Learning-based Methods for Occluder-aided Non-Line-of-Sight Imaging

by

Safa C. Medin

B.Sc. in Electrical and Electronics Engineering, Boğaziçi University (2019)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

Author ____________________

Department of Electrical Engineering and Computer Science
August 27, 2021

Certified by ____________________

Gregory W. Wornell
Sumitomo Professor of Engineering
Thesis Supervisor

Accepted by ____________________

Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Learning-based Methods for Occluder-aided Non-Line-of-Sight Imaging

by

Safa C. Medin

Submitted to the Department of Electrical Engineering and Computer Science
on August 27, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Imaging scenes that are not in our direct line-of-sight, referred to as non-line-of-sight (NLOS) imaging, has recently gained considerable attention from the computational imaging community. With a diverse set of potential applications in several domains, NLOS imaging is an emerging topic with many unanswered questions despite the progress made in the last decade. In this thesis, we aim to find answers to some of these questions by focusing on a popular NLOS imaging setting, namely occluder-aided imaging, which exploits occluding structure in the scenes to extract information from the hidden scenes. We do this by first focusing on the scene classification problem, where we study the problem of identifying individuals by exploiting shadows cast by occluding objects on a diffuse surface. In particular, we develop a learning-based method that discovers hidden cues in the shadows and relies on building synthetic scenes composed of 3D face models obtained from a single photograph of each identity. We transfer what we learn from the synthetic data to the real data using domain adaptation in a completely unsupervised way and report classification accuracies over 75% for a binary classification task that takes place in a scene with unknown geometry and occluding objects. Next, we focus on the problem of scene estimation, which aims to recover an image of the hidden scene from NLOS measurements. We present a learning-based framework that exploits deep generative models and demonstrate the promise of this framework via simulations.

Thesis Supervisor: Gregory W. Wornell
Title: Sumitomo Professor of Engineering

Acknowledgments

First and foremost, I would very much like to thank my advisor, Prof. Gregory Wornell, for giving me the opportunity to join his group and for the amazing guidance and supervision he has provided me over the last two years. My journey at MIT since 2019 has been certainly challenging with the pandemic covering most of it, and I could not ask for a better advisor during these difficult times. Greg has always been extremely patient and understanding with me, and I am beyond grateful to be one of his students. His exceptional intellectual ability and immense knowledge have shaped my career more than I could ever imagine, and his ideas have truly inspired me throughout my research that eventually led to this thesis.

Next, I would like to express my most sincere gratitude to Prof. Bill Freeman and Prof. Frédo Durand, with whom I had the privilege to collaborate. Bill and Frédo's exceptional guidance has helped me tremendously on my research and I will always be grateful for their generosity with their time and attention. Bill's frequent and detailed feedback on my work has been an invaluable resource, which I will forever appreciate.

Several inspirations I had for the initial parts of this thesis were thanks to the profound experience I have gained throughout my internships at the Mitsubishi Electric Research Laboratories (MERL), where I had an opportunity to work under the supervision of Dr. Tim Marks. I am beyond glad that Tim encouraged me to work on things that I had no familiarity with — his wonderful guidance made me a better researcher than I would have ever been otherwise. My time at MERL has been truly unforgettable thanks to my collaborators there: Prof. Xiaoming Liu, Dr. Bernhard Egger, Dr. Anoop Cherian, Dr. Ye Wang, and Prof. Josh Tenenbaum. I would specifically like to thank Xiaoming for expertly guiding me through the unfamiliar realms of computer vision and teaching me many things I found useful in my research. I am also very grateful to have collaborated with one of the most amazing people I have ever met, Bernhard, whose immense creativity, intellectual curiosity, and generosity helped me develop crucial skills in my career, and I feel very lucky to call him my friend.

I would also very much like to thank Prof. Vivek Goyal and Prof. John Murray-Bruce, with whom I had a chance to collaborate during my time at Boston University. I am very grateful to Vivek for introducing me to the exciting field of computational imaging, which later sparked my interest in non-line-of-sight imaging. Vivek is certainly one of the most influential people in my career, and I will forever be indebted to him for all of the guidance and support he has so generously offered me since 2016. I am also very glad to have met John during my undergraduate years, whose constant and insightful feedback on my work and sincere friendship helped me tremendously as I took my first steps into becoming a researcher.

Five amazing years I had at Boğaziçi University had a great influence on who I am today, and I would like to thank all of my professors there for their effort in my education and development. I would especially like to express my deep gratitude to Prof. Bülent Sankur, for his excellent supervision and mentorship during my undergraduate years. His unique way of teaching incited my passion for the field of computer vision, and I could not ask for a better research advisor at Boğaziçi — he helped me gain new perspectives on many research topics, which I found very useful in my later studies.

Next, I would like to thank to a few former and current members of the Signals, Information, and Algorithms lab at MIT: Tejas Jayashankar, Dr. Adam Yedidia, Gary Lee, Dr. Joshua Lee, Abhin Shah, and Tricia O'Donnell. Tejas has been one of my closest friends at MIT, and I am grateful for his support during challenging times. Adam helped me gain so much insight into the non-line-of-sight imaging, which I greatly appreciate. Gary, Joshua, and Abhin have been very generous with their time and I am very thankful for their continued advice. Last but certainly not least, I would also like to thank our amazing lab admin, Tricia, for her unending kindness and compassion since the day I joined the lab.

I would not be able to make it through the last couple of years at MIT without the continued support and encouragement of my friends, especially during the dreadful days of the pandemic. I would very much like to thank Doğa Doğan, Eren Kızıldağ, and Thomas Henzel for the treasured memories throughout our time in Cambridge. I would also like to thank my dearest friends from Boğaziçi, Ozan Yakar and Fatih Dinç, who never ceased their support despite living thousands of miles away from me.

Finally, I would like to express my heartfelt gratitude and deepest love for my parents, Emine and Yalçın, and my sister, Merve, for their unconditional love and endless support since the day I was born, and for doing their absolute best to help me achieve my dreams. A very special thanks goes to my little niece, Derin, for bringing much joy into my life over video calls. I am very happy to have finally met you in person!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Whenever we interact with images in our daily lives, we often care little about how these images are formed. Whether they are captured by an optical device or just form in our brains, these images are typically the result of a physical process referred to as *image formation*. When an image of a scene is formed, the light rays emitted from a light source hit the objects in the scene and are reflected, refracted, transmitted, or absorbed by these objects, eventually hitting a camera sensor, photographic film, or our retina. Although the appearances of physically distant objects might seem somewhat independent from each other, they are in fact connected through the light that propagates through the entire scene volume, creating a *light field*, which surrounds and binds all of the objects in the scene. Therefore, the appearance of each object in a given scene is effectively influenced by everything else in that scene.

When we observe a scene, we might sometimes be interested in having some knowledge about parts of the scene that are outside our field of view. Since the light field connects the hidden part of the scene to its visible part, at least some information about these hidden scenes is embedded in our observations. The study of extracting information from the hidden scenes based on the visible scenes that are in our direct line-of-sight is called *non-line-of-sight (NLOS) imaging*, and it is the primary focus of this thesis. NLOS imaging is currently an active area of research with a diverse set of potential applications in surveillance, search-and-rescue, robotic vision, and medical imaging.

Throughout the last decade, NLOS imaging has been applied to several different tasks such as recovering 2D images of the scene [1, 2], reconstructing videos of unknown scenes [3], and estimating the motion and the number of hidden objects [4]. While several methods aim to recover the whole hidden scene [1, 2, 3], often in accidental scenarios [5] where no prior assumptions can be made about the scenes, recovering certain *attributes* of the scene in such accidental scenarios can be useful in certain applications. For instance, deciding whether or not a non-visible scene includes a person could be potentially useful for autonomous driving [6], or determining whether there is hazardous activity in an unknown scene would be practical for security and surveillance applications. In this thesis, we explore both categories of applications, namely, we focus on both recovering certain attributes from the hidden scene (which we refer to as *scene classification*) and recovering the entirety of it (which we refer to as *scene estimation*). In both applications, we exploit occluding objects present in the scene, called *occluders*, which improve the conditioning of the imaging problem [5, 7]. We approach these problems from a learning perspective, where we leverage large amounts of image data to achieve robust and reliable NLOS imaging systems.

In Chapter 2, we first present a summary of the NLOS imaging literature by focusing on the methods that are most related to this thesis, and discuss how we model occluder-aided methods by describing the convolutional model of occlusion. Next, we provide a brief overview on 3D morphable face models by explaining how these models are used in different domains of application including our scene classification method. Finally, we focus on the field of domain adaptation by summarizing the most relevant unsupervised approaches, one of which we employ in our scene classification method.

In Chapter 3, we focus on the scene classification problem, where we study the problem of identifying individuals in a given room by only observing shadows cast by occluding objects on a blank wall. We present a learning-based framework that discovers hidden cues in the shadows and achieves promising classification accuracies in a two-person classification task that takes place in a scene with unknown geometry and occluding objects, and show that seemingly innocuous shadows arising all around us can be used to reveal at least some biometric information.

In Chapter 4, we explore the corresponding scene estimation problem, where we describe a learning-based methodology to recover images of hidden scenes. Our simulations suggest the potential of learning-based approaches to help build better NLOS imaging systems that are robust to several changes in the scenes of interest.

Finally in Chapter 5, we conclude by summarizing our findings and discussing the potential research directions for occluder-aided NLOS imaging.

Chapter 2

Background and Related Work

2.1 Imaging Beyond Line-of-Sight

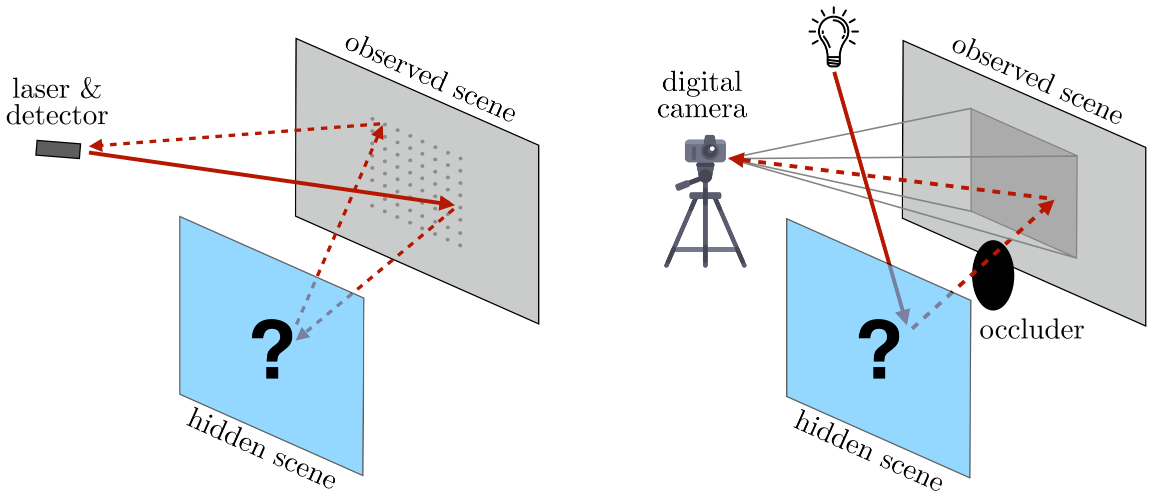
NLOS imaging has so far been explored in a variety of settings, with various scene geometries, data collection strategies, and imaging devices. In this section, we first present a short survey on NLOS imaging methods with more emphasis on occluder-aided approaches, which are the main focus of this thesis. Then, we focus on modeling the occlusion by describing the commonly used convolutional model, which we adopt in our scene estimation method presented in Chapter 4.

2.1.1 NLOS Imaging Methods

Based on how the observed data is collected, NLOS imaging methods can be divided into two categories: *active methods*, which typically involve an imaging device that consists of an coherent illumination source (such as laser) and a photon detector (such as single-photon avalanche diode), and *passive* methods, which do not require such specialized equipment and work under the ambient light from the scene. We illustrate typical configurations of active and passive methods in Figure 2-1.

Active Methods. In active imaging methods, several patches of the observed scene are illuminated so that the light pulses reflecting on these patches reach the hidden scene and are reflected back to the photon detector through the observed scene. The increasing availability of less expensive time-of-flight sensors has enabled the proliferation of active NLOS imaging methods over the last few years [8, 9, 10, 11, 12, 13, 14, 15, 16]. Due to the memory and computation requirements of active imaging systems, several methods focus on the development of faster and more accurate reconstruction algorithms under conventional scene geometries [9, 11, 13, 15], while [14] describes a novel acquisition geometry involving vertical structures in the scene, and [16] demonstrates NLOS imaging of hidden scenes over very large distances.

Passive Methods. Passive methods have been studied in a diverse set of scene geometries and imaging objectives, due to their wide applicability to different scenarios as they do not require specialized equipment. These methods typically work under



(a) active imaging (b) passive imaging

Figure 2-1: Typical scene configurations of active and passive imaging. (a) In active imaging, a coherent illumination source such as laser scans the observed scene and the light pulses respectively bounce from the observed scene, hidden scene, and observed scene again before reaching the detector. Most active methods do not rely on occluders in the scene. (b) Passive imaging methods typically work under the ambient light from the scene. The light rays reflecting from the hidden scene bounce from the observed scene and reach to an imaging device such as a digital camera. Many passive methods exploit occluders in the scene.

the ambient light from the scene, and the observed data can even be collected with an ordinary digital camera [1]. Although obtaining good reconstruction quality with passive methods is usually quite challenging as opposed to their active counterparts [7], promising results can be achieved by having some degree of control over the scene [17, 1]. In this thesis, we are primarily interested in achieving good reconstruction quality even when the control over the scenes is limited, and as observed in [5], accidental scene geometries that enable us to perform NLOS imaging arise around us more commonly than we think.

Passive NLOS imaging methods typically exploit structure present in the scenes that induces *occlusion*, and such structure has been historically used in imaging systems that use *coded apertures* [18, 19, 20, 21, 22], which rely on a known pattern of occlusion to recover the scenes of interest. These *occluders* improve the conditioning of the imaging problem [5, 7] and they have been recently exploited in several passive NLOS imaging methods [4, 17, 23, 24, 1, 2, 3, 25]. Among these methods, [4] shows that vertical occluder structure such as corners can be used to recover 1D projection of a moving scene, from which the number of people moving in the hidden scene, their sizes and speeds can be estimated. [24] and [25] extend this idea to image stationary objects and make 2D inferences about the hidden scenes, while [23] detects obstacles around the corners for autonomous driving applications..

In another line of work [17] proposes a method that infers 4D light fields of the hidden scenes from 2D shadows cast by a known occluder, even when the occluder has a

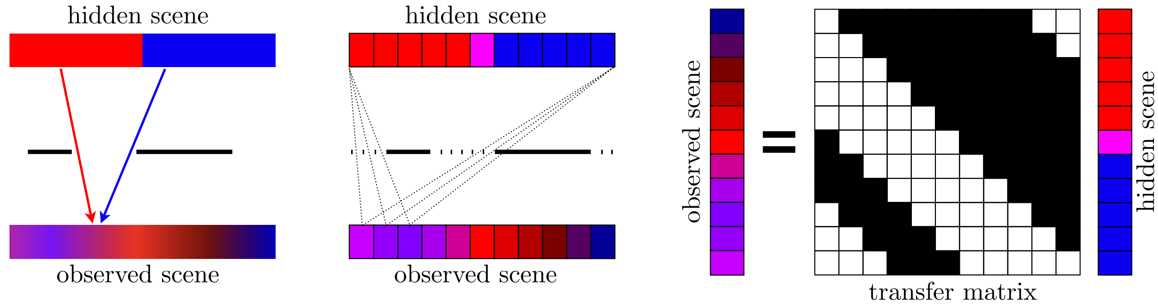


Figure 2-2: Convolutional model of occlusion. Under certain assumptions, we can model the power radiated from the observed scene as a convolution of the power radiated from the hidden scene and the occluder shape. Discretization allows us to represent this relation with a transfer matrix shown on the right. Figure adapted from [7].

complex structure. In a different setting, [1] uses a small, rectangular occluder with known shape but unknown position to recover 2D scenes, while [2] exploits motion in hidden scenes to recover the hidden scene without any assumptions about the occluder shape and position. In the latter method, however, the reconstruction quality remains limited in real-world applications. In a more unconstrained scene geometry, [3] studies the problem of recovering scenes by looking at a nearby visible region and formulates it as a matrix factorization problem. Although this method is able to reconstruct certain hidden scenes surprisingly well, it is not robust to changes in the hidden scenes and the parameterization of the neural network used in the pipeline.

2.1.2 Modeling Occluder-aided NLOS Imaging

Light Propagation Model. In this thesis as well as in preceding occluder-aided methods, the light propagation is described in terms of rays, also known as the *ray optics* or *geometrical optics model*. Under this model, the light moves in straight lines in a homogeneous medium, and it can be reflected and absorbed by the materials it interacts with [26]. Furthermore, it is commonly assumed that the ambient light sources in the scene generate light rays in random phases (also known as incoherence) and this allows us to assume that the light intensity is additive [27].

Scene Geometry. In the vast majority of the occluder-aided methods, it is assumed that the observations are made on a diffuse flat surface such as a flat wall. Under this assumption, the observations can be modeled as two-dimensional (2D) projections of three-dimensional (3D) hidden scenes onto a flat surface, which makes the problem poorly-conditioned as one 2D observation can be explained by multiple 3D hidden scenes. Therefore, it is commonly assumed that the hidden scene and the occluder lie in 2D planes that are parallel to each other as well as to the observation plane. Another common convention is to assume that the hidden scene, occluder and observed plane are sufficiently far away from each other relative to their sizes, which allows for neglecting the light attenuation over distance [2]. In particular, suppose that a point light source with intensity I illuminates a small flat surface dA with distance r

from the light source. If the angle between the incident light and the surface normal is θ, the intensity contribution of this light source to the surface is proportional to $I\,dA\,\cos(\theta)/r^2$ under the ray optics model [26]. Now suppose that the same light source is located at the origin and is incident on the plane $z = z_0$. Under the same model, the intensity contribution of the light source to a small surface patch dA located at (x, y, z_0) is proportional to $I\,dA\,z_0\,/\,(x^2 + y^2 + z_0^2)^{3/2}$ which simplifies to $I\,dA/z_0^2$ for all (x, y) such that $z_0 \gg \sqrt{x^2 + y^2}$, i.e., when the size of the scene of interest is sufficiently small compared its distance to the light source [2], the intensity contribution of the light source to any point in the scene is the same.

Convolutional Model of Occlusion. Under the light propagation model and the scene geometry we have introduced, we now show that the observations can be modeled as a 2D convolution of the hidden scene and the occluder [2], which has also been adopted in certain computer graphics applications [28, 29]. In particular, without loss of generality, assume that the hidden scene, occluder, and observed scene are all 1-dimensional (1D) and lie parallel to each other in a 2D plane as shown in Figure 2-2. Here, we denote the intensity of the hidden scene as $f(x)$, the intensity of the observed scene as $y(x)$ and the opacity of the occluder $\kappa(x)$ (the percentage of the light intensity blocked by the occluder) over space in one dimension $0 \leq x \leq L$.

Now suppose that we discretize the hidden and the observed scenes uniformly into n bins of size $\Delta = L/n$ each, and denote the centers of these bins as $x_1, x_1, \ldots, x_n$. Assuming the function f attains constant value at each bin (this is a valid assumption if the discretization is sufficiently fine), we can denote power radiated from bin i as $f_i = f(x_i) \cdot \Delta$, and similarly the measured power of the observed scene at bin i as y_i. Since we ignore the light attenuation over distance and assume that the light intensity is additive, the observed power at each bin can be written as a weighted linear combination of the radiated power from each bin of the hidden scene, where the weights are determined by the opacity of the occluder and the scene geometry. In particular, given a bin i in the hidden scene and a bin j in the observed scene, suppose that the line connecting the centers of these bins pass through the part of the occluder that has opacity $\kappa_{ij} \in [0, 1]$. In this construction, the observed power at bin j due to all bins in the hidden scene is simply $y_j = \sum_{i=1}^{n} \kappa_{ij} f_i$. Therefore, we can define a transfer matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with $\mathbf{A}_{i,j} = \kappa_{ij}$, which maps the hidden scene to the observed scene with the relation $\mathbf{y} = \mathbf{A}\mathbf{f}$ where $\mathbf{f} := [f_1, \ldots, f_n] \in \mathbb{R}^n$ and $\mathbf{y} := [y_1, \ldots, y_n] \in \mathbb{R}^n$. Under this model, we observe that an impulse in the hidden scene creates a shadow in the observed scene that exhibits a scaled and shifted pattern of the occluder determined by the scene geometry. Hence, the matrix $\mathbf{A}$ is simply a convolution operator that exhibits a Toeplitz structure as shown in Figure 2-2.

Under the convolution model, if both the hidden scene and the occluder are unknown, we can state the scene recovery problem as a *blind deconvolution* problem which is a well-studied problem for a diverse set of applications ranging from astronomical imaging to channel equalization [30, 31, 32, 33, 34, 35, 36, 37, 38]. Since the convolution operation is linear, the blind deconvolution problem is an instance of a *linear inverse problem*. Specifically, we aim to recover the scene of interest $\mathbf{f} \in \mathbb{R}^n$ from a

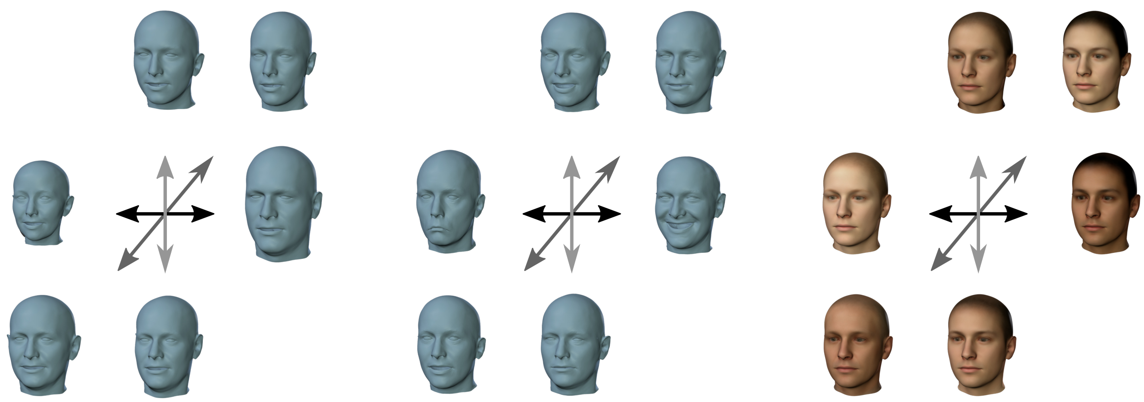


Figure 2-3: Variations in 3D shape, facial expression, and appearance for Basel Face Model 2019 [65]. Each attribute is represented by an individual PCA basis. Figure courtesy of [66].

set of measurements $\mathbf{y} \in \mathbb{R}^m$ with $\mathbf{y} = \mathbf{Af} + \mathbf{n}$ where $\mathbf{A} \in \mathbb{R}^{m \times n}$ captures the linear operation dictated by the convolution and $\mathbf{n} \in \mathbb{R}^m$ denotes the noise. Since there are infinitely many pairs of $(\mathbf{A}, \mathbf{f})$ that explain a given $\mathbf{y}$, the problem is inherently ill-posed. Traditionally, such problems are approached by imposing priors on the signals of interest to constrain the solution space. For natural images, promoting sparsity in wavelet domains or spatial gradients have been quite popular [39, 40, 41, 42, 43]. However, since these *hand-crafted* priors sometimes do not constrain the solution space sufficiently well, constructing stronger, more application-specific priors has motivated the use of data-driven approaches for popular vision problems such as superresolution [44, 45, 46], deblurring [47, 48, 49], inpainting [50, 51, 52], or for any linear inverse problem involving images [53]. Linear inverse problems can also be approached by relying on *deep generative models* [54, 55, 56] by constraining the solutions to be samples from an image distribution, which can be achieved by either estimating the distribution itself [57, 58, 59] or directly accessing samples from the distribution without explicitly constructing the distribution [60, 61, 62]. In Chapter 4 of this thesis, we will explore the latter approach by employing a conditional generative model [63, 64].

2.2 3D Morphable Face Models

3D morphable models (3DMMs) are statistical models of human faces [67, 68, 69, 65], which have been widely used in domains such as face recognition, entertainment, neuroscience and psychology for over 20 years [66]. Traditional 3DMMs were developed by constructing principal component analysis (PCA) bases of 3D shape and appearance of human faces, obtained from a collection of 3D scans. While early 3DMMs only modeled neutral faces, they were later extended to incorporate facial expressions as well [70, 65, 69], resulting in a full 3D model of human faces where 3D shape, facial

source domain target domain

Figure 2-4: An example of source and target domains for domain adaptation. Given a source domain and a target domain, domain adaptation seeks to modify the model trained with the source domain so that it performs well on the target domain. Samples from the MNIST [87] and SVHN [88] datasets for the source and target domains, respectively.

expression and appearance[1] are disentangled by design. In Figure 2-3, we illustrate the variations in 3D shape, facial expression, and appearance in one of the most commonly used 3DMMs, the Basel Face Model 2019 [65].

Since the traditional 3DMMs are linear models (based on PCA bases), they often have limited representation power, which recently has motivated the use of nonlinear 3DMMs [71, 72, 73] and deep neural networks for realistic face textures synthesis [74, 75, 76, 77, 78]. Over the last decade, advances in deep learning also allowed 3DMMs to achieve remarkable results in the challenging problem of recovering 3D faces from 2D images, commonly referred to as *3D face reconstruction* [79, 80, 81, 82], with more recent methods focusing on learning 3D face models without requiring explicit 3D shape labels [83, 84, 71, 85, 86]. Among these methods, Deng et al. [86] introduces an inverse graphics model that is trained in an end-to-end fashion. In this pipeline, a set of 3DMM parameters as well as lighting and pose parameters are estimated from a single 2D image, which are then used to render a 2D face image using a differentiable renderer. As we will elaborate in Chapter 3, we employ this reconstruction network to collect synthetic face data which we leverage in our identity classification method.

[1]Appearance is sometimes referred to as *albedo* or *texture*.

2.3 Domain adaptation

Over the last few years, there has been a significant amount of work in the area of domain adaptation [89, 90], which is the study of transferring knowledge learned from a source domain to a target domain. For example, suppose we are given a dataset of images of digits (shown as source domain in Figure 2-4), and suppose we learn a classifier from this data, which is able to identify which digit is displayed in a given test image. Since this classifier is trained on one particular dataset, we would expect it to perform poorly on a test image from another dataset (shown as target domain in Figure 2-4). The main objective of domain adaptation is to *adapt* the model learned from the source domain such that it performs well on the target domain.

Recent approaches in domain adaptation have been concentrated towards deep learning-based solutions and unsupervised methods where no labels from the target domain are used. These methods commonly rely on aligning the distributions of the source and target domains in feature spaces [91, 92, 93, 94, 95, 96, 97, 98]. Among these methods, Deep Domain Confusion [91] aims for learning domain-invariant representations by imposing a Maximum Mean Discrepancy loss [99], Deep Correlation Alignment [95] aligns the second-order statistics of the source and the target domains, while Adversarial Discriminative Domain Adaptation [97] employs an adversarial discriminator in order to make the representations of the two domains indistinguishable from each other. In another approach, Li et al. [100] shows that updating the batch normalization statistics [101] for the target domain can also be very effective, which we employ in our identity classification method presented in Chapter 3.

Chapter 3

Scene Classification

In this chapter, we focus on the scene classification problem, which we define as the problem of recovering certain attributes from the hidden scenes that are not in our direct line-of-sight. These attributes might include the number of people in the scene, speeds and sizes of the hidden objects, or 1D temporal summaries of activities around the corners as explored in the pioneering work of *corner camera* [4]. In this chapter, we introduce a novel task, namely, we study the problem of recovering the identities of people in a given room. We do this by observing shadows cast on a diffuse surface such as a blank wall, induced by the presence of an occluder.

We approach our scene classification task with a learning-based method that classifies identities by looking at images that contain shadows cast by occluding objects, where we rely on synthetically collected *labeled* data and real *unlabeled* data. In particular, we transfer what we learn from the synthetic data to the real data in a completely unsupervised way by using a domain adaption technique [100]. To minimize the domain gap between the real and synthetic domains, we employ a state-of-the-art 3D face reconstruction network [86] to obtain accurate 3D face models of the identities of interest using a single photograph of each identity. We show that our method is able to achieve surprisingly high classification accuracies in a two-person classification task.

While our work is focused on a methodology for identification from shadows, an important motivation stems from a desire to begin to understand whether otherwise benign images of shadow phenomema have the potential to leak at least some biometric information that could be of societal concern. Although it remains to be determined whether biometric cues we discover in shadows could be used to reliably distinguish large numbers of identities, these cues might potentially be used with malicious intent, e.g., to determine the presence of an individual in a room without their consent. Even if such technology do not reach the level of uniquely identifying an individual, it might reliably narrow the identity to within a group of individuals by extracting some amount of biometric information from shadows, which would still raise privacy concerns. At the same time, the extensions of our method could facilitate applications that would have positive societal impacts. For instance, such extensions would be useful in certain security and surveillance applications, or in identity recognition tasks that require no storage or observation of any sensitive information about the identities,

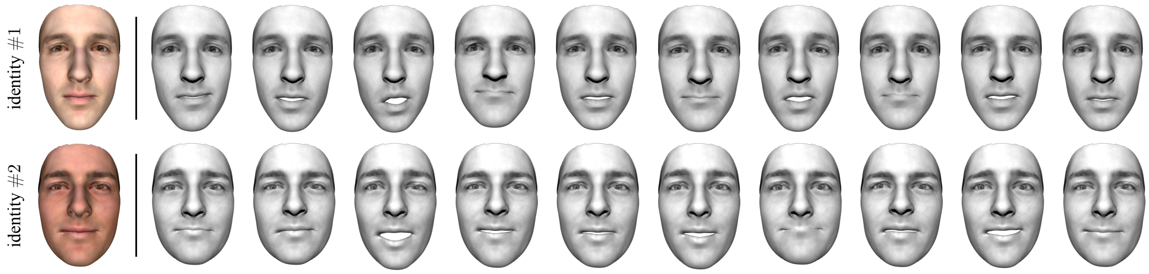


Figure 3-1: Face reconstructions of the two identities with varying expressions. Given RGB reconstructions of the faces, we first convert their textures to grayscale and match their average intensity levels. Expressions are randomly sampled and varied in the dataset.

enabling face recognition without taking any photographs of the individuals.

3.1 Overview

Suppose we are given K different identities who are individually present in a room with an unknown geometry, and suppose we observe shadows cast by an occluder in the room blocking the light reflected by each individual. Denoting each observation as $\mathbf{x} \in \mathbb{R}^{d\times d}$ (grayscale images of resolution $d \times d$) and its ground truth label as $y \in \mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$, we aim to learn a classifier given training data $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$. In this work, we restrict our attention to the case where $d = 256$ and $K = 2$, i.e., we focus on the problem of distinguishing two identities.

Since we follow a data-driven approach, representing possible variations such as occluder shape, lighting conditions, facial expressions, and head poses in the training data is crucial to achieve a robust classification system. Since collecting such data is highly impractical, we focus on a method that avoids such challenges. In particular, we use 3D graphics software to collect large amounts of training data, by placing 3D faces and objects into simulated scenes. Then, we transfer what we learn from these simulated scenes to the real settings by employing unsupervised domain adaptation.

3.2 3D Face Modeling

To minimize the discrepancy between the synthetic and real domains, we use a 3D face reconstruction network [86], which allows us to obtain a 3D model of an identity *from a single image*. The reconstructed faces in this work follow the Basel Face Model 2009 [68] with the neck and the ear regions excluded from the model, which enables us to ensure that the network trained with the synthetic data only relies on the identity information, i.e., trivial information such as the thickness of the neck or the contrast between the hair and skin intensities cannot be exploited in our method. The expression variations, on the other hand, are provided by the model constructed from the FaceWarehouse dataset [70], which we use to sample identities with varying expressions. Formally, given a number of vertices V, a face shape $\mathbf{S} \in \mathbb{R}^{3V}$ and its texture $\mathbf{T} \in \mathbb{R}^{3V}$ can be represented as

$$\begin{aligned} \mathbf{S} &= \bar{\mathbf{S}} + \mathbf{M}_{\text{id}}\boldsymbol{\alpha}_{\text{id}} + \mathbf{M}_{\text{exp}}\boldsymbol{\alpha}_{\text{exp}} \\ \mathbf{T} &= \bar{\mathbf{T}} + \mathbf{M}_{\text{tex}}\boldsymbol{\beta} \end{aligned} \tag{3.1}$$

where $\bar{\mathbf{S}} \in \mathbb{R}^{3V}$ and $\bar{\mathbf{T}} \in \mathbb{R}^{3V}$ are the mean shape and mean texture of the model; $\mathbf{M}_{\text{id}}$, $\mathbf{M}_{\text{exp}}$, $\mathbf{M}_{\text{tex}}$ are the identity, expression and texture bases; $\boldsymbol{\alpha}_{\text{id}} \in \mathbb{R}^{80}$, $\boldsymbol{\alpha}_{\text{exp}} \in \mathbb{R}^{64}$ and $\boldsymbol{\beta} \in \mathbb{R}^{80}$ are the identity, expression and texture coefficients. Here, $\bar{\mathbf{S}}, \bar{\mathbf{T}}, \mathbf{M}_{\text{id}}, \mathbf{M}_{\text{exp}}, \mathbf{M}_{\text{tex}}$ are all provided by the model whereas $\boldsymbol{\alpha}_{\text{id}}$ and $\boldsymbol{\beta}$ are provided by the face reconstruction. We create an expression variation in the dataset by sampling $\boldsymbol{\alpha}_{\text{exp}}$ from $\mathcal{N}(\mathbf{0}, 0.5\mathbf{I})$. Finally, we convert the reconstructed textures to grayscale to avoid potential reliance on color information, and scale the intensity levels of the two identities so that the average intensity of their textures are the same. We show the reconstructed faces and their grayscale versions with varying expressions in Figure 3-1.

3.3 Scene Geometry and Datasets

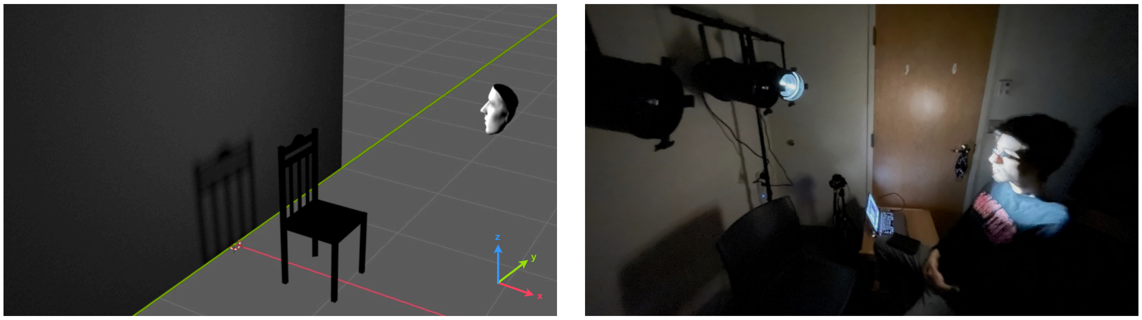
Our imaging configuration includes the following: a person whose identity is unknown, a light source that illuminates the face of this person, a blank wall where we make our observations, and an occluding object which creates shadows on this wall. In this work, for the purposes of illustration, we limit our attention to *chairs* as occluding objects, as they are one of the most common and diverse classes of indoor objects. We note that, however, our method can easily be extended to handle more classes of objects by incorporating them in the training set.

In our synthetic data collection, we use 3D chair models provided by ShapeNet [102], we use a white planar object as a wall, and a white spotlight as an illumination source. When we render these scenes, we cover as much variation as possible by changing the pose, position and expression of the faces and vary the illumination conditions by changing the position of the light sources, which we will elaborate on in the next section. A representative synthetic scene is shown in Figure 3-2a, where we also illustrate our coordinate convention.

In our real data collection, the two identities sit across a blank wall individually, where a chair is positioned between the identity and the wall. The identities are illuminated by spotlights in different positions while the expressions and poses of the subjects as well as the pose of the chair are varied during the data collection. We performed these experiments in a physical space shown in Figure 3-2b.

3.4 Domain Adaptation

Given two sets of data $\mathcal{S} = \{(\mathbf{x}_1^s, y_1^s), \dots, (\mathbf{x}_N^s, y_N^s)\}$ and $\mathcal{T} = \{(\mathbf{x}_1^t, y_1^t), \dots, (\mathbf{x}_N^t, y_N^t)\}$, which represent the source data and the target data, respectively, our objective is to



(a) Synthetic scene geometry (b) Real scene geometry

Figure 3-2: Scene geometries for synthetic and real settings. Both scenes consist of four main components: a person whose identity is unknown, an illumination source, a blank wall, and an occluding object that creates the shadows on the wall.

learn a classifier using the source data $\mathcal{S}$ such that it performs well on the target data $\mathcal{T}$. This can be achieved in a supervised manner by using very few labeled samples from $\mathcal{T}$, or in an unsupervised manner by using no labeled samples from $\mathcal{T}$. In this work we follow the latter, as we seek to ensure that the supervision signals coming from the target domain involves only identity information, i.e., these signals may depend on unintended cues from the real-world settings such as clothing, reflectance of the hair or other unintended phenomena.

Our method involves training a classification network that follows the ResNet-18 architecture [103], where we change the final classification layer so that it reflects the number of classes in our application. Initializing the feature extraction module with the pretrained weights, we first train the network on the synthetic data in a supervised manner. Then, we freeze the learned weights and update the running means and variances of each batch normalization layer in the network [100] by feeding the unlabeled target data $\mathcal{T} = \{\mathbf{x}_1^t, \dots, \mathbf{x}_N^t\}$ through the network. As we will show, the updated network generalizes reasonably well to the test samples from the target domain.

3.5 Experiments and Results

In this section, we describe our experiments in detail by elaborating on the collection of real and synthetic data, and provide classification accuracies obtained in different stages of our method.

3.5.1 Synthetic Data Collection and Training

We generate our synthetic data randomly, where we vary the pose, expression and the position of the face, the location of the light source, and the occluder shape. According to the coordinate definition shown in Figure 3-2a, we have the following configurations and variations in the dataset. Here, with a slight abuse of notation, we denote a point

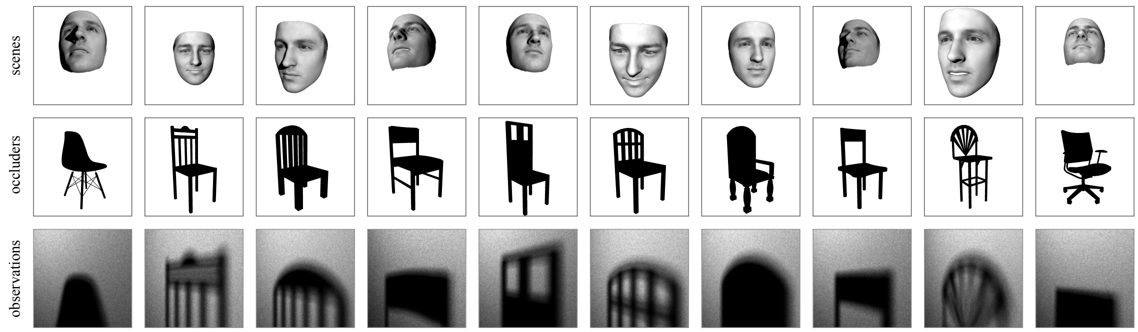
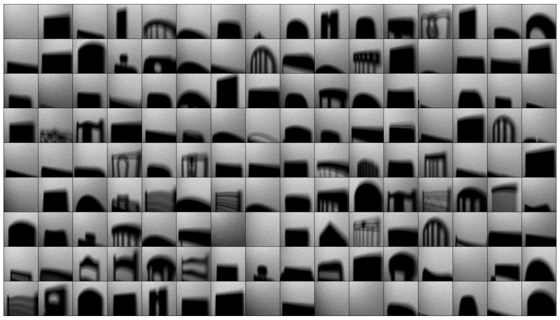


Figure 3-3: Representative samples from the dataset, where each column shows one sample. Our dataset covers a diverse set of head poses, facial expressions, and occluder shapes.

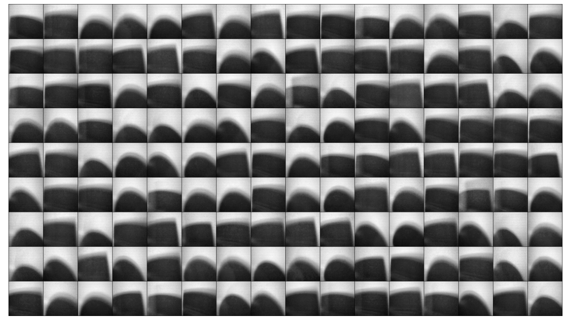
in 3D by (x, y, z) where the unit of measure is meters. We illustrate the variations in the dataset in Figure 3-3 by showing representative samples from the dataset.

- We vary the facial expressions by sampling the expression coefficients from $\mathcal{N}(\mathbf{0}, 0.5\mathbf{I})$, which changes the face shape according to Equation 3.1.
- We rotate the faces around y– and z–axes, which we refer to as elevation and azimuth. We sample both elevation and azimuth uniformly from $[-30, 30]$ degrees, where zero rotation means that the face is directly positioned towards the wall as shown in Figure 3-2a. Positive angles indicate clockwise rotations with respect to the xz– and xy–planes.
- We sample the position of the face uniformly along the line connecting $(1.55, 0.0, 1.15)$ and $(1.75, 0.0, 1.15)$, i.e., face position varies along the x–axis as variations in other axes are accounted for in the data augmentation step where the final images are randomly cropped.
- We use a white spotlight with a beamwidth of 15 degrees, directed to the face. We sample its location uniformly along the line connecting $(0.15, -1.0, 1.50)$ and $(0.15, 1.0, 1.50)$.
- Occluders are located 0.7 meters from the wall and situated on the ground, where we measure the distance from the center of mass of the occluder. We also render all occluders with black texture to eliminate the effect of the light bouncing off the occluder.

We collect our synthetic data using Mitsuba2 [104], with which we render 256×256 images of the observed wall using 50 000 samples per pixel. Rendering one image takes approximately 50 seconds on an NVIDIA GeForce RTX 2080 Ti GPU, and all images are normalized to $[0, 1]$ range after rendering. For each identity, we collect 4000 images which we split into train and test sets with $75\% - 25\%$ split, which gives us 6000 train and 2000 test samples. We illustrate random samples from the synthetic dataset in Figure 3-4a.



(a) Source images



(b) Target images

Figure 3-4: Random images from the source and the target datasets.

We train our classification network with the synthetic data for 30 epochs, using binary cross entropy loss and Adam optimizer [105] with a learning rate of 0.0001. We augment the training data by flipping the images randomly, resizing them to 280×280 resolution and randomly cropping a 224×224 patch from these images. At test time, we resize the images to 280×280 resolution and center-crop the 224×224 patch from them. In our experiments, we pick the epoch with the highest test accuracy, and use the network at that epoch as our baseline, on which we apply domain adaptation.

3.5.2 Real Data Collection and Domain Adaptation

To represent the typical use cases, we deliberately cover fewer variations in our real data compared to the synthetic data. In particular, we experiment with 4 light source locations by using 4 separate spotlights, and 2 different occluders which we repose in 5 different angles to increase the diversity in the dataset. Similar to what we have in the synthetic dataset, the identities also change their head poses and facial expressions while the data is collected. We collect 4000 samples for each identity, and we randomly split the whole dataset into train and test sets with $75\% - 25\%$ split. We illustrate random samples from the real dataset in Figure 3-4b.

We illustrate our results in Figure 3-5 where we visualize the feature distributions of the test samples before and after domain adaptation using t-SNE [106], where we extract these features from the final layer before classification. Before the domain adaptation (shown in the first row), we observe that the network trained on the source data produces two feature clusters for the source and the target domains. Furthermore, the ground truth labels of the source samples seem to be well-separated which allows the network to achieve a classification accuracy of 75.80% on the source domain, as illustrated in the predictions plot. Since the network has not seen any target samples before the domain adaptation, it performs poorly on the target domain, achieving 62.70% accuracy. After the domain adaptation (shown in the second row), we observe that the feature distributions of the source and the target data are well-aligned, and the ground truth labels for both domains seem to be well-separated which allows

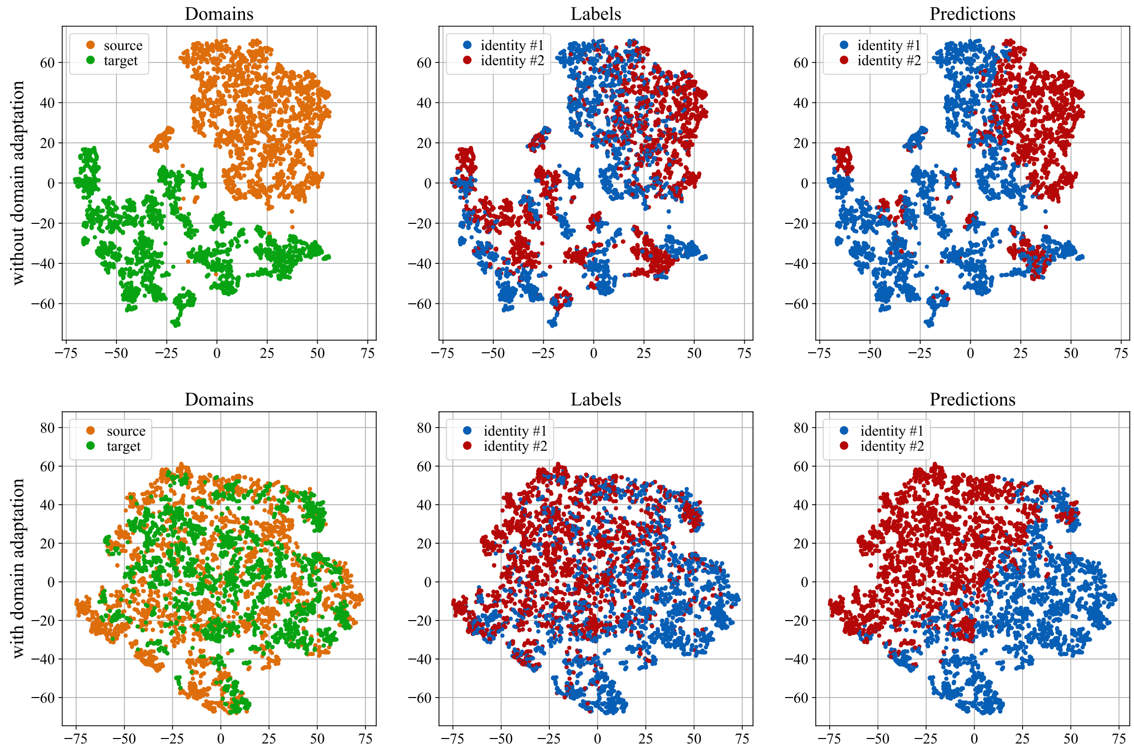


Figure 3-5: Summary of results. We illustrate feature distributions of the test data in 2D using t-SNE dimensionality reduction technique [106]. Feature distributions of source and target domains before domain adaptation are shown in the first row, where we observe that the network performs well on the source domain but not on the target domain. In the second row, feature distributions after domain adaptation are shown, which reflect that the network generalizes well to the target data as well.

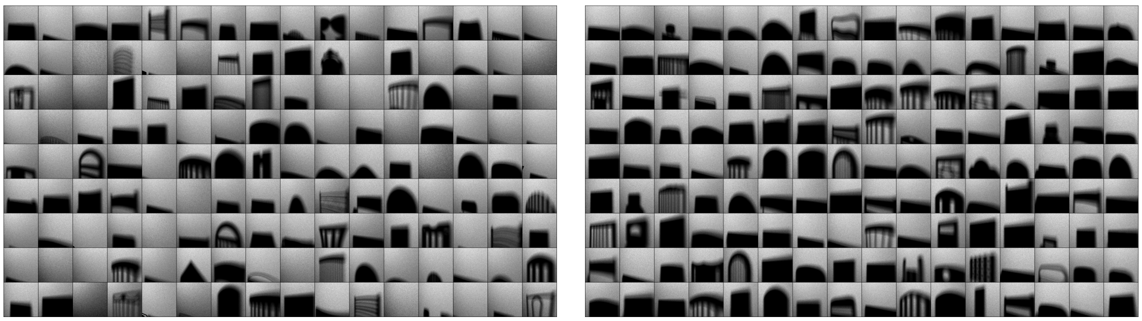
the network to achieve a classification accuracy of 76.35% on the target domain, as illustrated in the predictions plot. We also report average classification accuracies in Table 3.1 computed over 20 independent experiments using the same datasets, network architecture and hyperparameters.

3.6 Discussion and Analysis

In this section, we provide a detailed interpretation of our results where we seek to explain the behavior of our method in various scene configurations. To achieve this, we analyze our results on the synthetic images for which have access to the conditions under which they are rendered such as occluder shape, head pose, and light source location. In particular, we analyze the samples on which the network fails or performs well, and the regions of the input that the network relies on the most by using interpretable machine learning tools referred to as saliency methods.

Table 3.1: Average classification accuracies over 20 independent experiments. We report test accuracies on the source domain and on the target domain before and after adaptation.

source	**target (before adapt.)**	**target (after adapt.)**
74.57 ± 0.84	59.67 ± 9.26	77.08 ± 2.42



(a) Incorrectly classified images (b) Correctly classified images

Figure 3-6: Random samples from incorrectly and correctly classified images. We observe that incorrectly classified images usually lack shadows (hence penumbrae) where most useful information lies. In contrast, correctly classified images usually have large shadow areas.

In our first set of analyses, we investigate the influence of occluder shape and face appearance on the classification performance, where we compare all 484 fail cases (which gives us an accuracy of 75.80% on the source domain) with 484 of the correctly classified images with the highest softmax probabilities. For the occluder shape analysis, we illustrate random samples from the incorrectly and correctly classified images in Figure 3-6 where we observe that the incorrectly classified images usually lack shadows. In particular, defining black pixels (with zero intensity) in each image as *umbra*, the umbrae cover 12.11% of the incorrectly classified images on average, whereas they cover 21.95% of the correctly classified images.

The fact that the shadows appear to be crucial for inferring identities is consistent with the analysis of the resolving power of *single edge occluders* which are widely explored in the last few years [4, 24, 25]. In our case, we use the resolving power of the *edges of the occluder*, where the penumbra formed on the wall can be used to calculate 1D projections of the input face along the direction of the edges. In other words, our results suggest that the penumbrae contain the most useful information about the unknown scenes, and they are in fact where our network appears to rely on the most, as we will show in the saliency map analysis.

We now investigate the effect of the face appearance on the results by analyzing the impact of the head pose and light source location on the predictions. We illustrate our findings in Figure 3-7, where we show elevation-azimuth and light source position-azimuth plots for correctly and incorrectly classified examples. In the first plot, we

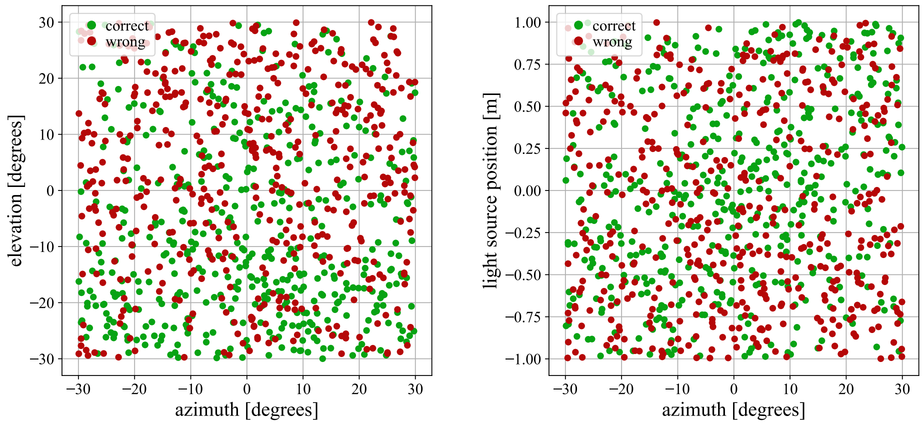


Figure 3-7: Correctly and incorrectly classified examples depending on azimuth, elevation and light source position. We observe that faces with lower elevations and fewer cast shadows are more likely to be classified correctly.

observe that the elevation has an evident impact on classification performance, where faces with higher elevation are more likely to be misclassified. This can be explained by our scene geometry shown in Figure 3-2a, where a more direct view of the face is reflected on the wall when the elevation is low, which makes the problem less challenging. Taking the averages over all samples shown in the plot, incorrectly classified examples have an average elevation of $+2.83$ degrees whereas correctly classified examples have an average of -6.73 degrees. In the second plot, we observe a positive correlation (a Pearson correlation of 0.22) between the light source position (measured along the y-axis) and the azimuth for correctly classified examples, for which the faces are illuminated with lower incidence angles. This means that the faces with fewer cast shadows are more likely to be predicted correctly, e.g., strong shadows cast by the nose on the cheek make the classification task more challenging.

Finally, we investigate which regions of the input images have more influence on the class predictions by employing a saliency method referred to as *integrated gradients* [107]. We illustrate several examples in Figure 3-8, where we show the original inputs and the image attributions for each input. We observe that the network is more sensitive to the penumbra regions compared to other parts of the image, which is in line with our previous observation that the penumbrae contain the most information about the identities.

3.7 Conclusion and Future Work

We show that it is possible to reliably identify individuals by looking at the shadows induced by their presence. We approach this problem as a domain adaptation problem, where we transfer what we learn from the synthetic data to the real data without using any labeled real data. Our synthetic data acquisition relies on a 3D face reconstruction

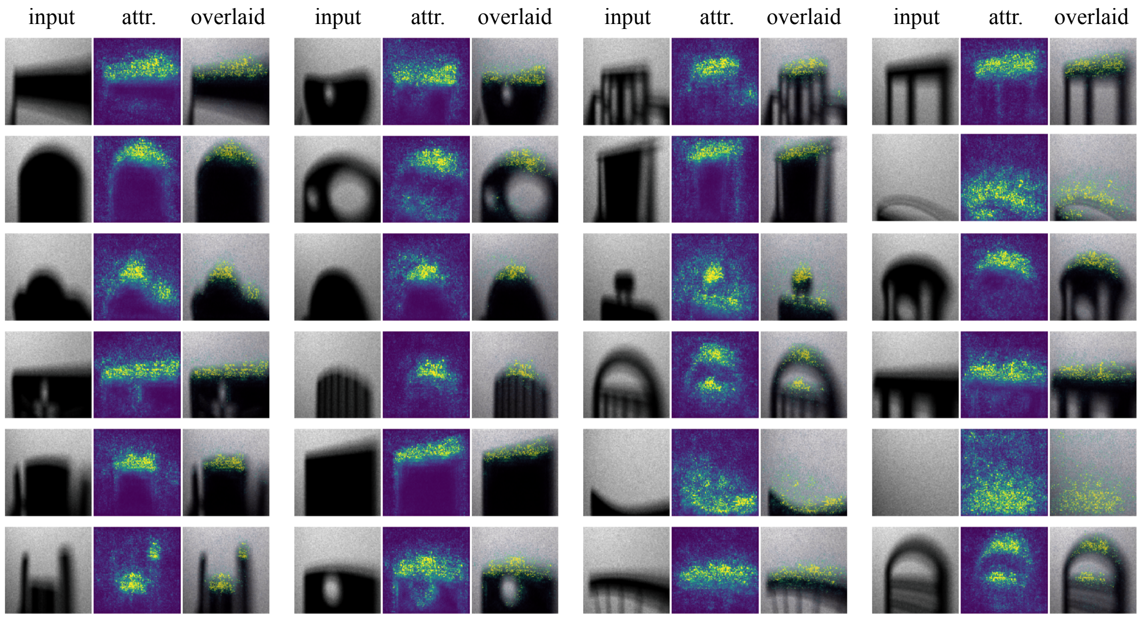


Figure 3-8: Image attributions extracted by the Integrated Gradients [107]. We observe that the network is mostly sensitive to the penumbra regions where most information lies.

network with which we obtain accurate 3D models of faces from only a single photo of each identity. We demonstrate that our method achieves surprisingly high classification accuracies in the real domain and is robust to several variations in the scene, such as occluder shape, lighting, head pose and facial expressions. Our results suggest that our network is sensitive to the penumbra portions of the shadows, which we explain with the resolving power of the occluding edges. Such shadows arise all around us in various scene geometries and we demonstrate the potential of turning these shadows into physical signatures. Although it remains to be seen whether our method could be extended to distinguish large numbers of identities, work under more extreme lighting conditions, or handle different classes of occluders, our results suggest that at least some biometric information is revealed by such shadows.

Chapter 4

Scene Estimation

We now focus on the problem of *scene estimation*, where we aim to recover the entire hidden scene based only on the measurements from the observed scene. To study this problem, we first summarize two of the more recent passive NLOS imaging methods, namely the blind scene recovery method in [2], and computational mirrors [3]. Based on the limitations of these methods, we then discuss how learning-based approaches could allow us to achieve better scene reconstruction quality and robustness to changes in the scene such as the occluder structure and the hidden scene content, which we support with our simulations.

4.1 Motivation

Blind scene recovery [2]. In [2], an occluder-aided NLOS imaging method is introduced that makes no prior assumptions about the occluder such as its shape and position. In this method, the hidden scene, occluder, and observed scene are assumed to lie in 2D planes that are parallel to each other, which, with an additional set of physical assumptions, gives rise to the convolutional model as described in Chapter 2. With this model, this scene reconstruction problem can be formulated as a blind deconvolution problem, which we previously established as an ill-posed linear inverse problem. Furthermore, if we constrain our scenes to include common indoor objects, many accidental scene geometries cause the effective kernel size induced by the occluder to be much larger than what the vast majority of image deblurring methods deal with [47, 48, 49]. This makes the blind occluder-aided scene recovery extremely challenging, and applying state-of-the-art image deblurring methods directly to NLOS imaging scenarios generally yield unsatisfactory results.

The blind scene recovery method in [2] consists of two steps. In the first step, the occluder shape is estimated from a video of the observed scene using an occluder recovery algorithm, which assumes that the hidden scene is slowly moving. Under this assumption, the differences between two consecutive time instances of the hidden scene are likely to be sparse signals, consisting of a superposition of impulses. Therefore, the difference frames of the observed video are likely to manifest a superposition of the shifted versions of the occluder shape, due to the linearity of convolution. The

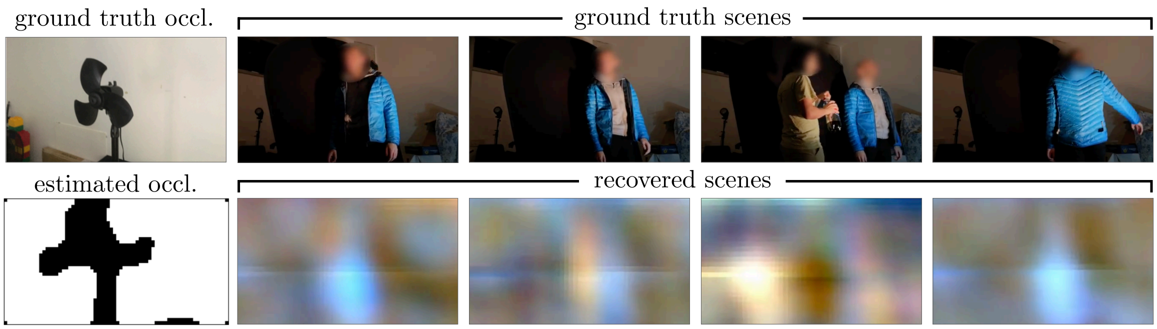


Figure 4-1: Occluder estimation and scene recovery results from [2]. Figure adapted from [2].

algorithm then simply registers these shifted versions of the occluder which yields the final estimate. In the second step, based on the estimated occluder, the hidden scene is recovered via the linear least squares estimator with Tikhonov regularization.

We illustrate a representative result obtained with [2] in Figure 4-1 where we observe a very limited reconstruction quality, which we attribute to two main factors. First, the occluder recovery relies on motions in the hidden scene and works under the assumption that the difference frames are sparse signals, which does not always happen in practice. Second, the hidden scene estimation is merely a least squares solution which does not promote natural image features sufficiently well. To circumvent these limitations, we: 1) explore a learning-based occluder estimation method that automatically recovers occluder shape from an observed video; and 2) develop a learning-based hidden scene estimation method that captures stronger image priors in a deep network.

Computational mirrors [3]. If the observed scene is not a flat surface, but rather an arbitrary scene, the hidden scene recovery problem can still be formulated as a linear inverse problem as demonstrated in [3], where the light transport matrix is not necessarily constrained to follow Toeplitz structure. In this more general scene geometry, illustrated in Figure 4-2, the observed scene $\mathbf{Z}$ can be written as a matrix product of the hidden scene $\mathbf{L}$ and the light transport matrix that defines a mapping from the hidden scene to the observed scene determined by the scene geometry and the objects in the scene. Therefore, the main objective in this method is to factorize the observed scene into the hidden scene and the light transport matrix.

The matrix factorization problem formulated in [3] is solved by parameterizing the hidden scene and the light transport matrix by two separate convolutional neural networks, which has been showed to impose natural image features and has been applied to several tasks in computer vision such as image denoising, superresolution, and inpainting [108]. The matrix factorization is then achieved by optimizing these two networks so that the product of their output gives the observed video, which yields the results shown in Figure 4-3. In these results, although we observe a promising reconstruction quality in two of the scenes (shown in the first row), the performance degrades notably when the scene is more complex (shown in the second row). Furthermore, the optimization stability and reconstruction quality of this method

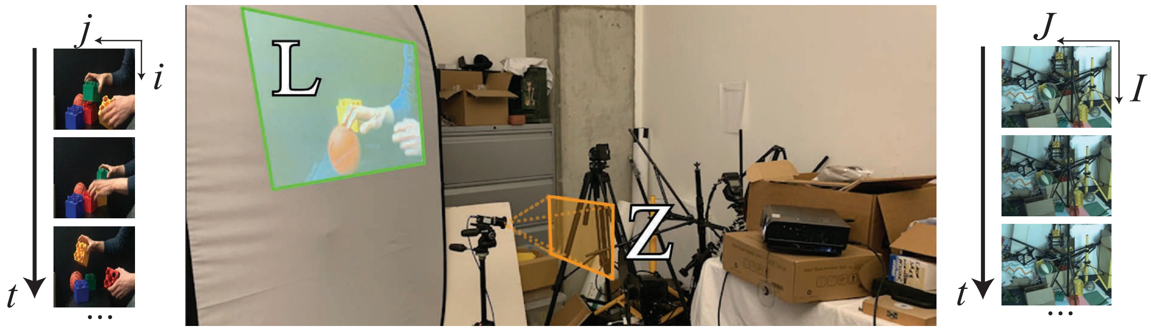


Figure 4-2: Scene geometry adopted in computational mirrors [3]. The observed scene can be written as a product of the hidden scene and the light transport. Figure courtesy of [3].

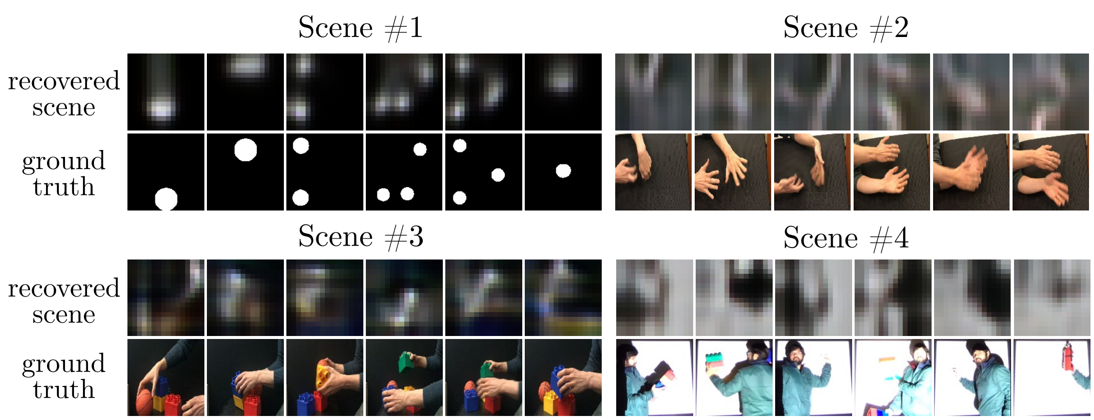


Figure 4-3: Scene estimation results of computational mirrors [3]. Although we observe a promising reconstruction quality in Scene #1 and #2, the performance degrades notably when the scene is more complex in Scene #3 and #4. Figure adapted from [3].

are also observed to be sensitive to several factors such as the network architecture, activation functions used, hyperparameter choices, and loss functions [3]. This further motivates us to develop a learning-based method that generalizes well over different hidden scenes and occluders, and is robust to the changes in the environment.

4.2 Learning-based Blind Scene Recovery

In this section, we develop an alternative approach to blind scene recovery. We begin by formally defining our objective and introducing the dataset we use in our simulations. We then describe a learning-based methodology for blind deconvolution. Finally, we present our preliminary results.

4.2.1 Problem Formulation

In the sequel, we assume that the convolution model is valid, i.e., the observed scene can be modeled as a convolution of the hidden scene and the occluder shape as

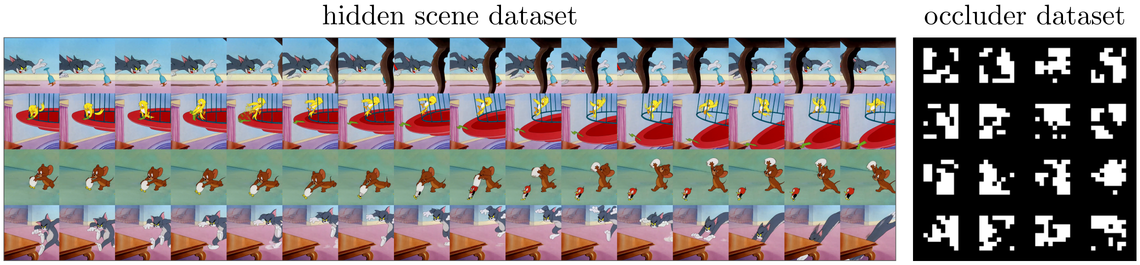


Figure 4-4: Representative samples from the dataset. We illustrate 4 videos and 16 occluders from the train set. Sizes of the frames and occluders are not drawn to scale.

described in Chapter 2. Furthermore, since the occluders are physical objects with opacity in the range $[0, 1]$, the observed scene is always a low-pass filtered version of the hidden scene. Therefore, the low-frequency content of the hidden scene is maintained in the observed scene, which motivates us to make our hidden scene estimations conditioned on the observed scene.

In our scene recovery method, we follow a two-step approach similar to what is proposed in [2]. First, we estimate the occluder shape based on a set of observed video frames, which is carried out by a standard learning-based approach. Then, we use the estimated occluder to recover the hidden scene, using a deep generative model trained with generative adversarial network (GAN) framework [60]. As mentioned in Chapter 2, we use these models to capture natural image statistics inside a neural network, which we optimize using a large amount of data.

Our dataset consists of the following: 1) a set of short videos consisting of 16 frames, with 24 frames-per-second frame rate and 256×256 resolution; 2) a set of randomly generated occluders with 64×64 resolution. We acquire our video data from two episodes of a cartoon that are publicly available (each containing ~ 25000 frames), and we construct the occluders such that they consist of 8×8 grid of binary-valued subblocks of size 5×5 pixels as shown in Figure 4-4. In this dataset, we obtain train samples from one episode while we obtain the test samples from the other one, so that there is no overlap between train and test datasets. To generate the occluders, we sample 64-dimensional binary vectors consisting of 32 ones and 32 zeros, which we reshape into an 8×8 grid and discard the samples that are not sufficiently smooth, which we determine by calculating the total variation of the samples. We generate 10000 occluders and use half of them for the train set and the other half for the test set.

In our scene estimation method, we assume that the hidden scene is dynamic, i.e., it changes over time whereas the occluder is static. Suppose we are given N RGB videos of resolution $R \times R$ and fixed length T each, and N occluders sampled from the training set, where we denote each video as $\mathbf{f}^i := [\mathbf{f}_1^i, \mathbf{f}_2^i, \ldots, \mathbf{f}_T^i] \in \mathcal{F}$ and each occluder as $\mathbf{k}^i \in \mathcal{K}$. Under the convolution model, we denote the observed video corresponding to these samples as $\mathbf{y}^i := [\mathbf{f}_1^i * \mathbf{k}^i, \mathbf{f}_2^i * \mathbf{k}^i, \ldots, \mathbf{f}_T^i * \mathbf{k}^i] \in \mathcal{Y}$. Our aim is then to learn a function that estimates the hidden scenes $\mathbf{f}^i$ from observed scenes $\mathbf{y}^i$. As mentioned previously, we approach this problem with a two-step method where we first estimate the occluder and then *deconvolve* the observed video with this occluder.

4.2.2 Occluder Estimation

As observed in [2], a slowly moving scene will have sparse difference frames $|\mathbf{f}_t^i - \mathbf{f}_{t-1}^i|$, and the difference observation frames $|\mathbf{y}_t^i - \mathbf{y}_{t-1}^i| = |\mathbf{f}_t^i * \mathbf{k}^i - \mathbf{f}_{t-1}^i * \mathbf{k}^i| = |\mathbf{f}_t^i - \mathbf{f}_{t-1}^i| * \mathbf{k}^i$ will manifest a superposition of the shifted versions of the occluder. Our occluder estimation method also builds on the idea that several observations in the presence of a static occluder should lead to a robust estimate of the occluder, but it does not necessarily require sparse difference frames.

In our method, we assume that each occluder in the dataset has a lower dimensional representation in some latent space $\mathcal{W}$. An obvious latent representation for the occluders in our dataset is the 64-dimensional binary vector used to generate these occluders as explained in Section 4.2.1, although different latent spaces can also be constructed or learned from the data. Assuming each occluder $\mathbf{k}^i$ in the dataset has a latent code $\mathbf{w}^i \in \{0, 1\}^{64}$, our objective is to learn an estimator $E : \mathcal{Y} \to \mathcal{W}$ that correctly estimates the latent codes of the occluders from the observed videos. Formally, given pairs of observed videos and latent codes $\{(\mathbf{y}^i, \mathbf{w}^i)\}_{i=1}^N$, we aim to solve the following optimization problem:

$$\arg\min_E \mathbb{E}_{\mathbf{y},\mathbf{w}} \left[\, \|E(\mathbf{y}) - \mathbf{w}\|_1 \, \right] \tag{4.1}$$

where E is a neural network that follows a ResNet-18 architecture [103], with the last layer reflecting the dimensionality of the latent codes. We minimize the above objective using stochastic gradient descent [105] with a minibatch size of 8 and a learning rate of 0.0001 for 2500 epochs. During inference time, we feed observed videos from the test dataset to the trained estimator E and threshold the output to obtain the estimated latent code of the occluder.

4.2.3 Non-Blind Deconvolution

Once the latent code of the occluder is estimated, we perform *non-blind* deconvolution on each frame of the observed video individually, which we achieve by training a deep generative model that is conditioned on both the observed frames and the estimated latent codes. In particular, our objective is to learn a generator $G : \mathcal{Y}_1 \times \mathcal{W} \times \mathcal{Z} \to \mathcal{F}_1$, where $\mathcal{W}$ denotes the set of occluder latent codes, $\mathcal{Z}$ denotes the set of noise vectors that inject stochasticity into the model, $\mathcal{F}_1$ and $\mathcal{Y}_1$ denote the sets of hidden and observed *frames*, respectively.

We train G with the generative adversarial network (GAN) framework [60, 63, 64] in which a generator G and an adversarial discriminator D play a two-player zero-sum game. In particular, the generator learns how to produce *fake* samples that fool the discriminator, which at the same time learns how to distinguish fake samples from *real* samples coming from the dataset. We define the GAN loss as

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{\mathbf{y}_1,\mathbf{w},\mathbf{f}_1}[\log D(\mathbf{y}_1, \mathbf{w}, \mathbf{f}_1)] + \mathbb{E}_{\mathbf{y}_1,\mathbf{w},\mathbf{z}}[\log(1 - D(\mathbf{y}_1, \mathbf{w}, G(\mathbf{y}_1, \mathbf{w}, \mathbf{z})))] \tag{4.2}$$

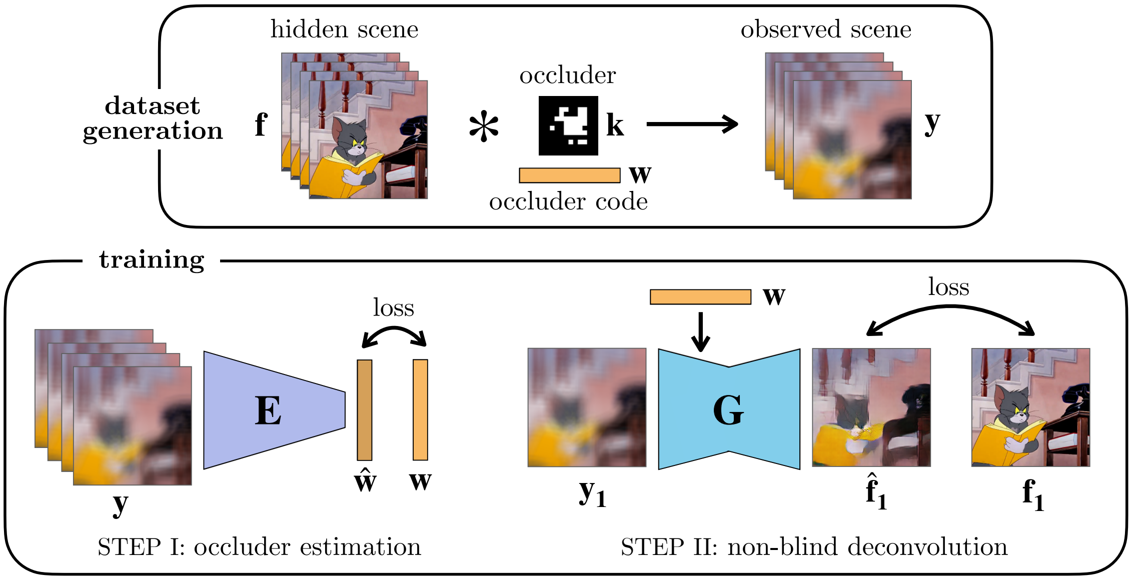


Figure 4-5: Summary of our scene estimation pipeline. Given a set of hidden scenes and occluders, we obtain the observed scenes by simply convolving them. During training, we estimate the occluders from observed videos using the estimator E, and the observed frames are deconvolved with the ground truth occluders using the conditional generator G. Both E and G are optimized with the loss functions indicated in Equations 4.1 and 4.5. At inference time, after estimating the occluder latent with E, we deconvolve each frame of the observed video with the estimated occluder using G.

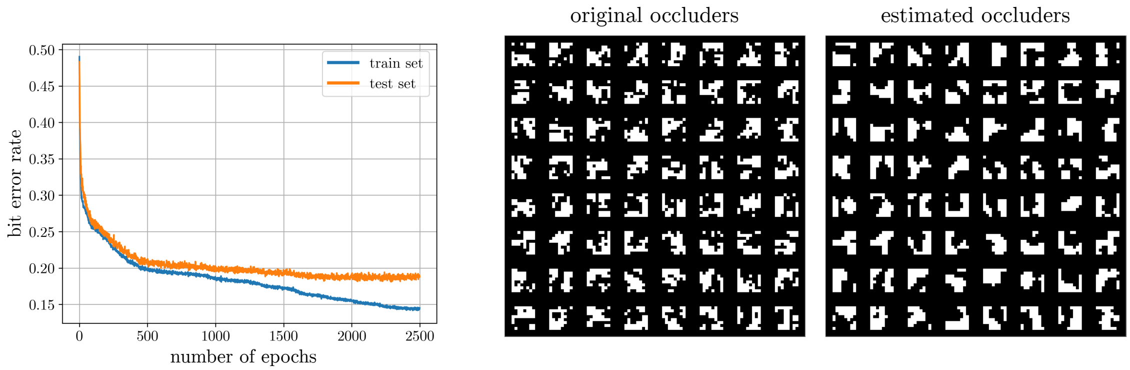
which is minimized over G and maximized over D. Here, $\mathbf{w} \in \mathcal{W}$ denotes the occluder latent code, $\mathbf{z} \in \mathcal{Z}$ denotes the random noise vector sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{f}_1 \in \mathcal{F}_1$ and $\mathbf{y}_1 \in \mathcal{Y}_1$ denote the hidden and observed video frames, respectively. Furthemore, we combine the GAN loss with two additional losses: perceptual loss [109], which measures the perceptual similarity between two images, and feature matching loss [110], which improves the training stability. Suppose we are given N tuples of observed frames, their occluder latents, and the ground truth hidden frames $\{\mathbf{y}_1^i, \mathbf{w}^i, \mathbf{f}_1^i\}_{i=1}^N$. We impose perceptual similarity between the ground truth hidden scene $\mathbf{f}_1$ and the estimated hidden scene $G(\mathbf{y}_1, \mathbf{w}, \mathbf{z})$ as follows:

$$\mathcal{L}_{\mathrm{P}}(G) = \mathbb{E}_{\mathbf{y}_1,\mathbf{w},\mathbf{f}_1,\mathbf{z}} \sum_{i \in \mathcal{I}_{\mathrm{P}}} \lambda_{\mathrm{P}}^{(i)} \left[\left\| V_i(\mathbf{f}_1) - V_i\big(G(\mathbf{y}_1, \mathbf{w}, \mathbf{z})\big) \right\|_1 \right] \tag{4.3}$$

where V_i denotes the ith layer feature extractor of a pretrained VGG-19 network [111], $\lambda_{\mathrm{P}}^{(i)}$ denotes the associated weighting factors, and $\mathcal{I}_{\mathrm{P}}$ denotes the index set of the feature extracted layers. Similarly, we match the features of the real and fake images extracted from multiple layers of the discriminator as follows:

$$\mathcal{L}_{\mathrm{FM}}(G, D) = \mathbb{E}_{\mathbf{y}_1,\mathbf{w},\mathbf{f}_1,\mathbf{z}} \sum_{i \in \mathcal{I}_{\mathrm{FM}}} \lambda_{\mathrm{FM}}^{(i)} \left[\| D_i(\mathbf{y}_1, \mathbf{w}, \mathbf{f}_1) - D_i(\mathbf{y}_1, \mathbf{w}, G(\mathbf{y}_1, \mathbf{w}, \mathbf{z})) \|_1 \right] \tag{4.4}$$

where D_i denotes the ith layer feature extractor of the discriminator D, $\lambda_{\mathrm{FM}}^{(i)}$ denotes



(a) (b)

Figure 4-6: Occluder estimation results. (a) Average BER after each epoch. We observe that the test BER plateaus slightly below 20%, reaching a minimum of 18.12%. (b) Occluder estimations on random pairs of observed videos and occluders from the test data.

the associated weighting factors, and $\mathcal{I}_{\text{FM}}$ denotes the index set of the feature extracted layers. Our final objective is to solve the following minimax problem:

$$\arg\min_G \left[\left[\max_D \mathcal{L}_{\text{GAN}}(G, D) \right] + \lambda \left[\mathcal{L}_{\text{P}}(G) + \mathcal{L}_{\text{FM}}(G, D) \right] \right] \tag{4.5}$$

where λ denotes the weighting factor for perceptual and feature matching losses. In our model, we use a U-Net architecture [112] for the generator and a patch-based fully convolutional network [64] for the discriminator, which we train in an alternating fashion using stochastic gradient descent [105] with a minibatch size of 16 and a learning rate of 0.0005 for both the generator and the discriminator for 50 epochs. We illustrate the two steps of our method in Figure 4-5.

4.3 Preliminary Results

We now present our preliminary results on the scene estimation problem. In particular, we first evaluate our occluder estimation network on the test data and provide both qualitative and quantitative results. Next, we present our non-blind deconvolution results on a set of observed videos with different choices of occluders.

Occluder estimation. Since each occluder $\mathbf{k}^i$ in our dataset has a latent representation $\mathbf{w}^i \in \{0, 1\}^{64}$ as a 64-dimensional binary vector, it is natural to adopt *bit error rate (BER)* as the error metric for our occluder estimation method, which is defined as the number of bit errors divided by the number of bits in a given code. We illustrate the average BER after each epoch in Figure 4-6(a), where we observe that the test BER plateaus slightly below 20%, reaching a minimum of 18.12%. In Figure 4-6(b), we show occluder estimations on random pairs of observed videos and occluders from the test data, which validates our quantitative evalutation. Finally, we illustrate the full pipeline of our occluder estimation method by presenting 4 test videos paired with random test occluders in Figure 4-7.

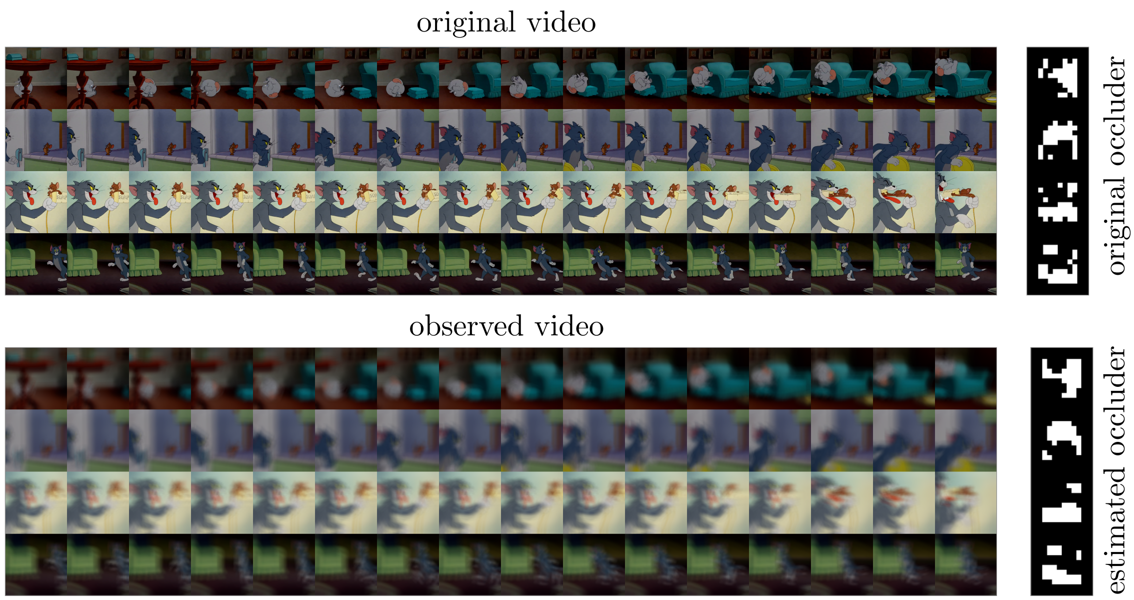


Figure 4-7: Full pipeline of our occluder estimation method. 4 test videos paired with random test occluders are illustrated. Each row shows a video and its corresponding occluder.

Non-blind deconvolution. We deconvolve the 4 test observed videos shown in Figure 4-7 with 3 sets of occluders as illustrated in Figure 4-8. In the first row, we deconvolve each frame with the original occluder, which shows the baseline performance for our method. In the second row, we use the estimated occluders to deconvolve the videos, and we observe only a slight degradation in the estimated videos compared to the baseline. Finally in the third row, we deconvolve the observed videos with randomly selected occluders which yields poor reconstruction quality since the occluder shapes do not match with the original occluders well.

Our occluder estimation results suggest that in the presence of a static occluder, multiple observations of the visible scene can be used to reliably estimate the occluder shape, even when we have a small number of observations. This further suggests that a more reliable and robust occluder estimation method can be achieved by simply increasing the number of observations, which might be practical in certain imaging settings, e.g., when a 5 minute video of a slowly moving visible scene can be collected by a camera with reasonable frame rate. Our scene reconstrucions, on the other hand, show that the generator indeed uses the occluder information to deconvolve the images, and that the reconstruction quality only slightly suffers from using the estimated occluders. This suggests that our blind scene recovery method is robust to imperfect estimations of the occluder shape.

4.4 Conclusion and Future Work

In this chapter, we have introduced a novel learning-based framework for occluder-aided NLOS scene estimation. We demonstrated that approaching the blind deconvolution

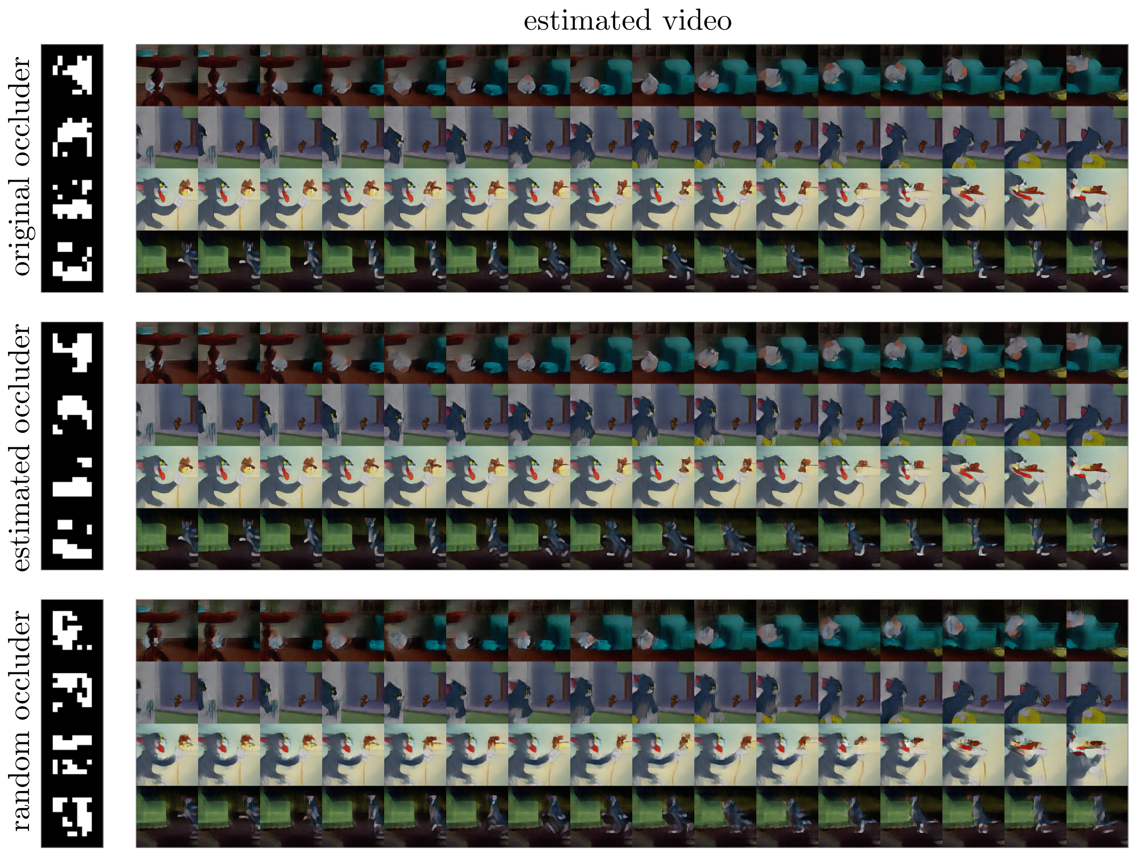


Figure 4-8: Non-blind deconvolution results. We deconvolve 4 observed videos shown in Figure 4-7 with 3 sets of occluders: original, estimated, and random. The generator uses the occluder information to deconvolve the images and the reconstruction quality only slightly suffers from using the estimated occluders.

problem with a two-step method can indeed be useful, and that the learning-based approaches have a potential to help build NLOS imaging systems that are robust to changes in the scene. We also showed that building a lower dimensional latent space for occluders can provide interpretable representations to the network, which raises the question of what the best representations are and how they can be constructed.

The immediate applicability of our scene estimation method to real-world tasks, however, still remains unclear. Since acquiring a large amount of real-world data with several variations in the scene is prohibitive, covering such variations with synthetic data might prove to be useful similar to what is demonstrated in Chapter 3. If the convolution model remains valid in the application of interest, we believe that our findings in this chapter might directly be applicable to such settings. In more unconstrained scene geometries, however, convolutional idealization might no longer apply, especially when the occluder cannot be approximated well with a 2D object. In such cases, we believe that a combination of 3D graphics software-aided data collection and domain adaptation is worth exploring.

Chapter 5

Concluding Remarks

In this thesis, we presented two methods for two different applications of occluded-aided non-line-of-sight (NLOS) imaging. In the scene classification part, we introduced a novel problem where we investigated whether seemingly innocuous shadows arising all around us can be used to reveal some identity information. We formulated this problem as an unsupervised domain adaptation problem, where we collected synthetic data comprising of shadow images under various scene geometries and configurations, and adapted what we learned from this data to the real data. Our results demonstrated the potential of exploiting an overlooked optical phenomenon to reveal useful biometric information, which we supported with our experiments.

Since it is yet unclear whether our identity classification method can be extended to handle multiple identities, we believe it is first worth evaluating the performance of the same method when more than two identities are of interest. Even if the overall accuracy of our method under such settings turns out to be not very promising, it might be used to reliably narrow identities within a group of individuals, which would still be of use in certain applications and at the same time raise privacy concerns. On the other hand, although we focused on a method that works well with an arbitrary occluder shape belonging to a specific class, it is also worth exploring an identity classification method in which the occluder shape is carefully chosen and the number of identities is arbitrary. This line of research would bring questions such as which occluder structures allow more identity information to leak into the shadows. In addition, one might also be interested in identity recognition that require no storage or observation of any sensitive information about the identities, for which such information leakage in shadows could be exploited. In this case, the amount of leakage should be sufficient to reliably distinguish one identity from the others while preventing it being used to reconstruct an image of the identity or reveal any other sensitive information.

In the scene estimation part of this thesis, we focused on a more classical passive NLOS imaging application, where we sought to determine whether learning-based approaches could bring more accuracy and robustness to the scene reconstructions in occluder-aided imaging. Motivated by the limitations of the state-of-the-art work, we proposed a two-step approach that first estimates the occluder shape and then reconstructs the scene based on these estimations. The preliminary results we obtained using our pipeline suggested the promise of learning-based approaches in occluder-aided imaging.

Although we suggest that the occluder-aided NLOS imaging might benefit greatly from the recent advancements in deep learning, we should also note that the lack of adequate data availability could delay progress. Therefore, exploring efficient data collection strategies or tailoring existing data for NLOS imaging applications would be another research direction that would contribute to the computational imaging community. In the absence of such data, however, generating synthetic data that are representative of the real world is crucial, and hence strong idealizations such as the convolutional model of occlusion might not be immediately used for synthetic data collection. To achieve more realistic synthetic data, it is worth further exploring 3D graphics software-based data collection and combining it with state-of-the-art domain adaptation methods or possibly developing novel domain adaptation techniques that are more suitable for NLOS imaging applications.

Non-line-of-sight imaging is an emerging topic with many exciting research directions that await investigation from the broader computational imaging community, and we anticipate that growing interest in learning-based methodologies will transform how we approach imaging problems in the next decade.

Bibliography

[1] Charles Saunders, John Murray-Bruce, and Vivek K Goyal. Computational periscopy with an ordinary digital camera. *Nature*, 565(7740):472–475, 2019.

[2] Adam B Yedidia, Manel Baradad, Christos Thrampoulidis, William T Freeman, and Gregory W Wornell. Using unknown occluders to recover hidden scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12231–12239, 2019.

[3] Miika Aittala, Prafull Sharma, Lukas Murmann, Adam Yedidia, Gregory Wornell, Bill Freeman, and Fredo Durand. Computational mirrors: Blind inverse light transport by deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 14311–14321, 2019.

[4] Katherine L Bouman, Vickie Ye, Adam B Yedidia, Frédo Durand, Gregory W Wornell, Antonio Torralba, and William T Freeman. Turning corners into cameras: Principles and methods. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2270–2278, 2017.

[5] Antonio Torralba and William T Freeman. Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 374–381. IEEE, 2012.

[6] Joshua Rapp, Julian Tachella, Yoann Altmann, Stephen McLaughlin, and Vivek K Goyal. Advances in single-photon lidar for autonomous vehicles: Working principles, challenges, and recent advances. *IEEE Signal Processing Magazine*, 37(4):62–71, 2020.

[7] Adam Yedidia, Christos Thrampoulidis, and Gregory Wornell. Analysis and optimization of aperture design in computational imaging. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4029–4033. IEEE, 2018.

[8] Mauro Buttafava, Jessica Zeman, Alberto Tosi, Kevin Eliceiri, and Andreas Velten. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Optics Express*, 23(16):20997–21011, 2015.

[9] Victor Arellano, Diego Gutierrez, and Adrian Jarabo. Fast back-projection for non-line of sight reconstruction. *Optics express*, 25(10):11574–11583, 2017.

[10] Chia-Yin Tsai, Kiriakos N Kutulakos, Srinivasa G Narasimhan, and Aswin C Sankaranarayanan. The geometry of first-returning photons for non-line-of-sight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7216–7224, 2017.

[11] Matthew O'Toole, David B Lindell, and Gordon Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature*, 555(7696):338–341, 2018.

[12] Felix Heide, Matthew O'Toole, Kai Zang, David B Lindell, Steven Diamond, and Gordon Wetzstein. Non-line-of-sight imaging with partial occluders and surface normals. *ACM Transactions on Graphics*, 38(3):1–10, 2019.

[13] Byeongjoo Ahn, Akshat Dave, Ashok Veeraraghavan, Ioannis Gkioulekas, and Aswin C Sankaranarayanan. Convolutional approximations to the general non-line-of-sight imaging operator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7889–7899, 2019.

[14] Joshua Rapp, Charles Saunders, Julián Tachella, John Murray-Bruce, Yoann Altmann, Jean-Yves Tourneret, Stephen McLaughlin, Robin MA Dawson, Franco NC Wong, and Vivek K Goyal. Seeing around corners with edge-resolved transient imaging. *Nature Communications*, 11(1):1–10, 2020.

[15] Sean I Young, David B Lindell, Bernd Girod, David Taubman, and Gordon Wetzstein. Non-line-of-sight surface reconstruction using the directional light-cone transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1407–1416, 2020.

[16] Cheng Wu, Jianjiang Liu, Xin Huang, Zheng-Ping Li, Chao Yu, Jun-Tian Ye, Jun Zhang, Qiang Zhang, Xiankang Dou, Vivek K Goyal, et al. Non–line-of-sight imaging over 1.43 km. *Proceedings of the National Academy of Sciences*, 118(10), 2021.

[17] Manel Baradad, Vickie Ye, Adam B Yedidia, Frédo Durand, William T Freeman, Gregory W Wornell, and Antonio Torralba. Inferring light fields from shadows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6267–6275, 2018.

[18] JG Ables. Fourier transform photography: A new method for x-ray astronomy. *Publications of the Astronomical Society of Australia*, 1(4):172–173, 1968.

[19] RH Dicke. Scatter-hole cameras for x-rays and gamma rays. *The Astrophysical Journal*, 153:L101, 1968.

[20] Edward E Fenimore and Thomas M Cannon. Coded aperture imaging with uniformly redundant arrays. *Applied optics*, 17(3):337–347, 1978.

[21] Adam Lloyd Cohen. Anti-pinhole imaging. *Optica Acta: International Journal of Optics*, 29(1):63–67, 1982.

[22] Ganesh Ajjanagadde, Christos Thrampoulidis, Adam Yedidia, and Gregory Wornell. Near-optimal coded apertures for imaging via Nazarov's theorem. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7690–7694. IEEE, 2019.

[23] Felix Naser, Igor Gilitschenski, Guy Rosman, Alexander Amini, Fredo Durand, Antonio Torralba, Gregory W Wornell, William T Freeman, Sertac Karaman, and Daniela Rus. ShadowCam: Real-time detection of moving obstacles behind a corner for autonomous vehicles. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 560–567. IEEE, 2018.

[24] Sheila W Seidel, Yanting Ma, John Murray-Bruce, Charles Saunders, William T Freeman, C Yu Christopher, and Vivek K Goyal. Corner occluder computational periscopy: Estimating a hidden scene from a single photograph. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2019.

[25] Sheila W. Seidel, John Murray-Bruce, Yanting Ma, Christopher Yu, William T. Freeman, and Vivek K Goyal. Two-dimensional non-line-of-sight scene estimation from a single edge occluder. *IEEE Transactions on Computational Imaging*, 7:58–72, 2021.

[26] Frank L Pedrotti, Leno M Pedrotti, and Leno S Pedrotti. *Introduction to optics*. Cambridge University Press, 2017.

[27] Adam Yedidia. *Analysis and optimization of occluder-based imaging*. PhD thesis, Massachusetts Institute of Technology, 2020.

[28] Cyril Soler and François X Sillion. Fast calculation of soft shadow textures using convolution. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, pages 321–332, 1998.

[29] Mehdi Askari, Seong-Bok Kim, Kwang-Soo Shin, Seok-Bum Ko, Sang-Hoo Kim, Dae-Youl Park, Yeon-Gyeong Ju, and Jae-Hyeung Park. Occlusion handling using angular spectrum convolution in fully analytical mesh based computer generated hologram. *Optics Express*, 25(21):25867–25878, 2017.

[30] Shun-ichi Amari, Scott C Douglas, Andrzej Cichocki, and Howard H Yang. Multichannel blind deconvolution and equalization using the natural gradient. In *First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*, pages 101–104. IEEE, 1997.

[31] Michael Cannon. Blind deconvolution of spatially invariant image blurs with phase. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(1):58–63, 1976.

[32] Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE Transactions on Image Processing*, 7(3):370–375, 1998.

[33] Fernando César Comparsi De Castro, Maria Cristina Felippetto De Castro, and Dalton S Arantes. Concurrent blind deconvolution for channel equalization. In *ICC 2001. IEEE International Conference on Communications. Conference Record (Cat. No. 01CH37240)*, volume 2, pages 366–371. IEEE, 2001.

[34] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 233–240, 2011.

[35] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1964–1971. IEEE, 2009.

[36] Timothy J. Schulz. Multiframe blind deconvolution of astronomical images. *J. Opt. Soc. Am. A*, 10(5):1064–1073, May 1993.

[37] Michael Hirsch, Suvrit Sra, Bernhard Schölkopf, and Stefan Harmeling. Efficient filter flow for space-variant multiframe blind deconvolution. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 607–614. IEEE, 2010.

[38] Charles L Matson, Kathy Borelli, Stuart Jefferies, Charles C Beckner Jr, E Keith Hege, and Michael Lloyd-Hart. Fast and optimal multiframe blind deconvolution algorithm for high-resolution ground-based imaging of space objects. *Applied Optics*, 48(1):A75–A92, 2009.

[39] David L Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[40] Javier Portilla, Vasily Strela, Martin J Wainwright, and Eero P Simoncelli. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image processing*, 12(11):1338–1351, 2003.

[41] Julien Mairal, Guillermo Sapiro, and Michael Elad. Learning multiscale sparse representations for image and video restoration. *Multiscale Modeling & Simulation*, 7(1):214–241, 2008.

[42] Tony F Chan, Jianhong Shen, and Hao-Min Zhou. Total variation wavelet inpainting. *Journal of Mathematical Imaging and Vision*, 25(1):107–125, 2006.

[43] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011.

[44] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1645, 2016.

[45] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.

[46] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[47] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018.

[48] Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo. Neural blind deconvolution using deep priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3341–3350, 2020.

[49] Phong Tran, Anh Tuan Tran, Quynh Phung, and Minh Hoai. Explore image deblurring via encoded blur kernel space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11956–11965, 2021.

[50] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In

Proceedings of the European Conference on Computer Vision (ECCV), pages 85–100, 2018.

[51] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5840–5848, 2019.

[52] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. Image inpainting: A review. *Neural Processing Letters*, 51(2):2007–2028, 2020.

[53] JH Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. One network to solve them all–solving linear inverse problems using deep projection models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5888–5897, 2017.

[54] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[55] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017.

[56] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. DeblurGAN: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018.

[57] Ruslan Salakhutdinov and Geoffrey Hinton. Deep Boltzmann machines. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 448–455, 16–18 Apr 2009.

[58] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

[59] Lucas Theis and Matthias Bethge. Generative image modeling using spatial LSTMs. *Advances in Neural Information Processing Systems*, 28:1927–1935, 2015.

[60] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014.

[61] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016*.

[62] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[63] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[64] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[65] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models – an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018.

[66] Bernhard Egger, William Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D morphable face models—past, present, and future. *ACM Transactions on Graphics*, 39(5):1–38, 2020.

[67] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999.

[68] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. IEEE, 2009.

[69] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.

[70] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013.

[71] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7346–7355, 2018.

[72] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3D face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019.

[73] Luan Tran and Xiaoming Liu. On learning 3D face morphable model from in-the-wild images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):157–171, 2019.

[74] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic facial texture inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5144–5153, 2017.

[75] Ron Slossberg, Gil Shamai, and Ron Kimmel. High quality facial surface and texture synthesis via generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[76] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. paGAN: Real-time avatars using dynamic textures. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018.

[77] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019.

[78] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. AvatarMe: Realistically renderable 3D facial reconstruction "in-the-wild". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2020.

[79] Elad Richardson, Matan Sela, and Ron Kimmel. 3D face reconstruction by learning from synthetic data. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 460–469. IEEE, 2016.

[80] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5908–5917, 2017.

[81] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5163–5172, 2017.

[82] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1259–1268, 2017.

[83] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *IEEE International Conference on Computer Vision, ICCV 2017*, pages 3735–3744.

[84] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018.

[85] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3D morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018.

[86] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019.

[87] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[88] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

[89] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[90] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

[91] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[92] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 2015.

[93] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[94] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France, 2015.

[95] Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[96] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[97] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[98] Geoffrey French, Michal Mackiewicz, and Mark H. Fisher. Self-ensembling for visual domain adaptation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018*.

[99] A. Gretton, AJ. Smola, J. Huang, M. Schmittfull, KM. Borgwardt, and B. Schölkopf. Covariate shift and local learning by distribution matching. In *Dataset Shift in Machine Learning*, pages 131–160, Cambridge, MA, USA, 2009. MIT Press.

[100] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

[101] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015.

[102] Angel X. hang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[103] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[104] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics*, 38(6):1–17, 2019.

[105] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*, 2015.

[106] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

[107] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, 06–11 Aug 2017.

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

[109] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[110] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. *Advances in Neural Information Processing Systems*, 29:2234–2242, 2016.

[111] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015.*

[112] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.