reflections
PLNet synthetic dataset	12585	flickr	$\tilde{T} = \alpha T, \tilde{R} =$ gaussian blurred and vignette applied R	defocused reflections
Convex blurring dataset	7642	PASCAL VOC 2012 [39]	$\tilde{T} = \alpha T, \tilde{R} = (1-\alpha)R \circledast G$	defocused reflections
Focused reflections dataset	7642	PASCAL VOC 2012 [39]	$\tilde{T} = T, \tilde{R} =$ gaussian blurred and gamma corrected R	focused reflections
Ghosting dataset	7642	PASCAL VOC 2012 [39]	$\tilde{T} = \alpha T, \tilde{R} = R \circledast K$	ghosted reflections

Table 3.1: Table describing the six datasets synthesized by us. Here $I, \tilde{T}$ and $\tilde{R}$ denote the mixed image, effective transmission layer and effective reflection layer respectively.

3.2.1 Linear mix dataset

The input mixed image I is synthesized as per Eqn. 3.1, with α drawn from $U(0.65, 0.90)$.

$$I = \alpha T + (1 - \alpha)R \tag{3.1}$$

7642 mixed input images are generated from 7642 real pairs of I, T taken from the PASCAL VOC 2012 dataset [39]. Images from the PASCAL VOC 2012 dataset [39] are randomly cropped to a size of 224×224 before using. Fig. 3.1 shows the effect of increasing α on the mixed image for the same T and R. Fig. 3.2 shows some examples of mixed images generated via this procedure for different T and R.

Figure 3.1: This figure shows the effect of increasing α in Eqn. 3.1 for the same T and R. From left to right : transmission layer T, reflection layer R, synthesized mixed image I with $\alpha = 0.5$, synthesized mixed image I with $\alpha = 0.7$ and synthesized mixed image I with $\alpha = 0.9$.

The major drawback of generating I via naive linear mixing of T and R is that many reflection removal networks when trained on synthetic images generated via linear mixing don't perform well on real test datasets. In typical mixed images R will only partially cover T. In fact, the visibility of R depends on the relative intensity between the transmitted light from T and the reflected light. Hence there are large regions in I where R is not visible at all. This model doesn't allow that. Also, scaling of both T and R is questionable. In real-world images, T and R contain various levels of luminance, from the darkest to the brightest color. Scaling the images not only constraints each layer within a relatively smaller color range, but also suppresses abrupt color transitions, especially for R. This doesn't model real-world mixed images well because reflection intensities can be arbitrarily large in some areas.

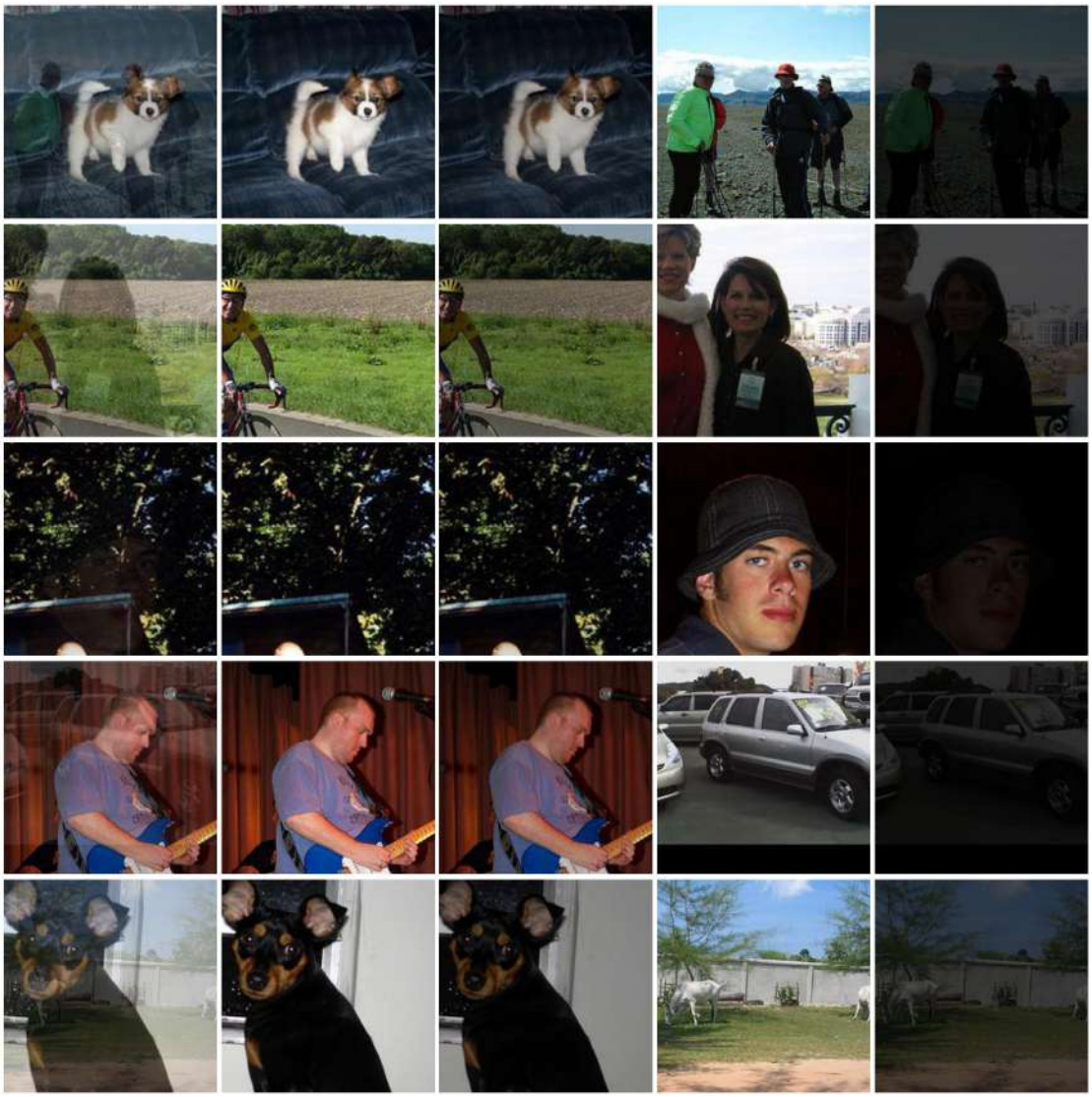

Figure 3.2: Examples of some synthesized mixed images via linear mixing of T and R. From left to right in each row : input mixed image I, transmission layer T, transmission layer T scaled by α, reflection layer R and reflection layer R scaled by $(1 - \alpha)$.

3.2.2 CEILNet synthetic dataset

In our thesis, we call the the dataset of mixed images synthesized via the procedure proposed in [1] as the CEILNet synthetic dataset. Two natural images are picked and normalized to $[0, 1]$ as transmission layer T and reflection layer R and then the input mixed image I is synthesized as shown below. (Eqns. 3.2-3.7)

$$1.\ \tilde{R} \leftarrow gauss\_blur_{\sigma}(R)\ with\ \sigma \sim U(2,5) \tag{3.2}$$

$$2.\ I \leftarrow T + \tilde{R} \tag{3.3}$$

$$3.\ m \leftarrow mean(I(x,c)|I(x,c) > 1, \forall x, \forall c = 1,2,3) \tag{3.4}$$

$$4.\ \tilde{R}(x,c) \leftarrow \tilde{R}(x,c) - \gamma(m-1), \forall x, \forall c;\ \gamma = 1.3 \tag{3.5}$$

$$5.\ \tilde{R} \leftarrow clip_{[0,1]}(\tilde{R}) \tag{3.6}$$

$$6.\ I \leftarrow clip_{[0,1]}(T + \tilde{R}) \tag{3.7}$$

One key difference from naive image mixing is that the brightness overflow issue is avoided not by scaling down the brightness, but by subtracting an adaptively computed value followed by clipping. In this way: **(i)** reflection-free regions are very likely to appear which is consistent with natural images, **(ii)** strong reflections can occur in other places, and **(iii)** the reflection contrast is better maintained. The subtraction and clipping operations allow reflection intensities to saturate and vanish. One drawback of this model is that it model doesn't account for ghosting effects of reflection i.e. disregards the thickness of glass. This model also inherently assumes that the reflection R is blurrier compared to the transmission layer T.

In our implementation, 7642 real images from PASCAL VOC 2012 [39] cropped randomly to a size of 224×224 are taken for T and R each. 7642 mixed images I are synthesized using these real pairs of (T, R). A gaussian blurring kernel of size 11×11 with σ drawn from $U(2, 5)$ is used. The Fig. 3.3 shows the synthesized mixed image I for the same T and R and increasing values of σ. Fig. 3.4 shows some examples of mixed images I generated via this procedure for different real T and R.

Figure 3.3: This figure shows the effect of the increasing the blurring of the reflection layer while synthesizing the mixed image I as per the CEILNet data synthesis procedure [1] in Eqns. 3.2 - 3.7. From left to right : transmission layer T, reflection layer R, synthesized mixed image I with $\sigma = 2$, synthesized mixed image I with $\sigma = 4$ and synthesized mixed image I with $\sigma = 5$.

3.2.3 PLNet synthetic dataset

In our thesis, we call the the dataset of mixed images synthesized via the procedure proposed in [2] as the PLNet synthetic dataset. This data synthesis procedure is similar to the CEILNet synthesis procedure described in Subsection 3.2.2 except for the following differences :

- The adaptive gamma correction is removed.
- Linear space is used to better approximate the physical formation of images.
- Instead of fixing the intensity decay on R, variation is applied to intensity decay since it's observed that reflection in real images could have comparable or higher intensity level than the transmission layer.
- Slight vignette is applied centered at a random position in R, which simulates the scenario when camera views the reflection from oblique angles.
- T is also scaled by α drawn randomly from $U(0.8, 1.0)$. R is not scaled any further after blurring and application of the vignette mask.

In our implementation, random image pairs (indoor-outdoor) are taken from flickr for T and R and are then resized to 260×260. The standard deviation σ of the gaussian blurring kernel is randomly chosen from 80 values uniformly spaced between 1 and 5 and the kernel size is set accordingly to $(2*(\lceil 2*\sigma \rceil + 1))$, between 3 and 17. The blurred R is multiplied element-wise with a vignetting mask centered at a random position. This resultant blurred and vignette R is then added to a scaled T to get the mixed image I. 12585 mixed images are generated in our dataset. Fig. 3.5 shows some examples of mixed images generated via this procedure for different T and R.

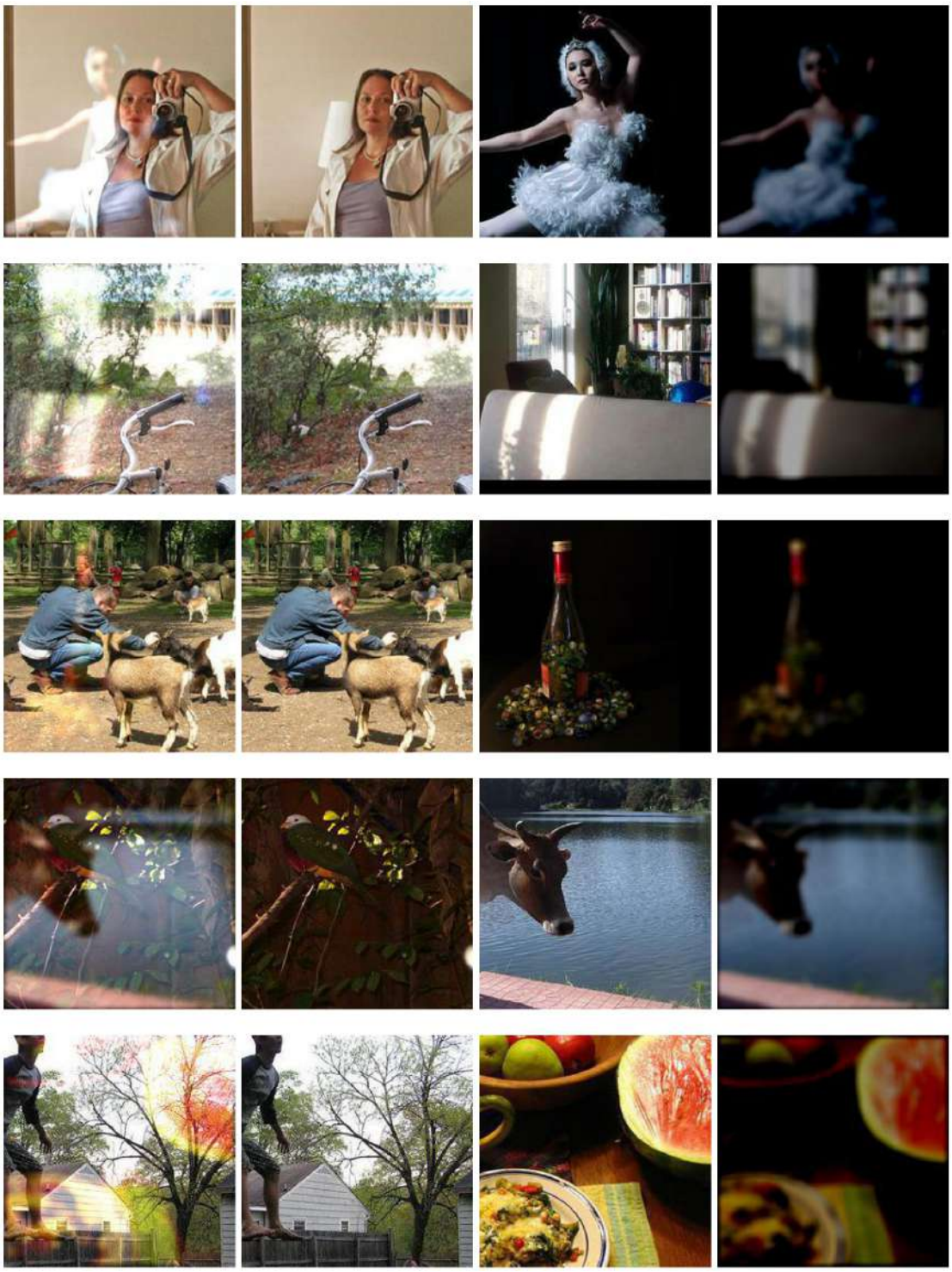

Figure 3.4: Examples of some synthesized mixed images via the CEILNet data synthesis procedure [1]. From left to right in each row : mixed image I, transmission layer T, reflection layer R and reflection layer R blurred by a gaussian blurring kernel of size 11×11 and σ drawn randomly from $U(2, 5)$.

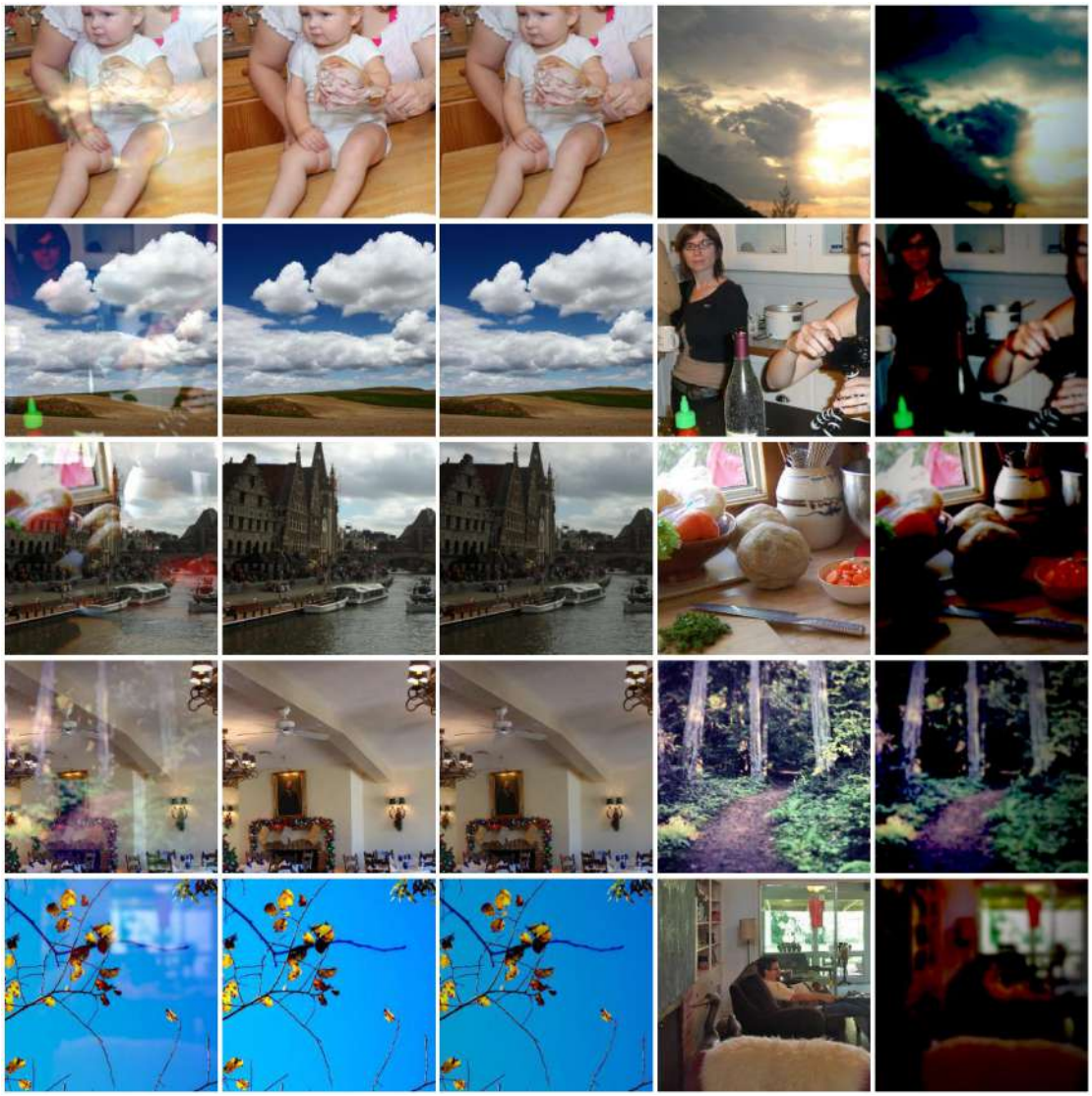

Figure 3.5: Examples of some synthesized mixed images via the PLNet data synthesis procedure [2]. From left to right in each row : input mixed image I, transmission layer T, the scaled transmission layer T, reflection layer R and reflection layer R after gaussian blurring and application of vignette mask.

3.2.4 Convex blurring dataset

The input mixed image is synthesized as per Eqn. 3.8, with $\alpha = 0.6$ and K set to a gaussian blurring kernel (for defocused reflection) of size 11×11 and standard deviation σ drawn randomly uniformly from $U(2, 5)$. In our implementation, 7642 real images from PASCAL VOC 2012 [39] cropped randomly to a size of 224×224 are taken for T and R each. 7642 mixed images I are synthesized using these real pairs of (T, R). Fig. 3.6 shows some examples of mixed images generated via this procedure for different T and R.

$$I = \alpha T + (1 - \alpha) R \circledast G \tag{3.8}$$

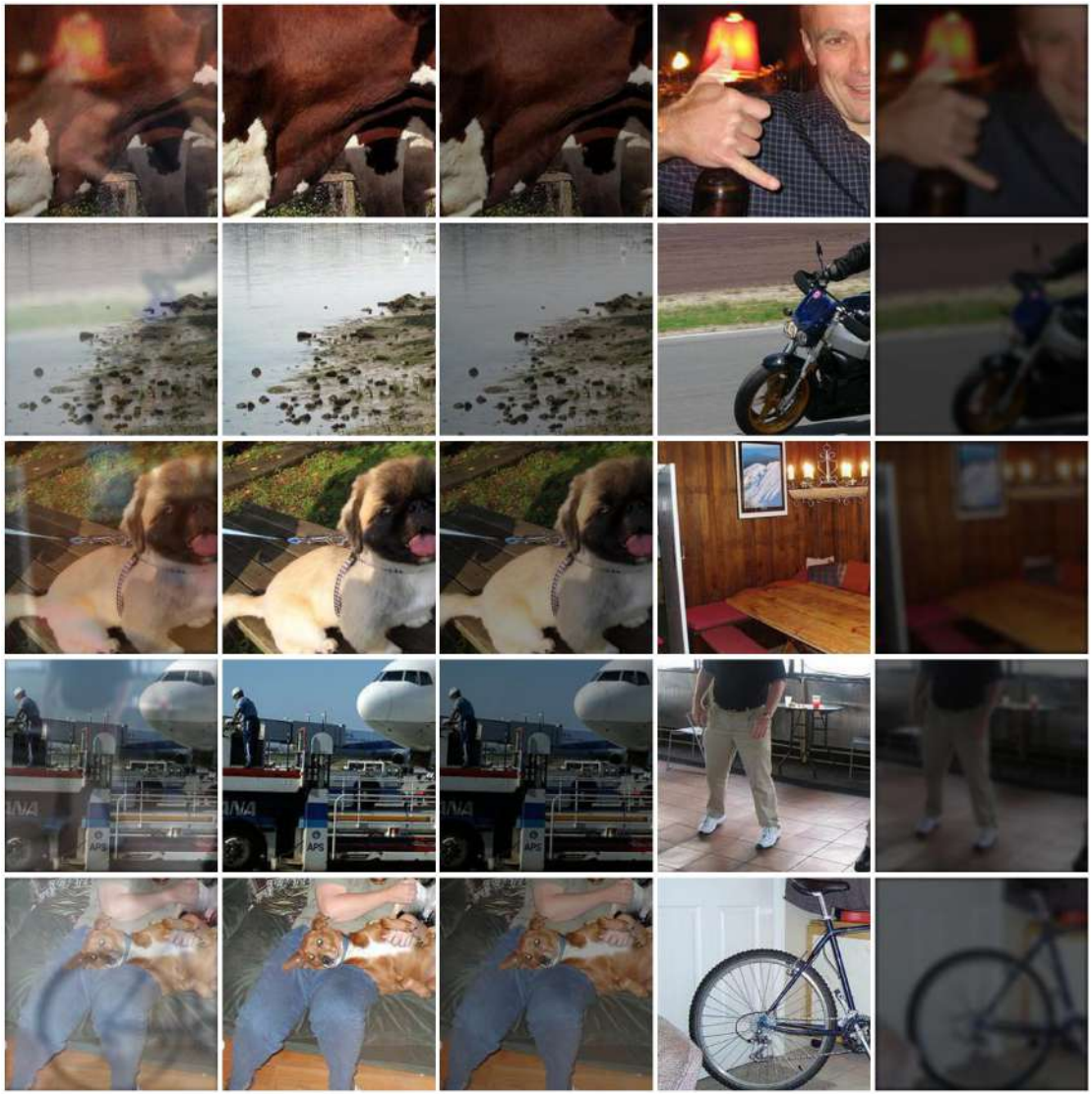

Figure 3.6: Examples of some synthesized mixed images via convex addition of T and blurred R, as mentioned in Subsection 3.2.4. From left to right in each row : input mixed image I, transmission layer T, the scaled transmission layer T, reflection layer R and reflection layer R after gaussian blurring and scaling.

3.2.5 Focused reflection dataset

In the previous subsections 3.2.2, 3.2.3 and 3.2.4, R was blurred by a gaussian blurring kernel with σ being drawn from either $U(2,5)$ or $U(1,5)$. Because of the uniform distribution, most of the mixed images had defocused reflections because of the larger proportion of high σ values. In this subsection, we implement a data synthesis procedure with focused reflections, where R is almost as sharp as T.

In our implementation, 7642 real images from PASCAL VOC 2012 [39] cropped randomly to a size of 224×224 are taken for T and R each. A synthesis procedure similar to the CEILNet data synthesis (as explained in Subsection 3.2.2) is utilized with the only difference being that σ is drawn from $U(0.5, 1.5)$ instead of $U(2,5)$. This allows for much lesser blurring of the reflection layer R and hence produces mixed images with focused reflections. Fig. 3.7 shows some examples of mixed images generated via this procedure for different T and R.

3.2.6 Ghosting dataset

In all the data synthesis methods discussed in the previous subsections, the glass thickness was assumed to be negligible. In this subsection, the glass thickness is not ignored, which results in two shifted reflections due to reflections from both sides of the glass surface. The mixed image I is synthesized as per Eqn. 3.9, with K denoting the ghosting kernel.

$$I = \alpha T + R \circledast K \tag{3.9}$$

In our implementation, 7642 images from PASCAL VOC 2012 [39] cropped randomly to a size of 224×224 are taken for T and R each. 7642 mixed images are synthesized via this procedure. K has a size of 20×20 and has two pulses of intensities $(1 - \sqrt{\alpha})$ and $(\sqrt{\alpha} - \alpha)$. Each of the coordinates $(x1, y1, x2, y2)$, where $(x1, y1)$ and $(x2, y2)$ denote the positions of the two pulses, are drawn from $U(5, 20)$. More is the distance between the two pulses, the greater is the the ghosting effect and the shift between the two reflections. This models different glass thickness values. α is drawn from $U(0.6, 0.9)$. Fig. 3.8 shows some some examples of mixed images generated via this procedure for different T and R.

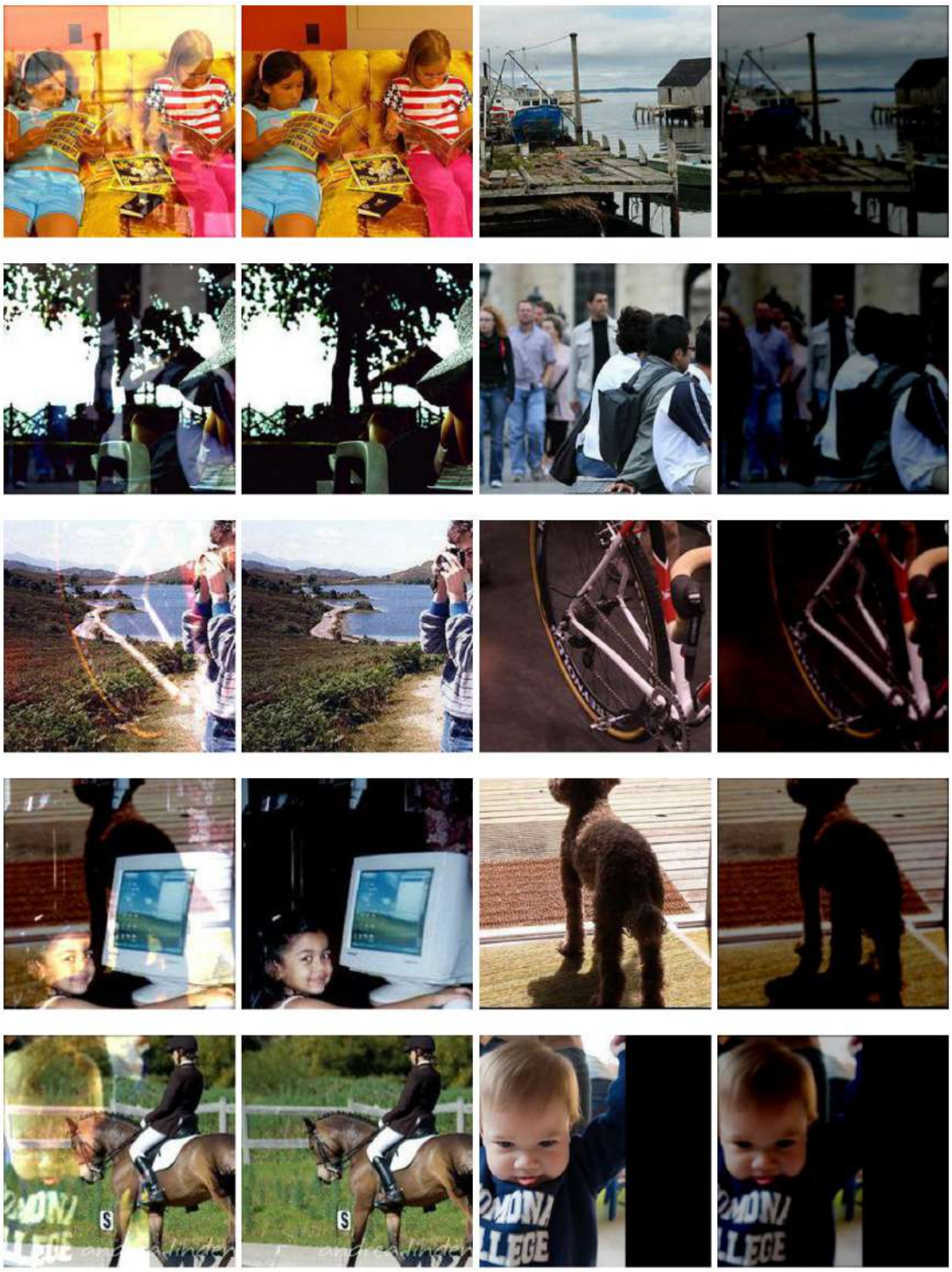

Figure 3.7: Examples of some synthesized mixed images with focused reflections. From left to right in each row : input mixed image I, transmission layer T, reflection layer R and reflection layer R blurred by a gaussian blurring kernel of size 11×11 and σ drawn randomly from $U(0.5, 1.5)$.

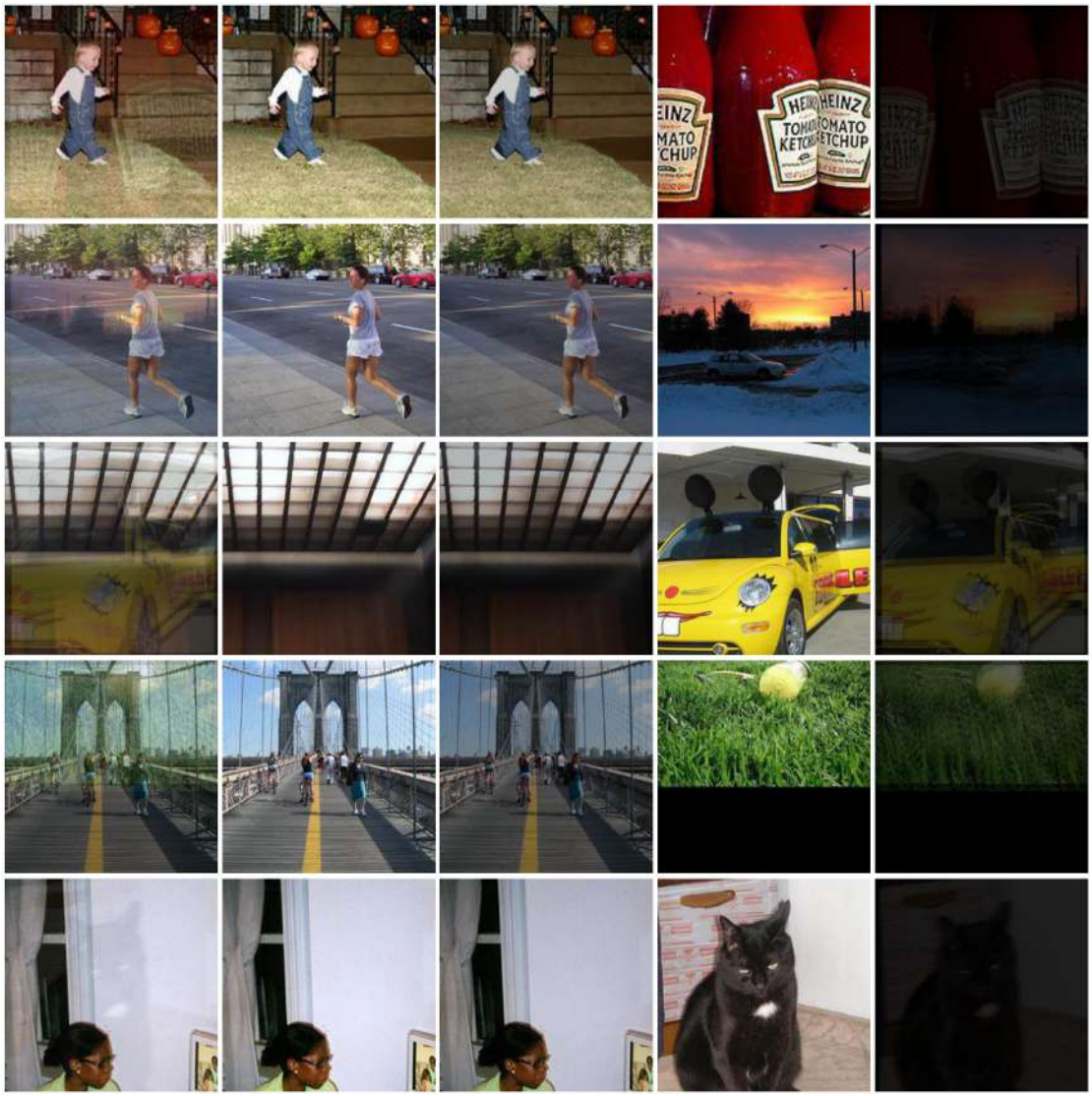

Figure 3.8: Examples of some synthesized mixed images with ghosting of reflections. From left to right in each row : input mixed image I, transmission layer T, transmission layer T scaled by α, reflection layer R and ghosted (or double shifted) reflection layer R.

3.3 Real datasets

In this subsection, we will describe the existing and proposed dataset of real world mixed images I containing undesirable reflection and the corresponding ground truth transmission layer T.

The existing publicly available datasets include **(i)** SIR^2 (Single Image Reflection Removal) dataset [3], **(ii)** CEILNet real dataset [1], **(iii)** PLNet real dataset [2], **(iv)** RID (Reflection Image Dataset) [4] and **(v)** ERRNet unaligned dataset [5]. Other real datasets (in [33, 34]) that are not publicly released at the time of writing this thesis are not described here. Fig. 3.9 illustrates the physical and mathematical image formation models of the three major types of reflection : focused, defocused and ghosting. Table 3.2 shows a comparison of the existing real datasets and the captured real dataset.

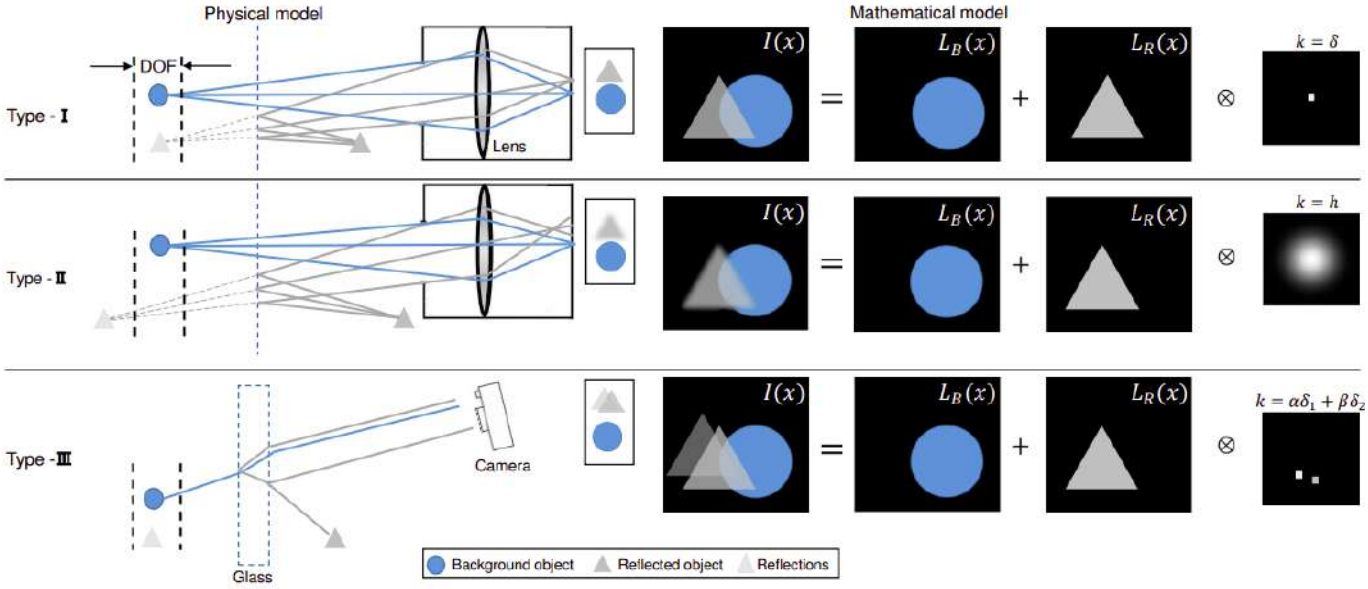

Figure 3.9: The physical and mathematical image formation models of the three major types of reflection : focused, defocused and ghosting.

The proposed dataset contains 622 (I, T) pairs captured using a smartphone placed on a tripod. Slight spatial misalignment is observable in some pairs of I and T due to refraction by glass. The dataset has both types of mixed images : the ones having localized reflection and the ones having reflection layer spread throughout the mixed image. Although the dataset has both indoor and outdoor scenes for R, T and R don't have varied illuminations. Only one glass thickness is used, and the glass is not thick enough to produce double shifted reflections (ghosting effect) in the mixed image. Fig.

Real dataset	Dataset size	Diversity	Post-capture spatial alignment of I and T
Our dataset	622 pairs of (I, T)	scenes, reflection location	yes
SIR^2 real dataset [3]	454 triplets of (I, T, R)	aperture size, glass thickness, illuminations, object reflectance, scenes, reflection location	yes
CEILNet real dataset [1]	45 I	scenes, reflection location	N/A
PLNet real dataset [2]	110 pairs of (I, T)	aperture size, illuminations, camera viewing angles, scenes, reflection location	no
Reflection Image Dataset [4]	3250 R	aperture size, illumination, scenes	N/A
ERRNet unaligned dataset [5]	450 pairs of (I, T)	reflection location	no

Table 3.2: A comparison of our real dataset and existing real datasets on the basis of dataset size, diversity and spatial alignment procedures implemented.

3.10 and Fig. 1.1 show examples of the (I, T) pairs in the captured dataset. Some experiments performed for spatial alignment of I and T in the captured dataset are described in detail in Section 4.2 .

The major limitation of all real datasets is that it is almost impossible to get a perfectly aligned pair of images I and T even with a tripod because of glass refraction. Due to refraction, the glass in front of the scene shifts the path of light transmitting through the glass and this leads to the spatial misalignment problem. Regional illumination changes between I and T can also cause misalignment. If not eliminate completely, it is possible to reduce these misalignment effects by carefully shooting only static scene from a stationary camera or using thinner glass with small refraction. The authors of [3] who propose the SIR^2 dataset perform processing on (I, T) pairs post-capture for spatial alignment. Specifically, they first extract SURF [40] feature points from I and T and then estimate the homographic transformation matrix using the RANSAC [41] algorithm. Finally, I is aligned to T with the estimated transformation. Even after this processing post-capture, slight spatial misalignments are visible.

Figure 3.10: Examples of some pairs of (I, T) in the proposed real dataset. In each row, the image at the left is the mixed image I and the image at the right is the transmission layer T.

CHAPTER 4

ESTIMATION OF GROUND TRUTH FOR REFLECTION SEGMENTATION

For reflection segmentation using any of the deep learning based approaches, we need a large amount of labelled training data consisting of mixed images with undesirable reflection and their corresponding pixel level reflection binary masks, with pixel values of the masks set to 1 if reflection is present at the pixel location, and 0 otherwise. As the pixel-level manual annotation of the reflection binary masks for a large number of images is very cumbersome and lengthy, we propose methods for estimation of ground truth of the reflection strength maps and reflection binary masks of synthetic mixed images. We also try our approach on real data, but the results are not very encouraging.

In Section 4.1, we present results of estimation of ground truth reflection strength maps and reflection binary masks on the synthetic datasets that are described in Section 3.2. In Section 4.2, we present the results of our experiments for estimation of ground truth reflection strength maps and reflection binary masks on real datasets that are described in Section 3.3.

4.1 Estimation of ground truth reflection strength maps and reflection binary masks for synthetic data

All the mixed image data synthesis procedures described in Section 3.2 essentially synthesize the mixed image as per

$$I = \tilde{T} + \tilde{R} \tag{4.1}$$

where $\tilde{T}$ and $\tilde{R}$ denote the effective transmission layer and reflection layer respectively. The effective transmission layer and reflection layer images are modified versions of the real transmission layer and reflection layer respectively, and they are modified differently for different data synthesis procedures.

The effective transmission and reflection layer images for each data synthesis procedure, as shown in Table 3.1, are listed below :

- Linear mix dataset : $\tilde{T} \leftarrow \alpha T, \tilde{R} \leftarrow (1-\alpha)R$
- CEILNet synthetic dataset : $\tilde{T} \leftarrow T, \tilde{R} \leftarrow gaussian\ blurred\ and\ gamma\ corrected\ R$
- PLNet synthetic dataset : $\tilde{T} \leftarrow \alpha T, \tilde{R} \leftarrow gaussian\ blurred\ and\ vignette\ applied\ R$
- Convex blurring dataset : $\tilde{T} \leftarrow \alpha T, \tilde{R} \leftarrow (1-\alpha)R \circledast G$
- Focused reflection dataset : $\tilde{T} \leftarrow T, \tilde{R} \leftarrow gaussian\ blurred\ and\ gamma\ corrected\ R$
- Ghosting dataset : $\tilde{T} \leftarrow \alpha T, \tilde{R} \leftarrow R \circledast K$

We will now describe the two proposed approaches to estimate the ground truth reflection strength maps and reflection binary masks for synthetic data of each type.

4.1.1 Proposed approach 1 : absolute differencing of the mixed image and transmission layer intensities

Both the RGB images, the mixed image I and the effective transmission layer $\tilde{T}$ are converted to YUV color space first. The reflection strength map is generated by subtracting the y channel of I and the y channel of $\tilde{T}$, as per Eqn. 4.2,

$$rsm[i][j] = |(y_I[i][j] - y_{\tilde{T}}[i][j])| \tag{4.2}$$

where rsm denotes the reflection strength map, y_I denotes the intensity channel of the mixed image I and $y_{\tilde{T}}$ denotes the intensity channel of the effective transmission layer $\tilde{T}$. The reflection strength map is a single channel image and has pixel values between $0-1$, indicating the absolute strength of reflection at that pixel location. A value of 0 indicates that $\tilde{R}$ has zero intensity at that location and a value of 1 indicates that $\tilde{T}$ has zero intensity at that location.

To generate the reflection binary mask, the reflection strength map is thresholded at pixel level for 3 values of thresholds : **(i)** 0.1, **(ii)** 0.2 and **(iii)** 0.3. As shown in Eqn. 4.3, the binary mask pixels are set a value of 1 where the strength map is above the

threshold and set a value of 0 otherwise. In Eqn. 4.3, rbm denotes the reflection binary mask and t denotes the threshold value chosen from amongst $0.1, 0.2$ and 0.3.

$$rbm[i][j] = \begin{cases} 1, & rsm[i][j] > t \\ 0, & otherwise \end{cases} \tag{4.3}$$

4.1.2 Proposed approach 2 : proportion of reflection intensity in the mixed image intensity

Both the RGB images, the mixed image I and the effective reflection layer $\tilde{R}$ are converted to YUV color space first. The reflection strength map is generated by dividing the y channel of $\tilde{R}$ and the y channel of I, as per Eqn. 4.4,

$$rsm[i][j] = (y_{\tilde{R}}[i][j]/y_I[i][j]) \tag{4.4}$$

where rsm denotes the reflection strength map, y_I denotes the intensity channel of the mixed image I and $y_{\tilde{R}}$ denotes the intensity channel of the effective reflection layer $\tilde{R}$. The reflection strength map is a single channel image and has pixel values between $0-1$ indicating the relative strength of reflection at that pixel location. A value of 0 indicates that $y_{\tilde{R}}$ contributes 0% to the total intensity at that location and value of 1 indicates that $y_{\tilde{R}}$ contributes 100% to the total intensity at that location.

To generate the reflection binary mask, the reflection strength map is thresholded at pixel level for 3 values of thresholds : **(i)** 0.1, **(ii)** 0.2 and **(iii)** 0.3. As shown in Eqn. 4.3, the binary mask pixels are set a value of 1 where the strength map is above the threshold and set a value of 0 otherwise.

4.1.3 Qualitative estimation results for synthetic data

In this subsection we provide some examples of ground truth reflection strength maps and ground truth reflection binary masks estimated for the following categories of synthetic data - **(i)** linear mix dataset : Fig. 4.1, **(ii)** CEILNet synthetic dataset : Fig. 4.2, **(iii)** PLNet synthetic dataset : Fig. 4.3, **(iv)** convex blurring dataset : Fig. 4.4, **(v)**

focused reflection dataset : Fig. 4.5 and **(vi)** ghosting dataset : Fig. 4.6.

4.1.4 Analysis of estimation results on synthetic data

Approach 1 is based on absolute differencing of the y channel values of I and $\tilde{T}$ and approach 2 is based on relative y channel values of $\tilde{R}$ and I. The ground truth reflection strength maps and reflection binary masks look good visually. Since the mixed image I is generated via $I = \tilde{T} + \tilde{R}$, the approach 1, which also subtracts the y channels of I and $\tilde{T}$, yields better reflection strength maps. The reflection strength maps generated via approach 1 also don't have any scene structures of T, unlike approach 2.

Since our reflection binary masks are generated from reflection strength maps via hard-coded thresholds, approach 2 can yield better reflection binary masks in certain situations. We can consider a example where $y_I[i][j] = 0.1$, $y_{\tilde{T}}[i][j] = 0.01$ and $y_{\tilde{R}}[i][j] = 0.09$ at a particular pixel location. Here, although $y_{\tilde{R}}[i][j]$ contributes 90% of the total intensity of $y_I[i][j]$, the binary reflection mask generated via approach 1 will not classify this pixel as reflection pixel because $y_I[i][j] - y_{\tilde{T}}[i][j] = 0.09$ which is lesser than the threshold values $0.1, 0.2$ and 0.3. Approach 2 will still classify it as a reflection pixel because $y_{\tilde{R}}[i][j]/y_I[i][j] = 0.9$, which is greater than the threshold values of $0.1, 0.2$ and 0.3. Similarly, some cases can be identified where reflection binary masks obtained via approach 1 are better than those obtained via approach 2.

4.2 Experiments for estimation of ground truth reflection strength maps and reflection binary masks for real data

For real data, since we do not have the ground truth reflection layer, only approach 1 in Subection 4.1.1 can be utilized to estimate reflection strength maps and reflection binary masks. Since the mixed image formation model is not known for given real T and R, theoretically estimating the reflection strength maps and reflection binary masks is a highly ill-posed problem. For estimating the reflection strength maps for real data, we use the approach 1 described in Subsection 4.1.1 with $\tilde{T} = \alpha T$ for five

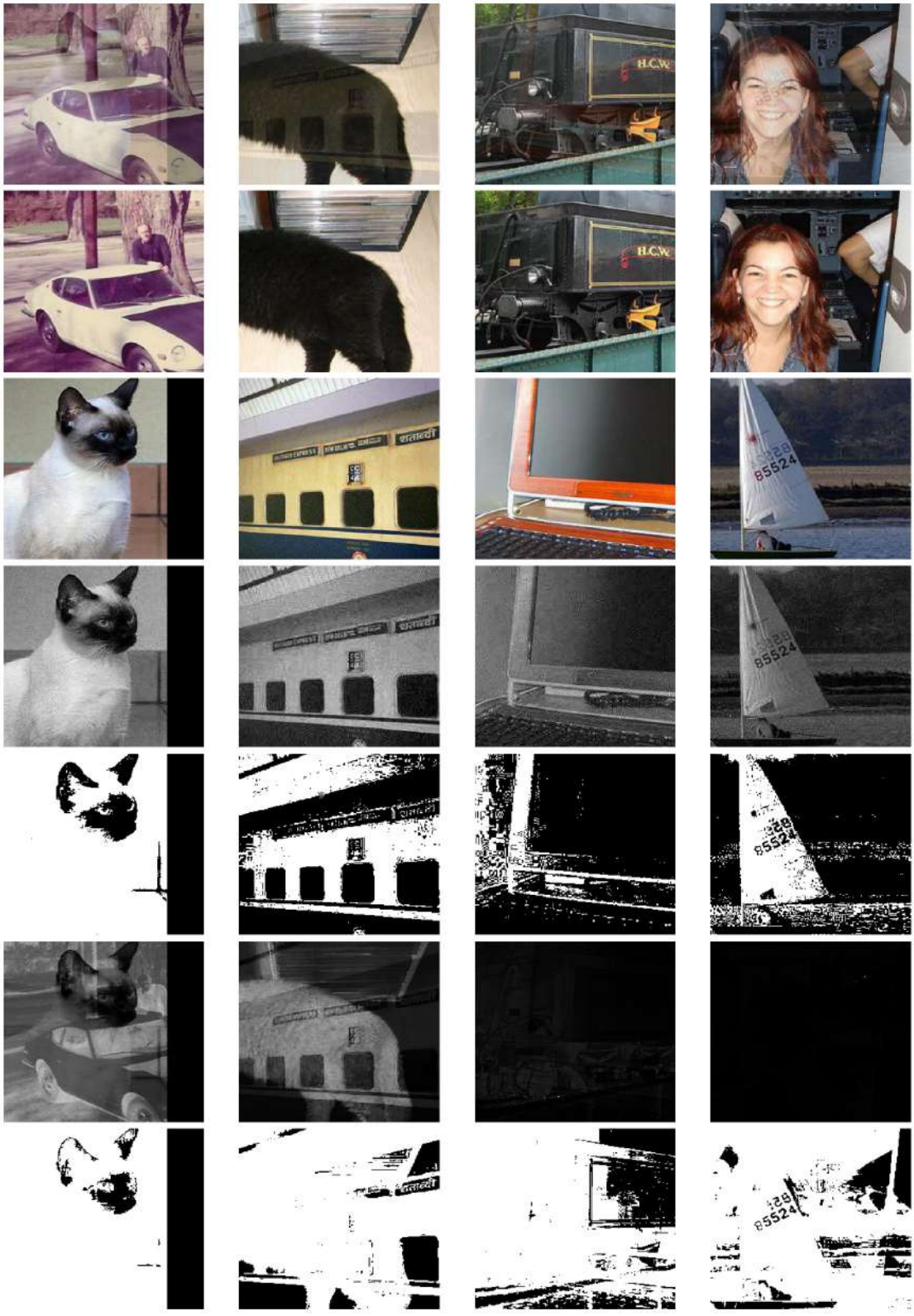

Figure 4.1: Some examples of reflection strength maps and reflection binary masks estimated for mixed images synthesized by linear mixing of T and R. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2.

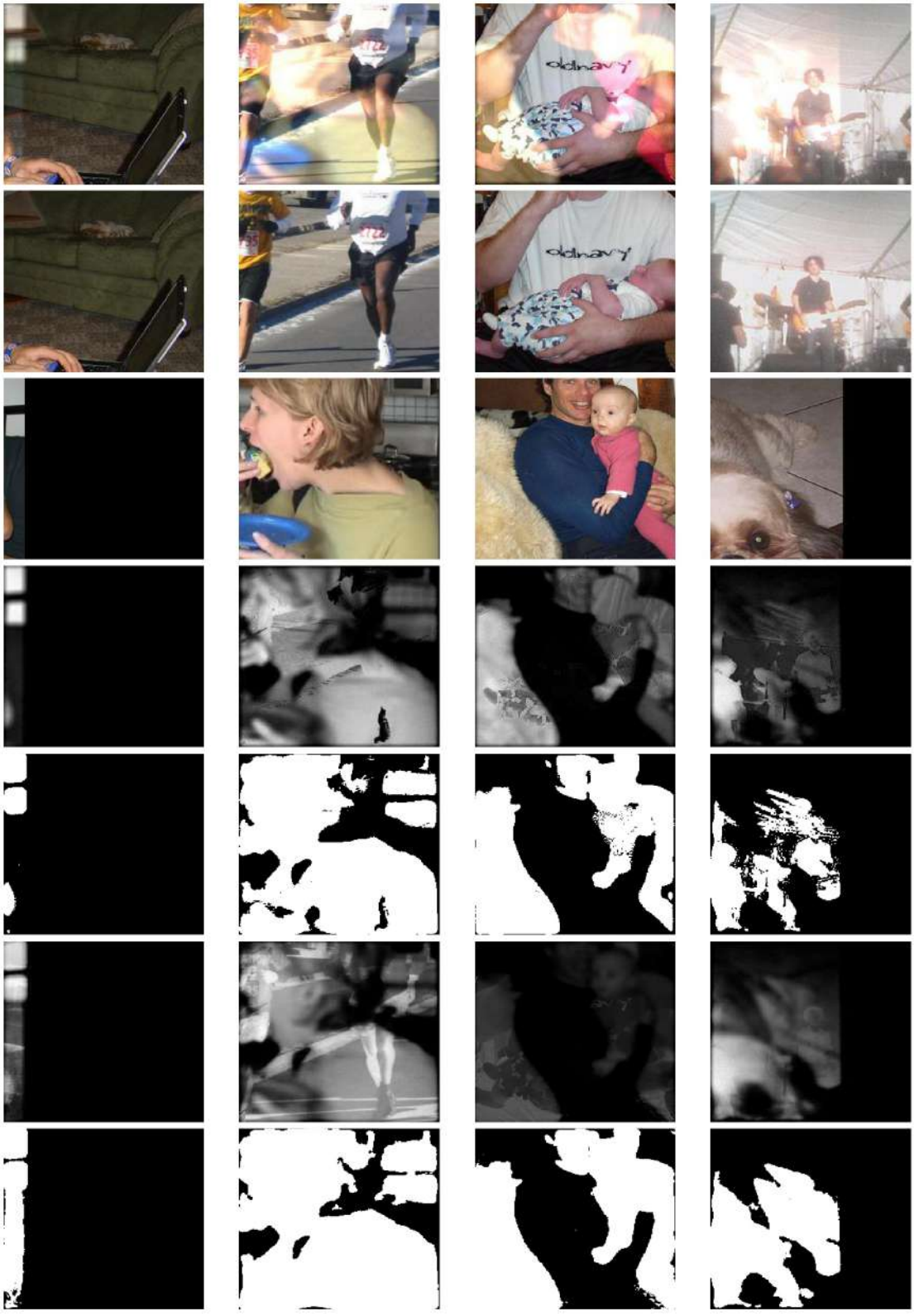

Figure 4.2: Some examples of reflection strength maps and reflection binary masks estimated for mixed images synthesized as per the CEILNet data synthesis procedure. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2.

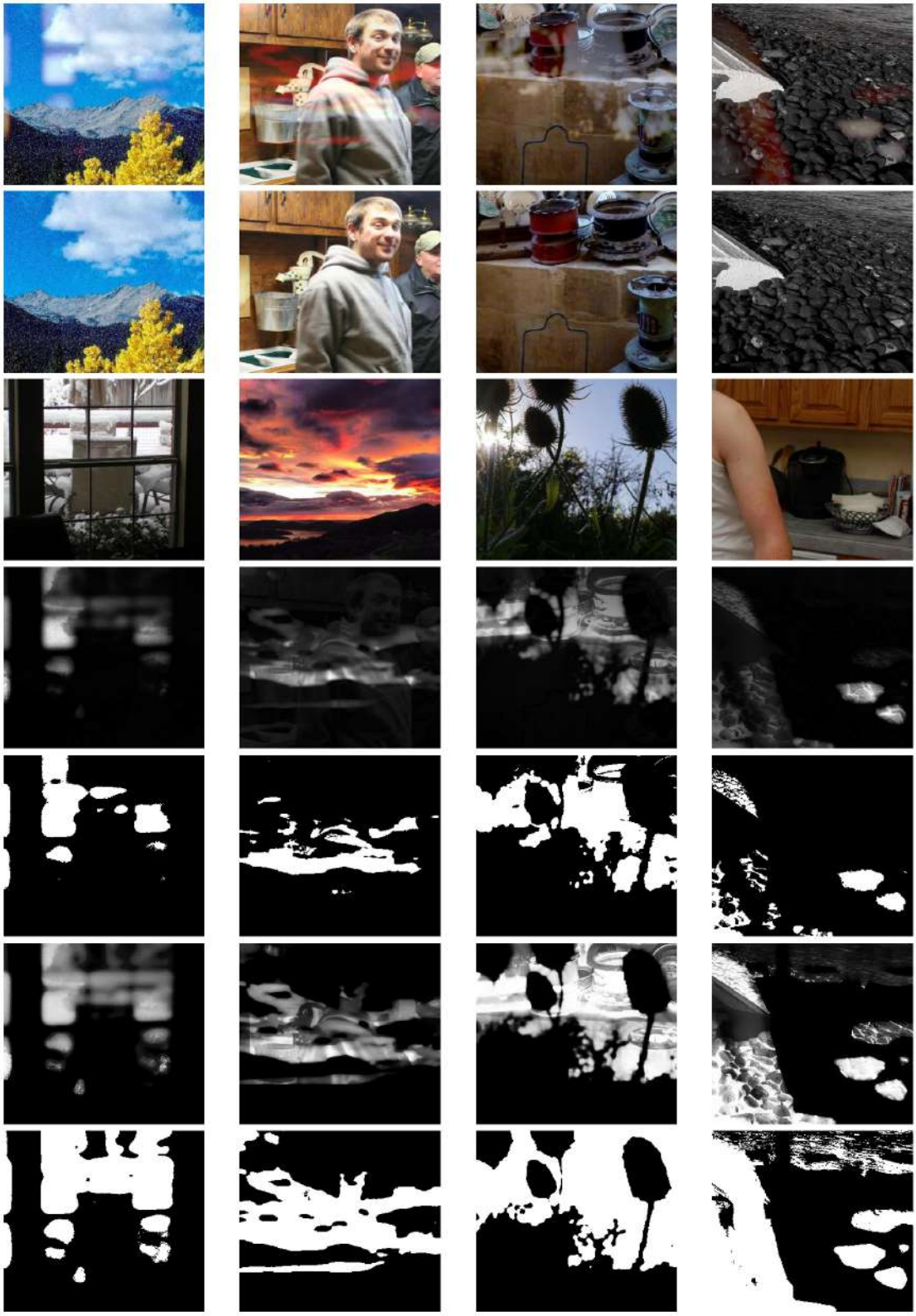

Figure 4.3: Some examples of reflection strength maps and reflection binary masks estimated for mixed images synthesized as per PLNet data synthesis procedure. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2.

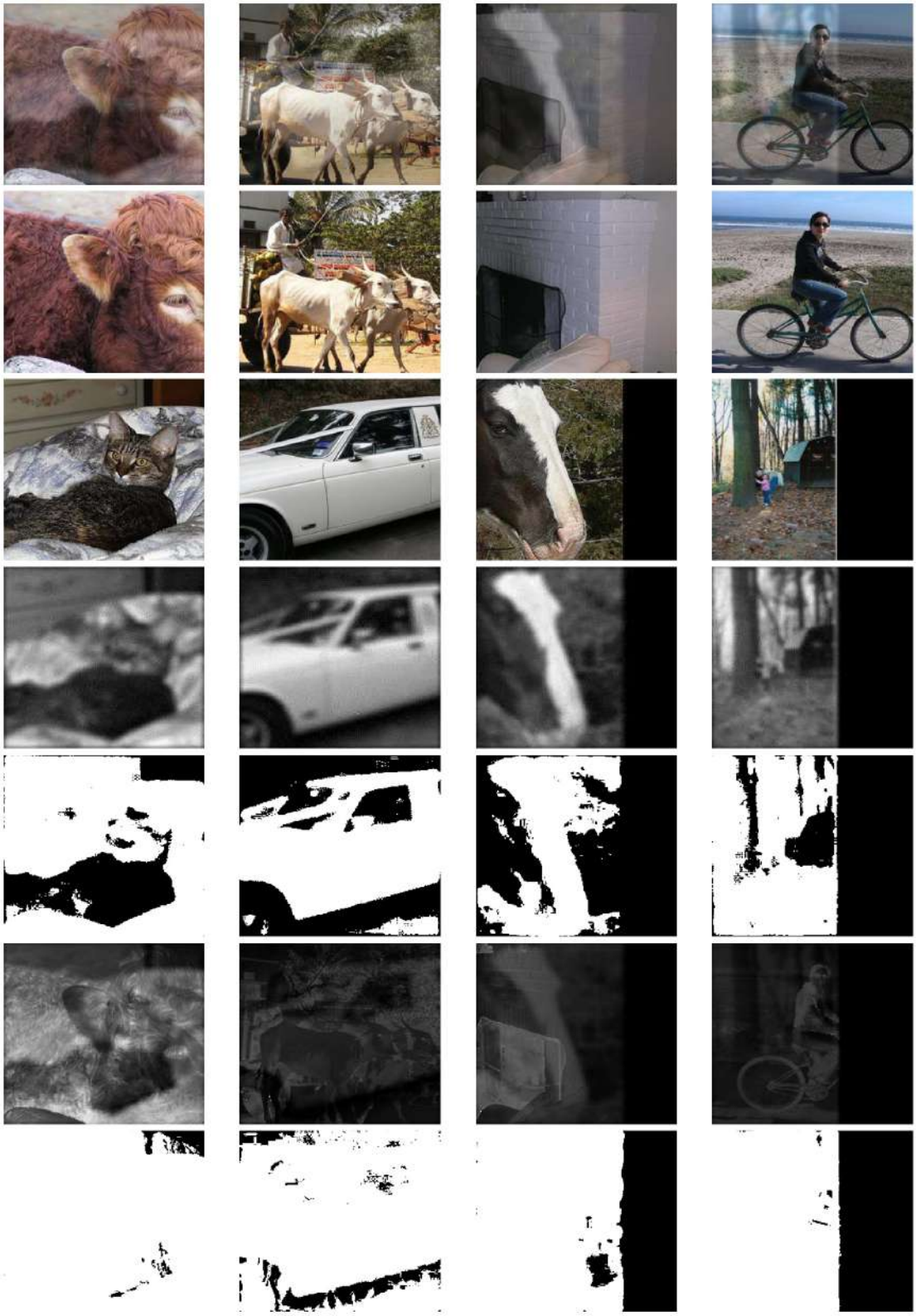

Figure 4.4: Some examples of reflection strength maps and reflection binary masks estimated for mixed images synthesized as per Subsection 3.2.4. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2.

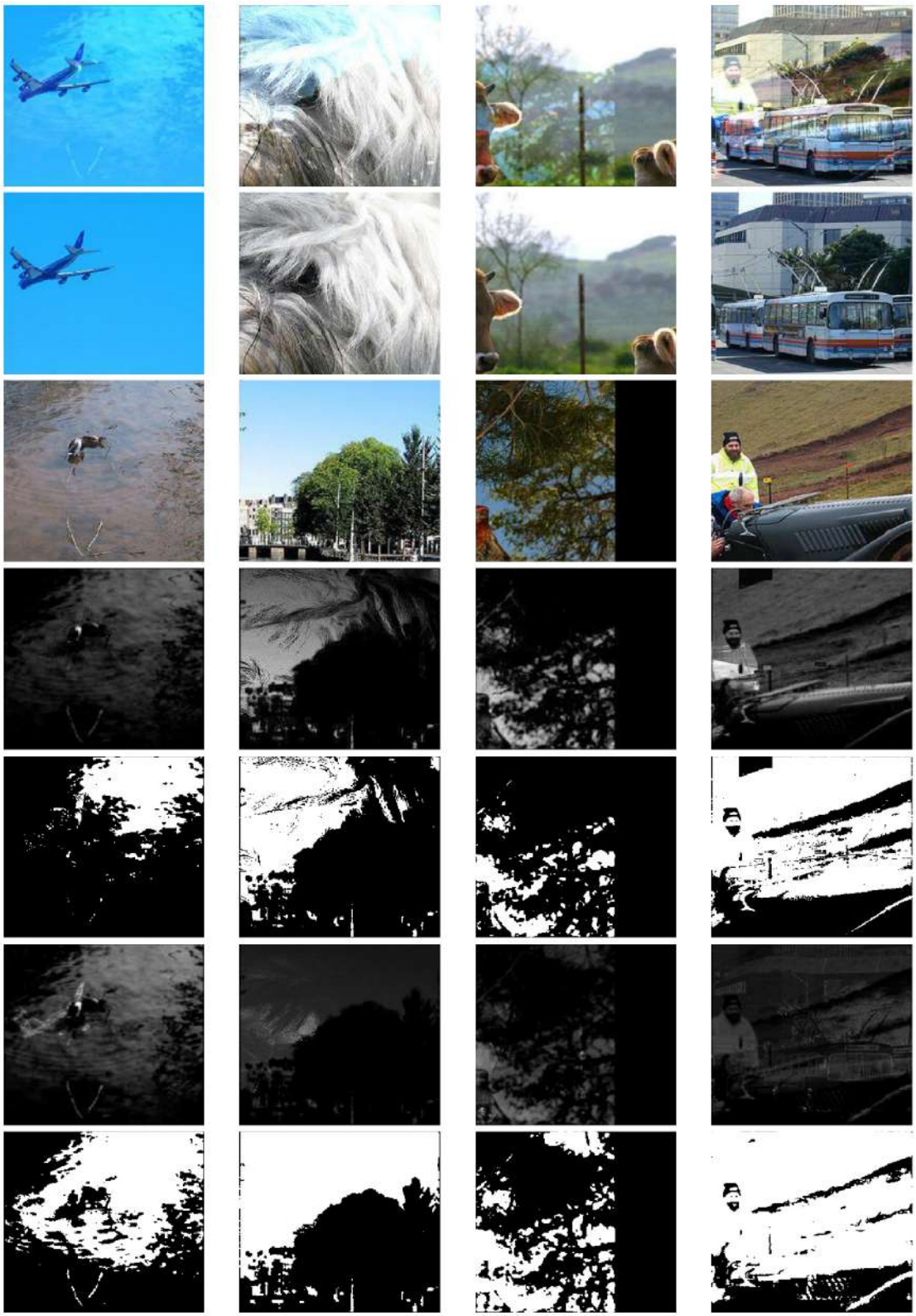

Figure 4.5: Some examples of reflection strength maps and reflection binary masks estimated for mixed images with focused reflections. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2.

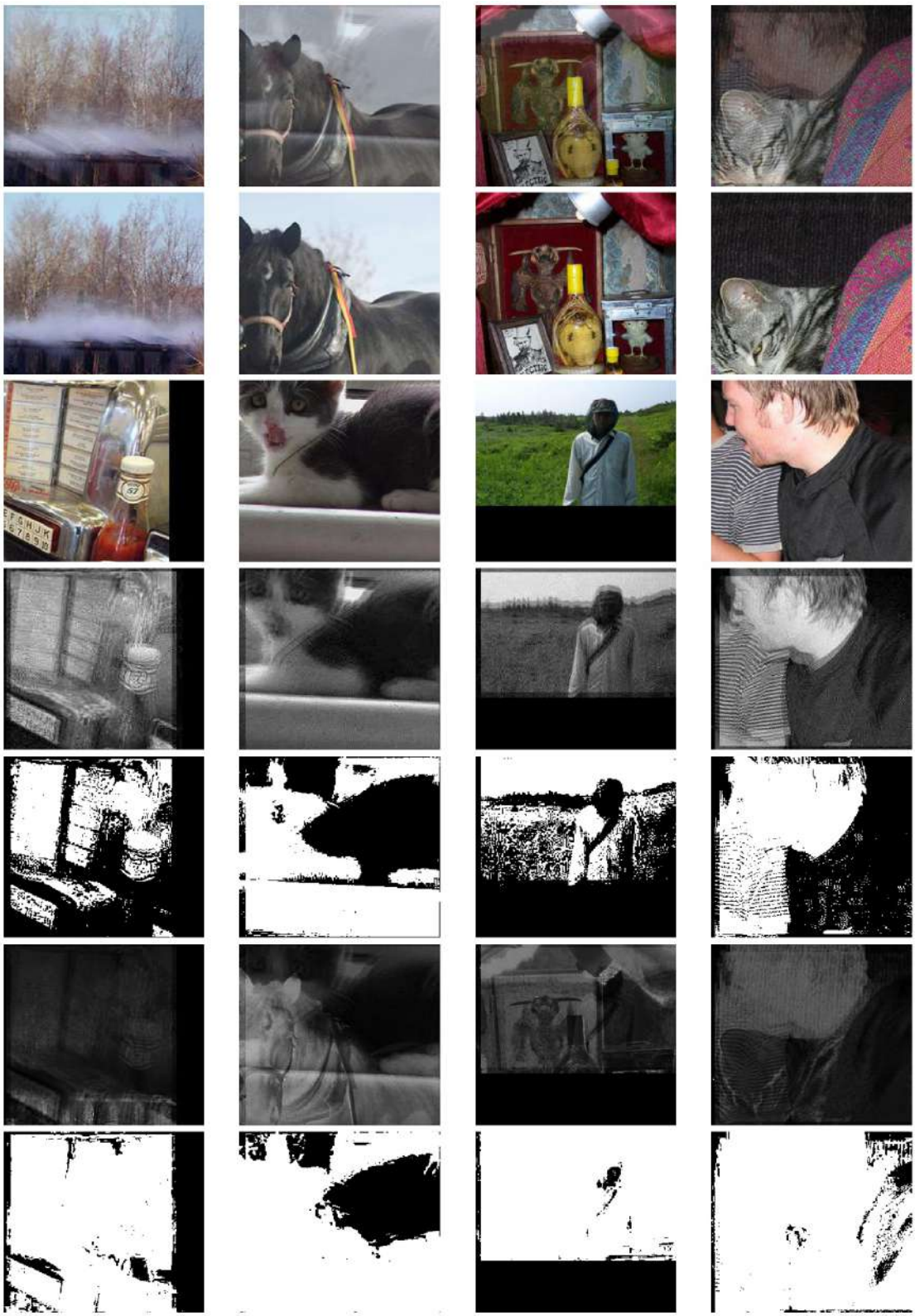

Figure 4.6: Some examples of reflection strength maps and reflection binary masks estimated for mixed images with ghosting effect. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1, rbm_1 obtained by setting a threshold of 0.1 on rsm_1, reflection strength map rsm_2 estimated via approach 2 and rbm_2 obtained by setting a threshold of 0.1 on rsm_2.

different values of α, with $\alpha = 0.6, 0.7, 0.8, 0.9$ and 1.0. For each reflection strength map, the corresponding reflection binary masks are subsequently generated by applying thresholds of $0.1, 0.2$ and 0.3.

The reflection strength maps and reflection binary masks don't look accurate for the proposed dataset and the SIR^2 dataset. Many pixel locations in the mixed image that are not affected by reflection are wrongly classified as affected by reflection. The reflection strength maps still have scene structures of T. This is primarily because of the slight spatial misalignment between I and T due to refraction by glass, because of which direct differencing of the intensities of the two images doesn't yield the desired residual reflection component. We reiterate the fact that this spatial misalignment between I and T exists not because of camera jerks during capture, but due to refraction by glass, which is unavoidable when capturing an image through glass. This spatial misalignment between I and T exists even if image capture is done in a completely static background and on a tripod. To reduce the spatial misalignment between I and T, we attempt the following approaches : **(i)** we extract ORB features [42] from I and T, estimate the homography transformation matrix using the RANSAC algorithm [41] and then align I and T using the estimated transformation matrix, **(ii)** perform gaussian blurring on I and T with kernel of sizes 3×3 to 17×17 with $\sigma = 2$ and **(iii)** downsample I and T from $(1960, 4032)$ to $(224, 460)$, with a gaussian anti-aliasing filter. Approach **(iii)** reduced the spatial misalignment in the proposed dataset to some extent but differencing still didn't yield the desired reflection components.

Even if I and T are spatially aligned (like in SIR^2), the estimation results are not accurate because our assumption of the linear image model formation with glass transmittance constant throughout the image doesn't hold true for real data. The reflection binary masks of different mixed images look best with different settings of α and threshold, because of which it is not feasible to use a single ground truth estimation approach with hardcoded thresholds. Some examples of the estimated ground truth reflection strength maps and reflection binary masks for real images are shown in Fig. 4.7.

We conclude this discussion by noting that for real data, estimation of ground truth reflection strength maps and reflection binary masks is an intractable problem and unlike synthetic data, pixel level manual annotation looks unavoidable for real data.

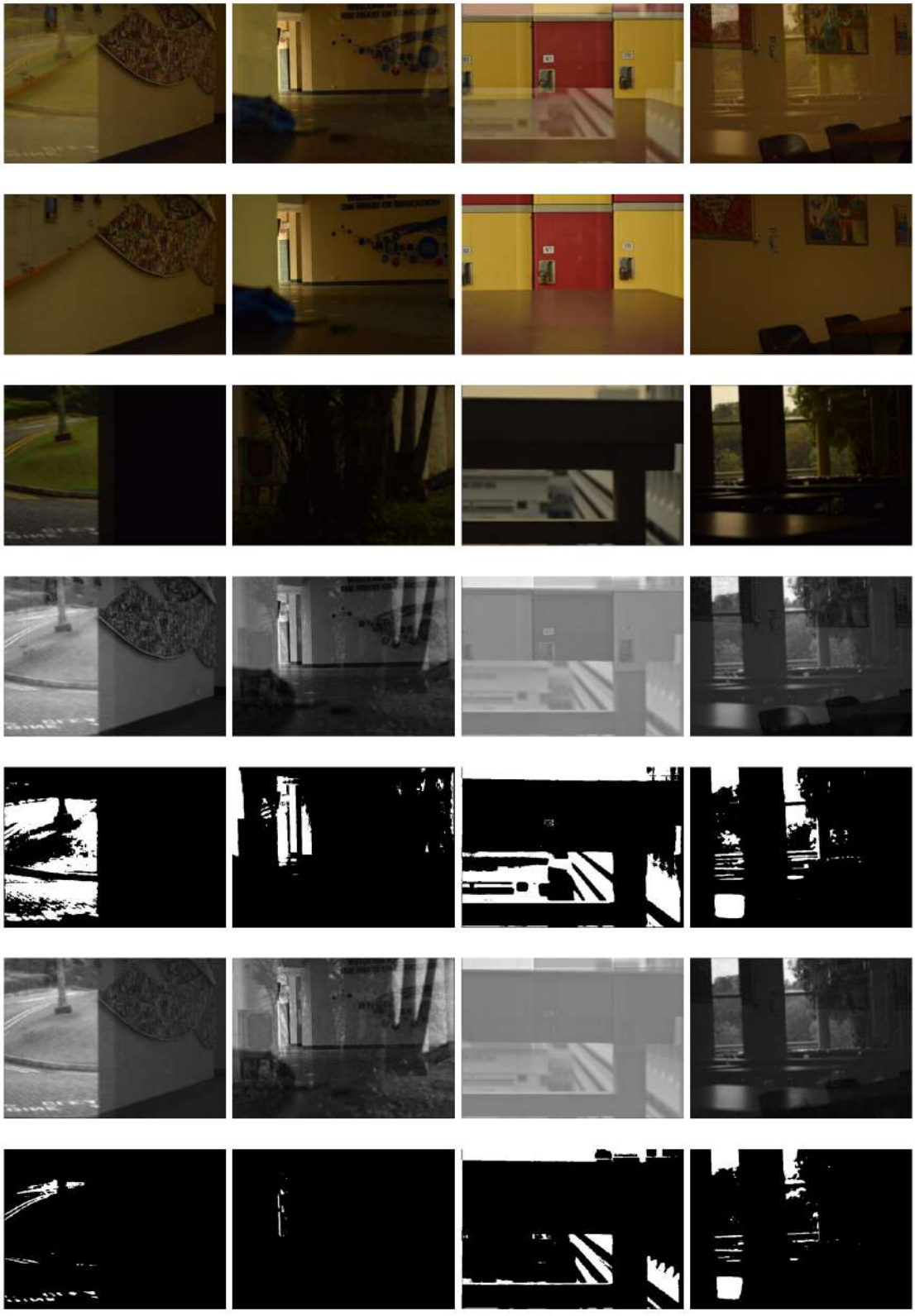

Figure 4.7: Some examples of reflection strength maps and reflection binary masks estimated for real mixed images from the SIR^2 dataset. In each column, from top to bottom : mixed image I, transmission layer T, reflection layer R, reflection strength map rsm_1 estimated via approach 1 with $\alpha = 0.6$, rbm_1 obtained by setting a threshold of 0.2 on rsm_1, reflection strength map rsm_2 estimated via approach 1 with $\alpha = 0.8$ and rbm_2 obtained by setting a threshold of 0.2 on rsm_2.

CHAPTER 5

REFLECTION SEGMENTATION

In this chapter, we provide the results of reflection segmentation using our network trained on synthetic pairs of the mixed images and their corresponding ground truth reflection binary masks estimated in Chapter 4. Section 5.1 provides the details of the network and the training procedure used for the task of reflection segmentation. This is followed by Section 5.2, which provides quantitative results of the proposed reflection segmentation network on synthetic images, and Section 5.3, which provides some visual results of the proposed reflection segmentation network on synthetic and real images.

5.1 Network and training details

In Chapter 4, we had proposed two approaches (Subsection 4.1.1 and Subsection 4.1.2) for estimation of ground truth reflection strength maps and reflection binary masks for synthetic data. We will use the reflection binary masks obtained via approach 1 (Subsection 4.1.1) as the ground truth images for training the network.

We train 4 different models for 4 different types of synthetic data - **(i)** CEILNet synthetic dataset (Subsection 3.2.2), **(ii)** PLNet synthetic dataset (Subsection 3.2.3), **(iii)** focused reflections dataset (Subsection 3.2.5) and **(iv)** ghosting dataset (Subsection 3.2.6). The training dataset for each model consists of synthetic pairs of mixed images and ground truth reflection binary masks estimated via approach 1, as described in Subsection 4.1.1. 20% of the training data is used as validation data. We call the networks trained on CEILNet synthetic dataset, PLNet synthetic dataset, focused reflection dataset and ghosting dataset as CNet, PNet, FNet and GNet respectively.

The network architecture is inspired from DeepLabv3 [43]. We modify the final layer in DeepLabv3 [43] so that the final output contains only two classes : **(i)** pixels having reflection and **(ii)** pixels not having reflection. We use a ResNet-101 [44] backbone for feature extraction and the weights are initialized as per pre-trained DeepLabv3

[43]. For training the network, we use the Adam optimizer [45] with a learning rate of 0.0001. MSE between the predicted and ground truth reflection binary masks is used as the training loss. We train the network for 10 epochs with a batch size equal to 4. For the training loss, BCE (Binary Cross Entropy) is also tried because MSE is not generally considered the best loss for classification problems. But we observed that the batch training loss isn't very stable on doing so. To improve the stability of the batch training loss, we tried the BCE with logits loss but the results were similar to the network trained with MSE as the loss function. The network outputs a grayscale image with pixel value equal to the probability that the pixel belongs to reflection class. For the final output reflection binary mask, the classification threshold value is set to 0.5. In this manner, we train 4 different models for the 4 types of synthetic data.

Fig. 5.1 shows the evolution of **(i)** train MSE loss, **(ii)** test MSE loss, **(iii)** F1 score on the training dataset (with a classification threshold of 0.5), **(iv)** train AUROC value, **(v)** F1 score on the test dataset (with a classification threshold of 0.5) and **(vi)** test AUROC value for the reflection segmentation model trained on the CEILNet synthetic dataset. Similarly, Fig. 5.2, Fig. 5.3 and Fig. 5.4 show the plots for networks trained on the PLNet synthetic dataset, focused reflections dataset and ghosting dataset respectively.

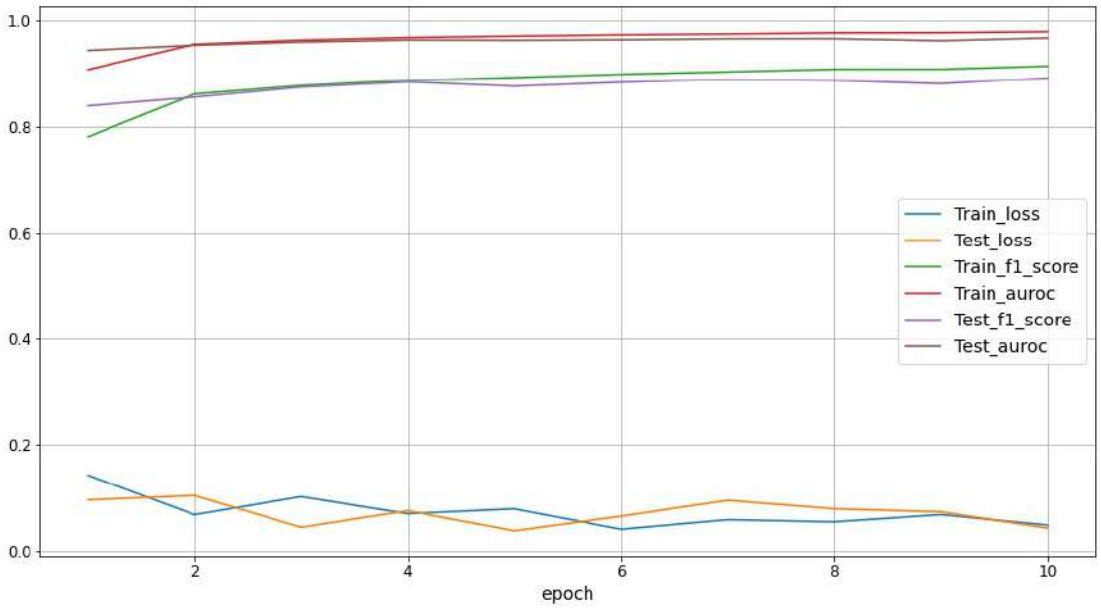

Figure 5.1: Plot showing the evolution of train MSE loss, test MSE loss, F1 score on the training dataset, train AUROC value, F1 score on the test dataset and test AUROC value for CNet.

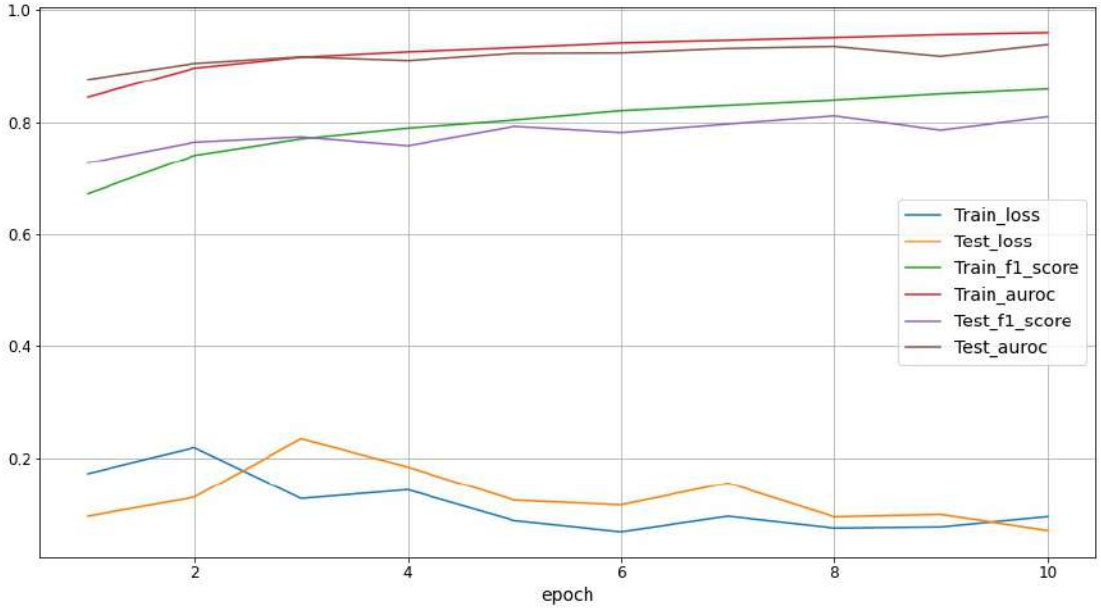

Figure 5.2: Plot showing the evolution of train MSE loss, test MSE loss, F1 score on the training dataset, train AUROC value, F1 score on the test dataset and test AUROC value for PNet.

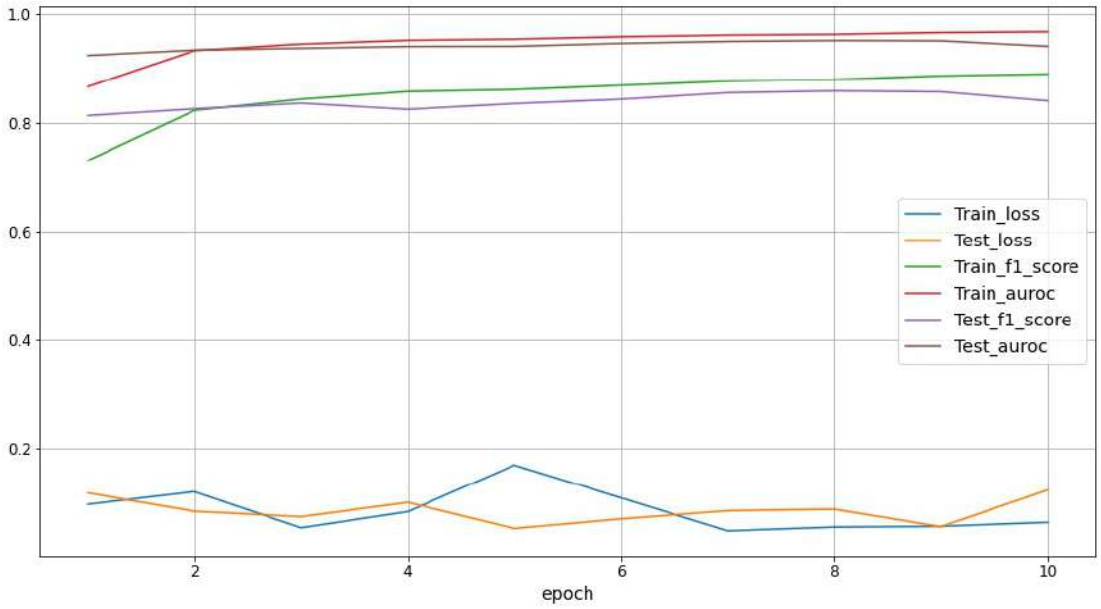

Figure 5.3: Plot showing the evolution of train MSE loss, test MSE loss, F1 score on the training dataset, train AUROC value, F1 score on the test dataset and test AUROC value for FNet.

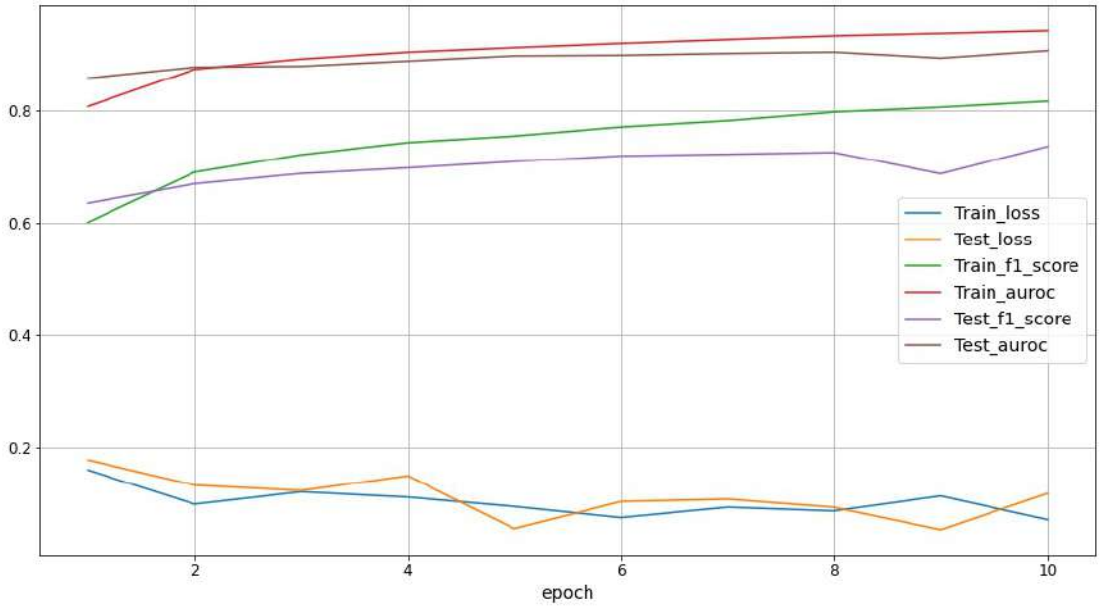

Figure 5.4: Plot showing the evolution of train MSE loss, test MSE loss, F1 score on the training dataset, train AUROC value, F1 score on the test dataset and test AUROC value for GNet.

5.2 Quantitative evaluation on synthetic data

The train and test AUROC values of the four trained models on their respective synthetic datasets are shown in Table 5.1, where the training dataset size of x means x pairs of mixed images and their corresponding ground truth reflection binary masks estimated via approach 1 (Subsection 4.1.1). Due to lack of ground truth reflection binary masks for real data (as explained in Section 4.2), we don't perform quantitative evaluation on real data.

For the CEILNet, focused reflection and ghosting synthetic datasets we use 4480 image pairs for training and 1120 image pairs for testing. For the PLNet synthetic dataset we use 6400 image pairs for training and 1600 pairs for testing. Higher the AUROC values, better the network is at predicting $0s$ as $0s$ and $1s$ as $1s$. As shown in Table 5.1, we achieve train and test AUROC values in excess of 0.9 for all models. These are outstanding results on synthetic data and the high test AUROC values suggest that all our 4 networks are able to distinguish between the two classes (reflection and no reflection) really well.

Synthetic dataset	Training dataset size	Testing dataset size	Train AUROC	Test AUROC
CEILNet synthetic dataset	4480	1120	0.97	0.96
PLNet synthetic dataset	6400	1600	0.95	0.93
focused reflection dataset	4480	1120	0.96	0.95
ghosting dataset	4480	1120	0.94	0.90

Table 5.1: The train AUROC and test AUROC values of the four different reflection segmentation networks trained on four synthetic datasets : **(i)** CEILNet synthetic dataset, **(ii)** PLNet synthetic dataset, **(iii)** focused reflection dataset and **(iv)** ghosting dataset

5.3 Qualitative evaluation on synthetic and real data

In this section, we show some visual results of the reflection segmentation networks on synthetic and real data. Fig. 5.5, Fig. 5.6, Fig. 5.7 and Fig. 5.8 show results of the four trained reflection segmentation networks tested on their corresponding synthetic datasets. Fig. 5.9, 5.10 and 5.11 show estimation results of the four trained reflection segmentation networks CNet, PNet, FNet and GNet on real data.

Based on the outputs of the reflection strength maps and reflection binary masks, we conclude that the networks achieve very good segmentation results on synthetic data. Of the pixel locations that are misclassified as not having reflection, most have very low intensity reflections. Most of the high intensity reflection regions are correctly classified as having reflection.

For qualitative evaluation on real data, we use 45 images from the CEILNet real dataset [1]. We observe that for almost all the 45 real images, PNet, the reflection segmentation network trained on the PLNet synthetic dataset, performs the best visually amongst all the four networks. CNet, trained on the CEILNet synthetic dataset, performs second best. Due to a lack of ghosted reflections in the CEILNet real dataset, GNet produces unsatisfactory outputs for most input images. In images with focused reflections, PNet still produces better results than FNet.

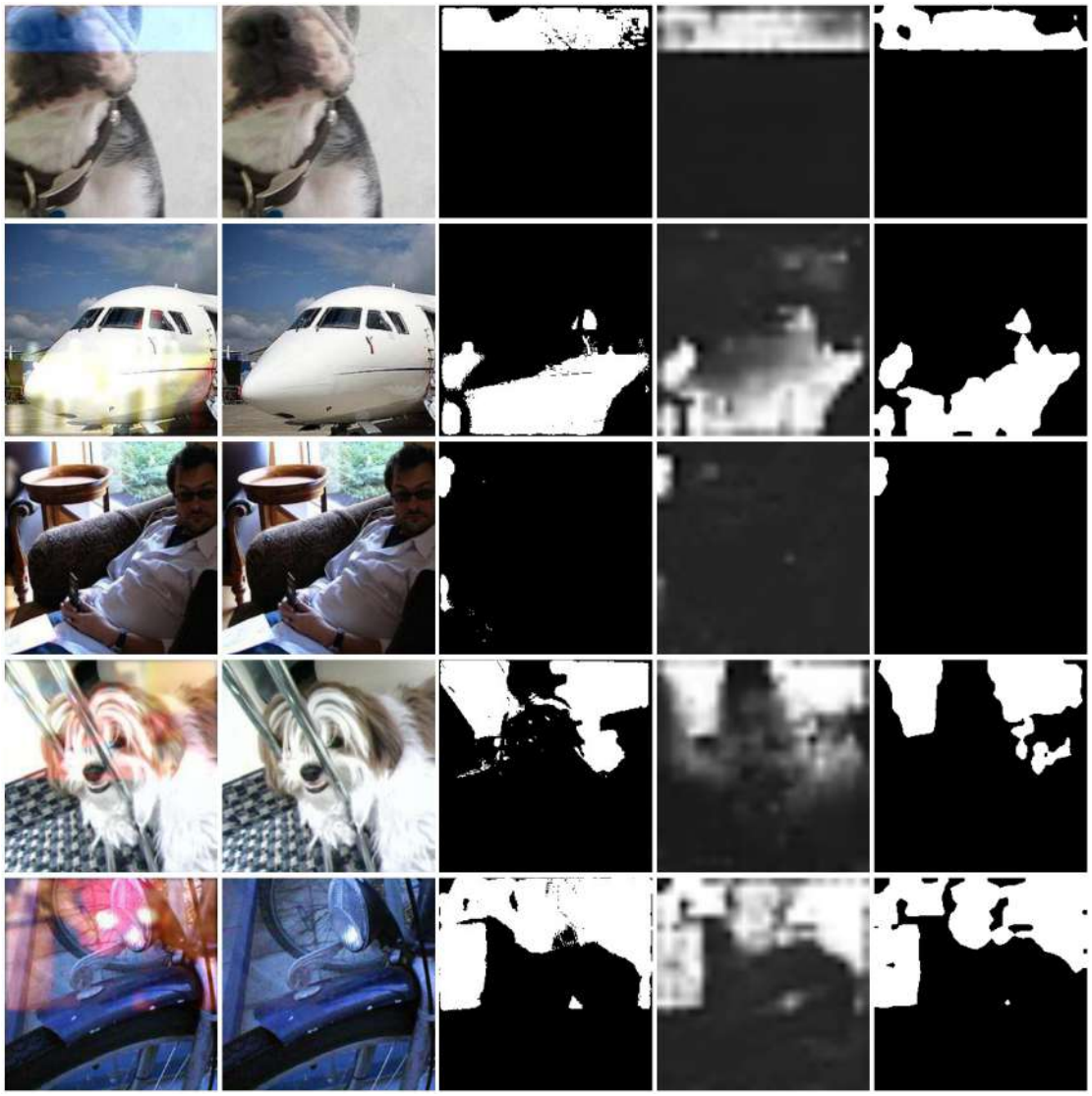

Figure 5.5: Some reflection segmentation results of CNet on the CEILNet synthetic dataset. In each row, from left to right : the mixed image I, transmission layer T, ground truth reflection binary mask as estimated in Subsection 4.1.1, output reflection strength map and output reflection binary mask

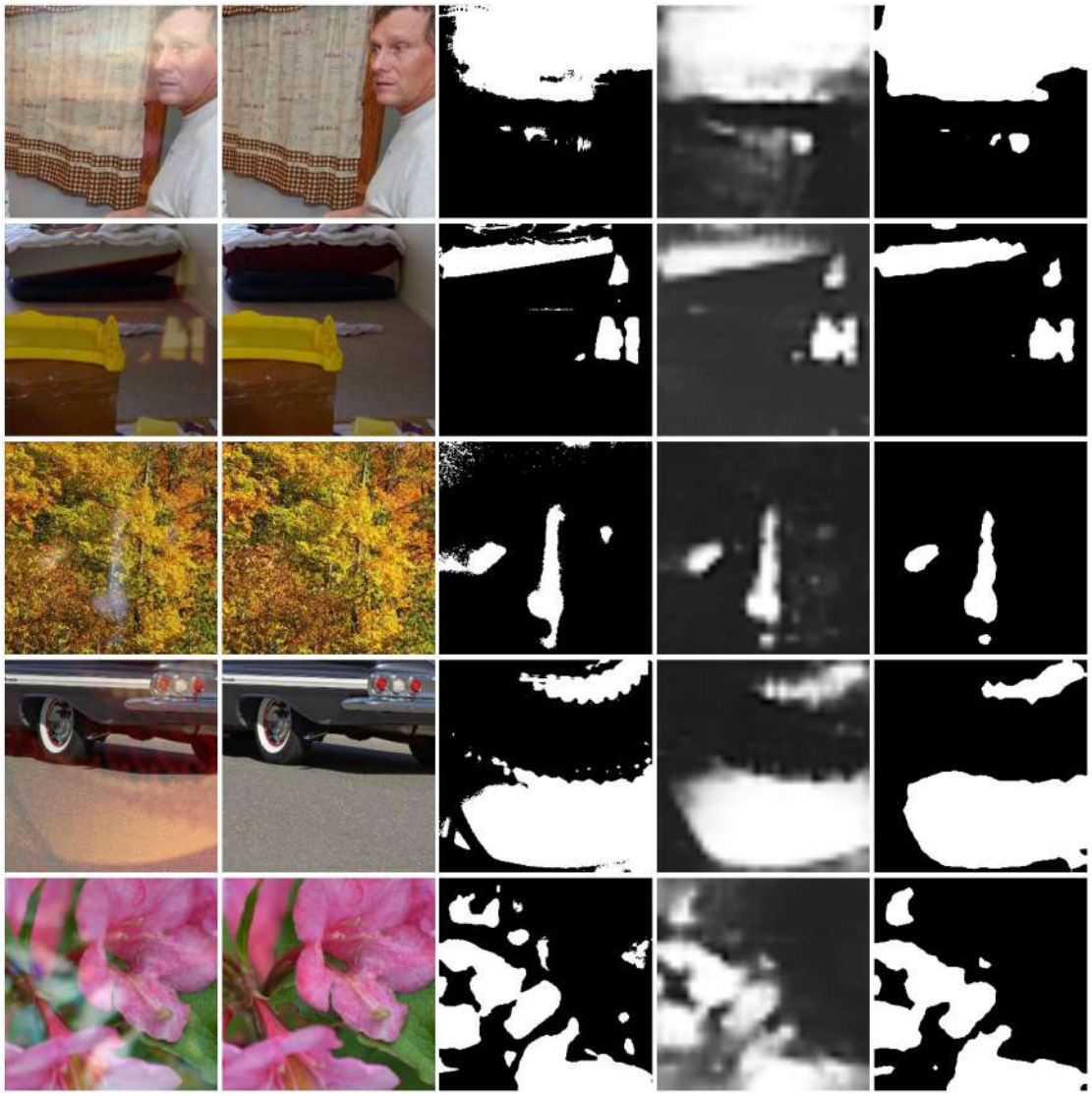

Figure 5.6: Some reflection segmentation results of PNet on the PLNet synthetic dataset. In each row, from left to right : the mixed image I, transmission layer T, ground truth reflection binary mask as estimated in Subsection 4.1.1, output reflection strength map and output reflection binary mask

Figure 5.7: Some reflection segmentation results of FNet on the focused reflection dataset. In each row, from left to right : the mixed image I, transmission layer T, ground truth reflection binary mask as estimated in Subsection 4.1.1, output reflection strength map and output reflection binary mask

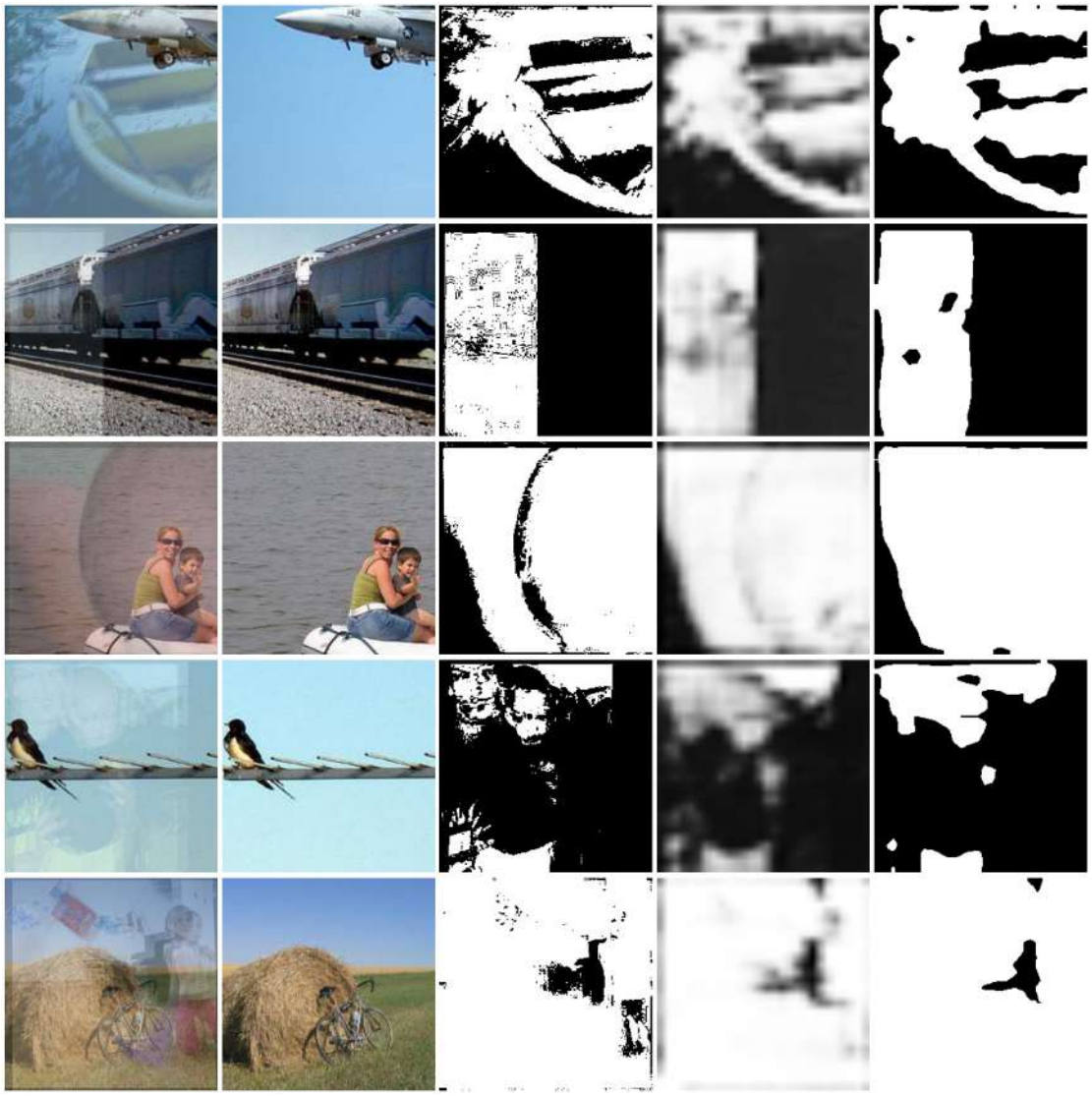

Figure 5.8: Some reflection segmentation results of GNet on the ghosting dataset. In each row, from left to right : the mixed image I, transmission layer T, ground truth reflection binary mask as estimated in Subsection 4.1.1, output reflection strength map and output reflection binary mask

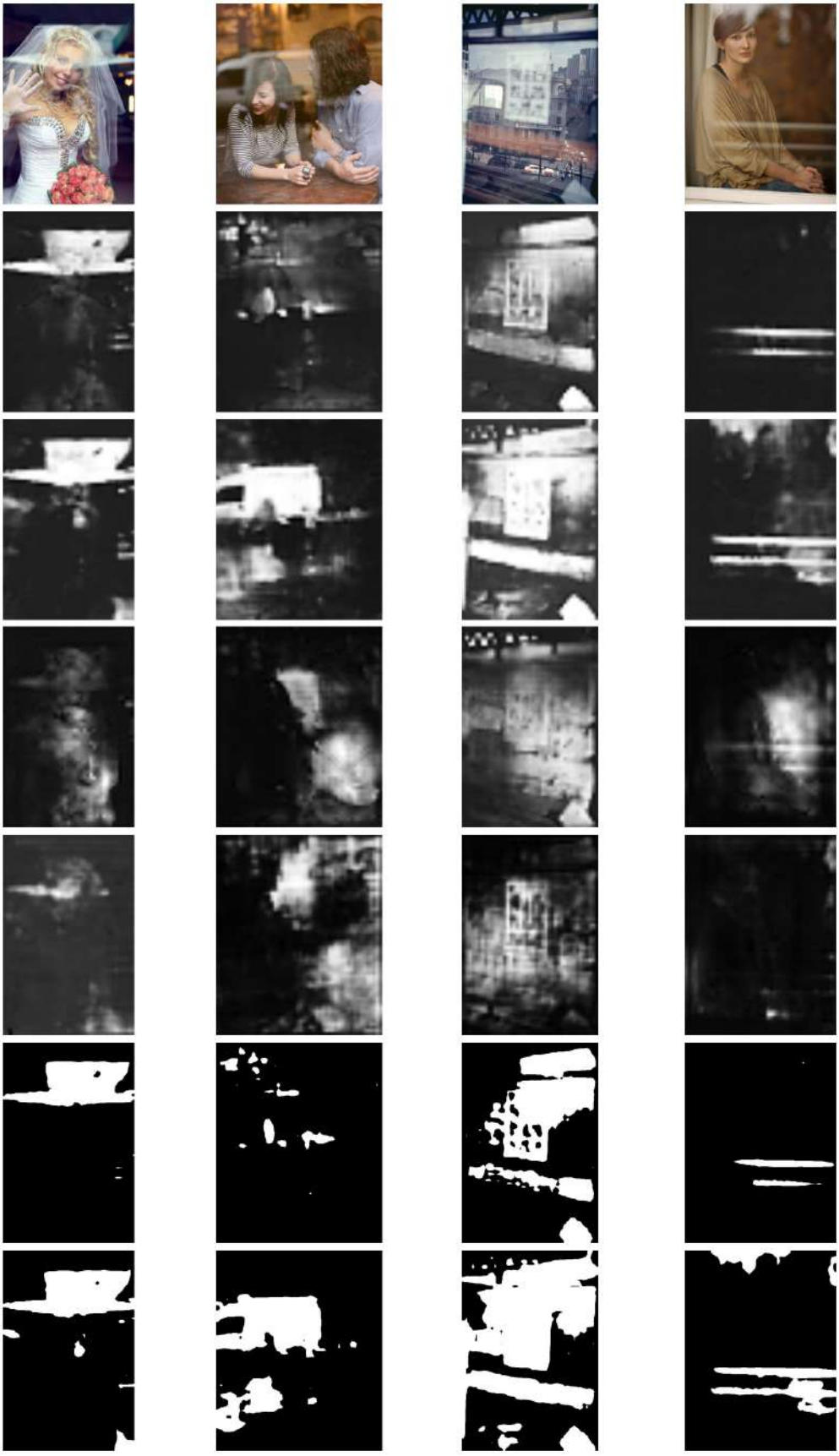

Figure 5.9: Some reflection segmentation results on real images from CEILNet real dataset [1]. In each column, from top to bottom : mixed image I, rsm output by CNet, rsm output by PNet, rsm output by FNet, rsm output by GNet, rbm output by CNet and rbm output by PNet.

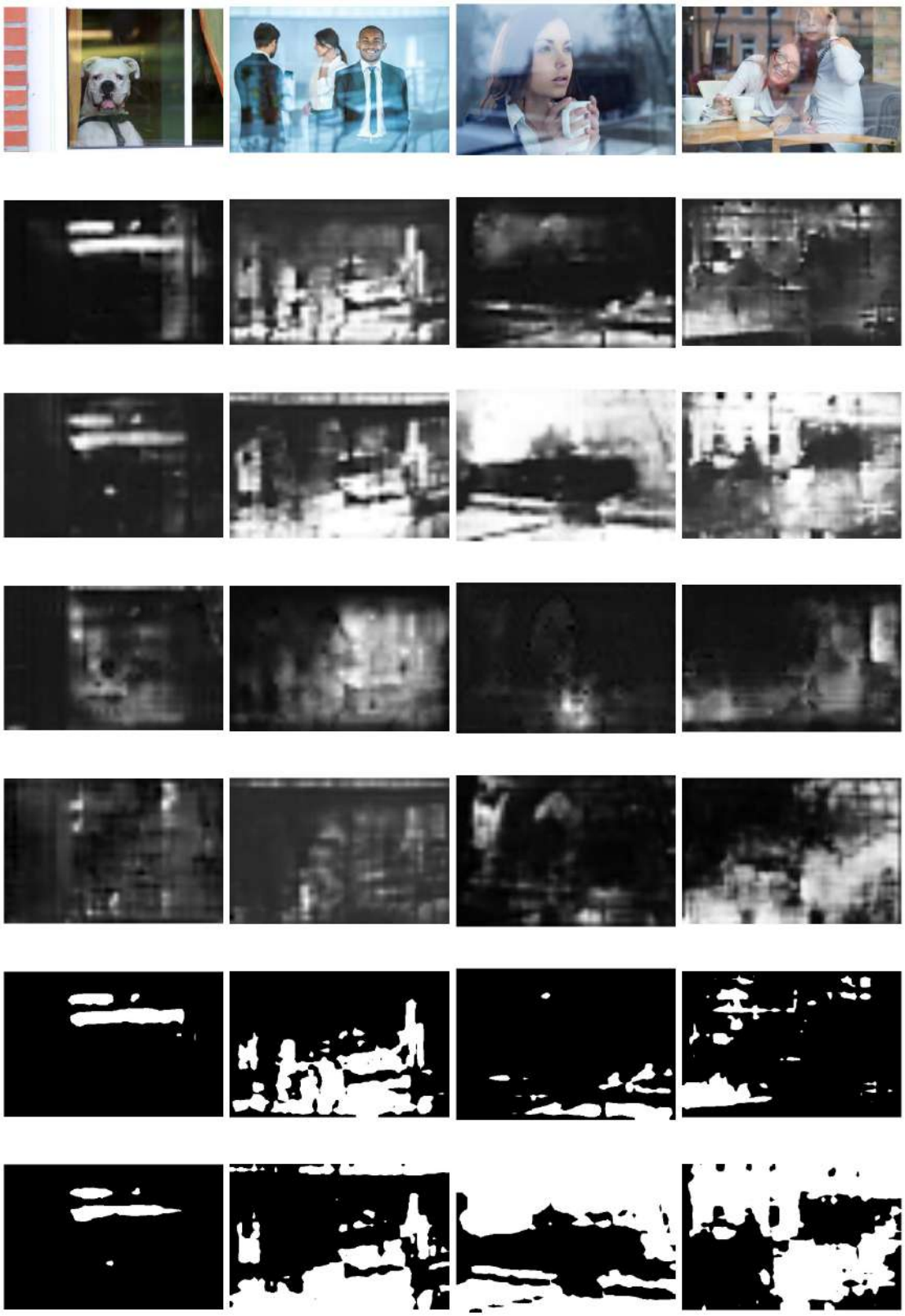

Figure 5.10: Some more reflection segmentation results on real images from CEILNet real dataset [1]. In each column, from top to bottom : mixed image I, rsm output by CNet, rsm output by PNet, rsm output by FNet, rsm output by GNet, rbm output by CNet and rbm output by PNet.

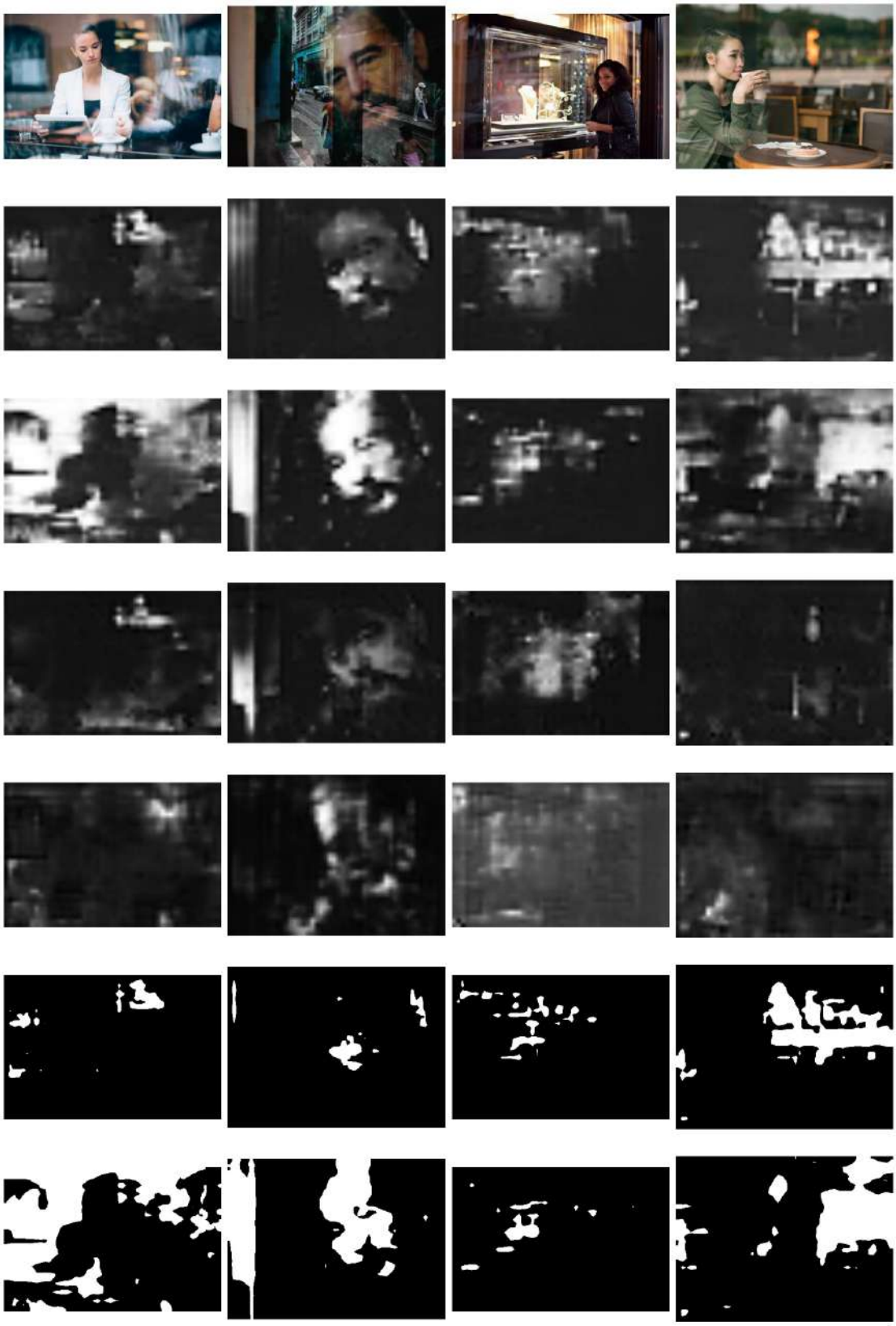

Figure 5.11: Some reflection segmentation results on real images from CEILNet real dataset [1] where none of the networks perform well. In each column, from top to bottom : mixed image I, rsm output by CNet, rsm output by PNet, rsm output by FNet, rsm output by GNet, rbm output by CNet and rbm output by PNet.

CHAPTER 6

KEY RESULTS and SUMMARY

The major outcomes of this thesis are :

- We have described the major challenges associated with reflection removal using single images in detail.
- We have provided a comprehensive overview of the existing deep learning based and traditional approaches for single image reflection removal.
- In order to model different kind of real-world reflections well, we synthesize a dataset of 50000 mixed images containing undesirable reflection from real pairs of the target transmission layer and the reflection layer.
- To deal with the acute shortage of real pairs of the mixed image and target transmission layer, we propose a real dataset with 622 pairs of mixed image and target transmission layer.
- We propose a novel approach for reflection removal, using a reflection binary mask as a prior image which can guide reflection removal from single images.
- We estimate the ground truth reflection masks for reflection segmentation on synthetic datasets to avoid cumbersome pixel level manual annotation of images.
- We achieve very accurate results for reflection segmentation on synthetic data. We also perform qualitative evaluation of the proposed reflection segmentation networks on real data.
- Though results for reflection removal are not provided here, we believe that accurate reflection segmentation of images can help achieve state-of-the-art results of reflection removal using single images.

CHAPTER 7

SCOPE FOR FUTURE WORK

This thesis provides accurate reflection segmentation results on mixed images affected by undesirable reflection. We hope that the accurate reflection segmentation maps can act as auxiliary information along with the input mixed image to help achieve state-of-the-art results for the task of single image reflection removal. An encoder-decoder architecture can be implemented to recover the transmission layer T using **(i)** the mixed image I and **(ii)** the corresponding reflection binary mask estimated via the proposed reflection segmentation network. The long-term extension of this thesis is thus, single image reflection removal.

For further improvements in the task of reflection segmentation using single image, we list down the following short-term steps that can be taken :

- We can manually annotate the reflection binary masks for real data at pixel-level for $100+$ mixed images. This will enable quantitative evaluation of the proposed reflection segmentation networks on real data. In this thesis, we have only performed qualitative evaluation of the reflection segmentation networks on real data due to lack of ground truth reflection binary masks for real data.
- We can segregate real mixed images having different types of reflections : **(i)** defocused, **(ii)** focused, **(iii)** ghosting and **(iv)** saturated. Performing quantitative evaluation of the proposed reflection segmentation networks on the segregated real datasets will provide us clarity about the four networks' performance on different types of reflections.
- We can train another network with the same architecture but with a combined training dataset consisting of **(i)** linear mix dataset, **(ii)** CEILNet synthetic dataset, **(iii)** PLNet synthetic dataset, **(iv)** convex blurring dataset, **(v)** focused reflection dataset and **(vi)** ghosting dataset. This might improve the reflection segmentation results because of increased robustness to different types of reflections.
- We can explore using the reflection binary masks estimated via approach 2 (Subsection 4.1.2) as ground truth for training the reflection segmentation network.
- We can train our reflection segmentation network using other state-of-the-art semantic segmentation network architectures : U-Net [36], DeepLabv3+ [46] etc.
- Most of the existing networks for single image reflection removal don't perform well in mixed images with low-light or weak backgrounds. A reflection segmentation network trained exclusively on mixed image with dark backgrounds can be trained. Mixed images with dark backgrounds can be synthesized if they are not available in sufficient quantity.

REFERENCES

[1] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David P. Wipf. A generic deep architecture for single image reflection removal and image smoothing. *CoRR*, abs/1708.03474, 2017.

[2] Xuaner Cecilia Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. *CoRR*, abs/1806.05376, 2018.

[3] R. Wan, B. Shi, L. Duan, A. Tan, and A. C. Kot. Benchmarking single-image reflection removal algorithms. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3942–3950, 2017.

[4] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. CRRN: multi-scale guided concurrent reflection removal network. *CoRR*, abs/1805.11802, 2018.

[5] Kaixuan Wei, Jiaolong Yang, Ying Fu, David P. Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. *CoRR*, abs/1904.00637, 2019.

[6] Amit Agrawal, Ramesh Raskar, Shree K. Nayar, and Yuanzhen Li. Removing photography artifacts using gradient projection and flash-exposure sampling. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, page 828–835, New York, NY, USA, 2005. Association for Computing Machinery.

[7] Naejin Kong, Yu-Wing Tai, and Joseph S. Shin. A physically-based approach to reflection separation: from physical modeling to constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(2):209–221, February 2014.

[8] Hany Farid and Edward H. Adelson. Separating reflections and lighting using independent components analysis. *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, 1:262–267 Vol. 1, 1999.

[9] K. Gai, Z. Shi, and C. Zhang. Blind separation of superimposed moving images using image statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):19–32, 2012.

[10] X. Guo, X. Cao, and Y. Ma. Robust separation of reflection from multiple images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2195–2202, 2014.

[11] Y. Li and M. S. Brown. Exploiting reflection change for automatic reflection removal. In *2013 IEEE International Conference on Computer Vision*, pages 2432–2439, 2013.

[12] Bernard Sarel and Michal Irani. Separating transparent layers through layer information exchange. In Tomás Pajdla and Jiri Matas, editors, *Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV*, volume 3024 of *Lecture Notes in Computer Science*, pages 328–341. Springer, 2004.

[13] Sudipta N. Sinha, Johannes Kopf, Michael Goesele, Daniel Scharstein, and Richard Szeliski. Image-based rendering for scenes with reflections. *ACM Trans. Graph.*, 31(4), July 2012.

[14] R. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 1, pages 246–253 vol.1, 2000.

[15] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T. Freeman. A computational approach for obstruction-free photography. *ACM Trans. Graph.*, 34(4), July 2015.

[16] Jiaolong Yang, Hongdong Li, Yuchao Dai, and Robby T. Tan. Robust optical flow estimation of double-layer images under transparency or reflection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1410–1419, 2016.

[17] A. Levin and Y. Weiss. User assisted separation of reflections from a single im-

age using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1647–1654, 2007.

[18] Anat Levin, Assaf Zomet, and Yair Weiss. Learning to perceive transparency from the statistics of natural scenes. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, page 1271–1278, Cambridge, MA, USA, 2002. MIT Press.

[19] Y. Li and M. S. Brown. Single image layer separation using relative smoothness. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2759, 2014.

[20] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot. Depth of field guided reflection removal. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 21–25, 2016.

[21] YiChang Shih, D. Krishnan, F. Durand, and W. T. Freeman. Reflection removal using ghosting cues. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3201, 2015.

[22] N. Arvanitopoulos, R. Achanta, and S. Süsstrunk. Single image reflection suppression. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1752–1760, 2017.

[23] Q. Zheng, B. Shi, X. Jiang, L. Duan, and A. C. Kot. Denoising adversarial networks for rain removal and reflection removal. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2766–2770, 2019.

[24] Renjie Wan, Boxin Shi, Haoliang Li, Ling yu Duan, Ah-Hwee Tan, and Alex Kot Chichung. Corrn: Cooperative reflection removal network. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[25] 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 (ECCV)*, pages 654–669, 2018.

[26] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Semantic guided single image reflection removal. *CoRR*, abs/1907.11912, 2019.

[27] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016.

[28] M. Jin, S. Süsstrunk, and P. Favaro. Learning to see through reflections. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12, 2018.

[29] Zhixiang Chi, Xiaolin Wu, Xiao Shu, and Jinjin Gu. Single image reflection removal using deep encoder-decoder network. *CoRR*, abs/1802.00094, 2018.

[30] Zhixin Xu, Xiaobao Guo, and Guangming Lu. Single image reflection removal based on deep residual learning. In *PRCV*, 2019.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[32] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. Generative single image reflection separation. *CoRR*, abs/1801.04102, 2018.

[33] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[34] Daiqian Ma, Renjie Wan, Boxin Shi, Alex C. Kot, and Ling-Yu Duan. Learning to jointly generate and separate reflections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[35] M. Heo and Y. Choe. Single-image reflection removal using conditional gans. In *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–4, 2019.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[37] Yakun Chang and Cheolkon Jung. Single image reflection removal using convolutional neural networks. *IEEE Transactions on Image Processing*, 28:1954–1966, 2019.

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

[39] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[40] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[41] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[42] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011.

[43] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[45] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[46] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018.