2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 978-1-6654-9346-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/WACV56688.2023.00093

Can Shadows Reveal Biometric Information?

Safa C. Medin, Amir Weiss, Frédo Durand, William T. Freeman, and Gregory W. Wornell

Massachusetts Institute of Technology

{medin, amirwei, fredo, billf, gww}@mit.edu

Abstract

We study the problem of extracting biometric information of individuals by looking at shadows of objects cast on diffuse surfaces. We show that the biometric information leakage from shadows can be sufficient for reliable identity inference under representative scenarios via a maximum likelihood analysis. We then develop a learning-based method that demonstrates this phenomenon in real settings, exploiting the subtle cues in the shadows that are the source of the leakage without requiring any labeled real data. In particular, our approach 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. Our model is able to generalize well to the real domain and is robust to several variations in the scenes. We report high classification accuracies in an identity classification task that takes place in a scene with unknown geometry and occluding objects.

1. Introduction

Imaging scenes that are not in our direct line-of-sight, referred to as non-line-of-sight (NLOS) imaging, has a diverse set of applications in several domains such as surveillance, search-and-rescue, robotic vision, and medical imaging [12]. NLOS imaging methods typically aim to extract information from the hidden scenes that are outside of our field of view based on the observations of a visible scene. In this work, we ask the question whether or not such observations can leak *sensitive* information about the hidden scenes. In particular, we introduce a novel problem where we seek to determine whether it is possible to extract *biometric* information of individuals present in a room by looking at the shadows on diffuse surfaces induced by their presence as shown in Figure 1. We investigate this problem in a *passive* NLOS imaging setting, meaning that we focus on scenarios where we rely on light sources naturally present in the scene.

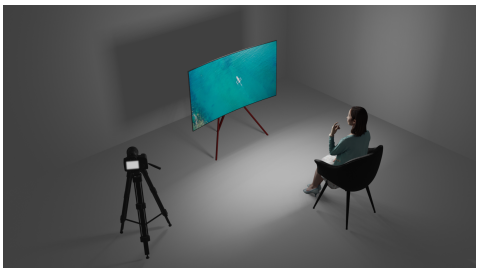


Figure 1: Consider an illustrative scenario where an individual sits across a TV screen in a given room, and suppose that the light reflected by the individual creates a shadow of the screen cast on a blank wall. We demonstrate that such shadows have a potential to leak biometric information in various scene configurations.

Passive NLOS imaging methods can address a number of tasks such as recovering 2D images of the scene [35, 56], reconstructing videos of unknown scenes [2], and estimating the motion and the number of hidden objects [5]. While several methods aim to recover the entire hidden scene [35, 56, 2], often in accidental scenarios [43] where no prior assumptions can be made about the scenes, recovering certain *attributes* of the scene in such scenarios can be useful in several applications. For instance, determining whether or not a non-visible scene includes a person could be potentially useful for autonomous driving, security, or search-and-rescue applications [32, 38]. Our focus, on the other hand, involves recovering biometric information of hidden individuals rather than merely detecting their presence. Here, we define the biometric information as any information that might be used to reveal an individual's identity, in whole or in part.

In this work, we aim to understand whether otherwise innocuous shadows can be used to reveal at least some biometric information by relying on existing contemporary learning tools. We approach this objective by focusing on a specific instance of biometric information extraction, namely, identity classification. In other words, we study the problem

This work was supported, in part, by NSF under Grant No. CCF-1816209 and the MIT-IBM Watson AI Lab under Agreement No. W1771646.

2642-9381/23/$31.00 ©2023 IEEE
DOI 10.1109/WACV56688.2023.00093

of recovering the identities of people in a given room by observing shadows of objects cast on a diffuse surface such as a blank wall. We emphasize that our approach does *not* focus on shadows cast by the individuals but the shadows cast by the objects. We first carry out a maximum likelihood (ML) analysis of this task, characterizing performance under varying number of identities and noise levels by leveraging synthetic face data. Our results suggest that the information leakage from shadows can be significant under representative scenarios. Next, we investigate whether our findings are accurate predictions of what an adversary may be able to accomplish in practice by developing a learning-based method that discovers hidden biometric cues in the shadows without relying on any labeled real data. In particular, we build synthetic scenes composed of 3D face models obtained from a single photograph of each identity of interest [9], and we transfer what we learn from this data to the real data in a completely unsupervised way by leveraging unsupervised domain adaptation techniques. Our method generalizes well to the real domain and is robust to several variations in the scene, such as the shape of occluding objects, lighting conditions, head poses, and facial expressions. We report high classification accuracies in an identity classification task that takes place in a scene with unknown geometry and occluding objects, suggesting that, indeed, there is a significant biometric leakage phenomenon.

This work can be viewed as a first step towards understanding the degree to which seemingly benign images of shadow phenomema have the potential to leak at least some biometric information that could be of societal concern. Such information leakage might potentially be used with malicious intent, e.g., to determine the presence of an individual in a room without their consent. We emphasize that we deliberately do not seek to design an optimized identity classification system (such as a sophisticated adversary might want to). Rather, our methodology serves to demonstrate and characterize the biometric information leakage phenomenon to raise awareness to an overlooked privacy concern. Our results suggest that the biometric cues we discover in shadows could be used to distinguish identities as well as to reliably narrow the identity to within a group of individuals by extracting some amount of biometric information.

The main contributions of this work are as follows:

- We introduce a timely biometric leakage question, which we formulate as a novel NLOS imaging problem of extracting an individual's identity from subtle, indirect shadow phenomena.
- Via a maximum likelihood analyis, we show that such shadows have a significant potential to leak sensitive information under representative scenarios.
- By combining existing learning tools, we develop a methodology that discovers biometric cues in the shadows without relying on any labeled real data, and report nontrivial accuracies in an identity classification task that takes place in a scene with unknown geometry and occluding objects.

2. Background and Related Work

We now summarize the key background concepts and methodologies from 3D face modeling in computer graphics and domain adaption in machine learning, which we will leverage in our work. We also include a brief summary of related work within NLOS imaging—albeit with different objectives—as additional context for our contributions.

Non-line-of-sight imaging. 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 a coherent illumination source (laser) and a photon detector, and *passive methods*, which do not require such specialized equipment. Passive methods have been explored in a variety of scene configurations and imaging objectives, and they typically exploit structure present in the scenes that induces *occlusion*, which improves the conditioning of the imaging problem [55]. Among these methods, Bouman et al. [5] 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. Seidel et al. extend this idea to image stationary objects and form 2D reconstructions of the hidden scenes [37, 36], while Naser et al. [26] detects obstacles around the corners for autonomous driving applications. However, none of these methods explore extracting biometric information from NLOS measurements. In a different setting, [35] uses a pinspeck occluder with known shape but unknown position to recover 2D scenes while [56] exploits motion in hidden scenes to recover the scene without any assumptions about the occluder shape and position. More recently, [51] and [38] study classification tasks from NLOS measurements. Unlike our more practical *unsupervised* approach, however, these methods use supervised learning tools and do not focus on identity classification. In active imaging methods, on the other hand, 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 [6, 28, 30, 19, 31, 54].

3D morphable face models. 3D morphable face models are statistical models of human faces [4, 29, 22, 16], which have been widely used in domains such as face recognition, entertainment, neuroscience and psychology [11]. Over the last decade, advances in deep learning allowed these models to achieve remarkable results in the challenging problem

of recovering 3D faces from 2D images [33, 10, 45, 34], where some of the more recent approaches do not require explicit 3D shape labels [42, 41, 44, 15, 9]. Among these methods, Deng et al. [9] develops a reconstruction network that recovers accurate 3D faces from a single image, which we use in our synthetic data collection.

Domain adaptation. Over the last few years, there has been a significant amount of work in the area of domain adaptation [50, 53], which is the study of transferring knowledge learned from a source domain to a target domain. More recent approaches in domain adaptation are more 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 [14, 46, 48, 25, 39, 24, 47, 13]. Among these methods, DDC [48] aims for learning domain-invariant representations by imposing a maximum mean discrepancy loss [17], Deep CORAL [39] aligns the second-order statistics of the source and the target domains, while ADDA [47] employs an adversarial discriminator to make the representations of the two domains indistinguishable from each other. In another approach, Li et al. [23] shows that updating the batch normalization statistics [20] for the target domain can also be very effective, which we employ in our method.

3. Methodology

Suppose we are given M 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 $\mathbf{x} \in \mathbb{R}^n$ (grayscale images of resolution $\sqrt{n} \times \sqrt{n}$, with $\sqrt{n} \in \mathbb{N}$) and its ground truth identity label as $y \in \mathcal{Y} = \{1, 2, \ldots, M\}$, our objective is to learn a classifier that is capable of reliably inferring identities from shadows given training data $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$.

We begin presenting our approach by first describing our 3D face model. Then, we follow an ML analysis to understand how much information is leaked by shadows under varying numbers of identities and noise levels. Inspired by our findings, we develop our learning-based methodology that could be employed to distinguish identities in practice.

3.1. 3D Face Modeling

In our approach, we make use of synthetic face models which represent faces as triangular meshes. Given a number of vertices V in a mesh, we represent a face shape $\mathbf{s} \in \mathbb{R}^{3V}$ (3D coordinates for each vertex) and its texture $\mathbf{t} \in \mathbb{R}^{3V}$ (RGB colors for each vertex) with the following linear 3D morphable model [29, 7]:

$$\begin{aligned} \mathbf{s} &= \bar{\mathbf{s}} + \mathbf{M}_{\text{id}}\boldsymbol{\alpha}_{\text{id}} + \mathbf{M}_{\text{exp}}\boldsymbol{\alpha}_{\text{exp}} \\ \mathbf{t} &= \bar{\mathbf{t}} + \mathbf{M}_{\text{tex}}\boldsymbol{\alpha}_{\text{tex}} \end{aligned} \tag{1}$$

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; $\mathbf{M}_{\text{id}} \in \mathbb{R}^{3V \times k_{\text{id}}}$, $\mathbf{M}_{\text{exp}} \in \mathbb{R}^{3V \times k_{\text{exp}}}$ and $\mathbf{M}_{\text{tex}} \in \mathbb{R}^{3V \times k_{\text{tex}}}$ are the identity, expression and texture bases; and $\boldsymbol{\alpha}_{\text{id}} \in \mathbb{R}^{k_{\text{id}}}$, $\boldsymbol{\alpha}_{\text{exp}} \in \mathbb{R}^{k_{\text{exp}}}$ and $\boldsymbol{\alpha}_{\text{tex}} \in \mathbb{R}^{k_{\text{tex}}}$ are the identity, expression and texture coefficients. Here, $\bar{\mathbf{s}}, \bar{\mathbf{t}}, \mathbf{M}_{\text{id}}, \mathbf{M}_{\text{exp}}$, and $\mathbf{M}_{\text{tex}}$ are all provided by the model, whereas $\boldsymbol{\alpha}_{\text{id}}$ and $\boldsymbol{\alpha}_{\text{tex}}$ are fixed and known vectors obtained via sampling from a Gaussian prior [11] or by reconstructing 3D faces from 2D images of the identities of interest.

3.2. Maximum Likelihood Analysis

To determine whether and how much biometric information leaks from the shadows, we simulate a representative scene in the synthetic domain. For this, we first describe our data generation, and in particular the creation of 3D faces using the model described in (1) and the convolutional model of occlusion. Next, we present our ML-based learning algorithm with an objective to obtain lower bound on the classification accuracy with respect to numbers of identities and noise levels, where we make certain assumptions regarding the data distribution to allow for ease of analysis.

3.2.1 Observation Model

Let $(\boldsymbol{\alpha}_{\text{id}}^m, \boldsymbol{\alpha}_{\text{tex}}^m)$, $m = 1, 2, \ldots, M$, denote *fixed* and *known* identity and texture coefficients of M identities. For simplicity, suppose that the faces are sufficiently far away from the observation surface so that they can be represented as 2D rendered images of the 3D face objects, and suppose that we observe grayscale images of shadows, $\mathbf{x}^m \in \mathbb{R}^n$ for each identity m, according to the following data model:

$$\mathbf{x}^m = \mathbf{A}\underbrace{\mathbf{R}^m\widetilde{\mathbf{M}}_{\text{tex}}\boldsymbol{\alpha}_{\text{tex}}^m}_{\triangleq\, \mathbf{r}^m} + \mathbf{z} = \mathbf{A}\mathbf{r}^m + \mathbf{z}, \quad m = 1, \ldots, M \tag{2}$$

where $\widetilde{\mathbf{M}}_{\text{tex}} \triangleq [\mathbf{M}_{\text{tex}} \quad \bar{\mathbf{t}}] \in \mathbb{R}^{3V \times (k_{\text{tex}}+1)}$ is the augmented texture basis that generates a texture map from a texture code $\boldsymbol{\alpha}_{\text{tex}}^m \in \mathbb{R}^{(k_{\text{tex}}+1)}$; the random matrix $\mathbf{R}^m \triangleq \mathbf{R}(\boldsymbol{\alpha}_{\text{id}}^m, \boldsymbol{\alpha}_{\text{exp}}^m, \boldsymbol{\theta}^m, \boldsymbol{\gamma}^m) \in \mathbb{R}^{n \times 3V}$ denotes the rendering operation that maps vertex colors of the mesh to grayscale image pixels as a deterministic nonlinear function of the fixed and known identity vector $\boldsymbol{\alpha}_{\text{id}}^m \in \mathbb{R}^{k_{\text{id}}}$, random expression vector $\boldsymbol{\alpha}_{\text{exp}}^m \in \mathbb{R}^{k_{\text{exp}}}$, random pose vector $\boldsymbol{\theta}^m \in \mathbb{R}^{k_\theta}$ and random lighting $\boldsymbol{\gamma}^m \in \mathbb{R}^{k_\gamma}$; $\mathbf{A} \in \mathbb{R}^{n \times n}$ is an unknown light transport matrix that maps a face image to a shadow image due to the presence of an occluder; and finally $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n) \in \mathbb{R}^n$ denotes the additive noise due to thermal noise and shot noise [55], assumed to be statistically independent of the signal-related term $\mathbf{A}\mathbf{r}^m$, where $\mathbf{I}_n$ is the n-dimensional identity matrix.

3.2.2 3D Face Generation

We generate faces using the Basel Face Model 2009 [29] with the neck and the ear regions excluded. We randomly generate

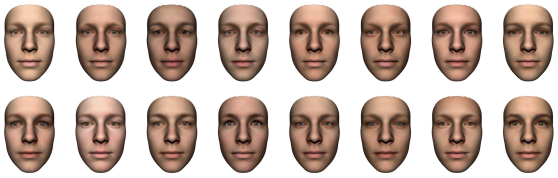


Figure 2: $M = 16$ randomly sampled identities we use in our experiments, rendered with neutral expressions (i.e., $\boldsymbol{\alpha}_{\text{exp}} = \mathbf{0}$), same head poses and lighting conditions. Identities differ only with respect to their textures. In the experiments, textures are converted to grayscale to avoid potential reliance on color information.

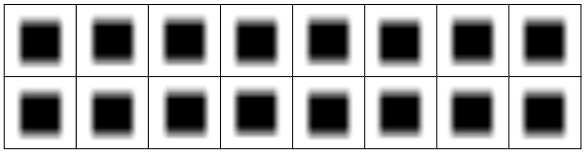


Figure 3: Representative shadow images for each identity, normalized to the range $[0, 1]$ for illustration purposes.

$M = 16$ identities by sampling $\boldsymbol{\alpha}_{\text{tex}}$ from $\mathcal{N}(\mathbf{0}_{k_{\text{tex}}}, \mathbf{I}_{k_{\text{tex}}})$ and setting $\boldsymbol{\alpha}_{\text{id}} = \mathbf{0}_{k_{\text{id}}}$, i.e., the identities have the same face shape under the same facial expression. Hence, the differences between identities only result from the texture, as we seek to show that subtle differences in texture alone can be sufficient to distinguish identities from each other. The expression basis is provided by the model constructed from the FaceWarehouse dataset [7], which we use to sample identities with varying expressions. Finally, we render the faces with random pose and lighting parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$. We illustrate the 16 identities under neutral expressions, same head poses, and lighting conditions in Figure 2.

3.2.3 Convolutional Model of Occlusion

In this part of our analysis only, we assume that the face and the occluder lie in 2D planes that are parallel to each other as well as to the observation plane. This gives rise to the convolutional model of occlusion, a model commonly adopted in passive NLOS imaging applications [21, 56]. Under this model, assuming the occluder is an opaque object that completely blocks the light, the light transport matrix $\mathbf{A}$ can be defined as $\mathbf{A}_{ij} = 0$ if the occluder blocks the light coming from i-th pixel of the face image to j-th pixel of the observation plane, and $\mathbf{A}_{ij} = 1/n$ otherwise. Here, the $(1/n)$-scaling ensures that the observed total power does not exceed the total power reflected from the face [55, 3]. We illustrate representative images of shadows obtained with this model in Figure 3, where a rectangular occluder is simulated.

3.2.4 Learning Algorithm

Let $\mathbf{r}^m$ be the random vector representing random face images for identity m as defined in (2). Given a fixed but unknown $\mathbf{A}$, we assume that the *noiseless* images of shadows $\mathbf{A}\mathbf{r}^m$ are normally distributed with mean $\boldsymbol{\mu}^m$ and covariance $\boldsymbol{\Sigma}^m$, i.e., $\mathbf{A}\mathbf{r}^m \sim \mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)$. Since we assume that the noise $\mathbf{z}$ is statistically independent of these images, the labeled examples $\mathbf{x}^m$ for each identity m are distributed according to $\mathcal{N}(\boldsymbol{\mu}^m, \mathbf{Q}^m)$, where $\mathbf{Q}^m \triangleq \boldsymbol{\Sigma}^m + \sigma^2\mathbf{I}_n$. Given training data $\{\mathbf{x}_1^m, \ldots, \mathbf{x}_N^m\}$ for each identity m, we first compute the sample means and covariances as

$$\widehat{\boldsymbol{\mu}}^m \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^m \quad \widehat{\mathbf{Q}}^m \triangleq \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^m - \widehat{\boldsymbol{\mu}}^m)(\mathbf{x}_i^m - \widehat{\boldsymbol{\mu}}^m)^{\mathrm{T}} \tag{3}$$

At test time, given a test observation $\mathbf{x}$, assuming each identity is equally likely and that the determinant of the covariance matrices of each identity are equal to each other, the ML estimation rule is given by

$$\widehat{m} = \underset{m=1,2,\ldots,M}{\arg\min}\ (\mathbf{x} - \widehat{\boldsymbol{\mu}}^m)^{\mathrm{T}}\big(\widehat{\mathbf{Q}}^m\big)^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}^m) \tag{4}$$

In practice, since n is typically very large, we cannot assume that $N \gg n$. Therefore, inverting the sample covariance matrices obtained by a finite number of samples N does not yield a robust classifier. Assuming $\operatorname{rank}(\boldsymbol{\Sigma}^m) = r$ for any m, and denoting the eigenvalue decomposition of the m-th sample covariance matrix $\widehat{\mathbf{Q}}^m \triangleq \mathbf{U}^m\boldsymbol{\Lambda}^m(\mathbf{U}^m)^{\mathrm{T}}$ with $\boldsymbol{\Lambda}^m = \operatorname{diag}(\lambda_1^m, \ldots, \lambda_n^m)$ such that $\lambda_1^m \geqslant \cdots \geqslant \lambda_n^m > 0$, we propose the following improved procedure. Since the noise is assumed to be spatially white, we first estimate the noise variance for each identity as [52]

$$\widehat{\sigma}^2 = \frac{1}{n-r}\sum_{i=r+1}^{n}\lambda_i^m \tag{5}$$

Then, we set the refined estimate $\widehat{\mathbf{Q}}^m = \mathbf{U}^m\bar{\boldsymbol{\Lambda}}^m(\mathbf{U}^m)^{\mathrm{T}}$, where $\bar{\boldsymbol{\Lambda}}^m = \operatorname{diag}(\lambda_1^m, \ldots, \lambda_r^m, \widehat{\sigma}^2, \ldots, \widehat{\sigma}^2)$, and adopt the ML estimator defined in (4).

3.3. Neural Network Classifier

As an example of the kind of processing an adversary seeking to extract biometric information might use in practice, we now develop a learning-based framework suitable for identity classification in real settings, where we assume that the occluder shape is not fixed but arbitrary. Since we follow a data-driven approach, representing possible variations such as the occluder shape, lighting conditions, facial expressions, and head poses in the training data is crucial to achieve a robust classification system. Because collecting such data is highly impractical, we develop a method that avoids such challenges. In particular, we use 3D graphics software to collect large amounts of training data by placing

In fact, this is only *asymptotically* an ML classifier, since we use the sample means and covariances obtained from a finite number of samples.

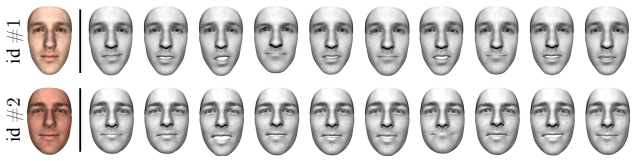


Figure 4: Face reconstructions of $M = 2$ identities we use in our experiments, rendered with varying expressions. Given RGB reconstructions of the faces, we first convert their textures to grayscale and match their average intensity levels. Expressions are randomly sampled and varied in the dataset.

3D faces and objects into simulated scenes. Then, we transfer what we learn from these scenes to the real settings by employing unsupervised domain adaptation techniques.

3.3.1 3D Face Generation

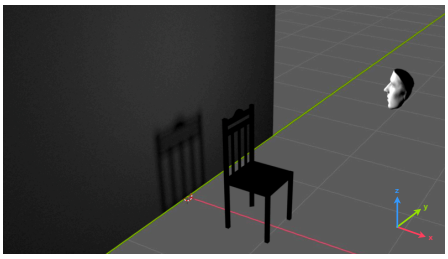
To minimize the gap between the synthetic and real domains, we use a 3D face reconstruction network [9], which allows us to obtain a 3D model of an identity *from a single image*. The reconstructed faces in this work also follow the Basel Face Model 2009 [29] with the neck and the ear regions excluded, which ensures that the network trained with the synthetic data relies only on the identity information, i.e., information such as the thickness of the neck or the contrast between the hair and skin intensities cannot be exploited in our method. As before, we create variations in facial expressions using the model obtained from [7]. 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 with varying expressions in Figure 4.

3.3.2 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 that creates shadows on this wall. For illustration purposes, we focus on *chairs* as occluding objects, as they are one of the most common and diverse classes of indoor objects. However, we emphasize that 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 [8]. We use a white planar object as a wall and a white spotlight as an illumination source. When we render the scenes, we cover as much variation as possible by changing the pose, position and expression of the faces, and vary the position of the light sources. We illustrate a representative synthetic scene in Figure 5a.

In our real data collection, the individuals sit across a blank wall individually, where a chair is positioned between



(a) Synthetic scene geometry

(b) Real scene geometry

Figure 5: Scene geometries for (a) synthetic and (b) 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. The light reflecting from the face creates shadows of objects on the wall.

the identity and the wall. The faces 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 process. We performed these experiments in a physical space shown in Figure 5b.

3.3.3 Domain Adaptation

Given two sets of data $\mathcal{S} = \{(\mathbf{x}_1^s, y_1^s), \ldots, (\mathbf{x}_N^s, y_N^s)\}$ and $\mathcal{T} = \{(\mathbf{x}_1^t, y_1^t), \ldots, (\mathbf{x}_N^t, y_N^t)\}$, which represent the source data and the target data, respectively, our objective is to 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 involve only identity information, i.e., these signals may depend on unintended cues from the real 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 [18], 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

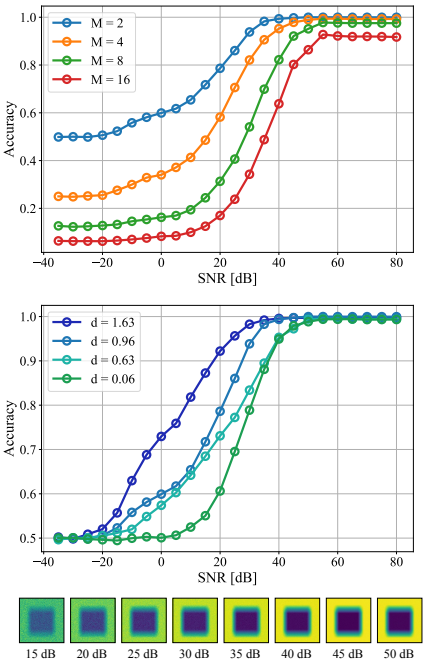


Figure 6: Empirical results for maximum likelihood classification. *Top*: Classification accuracies with respect to varying number of identities and signal-to-noise ratio (SNR). *Bottom*: Classification accuracy with respect to the distance between the estimated means of the identities and SNR.

the network [20, 23] by feeding the unlabeled target data $\mathcal{T} = \{\mathbf{x}_1^t, \ldots, \mathbf{x}_N^t\}$ through the network. As we will show in the next section, the updated network generalizes well to the test samples from the target domain.

4. Experiments and Results

We describe our experiments and present their results by first focusing on our ML classifier and characterizing its performance with respect to the number of identities and noise levels. Next, we focus on our neural network classifier by elaborating on the real and synthetic data colletion, and provide accuracies obtained in different stages of the method.

4.1. Maximum Likelihood Analysis

Following the data generation pipeline we described in Section 3.2, we generate 20000 samples for each of the 16 identities, where we randomly change the head poses, facial expressions, and lighting conditions (we include the details of this process in the supplementary material). We split this data into train and tests sets with 90%–10% split, which gives us 18000 train and 2000 test samples for each identity. Next, we pick the first 2, 4, 8, and then all 16 identities from this dataset, and apply our ML-based learning algorithm described in Section 3.2.4, where we calculate classification accuracies on the test data under varying noise levels. We summarize our results in Figure 6 (*top*), where we observe

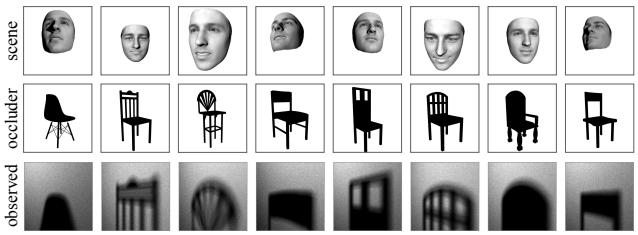


Figure 7: Representative samples from the dataset, where each column shows one sample. Our dataset covers a diverse set of head poses and facial expressions as well as occluder shapes.

high accuracies for all numbers of identities at moderate-to-high signal-to-noise ratio (SNR) levels.

Although the results we illustrate in Figure 6 (*top*) are representative under the face model described in Section 3, we note that the similarities between the identities have a natural influence on the accuracy, since it is more difficult to distinguish two identities that have very similar face shapes and textures (e.g., two identities might include identical twins in practice). To investigate this, we pick 4 pairs of identities from our dataset with varying distances $d(i, j)$ between their estimated means, where $d(i, j) \triangleq \|\hat{\boldsymbol{\mu}}^i - \hat{\boldsymbol{\mu}}^j\|_2 \in \mathbb{R}_+$ for an identity pair (i, j). We report accuracy curves for these pairs in Figure 6 (*bottom*), where we observe that the distance between the means has a direct impact on the performance, although almost perfect classification is still possible when the SNR is sufficiently high. This suggests that the variabilities of shadow images associated with each identity (determined by the respective covariances) are almost negligible with respect to the distances between the means when the noise is sufficiently small.

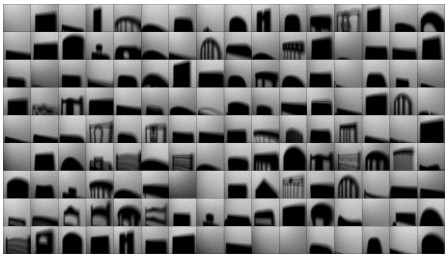
4.2. Neural Network Classifier

We now present our experiments and results for the neural network classifier, and show that our method is effective in extracting subtle biometric cues from shadows in real settings, consistent with our ML analysis predictions.

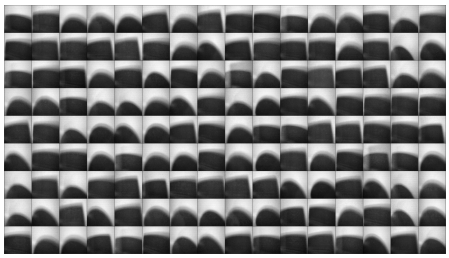
4.2.1 Synthetic Data Collection and Training

We generate the synthetic data for our network randomly, where we vary the pose, expression and position of the face, the location of the light source, and the occluder shape. We illustrate representative samples from the dataset in Figure 7.

We collect our synthetic data using Mitsuba2 [27], with which we render 256×256 images of the observed wall using 50000 samples per pixel. Rendering one image takes ≈ 50 seconds on an NVIDIA GeForce RTX 2080 Ti GPU, and the intensities of all images are normalized to $[0, 1]$ range after rendering. For each identity, we collect 4000 images that we split into train and test sets with 75%–25% split, which gives us a total of 6000 train and 2000 test samples. We illustrate random samples from this dataset in Figure 8a.



(a) Source images



(b) Target images

Figure 8: Random images from the source and the target datasets.

Details of the dataset generation and training procedure are included in the supplementary material.

4.2.2 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 (e.g., collecting data in a very diverse set of scene configurations may not be feasible or practical for an adversary). In particular, we experiment with 4 light source locations by using 4 separate spotlights (which are individually lit during the data collection), 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 8b.

We show our results in Figure 9 where we visualize the feature distributions of the test samples before and after domain adaptation using t-SNE [49]. 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 target domains. Furthermore, 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 does not see any target samples before the domain adaptation, it performs rather poorly on the target domain, achieving 62.70% accuracy. After the domain adaptation (shown in the second row), we observe

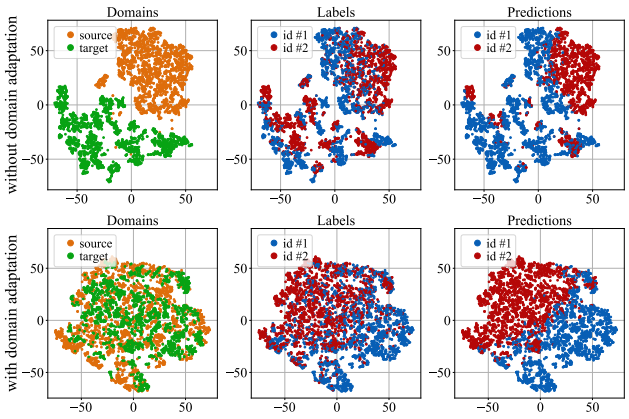


Figure 9: Illustration of results: feature distributions of the test data (extracted from the final layer before classification) in 2D using t-SNE dimensionality reduction technique [49]. *First row:* Feature distributions of source and target domains before domain adaptation, where we observe that the network performs well on the source domain but not on the target domain. *Second row:* Feature distributions after domain adaptation, which reflect that the network generalizes well to the target data as well.

source	target (before adaptation)	target (after adaptation)
74.57 ± 0.84	59.67 ± 9.26	77.08 ± 2.42

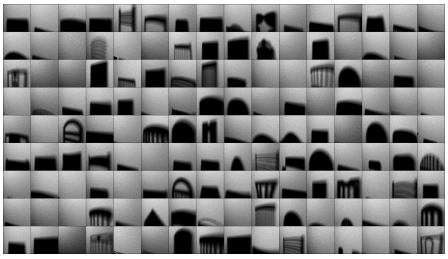
Table 1: Average classification accuracies (in percentage) at different stages of our method, computed over 20 independent trials.

that the feature distributions of the source and target data are well-aligned, and ground truth labels for both domains seem to be well-separated, which allows 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 1 computed over 20 independent trials using the same network and hyperparameters.

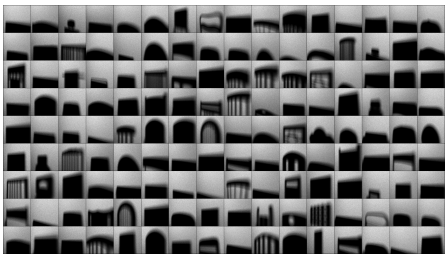
5. Discussion and Analysis

Having demonstrated that biometric information leakage *can* occur, we now turn to understanding aspects of *how* it occurs, by interpreting and analyzing the behavior of our neural network classifier in various scene configurations. To achieve this, we analyze the results on the synthetic images for which we have access to the conditions under which they were rendered, such as occluder shape, head pose, and light source location. 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 [1].

We first investigate the influence of occluder shape and face appearance on the performance, where we compare



(a) Incorrectly classified images



(b) Correctly classified images

Figure 10: Random samples of incorrectly and correctly classified images. Incorrectly classified images usually lack shadows (hence penumbrae) where most information appears to lie. In contrast, correctly classified images usually have large shadow areas.

all 484 fail cases (which gives us 75.80% accuracy 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 10. Here, it is observed that the incorrectly classified images usually lack shadows. Specifically, 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* [5, 37, 36]. 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 identity present in the scene, and they are in fact where our network appears to rely on the most. In particular, 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 [40]. We illustrate several examples in Figure 11, 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 leak most of the sensitive information.

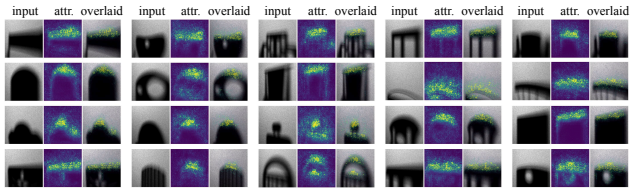


Figure 11: Image attributions extracted by the Integrated Gradients method [40]. We observe that the network is mostly sensitive to the penumbra regions, where most biometric information seems to lie.

6. Concluding Remarks

We show that it is possible for biometric information of individuals to be inferred from indirect shadows cast by objects on diffuse surfaces. We analyze this largely overlooked optical phenomenon first via a maximum likelihood analysis, which shows that otherwise innocuous shadows can be exploited for reliable identity inference under representative scenarios. We further construct a method—representative of one that might be used in practice by an adversary—that demonstrates these vulnerabilities in real settings. In particular, we use a learning-based approach that discovers hidden biometric cues in the indirect shadows by combining synthetic data training with unsupervised domain adaptation. Our synthetic data acquisition relies on a state-of-the-art 3D face reconstruction network, with which we obtain accurate 3D face models from only a single photograph of each identity. We show that our method achieves high accuracies in an identity classification task in real settings, and is robust to several variations in the scene, such as the shape of the occluding objects, lighting, head pose, and facial expressions. Our results suggest that the primary source of the biometric information leakage is the penumbra portions of the shadows, which we explain with the resolving power of occluding edges. Although the degree to which larger numbers of identities can be distinguished—and different types of biometric information can be extracted—remains to be investigated, our results make clear that biometric leakage occurs and that the information can be extracted by using existing tools and learning methodologies. Given that indirect shadow phenomena is omnipresent, our results make a case for further investigation of the risks and an exploration of approaches to their mitigation.

At the same time, the extensions of our methodology could, in principle, 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, enabling face recognition without taking any photos of the individuals. These, too, warrant further investigation.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018.

[2] Miika Aittala, Prafull Sharma, Lukas Murmann, Adam B Yedidia, Gregory W Wornell, William T Freeman, and Fredo Durand. Computational mirrors: Blind inverse light transport by deep matrix factorization. *arXiv preprint arXiv:1912.02314*, 2019.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[9] 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.

[10] 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.

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

[12] Daniele Faccio, Andreas Velten, and Gordon Wetzstein. Non-line-of-sight imaging. *Nature Reviews Physics*, 2(6):318–327, 2020.

[13] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

[14] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015.

[21] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR 2011*, pages 233–240. IEEE, 2011.

[22] 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.

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

[24] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *arXiv preprint arXiv:1606.07536*, 2016.

[25] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105. PMLR, 2015.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] Joshua Rapp and Vivek K Goyal. A few photons among many: Unmixing signal and noise for photon-efficient active imaging. *IEEE Transactions on Computational Imaging*, 3(3):445–459, 2017.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] Sheila Patricia 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*, 2020.

[37] 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.

[38] Prafull Sharma, Miika Aittala, Yoav Y Schechner, Antonio Torralba, Gregory W Wornell, William T Freeman, and Fredo Durand. What you can learn by staring at a blank wall. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2330–2339, 2021.

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

[40] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[41] 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.

[42] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

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

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

[51] Yangyang Wang, Yaqin Zhang, Meiyu Huang, Zhao Chen, Yi Jia, Yudong Weng, Lin Xiao, and Xueshuang Xiang. Accurate but fragile passive non-line-of-sight recognition. *Communications Physics*, 4(1):1–9, 2021.

[52] Mati Wax and Thomas Kailath. Detection of signals by information theoretic criteria. *IEEE Transactions on acoustics, speech, and signal processing*, 33(2):387–392, 1985.

[53] 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.

[54] 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.

[55] 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.

[56] 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.