Fairness Properties of Face Recognition and Obfuscation Systems

Harrison Rosenberg, *University of Wisconsin–Madison;*
Brian Tang, *University of Michigan;* Kassem Fawaz and
Somesh Jha, *University of Wisconsin–Madison*

https://www.usenix.org/conference/usenixsecurity23/presentation/rosenberg

This paper is included in the Proceedings of the 32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Fairness Properties of Face Recognition and Obfuscation Systems

Harrison Rosenberg
University of Wisconsin–Madison

Brian Tang
University of Michigan

Kassem Fawaz
University of Wisconsin–Madison

Somesh Jha
University of Wisconsin–Madison

Abstract

The proliferation of automated face recognition in the commercial and government sectors has caused significant privacy concerns for individuals. One approach to address these privacy concerns is to employ evasion attacks against the metric embedding networks powering face recognition systems: Face obfuscation systems generate imperceptibly perturbed images that cause face recognition systems to misidentify the user. Perturbed faces are generated on metric embedding networks, which are known to be unfair in the context of face recognition. A question of demographic fairness naturally follows: *are there demographic disparities in face obfuscation system performance?* We answer this question with an analytical and empirical exploration of recent face obfuscation systems. Metric embedding networks are found to be demographically aware: face embeddings are clustered by demographic. We show how this clustering behavior leads to reduced face obfuscation utility for faces in minority groups. An intuitive analytical model yields insight into these phenomena.

1 Introduction

Automated face recognition has proliferated in various commercial and government sectors. Face recognition systems can identify users on social media, search for missing persons, aid law enforcement and surveillance, and verify identities of individuals [1, 2]. The widespread adoption of face recognition systems has been swift with the emergence of metric embedding networks such as FaceNet [3] and ArcFace [4] as well as the abundance of labeled face data [5, 6].

Recent coverage of data breaches, privacy law violations, and the adoption of face recognition by law enforcement entities have shed light on the significant privacy issues with face recognition systems. To mitigate growing privacy concerns, face obfuscation systems have been proposed to hide user identities. Several of these systems, such as Face-Off [7], LowKey [8], and Foggysight [9], leverage properties of evasion attacks against machine learning models [10–12]. By introducing small, structured, and imperceptible perturbations to their face, a user can evade identification by a face recognition system. Such systems are attractive for an end-user: perturbations are often visually acceptable to the user; features of social media applications, such as face-augmenting filters, do not suffer; and the obfuscation mechanism may run locally, without access to target face recognition systems.

However, face obfuscation systems suffer major shortcomings. Researchers have identified the ability of face recognition systems to adaptively learn from perturbed faces [13] and re-identify them in the future. In this work, we uncover another shortcoming of such systems: *the presence of performance disparities with respect to demographics.* This disparity leads to the following research questions:

- Are the metric embedding networks underlying face obfuscation systems aware of demographic attributes in faces?
- How does the behavior of face obfuscation systems depend on the demographic attributes of faces?
- Do bias mitigation strategies for face recognition systems also mitigate bias in face obfuscation systems?

This paper characterizes demographic disparities of face obfuscation systems and their underlying metric embedding networks[1]. Previous research has studied the fairness and robustness properties of face recognition [14, 15][2] in the classification setting. However, we study the fairness properties of face recognition and obfuscation in the context of the metric embedding networks — the real-world setting for such systems. Our work yields the following insights into fairness implications of face recognition and obfuscation.

Are the metric embedding networks underlying face obfuscation systems aware of the demographic attributes of faces? We observe face recognition systems are better at discerning individuals in different demographic groups than discerning individuals within the same demographic. Without explicit access to demographic information, face recognition systems still learn to differentiate demographic groups.

How does the behavior of face obfuscation systems depend on the demographic attributes of faces? We analyze two recent face obfuscation systems. Face-Off [7] and LowKey [8] serve as proxies for targeted face obfuscation and untargeted face obfuscation, respectively. Minority groups require stronger perturbations to successfully obfuscate a face. This is especially true in a black-box setting, such as the Face++, Azure, and AWS Rekognition face recognition APIs. We also show that faces perturbed by untargeted attacks often retain their original demographic attributes. We conclude that larger, more visible perturbations are necessary to successfully target identities in demographic groups different from the original image.

---

[1]Code repo: github.com/wi-pi/fairness_face_obfuscation
[2]See section 6 for further discussion.

How does the training regime affect the behavior of face obfuscation? We compare the effects of a standard face recognition network training method with two alternative training regimes designed to mitigate bias. The first regime is the training procedure defined by Xu et al [16]. The second regime is the training of models on demographically balanced datasets. We show these techniques do not entirely eliminate demographic-wise performance disparities.

To aid our response to these three questions, we devise an analytical model based on a mixture of Gaussian distributions and Principal Component Analysis (PCA). This analytical model allows us to formalize the apparent behavior of face recognition and obfuscation when conditioning on the demographic group. Our model reveals that obfuscated faces are more likely to belong to their original demographic group.

2 Background

The terminology with respect to population demographics used in this paper follows that of Buolamwini and Gebru [14] and Nanda et al. [15], two leading works on face recognition fairness. In the dataset annotations, there are only two sexes, hence we use "male" and "female." As for ethnicity, previous literature utilizes terms such as "White" "Black," "Asian," and "Indian" within their attribute annotations [17]. We find it more accurate to refer to these demographic labels as "race." For consistency, we use the same demographic attribute labels in the VGGFace2 dataset in our face recognition and face obfuscation performance evaluations.

2.1 Notation

We consider the setting in which there exists an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and a discrete label set $\mathcal{Y}$. A subset of examples, also referred to as a dataset, is denoted as $S \subseteq \mathcal{X} \times \mathcal{Y}$. Sometimes we abuse notation and let S contain only unlabeled examples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots\}$. A sample $\boldsymbol{x}$ is a d-dimensional real-valued vector. Often, $\boldsymbol{x}$ refers to a cropped face, d is the number of pixels in the cropped face multiplied by three (the RGB channels), and $\mathcal{Y}$ to the set of identities. Given a vector $\boldsymbol{z}$, z_j denotes its jth entry. Calligraphic capital letters denote probability distributions. $\mathcal{D}$ represents the distribution from which training data S are drawn. $\mathbb{1}$ denotes the indicator function. Let z be a Boolean expression. $\mathbb{1}[z]$ evaluates to 1 if z is true, otherwise $\mathbb{1}[z]$ evaluates to 0. A metric is denoted by $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$.

2.2 Face Recognition System

The core component of a face recognition system is a *metric embedding network*. The metric embedding network, denoted by $f_k : \mathbb{R}^d \to \mathbb{R}^k$, is a neural network which takes an RGB face image $\boldsymbol{x}$ as input, and returns a k dimensional embedding. Thus, the term *embedding* refers to the k-dimensional face representation output by the metric embedding network. We
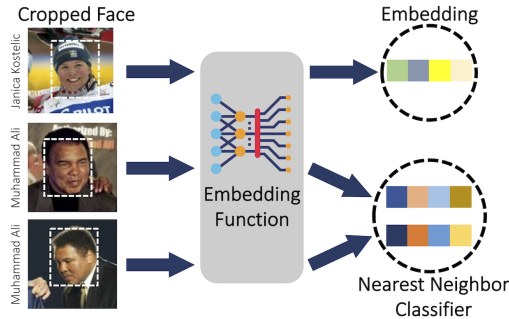

Figure 1: For well-trained embedding functions, embeddings of images belonging to the same identity will have smaller pairwise distances than the pairwise distances between embeddings of different identities.

sometimes omit the subscript k when referring to a generic embedding function. The goal of a metric embedding network is to map high dimensional images into an embedding space such that any two images belonging to the same identity have lower pairwise distance than any two images with different identities. Metric embedding networks are typically trained with one of two classes of loss functions: contrastive loss [18] and triplet loss [3]. The functionality of a face recognition system is depicted in fig. 1.

Upon obtaining an embedding, tasks such as clustering, matching, or classification may be performed. A more performant metric embedding network is one which only matches embeddings of faces belonging to the same identity. A match occurs when two embeddings are sufficiently close: Given a non-negative, real-valued threshold τ and two examples $\boldsymbol{x}, \boldsymbol{x}'$, a match occurs when $\rho(\boldsymbol{x}, \boldsymbol{x}') \leq \tau$. Often, ρ is the ℓ_2 norm.

Matching performance is measured by TPR_z which is a parametrized notion of true positive rate. For choice of metric ρ, the match threshold threshold τ is chosen such that it satisfies a false acceptance rate upper-bounded by z.

$$\text{TPR}_z(S, S') \triangleq \sum_{(\boldsymbol{x}_i, y_i, \boldsymbol{x}'_i, y'_i) \in \{S, S' | y_i = y'_i\}}^{m} \frac{\mathbb{1}\left[\rho\left(f(\boldsymbol{x}), f(\boldsymbol{x}')\right) < \tau\right]}{\left|\{S, S' \mid y_i = y'_i\}\right|} \quad (1)$$

where threshold τ is defined as

$$\max \tau \quad \text{s.t.} \quad z > \sum_{(\boldsymbol{x}, y, \boldsymbol{x}', y') \in \{S, S' | y_i \neq y'_i\}} \frac{\mathbb{1}\left[\rho\left(f(\boldsymbol{x}), f(\boldsymbol{x}')\right) < \tau\right]}{\left|\{S, S' \mid y_i \neq y'_i\}\right|} \quad (2)$$

and the right side of inequality (2) is false acceptance rate.

2.3 Face Obfuscation Systems

Face recognition systems are vulnerable to adversarial examples. Such vulnerabilities were famously discovered by Szegedy et al [19]. Small, structured perturbations, imperceptible to humans, may cause networks to misclassify given inputs. This branch of machine learning research has led to the design of many attack algorithms, of which the most prominent are so-called evasion attacks. Examples of evasion attacks include the Projected Gradient Descent (PGD) [11]

and Carlini-Wagner (CW) [12] attacks. The bulk of study focuses on algorithmic generation of ℓ_p norm-bounded perturbations via noisy gradient-based optimization.

Instead of seeing adversarial examples as a challenge, privacy researchers have demonstrated that systems which leverage such structured perturbations can provide users with privacy utility against face recognition. These privacy benefits are encapsulated in systems known as face obfuscation systems. In this section, we present notation and discuss work relevant to such face obfuscation systems.

2.3.1 Threat Model

In the face obfuscation setting, an end user considers the machine learning provider to be the main threat. Such threats include data breaches [20], cyber-stalkers [21], web scrapers, big government entities [22], and more. Face obfuscation systems counter such threats: Prior to upload, the user applies (imperceptible) perturbations to their face images. Perturbations are designed so the face recognition system is aware of the presence of the face, but the predicted identity of the face is incorrect. In other words, perturbations are constructed so that, from the perspective of the face recognition system, the user impersonates another identity. This is the *white-box* setting: examples are generated on the target model directly.

We also consider the black-box setting. Perturbations applied in the black-box setting leverage an important property of adversarial examples: their ability to transfer across models [23, 24]. Transferability allows users to impersonate identities without accessing a target model. In the *black-box* setting, face obfuscation systems leverage transferability by querying a surrogate model to generate perturbed faces.

2.3.2 Overview of Face Obfuscation Systems

The earliest works in face obfuscation explore physically perturbed examples. For example, Sharif et al. [25] present physically realizable glasses that allow a human user to impersonate a target individual. With the recent societal focus of online privacy, researchers have shifted their focus to digital face obfuscation systems. Digital face obfuscation systems are better suited for social media and internet applications. Examples of such systems include FAWKES [26], Face-Off [7], Low-Key [8], and FoggySight [9]. With the exception of FAWKES, which leverages data poisoning attacks, these face obfuscation systems utilize evasion attacks in an attempt to hide the user's identity.

We now describe face obfuscation systems more formally. Let Δ denote an evasion attack function (e.g. PGD, CW), and let δ denote a generic perturbation output by the evasion attack function, i.e. $\Delta(x)$. A face obfuscation system feeds $x + \delta$ into the face recognition system. Note that Δ may include dependence on the metric embedding network, the underlying dataset, or some other surrogate model.

2.3.3 Evasion Attacks on Metric Embedding Networks

There are two types of evasion attacks: targeted and untargeted. We first describe the embedding centroid before defining the attacks: Given dataset S, denote by $c_{f,y}$ the embedding centroid of identity y as computed on embedding function f:

$$c_{f,y} \triangleq \frac{1}{|\{(x',y') \in S \mid y' = y\}|} \sum_{(x',y') \in S} f(x') \cdot \mathbb{1}\left[y' = y\right] \quad (3)$$

Untargeted Attacks: Given a labeled example (x, y), untargeted attacks find a perturbation δ for which obfuscated face $x + \delta$ impersonates some identity y_0, where $y_0 \neq y$. Given a metric embedding network $f : \mathcal{X} \to \mathcal{Y}$, and metrics ρ_1, ρ_2 an *untargeted attack* is formulated as:

$$\min_{\delta \mid \{x + \delta \in \mathcal{X}\}} \rho_1(\delta, 0)$$

$$\text{s.t.} \quad \underset{y' \in \mathcal{Y}}{\arg\min} \, \rho_2(c_{f,y'}, f(x)) \neq \underset{y' \in \mathcal{Y}}{\arg\min} \, \rho_2(c_{f,y'}, f(x + \delta))$$

The problem with the formulation above is its intractable constraint. To make the attack implementable in practice, the constraints must be relaxed. For a labeled example (x, y), the optimization objective which yields an untargeted perturbation can be written as:

$$\underset{\delta \mid \{x' + \delta \in \mathcal{X}\}}{\arg\max} \, \rho_2(f(x' + \delta), c_{f,y}) \quad \text{s.t.} \quad \rho_1(\delta, 0) \leq \varepsilon \quad (4)$$

In this paper, we use the LowKey attack [8] to instantiate the attack described in eq. (4). Within LowKey, PGD is used to solve eq. (4), ρ_1 is the Learned Perceptual Image Patch Similarity (LPIPS) metric [27], and ρ_2 is the distance between the original face and perturbed face in the embedding space. Note that the latter distance is averaged through an ensemble of models and after applying Gaussian smoothing.

Targeted Attacks: Given a labeled example (x, y), and target identity y_0, targeted attacks find perturbation δ so that obfuscated face $x + \delta$ impersonates y_0. Given a metric embedding network $f : \mathcal{X} \to \mathbb{R}^k$, and metrics ρ_1, ρ_2 a *targeted attack* may be formulated as:

$$\min_{\delta \mid \{x + \delta \in \mathcal{X}\}} \rho_1(\delta, 0) \quad \text{s.t.} \quad y_0 = \underset{y \in \mathcal{Y}}{\arg\min} \, \rho_2(c_{f,y}, f(x + \delta)) \quad (5)$$

Similar to the untargeted case, targeted attacks on face recognition systems relax the above constraints to arrive at a more tractable optimization formulation. Multi-class hinge loss, as used by FaceOff [7], provides the desired relaxation of the constraints in the targeted attack eq. (5). Given a perturbed example $x + \delta$, target label y_0, and positive real number κ, the multi-class hinge loss is denoted by $G_\kappa(x + \delta, y_0)$ where:

$$G_\kappa(x + \delta, y_0) \triangleq \max\left\{0, \kappa + \rho_2(x + \delta, c_{f,y_0}) - \max_{y' \neq y_0} \rho_2(x + \delta, c_{f,y'})\right\}$$

For a labeled example (x, y), and target label $y_0 \neq y$, the CW attack can minimize the following optimization objective to yield a targeted perturbation:

$$\underset{\delta \mid \{x + \delta \in \mathcal{X}\}}{\arg\min} \, \rho_1(\delta, 0) \quad \text{s.t.} \quad \rho_1(\delta, 0) \leq \varepsilon \text{ and } G_\kappa(x + \delta, y_0) \leq 0 \quad (6)$$

2.4 Analytical Techniques

We utilize a combination of established data analysis techniques to assess the fairness of both face recognition and face obfuscation systems.

PCA: Neural networks with nonlinearities are notoriously difficult to analyze. To gain intuition for face obfuscation, we analyze PCA. PCA is a theoretically tractable embedding function which can be represented as a neural network. Hence, we use PCA as a proxy for non-linear embedding functions.

t-SNE: The t-Distributed Stochastic Neighbor Embedding (t-SNE) [28] is a dimensionality reduction technique useful for visualizing high dimensional embedding spaces and image datasets. Visualizations rendered by t-SNE are two or three dimensional. t-SNE is a variant of Stochastic Neighbor Embeddings [29] which avoids crowding data points and can capture the implicit structure of data. t-SNE plots aid our discussion of the embedding space geometry.

TCAV: Testing with Concept Activation Vectors (TCAV) [30] is a tool used to interpret deep neural networks. Given user-specified high-level concepts, such as patterns or colors, linear classifiers are trained on the neural network's activations for those concepts. Concept Activation Vectors (CAVs) are extracted from the vector orthogonal to the linear classifier's decision boundary, and a statistical significance test is performed on the TCAV score generated from the dot product of each CAV and the model's gradients.

Fairness Definitions: There is no single definition of fairness, mathematical or otherwise [31]. Hence, we study several quantities which have an intuitive connection to both face recognition and face obfuscation systems. These quantities include a comparison, by demographic, of face obfuscation success. When success rates are equal across demographics, the fairness constraint known as statistical parity [32] is satisfied. We also compare, by demographic, the strength of perturbations necessary to yield a successful face obfuscation. For targeted obfuscation, we study the strength of such perturbations necessary to impersonate identities within both the inter-demographic group and intra-demographic group settings. We also study how likely untargeted perturbations are to change the perceived demographic of an image.

True positive rate balancing is another notion of fairness which appears in machine learning literature. In section 4.3.1, we see that a specialized training procedure by Xu et al. [16] yields a metric embedding network for which demographic groups are approximately equal in their matching performance TPR_z. An optimization constraint explicitly enforcing equal true positive rates between groups is the fairness metric known as Equalized Odds [33].

Statistical Significance: We use statistical significance testing to draw conclusions from experiments. The most general two-sample t-test, used to determine if two distributions have unequal means, is Welch's t-test [34]. Welch's t-test is applicable when population variances are nonequal and/or the number of elements in each sample differs. For statistical tests on more than two samples, we apply the Alexander-Govern test [35], a multi-sample generalization of Welch's t-Test. Each statistical test is designed to determine if any two or more samples are drawn from the same distribution. In all our statistical tests, we consider a p-value less than 0.05 to be significant. For the remainder of the paper, we will omit the specific t-test used to obtain p-values. p-values are implicitly assumed to have been obtained by either Welch's t-test or the Alexander-Govern test. Because we have only applied one null hypothesis per statistical test, our statistical tests do not suffer from the multiple testing problem.

3 Experiment Motivation and Overview

From our experiments and analysis, we wish to understand how biases inherent in both face recognition datasets and metric embedding networks impact the performance of face obfuscation systems. We answer three questions:

1. *Bias in Face Recognition*: **What is the baseline bias present in face recognition systems?** Our experiments concur with existing literature: when conditioning by demographic, there is a disparity face recognition system performance. Further, we identify that networks learn to identify skin-tone in early layers of the network.

2. *Effectiveness of Obfuscation*: **How does the strength of perturbation necessary to obfuscate a face depend on a face's demographic?** Our findings indicate face recognition systems are less robust to perturbations applied to faces from minority demographic groups. For minority demographic groups, the perturbation strength necessary to impersonate an identity is smaller compared to majority demographics. Consequently, obfuscated faces tend to remain classified as a member of the same demographic group. Performance disparities are evident between demographics in both targeted and untargeted obfuscation.

3. *Bias Mitigation*: **How do bias mitigation strategies applied during training affect the utility of face obfuscation?** We apply two bias mitigation strategies to the training of metric embedding networks. The first strategy is a training procedure defined by Xu et al. [16]. The second strategy is training on demographically balanced datasets. With such training, we see the resulting embeddings are less clustered. Performance disparities of the face obfuscation system with respect to demographics are also attenuated. The benefits of bias mitigation are not free: overall model accuracy is reduced.

3.1 Experimental Setup

Datasets and Models: We utilize the LFW and VGGFace2 datasets in our evaluation. In the white-box setting, we generate perturbed faces on the FaceNet model. In the black-box

setting, we test perturbed faces for transferability by evaluating them on seven different models. Three of the seven models are commercial face recognition APIs: Face++ [36], Azure Face [37], and AWS Rekognition [38]. The remaining four are pre-trained open source models: OpenFace [39], DeepFace [40], ArcFace [4], and VGGFace [6]. Each model uses convolutional neural network architecture similar to FaceNet. **Fundamental Comparison:** To determine the relationship between demographics and face obfuscation performance, we consider the performance of obfuscation targeting identities in the same demographic group as the source identity, and obfuscation targeting identities in demographic groups different from that of the source identity. In particular, we consider six demographic attributes, four for race (Black, White, Indian, Asian) and two for sex (female, male). 50 identities per attribute from LFW are sampled[3]. These images are referred to as source images. Source images are inputs to the untargeted attack. Two targeted attacks scenarios are considered:

- **Same demographic:** We choose 49 pairwise combinations of target identities from the same race/sex for each source identity. This sampling leads to 2450 source-target pairs of the same demographic.
- **Different demographic:** We subsample, uniformly at random, 15 target identities from each race group of which the source identity is not a member for a total of 45 target identities. For the sex demographic, we assemble 50 target identities from the opposite sex. This sampling leads to 2250 source-target pairs of the different races and 2500 source-target pairs of the different sex.

We generate untargeted adversarial examples for each of the 5,749 identities in the LFW dataset and their associated images. We generate targeted adversarial examples for earlier scenario's 300 identities and 80,000+ targeted examples corresponding to the 28,700 pairs of identities. If unstated within a caption, the chosen dataset is LFW and the chosen metric embedding network is FaceNet.

Obfuscation Techniques: We perform our evaluation using untargeted and targeted variants of face obfuscation systems, as described in section 2.3. Utilizing the Face-Off face obfuscation system [7] and the FaceNet model [3], we generate adversarial examples on a subset of LFW [5]. For sections 4.1 and 4.2 we use 0, 5, and 10 as our margin values (κ in eq. (6)). We also generate untargeted perturbations with the LowKey [8] procedure described in section 2.3.3. The ensemble used by LowKey includes two pre-trained ArcFace models [4] and two pre-trained Cosface models [41]. We report the obfuscation success rate as an indicator of perturbation effectiveness. In the targeted case, obfuscation success rate measures the proportion of perturbed faces which match their intended targets, so we expect targeted obfuscation success rate to decrease as a function of the threshold τ. Furthermore,

---

[3]The LFW demographic attributes were annotated using attribute classifiers described by Kumar et al. [17].

the distance between an embedding and its target directly controls obfuscation success. In the untargeted case, obfuscation success rate measures the proportion of perturbed faces that evade their source identity. We expect the untargeted obfuscation rate to increase as a function of the threshold τ.

3.2 An analytical model for face obfuscation

Consistent with usage of toy models in previous machine learning literature [42–44], we devise a hierarchical Gaussian distribution with which we explore fair face obfuscation. While our model does entirely capture neural network behavior, the model conveys intuition on how how discrepancies in demographic sampling lead to disparities we observe in face obfuscation utility. To this end, the analytical model consists of a k-component PCA and hierarchical Gaussian distribution. The k-component PCA, which utilizes projections onto the leading k principal components to perform its dimensionality reduction, is the theoretically tractable proxy for a nonlinear metric embedding network. Use of the hierarchical Gaussian is inspired by both the hierarchical nature of popular face recognition datasets and the embedding space geometry. Though all samples drawn from our simplified model are vectors, we use terms *identity* and *image* to draw parallels between the hierarchical nature of our probabalistic model and the hierarchical structure of existing face recognition datasets.

Within the hierarchical Gaussian are two mutually exclusive groups: group a and group b. Sometimes a placeholder g is used to represent a group $g \in \{a, b\}$. $\boldsymbol{\mu}_a \in \mathbb{R}^d$ and $\boldsymbol{\mu}_b \in \mathbb{R}^d$ denote the mean vectors for population groups a and b, respectively. Moreover, $\boldsymbol{\mu}_a = -\boldsymbol{\mu}_b$, $\|\boldsymbol{\mu}_a\|_2 = 1$, and $\|\boldsymbol{\mu}_b\|_2 = 1$. Figure 2 is a visual depiction of our analytical model.

The ith identity in group g is denoted by $\mathbf{v}_{g,i} \in \mathbb{R}^d$. The jth image representing identity $\mathbf{v}_{g,i}$ is denoted by $\boldsymbol{x}_{g,i,j} \in \mathbb{R}^d$. $\boldsymbol{\Sigma}_a \in \mathbb{R}^{d \times d}$ and $\boldsymbol{\Sigma}_b \in \mathbb{R}^{d \times d}$ are diagonal covariance matrices. Furthermore, $\boldsymbol{\Sigma}_a = \gamma \boldsymbol{\Sigma}_b$ where $\gamma \in \mathbb{R}^+$.

By construction, the γ parameter captures demographic imbalance in sampling. Let us assume, that if both groups a and b are sampled equally in a manner matching the natural distribution, then $\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}_b$. Without loss of generality, let us assume $\gamma \in (0, 1]$. As γ decreases, the sampling from the natural distribution deviates further from the underlying distribution. Disparities in embedding density become apparent: When group a has few samples, embeddings of identities within group a are close together, similar to how minority groups are depicted in fig. 3.

An identity $\mathbf{v}_{g,i}$ is drawn from the identity distribution $\mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. The identity distribution may be thought of as a hyperprior on images. An image $\boldsymbol{x}_{g,i,j}$ is drawn from $\mathcal{N}(\mathbf{v}_{g,i}, \beta \boldsymbol{I})$ where β is a positive, real-valued number. For each identity $\mathbf{v}_{g,i}$, exactly m images are drawn from $\mathcal{N}(\mathbf{v}_{g,i}, \beta \boldsymbol{I})$. Lastly, we denote by $\mathcal{D}_g$ the distribution of images in group g. Thus, given $\alpha \in (0, 1)$, we can represent, the hierarchical Gaussian distribution as $\mathcal{D} = \alpha \mathcal{D}_a + (1 - \alpha) \mathcal{D}_b$.
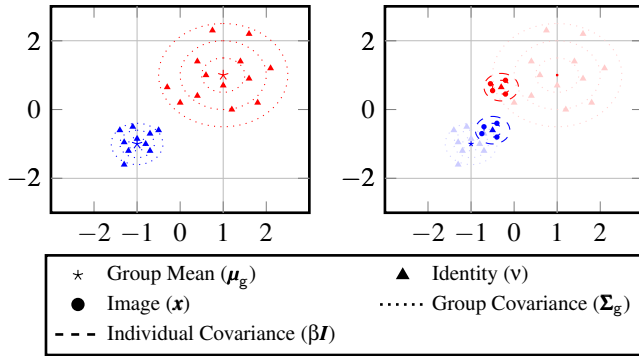
Figure 2: The hierarchical Gaussian distribution models patterns we observe with respect to faces in the embedding space. In the left plot, identities are drawn. Identities, μ_a and μ_b are depicted with a red star and a blue star, respectively. Identities $\mathbf{v}_{g,i}$ in group g are sampled from $\mathcal{N}(\mu_g, \sigma_g)$. Identities are depicted as triangles. In the right plot, images are drawn. These images $x_{g,i,j}$, depicted as circles, are drawn from $\mathcal{N}(\mathbf{v}_{g,i}, \beta I)$.

In the context of fair face obfuscation, we concern ourselves with how well the embedding represents a particular group from the lens of that group. This is captured by relative projection distance. The relative projection distance measures how well the *k*-component PCA embedding function represents a particular group, from the lens of that group:

Definition 3.1. Let S be a sample of images drawn from the overall synthetic distribution $\mathcal{D}$. The relative projection distance of a point x, a member of group g, with respect to the leading k principal components of a dataset S is denoted by $\rho_{\mathrm{rp},S,\mathrm{g},k}: \mathcal{X} \to \mathbb{R}^+$. More precisely:

$$\rho_{\mathrm{rp},S,\mathrm{g},k}(x) \triangleq \frac{\left\| \left(x - \sum_{i=1}^{k} \left\{ \frac{q_i^\top x}{\|q_i\|_2 \|x\|_2} q_i \right\} \right) \right\|_2}{\sum_{j=1}^{d} (\Sigma_\mathrm{g})_{jj}}, \qquad (7)$$

where the eigen-decomposition of the covariance of overall synthetic data distribution $\mathcal{D}$ is $\Sigma = Q \Lambda Q^\top$. Furthermore, Q may be decomposed as $Q = [q_1, \ldots, q_d]^\top$.

The numerator of relative projection distance is the norm of the portion of sample x which is not captured by f_k. The denominator represents a group specific normalization factor representing the overall variance within all identity vectors $\mathbf{v}_{g,i}$. It is this normalization factor which allows us to show how error can be measured in the context of group g.

4 Fairness on Face Obfuscation Systems

Utilizing the experimental procedure from section 3.1, we answer the three questions appearing at the beginning of section 3. Our findings highlight demographic disparities in face obfuscation systems. In particular, we show that it is easier to impersonate identities in the same demographic than it is to impersonate identities in a different demographic.

	Pre-trained Facenet					
$\mathrm{TPR}_{0.001}$	.9618	.8516	.9594	.8536	.9242	1.000
AUC	.9994	.9977	.9996	.9981	.9995	1.000
$\mathrm{TPR}_{0.001}$	.9678	.9436	.9732	.9448	.9742	1.000
AUC	.9995	.9988	.9993	.9998	.9999	1.000
N	10000	5000	10000	2500	1240	20
	Male	**Female**	**White**	**Asian**	**Black**	**Indian**

▨ Same Demographic ☐ Any Demographic

Table 1: Matching accuracy on LFW embeddings generated by FaceNet. Pairs within only the same demographic have lower accuracy compared to pairs matching any demographic. *N*: number of pairs.

4.1 Error Disparity Among Groups

Face recognition systems are known to be biased with respect to population demographics. We show such demographic disparities exist in the embedding space, and can be traced through the learning process of the embedding network.

Face Recognition Empirical Performance: Table 1 demonstrates the demographic disparities in face recognition. When evaluating matching performance $\mathrm{TPR}_{0.001}$ on the pretrained FaceNet model, pairs of identities selected within the same demographic group perform worse when compared to pairs selected without any such demographic restriction (table 1). Because identities in the same demographic group are closer together, as seen in fig. 3, false accepts are more likely thereby lowering $\mathrm{TPR}_{0.001}$. This idea also explains demographic disparities in matching performance. With the exception of the Indian demographic group, which is too small to source any significant conclusions, minority groups have lower matching performance. This is attributed to the tightly clustered embeddings we observe for minority demographic groups.

Intuition from Analytical Model: Understanding why embedding networks have disparate demographic-wise behavior is intractable given the state of current literature in neural network analysis, so we turn to our analytical model and PCA for intuition. Given our interest in studying the impact of the frequency of each group in a training set on the efficacy of the learned embedding network, we formulate a proposition which relates a relative projection distance, our experiments, and our analytical model:

Proposition 4.1. For fixed μ_a and fixed Σ_a, as γ approaches 0, the relative projection distance (defined in eq. (7)) of examples in group a increases.

This proposition suggests minority demographic groups are poorly represented by metric embedding networks. The magnitude of performance disparities can be explained by γ, the parameter capturing demographic imbalance in sampling. For smaller values of γ, the sampled distribution does not represent the natural distribution of the minority group. For the metric embedding network, such poor representation leads the embedding network to have trouble discerning identities in
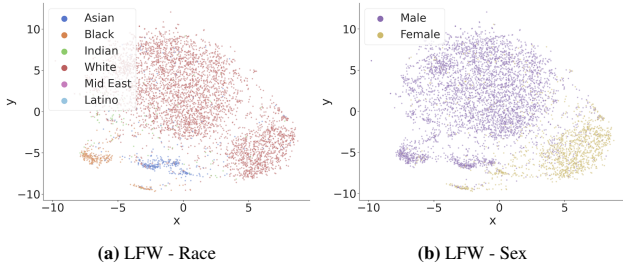
(a) LFW - Race

(b) LFW - Sex

Figure 3: t-SNE [28] of LFW embeddings generated by FaceNet.

(a) Block 35 Activation 3

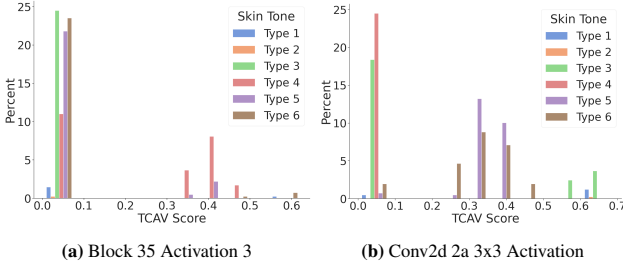**(b)** Conv2d 2a 3x3 Activation

Figure 4: TCAV score distribution of 408 identities in VGGFace2. Figures 4a and 4b learn concepts for lighter and darker skin tones, respectively.

the minority demographic group. Hence, we see more errors and ease of impersonation for members of the minority group. Proposition 4.1 is discussed further in our long-form report.[4]

What the Metric Embedding Network Learns: To understand *what* the metric embedding network architecture learns, we plot the resulting embedding structure of FaceNet using t-SNE [28] in fig. 3. Embedding networks trained without explicit demographic information can discern demographic groups in the embedding space. For each demographic group, separate clusters appear within the embedding space. The male and female clusters appear almost linearly separable.

Where the Metric Embedding Network Learns: Given the demographic-wise clustering behavior we see in fig. 3, we determine the degree to which the network itself contributes to observed clustering behavior. We utilize TCAV [30] to investigate if intermediate layers of the network learn to distinguish demographic attributes. As it was originally defined for classification networks, we retrofit the TCAV framework to metric embedding networks; we associate each identity with an anchor embedding, corresponding to its embedding centroid. We then use the ℓ_2 distance between the embedding of an input face and its anchor embedding to estimate the gradient from the output layer to the relevant activation layer.

We use skin tones as the demographic concepts. We annotate the skin tones with the Fitzpatrick scale [45], depicted in table 3. At each activation layer, we train linear classifiers that distinguish the layer activation according each skin tone. Each classifier is trained on 6 skin tones, each containing 75 images. The TCAV score associated with each identity is the

proportion of images for which the dot product of this linear classifier and gradient is positive. Nonzero TCAV scores indicate a concept heavily utilized in embedding construction.

Our evaluation involves annotated subsets of the VG-GFace2 dataset, one containing 408 identities and another with 4102 identities, with each identity containing 100 images of a person's face. The subset with 408 identities contains a balanced subset of skin tones, whereas the subset with 4102 mainly consists of faces with type 4 and type 5 skin tones. Only the TCAV scores with exactly matching skin-tone annotations are reported in fig. 4. We see high utilization of the skin tone concepts by two early layers in the network. We see that darker skin tones are learned separately (Conv2d 2a) than lighter skin tones (Block 35). This suggests metric embedding networks differentiate between skin tones in early layers.

4.2 Effectiveness of Obfuscation

After characterizing the baseline disparities in face recognition systems, we study the subsequent disparities in face obfuscation. We assess, for each demographic group, the difficulty with which identities may be impersonated. We show embedding space geometry induces demographic disparities in the perturbation norm, $\|\boldsymbol{\delta}\|_2$, necessary to successfully impersonate an identity. Further, we study the demographic disparities in the obfuscation success rates. These disparities are present in both the white-box and black-box settings.

Perturbation Norms: To detect disparities in the difficulty of face obfuscation, we examine the norms of perturbations $\|\boldsymbol{\delta}\|_2$, needed to successfully impersonate an identity. Using the adversarial examples generated by targeted attacks, we depict in fig. 5 the distribution of perturbation norms conditioned by demographic. We observe that the perturbation strength necessary to successfully impersonate an identity depends on demographic. Further, we observe that for each demographic group, the strength of the perturbation necessary to impersonate an identity is larger if the target identity is in a demographic different from the source identity. We put forth null hypothesis 4.2 which formally addresses demographic disparities in perturbation strength.

Null Hypothesis 4.2.

For each demographic group, the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary for targeted obfuscation is identical.

A formal analysis rejects null hypothesis 4.2: Differences in perturbation strength for impersonation of identities in the same and different sex demographic group, and the same and different race demographic group are statistically significant: p-values do not exceed 3.34×10^{-28}. Furthermore, there is a statistically significant difference between perturbations targeted within the same demographic and perturbations targeted outside the demographic for the Male, Female, Asian and White population groups with p-values of 6.61×10^{-24}, 2.00×10^{-26}, 0.00110, and 0.00180, respectively. These are
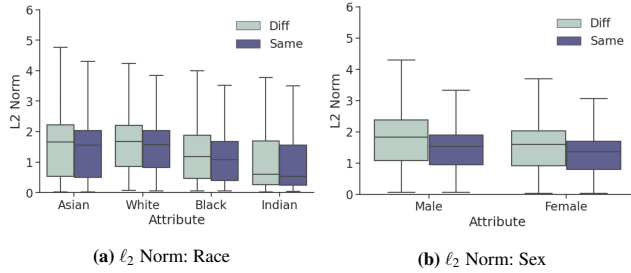
---

[4]Long-form report: arxiv.org/abs/2108.02707

(a) ℓ_2 Norm: Race

(b) ℓ_2 Norm: Sex

Figure 5: Distribution of perturbation norms generated by CW [46] on LFW.

(a) Targeted: Different Race (Left) vs. Same Race (Right)

(b) Targeted: Different Sex (Left) vs. Same Sex (Right)

Figure 6: Targeted obfuscation success on FaceNet in a white-box setting.

the largest four demographic groups in our dataset. Compared to the Black and Indian minority groups, we expect a more significant difference in perturbation strength necessary to impersonate identities in the same demographic and that which is needed to impersonate identities in different demographics.

Obfuscation Success: Figure 6 shows that targeted perturbations' success rates are dependent on the demographic attribute. Faces with the White race attribute exhibit a higher obfuscation success rate. These results also hold in the black box settings: similar trends exist for embeddings generated on OpenFace, Deepface, ArcFace, and VGGFace models. Success rates for the OpenFace model are shown in fig. 8, and success rates for the remaining black box models may be found in our long-form report. We observe examples generated on faces from the majority group transfer better than those of minority groups. We conjecture that the larger perturbation norms of such faces contribute to improved transferability rates. This observation is consistent with an observation from Face-off [7]; increasing the norm of perturbations improved transferability to black-box models.

In addition to offline models, this trend also holds for commercially available online face recognition APIs. We test the success of the perturbed faces against the Face++, Amazon AWS Rekognition, and Microsoft Azure Face APIs. In fig. 7, we present the CCDF of confidence scores for untargeted examples generated tested by Azure. For race, on a fixed obfuscation success rate we observe a generally higher confidence in impersonation for White faces. In contrast, Black faces see a generally lower confidence in impersonation for that same obfuscation success rate. Interestingly, for the sex attribute, we generally observe a slightly higher confidence for females. By examining the distance between an embedding and its target, we gain insight into the factor which directly controls obfuscation success rate. We put forth a null hypothesis to formally test our observations.

Null Hypothesis 4.3. Across all source demographic groups, the distribution of distances between obfuscated embeddings and their targets is identical.

This null hypothesis is easily rejected: Differences in distance between face embeddings and targets are statistically significant regardless of target demographic: The *p*-values do
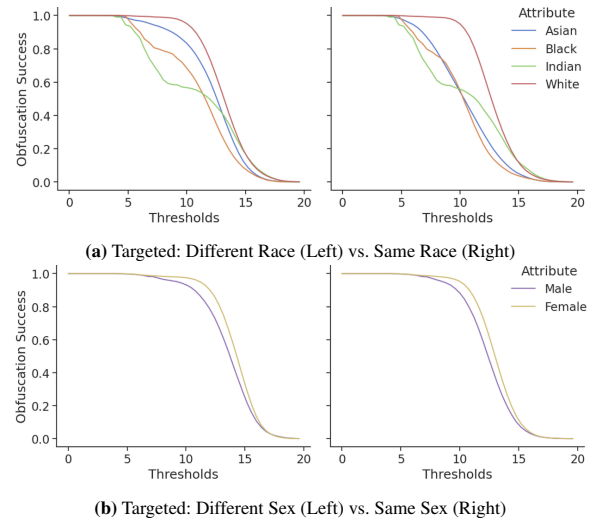
not exceed 0.0395. These results confirm the bias in obfuscation performance previously identified. From a practical perspective, our results may suggest that users should select an identity to impersonate of the same race or sex in order to optimize face obfuscation system utility. This is counterproductive to the user's privacy for two reasons: 1) adversarial perturbations with smaller ℓ_p norms will struggle in transferring to other models and 2) a user may leak demographic information to an adversary.

Intuition from Analytical Model: From fig. 5, we observe demographic disparities in the strength of perturbation necessary to successfully obfuscate an image. We also showed it is generally easier to impersonate identities in the same demographic group than it is to impersonate identities in different demographic groups. This translated into the disparate success rates observed in figs. 6 to 8. We use our analytical model to provide some mathematical intuition on this phenomenon.

Given an example image $\boldsymbol{x}$ in group g, we study how difficult it is to construct a perturbation $\boldsymbol{\delta}$ such that $\boldsymbol{x} + \boldsymbol{\delta}$ successfully impersonates an identity outside of group g. The analytical model is a natural medium in which quantifying the necessary strength of such a perturbation $\boldsymbol{\delta}$ may occur.
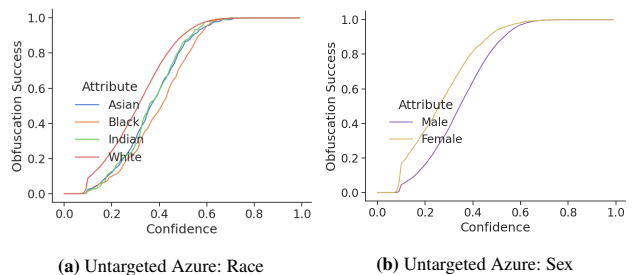

(a) Untargeted Azure: Race

(b) Untargeted Azure: Sex

Figure 7: Untargeted obfuscation success on Microsoft Azure Face API.

(a) OpenFace: Different Race (Left) vs. Same Race (Right)

(b) OpenFace: Different Sex (Left) vs. Same Sex (Right)
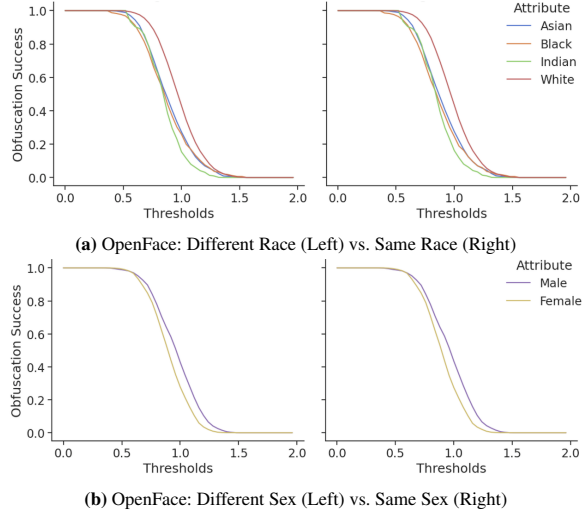
Figure 8: Targeted obfuscation success on OpenFace in a black-box setting.

For the sake of simplicity, consider the 1-PCA embedding function $f_1 : \mathbb{R}^d \to \mathbb{R}$. Let sample $\boldsymbol{x}$ be drawn from the overall synthetic data distribution $\mathcal{D}$ and without loss of generality assume this $\boldsymbol{x}$ is a member of group a. Denote by p_Q the probability density function for a probability distribution Q. We will assume group a is the minority group and so we assume that $\gamma \le 1$, $p_{\mathcal{D}_a}[\boldsymbol{\mu}_a] > p_{\mathcal{D}_b}[\boldsymbol{\mu}_a]$, and $p_{\mathcal{D}_b}[\boldsymbol{\mu}_b] > p_{\mathcal{D}_a}[\boldsymbol{\mu}_b]$. Given a perturbation $\boldsymbol{\delta}$, we assume that perturbation $\boldsymbol{\delta}$ is norm-bounded and in the direction of group b, i.e. $\frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2}$. That is, we assume $\|\boldsymbol{\delta}\|_2 \le \varepsilon$ where ε is a non-negative real number. We quantify the values of ε for which the following optimization problem is infeasible, thereby guaranteeing $\boldsymbol{x} + \boldsymbol{\delta}$ impersonates[5] an identity in group a:

$$\min_{\boldsymbol{\delta}: \|\boldsymbol{\delta}\|_2 \le \varepsilon} \|\boldsymbol{\delta}\|_2 \quad \text{s.t.} \quad \mathbb{P}_{\boldsymbol{x} \sim \mathcal{D}_b}[f(\boldsymbol{x} + \boldsymbol{\delta})] > \mathbb{P}_{\boldsymbol{x} \sim \mathcal{D}_a}[f(\boldsymbol{x} + \boldsymbol{\delta})] \quad (8)$$

$$\boldsymbol{\delta} = \eta \cdot \frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2} \text{ where } \eta \in \mathbb{R} \quad (9)$$

For notational compactness, we denote:

$$a = f(\boldsymbol{x}) \text{ and } b = f\left(\frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2}\right).$$

This optimization objective in eq. (9) guaranteed to be infeasible when:

$$\varepsilon < \max\left\{0, \frac{2b(\gamma - 1)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}} + a\gamma + a + f(\boldsymbol{\mu}_b)(1 - \gamma)}{b(\gamma - 1)}\right\} \quad (10)$$

This result is further detailed in appendix A.

Therefore we conclude that $\boldsymbol{x} + \boldsymbol{\delta}$ impersonates an identity in a when inequality (10) holds. Furthermore, note that the bound in inequality (10) is not tight. The bound loosens as

---

[5] Here, we assume it is possible for an example to impersonate itself.

γ approaches 0. Within inequality (10), we notice that for a fixed Σ_a, $\boldsymbol{\mu}_a$, and $\boldsymbol{x}$ which impersonates an identity in group a, as γ approaches 0, the set of perturbations $\boldsymbol{\delta}$ for which $\boldsymbol{x} + \boldsymbol{\delta}$ still impersonates an identity in group a, decreases in size.

Relating to experiments in this section on existing face recognition datasets, it is the disparities in sampling which affect the strength of the perturbation necessary to impersonate an identity in a different demographic group. These disparities are captured by γ. This analysis agrees with perturbation norms in fig. 5: impersonating an identity in a different demographic is more difficult than impersonating an identity in the same demographic group.

Stability Properties of Face Obfuscation: We further study the impact of γ on the stability of networks. By examining estimates of the local Lipschitz constants, we investigate the stability of metric embedding networks in relation to the demographic distribution of their training sets. A classifier's margin scales inversely with the Lipschitz constant, making classifiers with high local Lipschitz constants less stable and easier to attack [47–51]. We use the RecurJac [52] and Fast-Lin [53] bound algorithms to upper-bound local Lipschitz constants within small neural networks trained on datasets with uniform and non-uniform distributions of demographic groups. The level of uniformity in the demographic distribution is captured by the parameter γ. When the demographics are sampled uniformly, then $\gamma = 1$. If we assume the minority demographic group is a, then greater disparities in sampling mean γ tends to 0. Figure 16 in appendix B.2 shows the distributions of the upper bounds on the estimated local Lipschitz constants for the non-uniform and uniform classifiers trained on 600 identities; this distribution is plotted for the training set. We observe larger upper bounds in non-uniform sampling of each demographic group. Further, identities corresponding to minority demographic groups have larger upper bounds on local Lipschitz constants than do majority identities. These results suggest networks generalize worse for demographic groups which are a minority in the training set, thus networks are less robust to perturbation for certain demographics.

Nearest Neighbors: In addition to performance of source-target matching, we consider the generalized setting in which a source image must be matched to the best candidate among multiple targets. This generalized scenario is studied by measuring accuracy in a nearest neighbors model. In the experiment, accuracy refers to the rate at which embeddings designed to impersonate are still classified as the original source identity. The training points for the nearest neighbor model are embedding centroids for each identity.

For each demographic group, when impersonating identities in the same demographic group, generated embeddings are accurate on at least 66.7% of examples. For each demographic group, when impersonating identities in different demographic groups, the generated embeddings are accurate on at least 62.9% of examples.

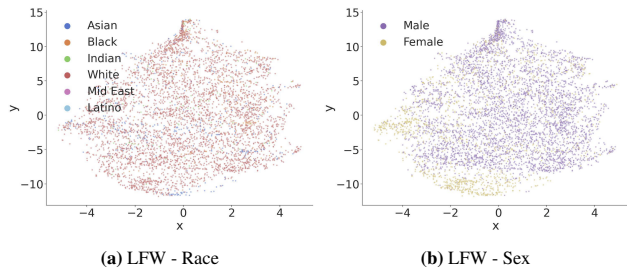As expected, the accuracy decreases or remains approxi-

(a) LFW - Race   **(b)** LFW - Sex

Figure 9: t-SNE of the embeddings for the LFW dataset. Embeddings are generated on a FaceNet model trained by Xu et al. procedure.

(a) ℓ_2 Norm: Race   **(b)** ℓ_2 Norm: Sex
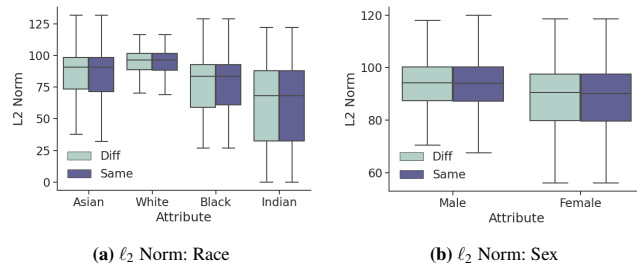
Figure 10: Distribution of perturbation norms generated by the CW attack on FaceNet model trained with Xu et al. procedure.

mately the same for all demographic groups when impersonating identities from different demographics. The general trend is a decrease in accuracy.

4.3 Bias Mitigation

Finally, we examine two bias mitigation techniques and study their impact on the demographic disparities of face obfuscation. The first technique is a training procedure designed by Xu et al. [16] to promote fairness, and the second technique is dataset balancing by demographic. Models of the FaceNet architecture, outputting embeddings in $\mathbb{R}^{512}$ are trained. Because we are interested in characterizing the best-case obfuscation performance in the presence of bias mitigation strategies, we focus exclusively on the white-box setting.

4.3.1 Training Procedure of Xu et al.

Xu et al. propose a technique, that within the context of face obfuscation, is designed to mitigate disparate susceptibility of individuals to impersonation. The training technique decomposes overall error into the natural error and boundary error. *Natural error* refers to the error on examples prior to the addition of a perturbation, and *boundary error* refers to the net increase in error induced when perturbing the natural example. The Xu et al. training procedure incentivizes training of an accurate model such that natural error and boundary error are each roughly equivalent across all identities.

Visualizing Embeddings: Upon examining the t-SNE plots in fig. 9, we observe the training procedure of Xu et al. does

(a) Different (Left) Race vs. Same (Right) Race

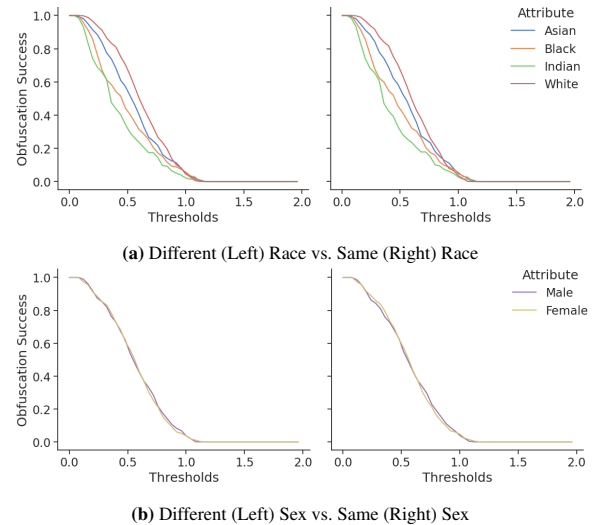**(b)** Different (Left) Sex vs. Same (Right) Sex

Figure 11: Targeted white-box obfuscation success on a FaceNet model trained with Xu et al. procedure.

mitigate sources of bias in the embedding space geometry: Demographic clusters are less distinct and embeddings for each demographic group are somewhat interspersed amongst each other. However, clustering is not eliminated entirely; there is a cluster of females in fig. 9b, though many female embeddings are interspersed among embeddings corresponding to males.

Analyzing Obfuscation: Embedding space geometry impacts the perturbation norms needed for impersonation: Compared to the pre-trained model, we see less disparity between the perturbation norms needed to impersonate identities in the same and different demographic groups in fig. 10. We put forth a formal null hypothesis to assess our observations.

Null Hypothesis 4.4. For the model trained by the Xu et al. procedure, the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity in the same demographic is identical to the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity in a different demographic group.

Though we perceive a significant reduction in bias due to the Xu et al. training procedure, we can partially reject null hypothesis 4.4. There is a statistically significant difference between perturbations targeted within the same demographic and perturbations targeted outside the demographic for only the White demographic, with a *p*-value of 9.61×10^{-78}.

We also observe that the perturbation strength necessary to impersonate an identity has a dependence on demographic. We put forth a formal null hypothesis to test this observation:

Null Hypothesis 4.5. For the model trained by the Xu et al. procedure, the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity is identical for each source demographic group.

As anticipated, we can reject this null hypothesis. Differences in perturbation strength between demographic groups

are statistically significant for impersonation targeting identities in the same and different sex demographic groups, and the same and different race demographic groups. The *p*-values do not exceed 2.15×10^{-90}.

The significance of such disparities manifests itself in obfuscation obfuscation success rates as depicted in fig. 11, especially for success rates conditioned by race. Much like the pre-trained FaceNet, the White race group tends to have the highest obfuscation success rate. Looking at the sex demographic groups, obfuscation success rates appear to be similar. It is unclear what structurally about Xu et al procedure yields a model with obfuscation success rates disparate with respect to race, yet obfuscation success rates conditioned by sex are similar. For completeness, we put forth a null hypothesis to test the significance of this observation:

Null Hypothesis 4.6. For the model trained by the Xu et al. procedure: Across all source demographic groups, the distribution of distance between each obfuscated face embedding and its target is identical.

We partially reject this null hypothesis: Only when targeting identities outside the identity of source identity demographic group, differences in distance between face embeddings and targets are statistically significant. The *p*-values are 2.51×10^{-24} and 2.39×10^{-7} when targeting identities in a different race and sex, respectively. Given these results, it is apparent that the Xu et al. training procedure does mitigate bias to some extent, but is far from a perfect solution.

Accuracy Trade-off: The observed bias mitigation also comes at the cost of model accuracy. We measure these costs by examining the matching performance and through accuracy of a nearest neighbors model. The matching performance, $\text{TPR}_{0.05}$, of a model trained with the Xu et al. procedure is depicted in the middle subtable of table 2. Matching performance of the reference FaceNet is depicted in the upper subtable of table 2. The matching performance across most demographics is reduced by 1-4% compared to the reference model. When measuring training accuracy of a nearest neighbors classifier, conditioned by demographic, the lowest accuracy in a model trained with the Xu et al. procedure is 55.2%. This compares poorly to the reference FaceNet model, for which the worst natural accuracy, when conditioning by demographic, is 95.4%. Full results for nearest neighbors accuracy can be found in our long-form report.

4.3.2 Dataset Balancing

The second bias mitigation strategy we study is that of training face recognition models on demographically balanced datasets. To do so, we train three metric embedding networks of the FaceNet architecture. Each training procedure is identical, with the exception of the training dataset. The network trained on the first dataset acts as the reference: It is trained on the entirety of the VGGFace2 [6] training split, a total of 8631 identities. Experiments for this model appear in the appendix.
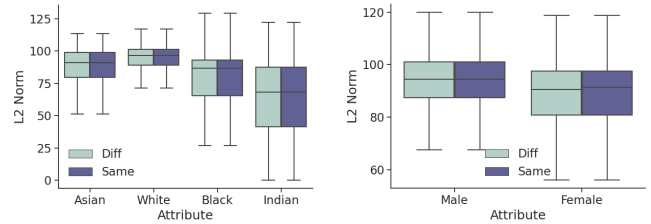
(a) Race-Balanced FaceNet on race data **(b)** Sex-Balanced FaceNet on on sex data

Figure 12: The distribution of adversarial perturbation sizes generated using the CW attack on FaceNet trained on demographically balanced datasets.

The remaining two datasets are used to train comparison models. One such dataset is Sex-Balanced VGGFace2, which contains a total of 4866 identities. The Sex-Balanced VGGFace2 dataset was created by removing original identites to create a training set which contains an equal number identities for the demographic of interest. Instead of performing data augmentation, we remove examples from VGGFace2 so as to not introduce any artifacts. Similarly, we construct a third dataset, which consists of 307 identities for each of the race subgroups. We call this dataset, containing 1842 total identities, Race-Balanced VGGFace2. Race-Balanced VGGFace2 and Sex-Balanced VGGFace2 are collectively referred to as "(demographically) balanced datasets". Race-Balanced VGGFace2 and Sex-Balanced VGGFace2 are the training sets used to obtain Race-Balanced FaceNet and Sex-Balanced FaceNet, respectively. We observe that the Balanced FaceNets have less demographic-wise disparity in both face recognition and face obfuscation performance.

Visualizing Embeddings from Models: Upon examining the t-SNE plots in fig. 13, we notice that minority demographic groups have larger clusters than do the clusters for the same demographic groups generated on the reference model. Compared to embeddings generated by the model trained with the Xu et al. procedure, demographic clusters generated by Sex-Balanced and Race-Balanced are more separate and distinct. Further, each demographic group generally appears more distinct.

Analyzing Obfuscation: Given the distinct demographic clusters we see in balanced embedding space geometry, we do not expect to see much bias mitigation. Before drawing such a conclusion, we formally test a null hypothesis:

Null Hypothesis 4.7. For models trained on balanced datasets, the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity in the same demographic is identical to the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity in a different demographic group.

We are able to partially reject null hypothesis 4.7. There is a statistically significant difference between perturbations targeted within the same demographic and perturbations targeted outside the demographic for the only male, female and white demographic groups with *p*-values of 3.20×10^{-6},

5.28×10^{-5}, and 1.65×10^{-30}, respectively.

We observe disparities in the perturbation norms required for faces in each demographic group to be successfully obfuscated. We propose a null hypothesis to test this observation:

Null Hypothesis 4.8. For models trained on balanced datasets, the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity is identical for each source demographic group.

We are also able to reject this null hypothesis: For models trained on both sex-balanced and race-balanced data, differences in perturbation strength between demographic groups are statistically significant for the same sex demographic group, different sex demographic group, same race demographic group, and different race demographic group. The p-values do not exceed 2.04×10^{-44}.

The disparities in perturbation norm also manifest themselves in obfuscation success rates as depicted in fig. 14. Such success rates are disparate when conditioned by both sex and race. Much like the pre-trained FaceNet model, when conditioned by race, the White demographic group has the highest success rate. Unlike the pre-trained FaceNet, the Female demographic group has higher obfuscation success rate than the male demographic group. It is unclear why the Female demographic group now has higher obfuscation success rates. We put forth a null hypothesis to formally test our observations:

Null Hypothesis 4.9. For models trained on balanced data: Across all source demographic groups, the distribution of distance between each obfuscated face embedding and its target is identical.

The null hypothesis is easily rejected: Differences in distance between face embeddings and targets are statistically significant in all settings. The p-values do not exceed 1.13×10^{-17}. Given the significant disparities in face obfuscation performance, it is apparent dataset balancing is ineffective at mitigating bias in face obfuscation.

Natural Accuracy Trade-off: The matching performance, $\text{TPR}_{0.05}$, of a models trained on balanced datasets is the lowest sub-table in table 2. The matching performance across most demographics is reduced by 0.5-3% compared to the reference model. When measuring training accuracy of a nearest neighbors classifier, conditioned by demographic, the lowest accuracy in a model trained with demographically balanced training data 94.5%. This is only a very slight performance reduction relative to the reference FaceNet model.

5 Discussion

Potential Remedies: In section 4.3, we discussed the impact of two bias mitigation strategies on face obfuscation. First, we examined a model training procedure by Xu et al. This procedure does not explicitly optimize for fairness, but instead implicitly relies on a documented connection between adversarial learning and fairness [43]. The second bias mitigation procedure involved training FaceNet models on balanced

Reference FaceNet Model					
.9046	.8620	.9164	.9152	.8565	.9000
.9808	.9864	.9826	.9820	.9711	1.000
.9164	.9328	.9294	.9720	.9919	.8000
.9828	.9864	.9847	.9938	.9967	1.000
Training Procedure of Xu et al.					
.8984	.8180	.8952	.8776	.8258	.6000
.9774	.9631	.9782	.9796	.9644	1.000
.9036	.9408	.9104	.9768	.9806	.9000
.9806	.9866	.9816	.9945	.9955	1.000
Balanced Training					
Sex-Balanced		**Race-Balanced**			
.9184	.8804	.8716	.9160	.8726	.8000
.9837	.9733	.9719	.9825	.9738	.9800
.9330	.9244	.8978	.9672	.9823	.7000
.9867	.9842	.9772	.9938	.9950	1.000
10000	5000	10000	2500	1240	20
Male	**Female**	**White**	**Asian**	**Black**	**Indian**

(Row labels: $\text{TPR}_{0.05}$, AUC alternating; bottom row N)

☐ Same Demographic ☐ Any Demographic

Table 2: The matching performance on LFW on a reference FaceNet model (top), FaceNet trained with Xu et al. procedure (model), and FaceNet models on demographically balanced data.
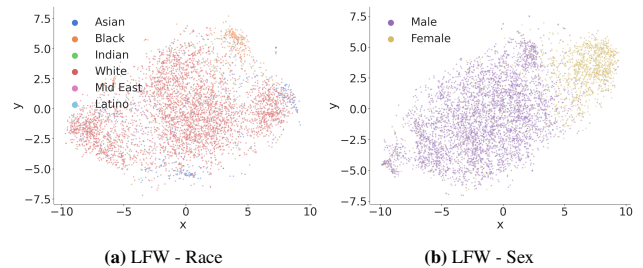

(a) LFW - Race

(b) LFW - Sex

Figure 13: t-SNE of embeddings for the LFW dataset. Figure 13a is generated on Race-Balanced FaceNet. fig. 13b is generated on Sex-Balanced FaceNet.

datasets. The Xu et al. procedure mitigated more bias than the balanced Facenet models, but neither procedure completely mitigated disparities in face obfuscation performance. A training procedure which better mitigates performance disparities in face obfuscation efficiently, is a topic for future work.

Beyond bias mitigation in the manner we studied in this paper, balancing performance across combinations of multiple demographic attributes might be of interest. There exists work addressing this problem: Serna et al. propose Sensitive Loss, a "discrimination-aware" Triplet Loss tuning procedure for pre-trained models [54]. This tuning procedure involves adding a layer to the network, then tweaking this layer using only identities drawn from the same demographic group.

Threats to Validity: The intent of section 3.2 is the creation of a tractable analytical model for embedding spaces and dimensionality reduction in general. Perhaps the most significant limitation of this analytical model is that PCA is a linear embedding function and is incapable of capturing non-linear effects present in metric embedding networks. Another threat to validity is the scarcity of publicly available face recogni-
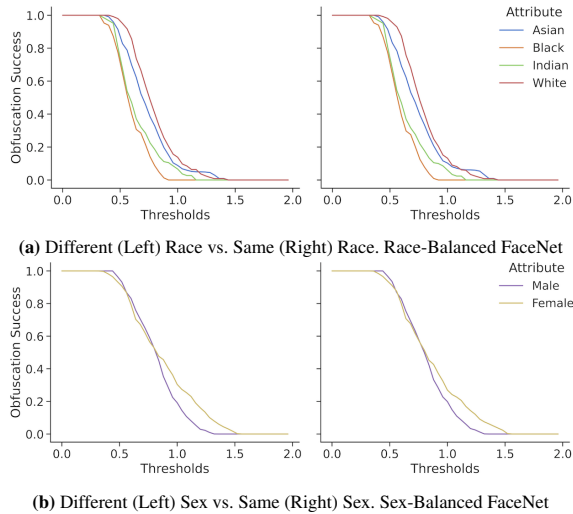
(a) Different (Left) Race vs. Same (Right) Race. Race-Balanced FaceNet

(b) Different (Left) Sex vs. Same (Right) Sex. Sex-Balanced FaceNet

Figure 14: Targeted obfuscation success evaluated on Demographically Balanced FaceNets in a white-box setting.

tion datasets; datasets with labeled demographic attributes are even more rare. Such datasets are rather small, as such, our results may not generalize to datasets orders of magnitude larger than the available face recognition datasets.

6   Related Work

Buolamwini and Gebru study commercially available face recognition datasets and classifiers [14]. Their findings indicate that prominent commercial face classifiers exhibit disparate performance across demographic groups. Follow-up research has attempted to address these demographic disparities through balanced datasets [55, 56]. Such datasets improve model generalization but do not resolve disparities completely. These findings are consistent with results in appendix B.

Demographic disparities seem to contribute to the phenomenon of overlearning, a term coined by Song and Shmatikov [57]. Overlearning refers to the phenomenon where models implicitly learn to recognize sensitive patterns not part of the original learning objective. Metric embedding networks trained on faces do overlearn; they learn a demographically-aware dimensionality reduction on faces.

Cherepanova et al. [58] study the relationship between training data, test data, population demographics and face recognition performance. The authors, independent of our results, observe that models trained on demographically balanced datasets are not bias-free. They discuss neither face obfuscation, nor impacts on privacy and security.

Nanda et al. discuss the relationship between fairness and robustness of face recognition [15]. Our work differs in several respects: First, there is a distinction between the study of robust metric learning. Our paper explores bias present in the embedding space and how it affects recently proposed face obfuscation systems. Through an analytical model in

section 3.2, we show the amount of tolerable perturbation in a face recognition system depends on class imbalances. Second, we examine face obfuscation in the black-box setting; these experiments are performed on both open-source and commercial models as depicted in fig. 7. The results demonstrate transferability of biases in adversarial perturbations; we note the negatively-affected demographics can differ between the white-box and black-box settings depending on the perturbation strength. Third, we investigate root causes for demographic-wise performance disparities via TCAV. We also explore dataset balancing and the Xu et al. training procedure as bias mitigation techniques.

Cilloni et al. study a targeted face obfuscation which attempts to find the minimum strength successful perturbation via optimization procedure regularized by the a well-established image similarity scoring function known as DSSIM. Modulo the regularizer, our Carlini-Wagner based experiments achieve the same purpose. Rajabi et al [59]. consider two methods of face obfuscation: the first is a universal ensemble perturbation, an untargeted obfuscatory method designed to be transferable; we explore transferability of untargeted perturbations in section 4.2 and fig. 8. The second technique is cryptographic in nature. Finally, a Master's thesis by Qin [60] evaluates discrepancies in a face obfuscation system called FAWKES [26], conditioned on skin tones. Qin finds differences in perturbation visibility for certain demographics. In comparison, our work characterizes these discrepancies and their impact on face obfuscation systems.

7   Conclusion

Face recognition systems have seen increased usage in online settings at the cost of heightened privacy concerns. Researchers have proposed face obfuscation systems that leverage evasion attacks against metric embedding networks. Our results show that, in an effort to mitigate such privacy concerns, face obfuscation systems have performance characteristics that depend on demographic information, thereby creating a new privacy incursion. Such performance characteristics can not only leak demographic membership information and decrease the performance of face obfuscation among underrepresented demographic groups. Imbalances in training set demographics are only partially to blame for this privacy leak: remaining causes are yet to be discovered. To mitigate the effects of this incidental privacy leak, we must develop both loss functions for training fair metric embedding networks and techniques to characterize if such privacy leaks will occur.

Acknowledgement

References

[1] L. Feiner and A. Palmer, "Rules around facial recognition and policing remain blurry," *CNBC Tech*, 2021.

[2] A. Roussi, "Resisting the rise of facial recognition," *Nature*, 2020.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[4] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[5] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[7] V. Chandrasekaran, C. Gao, B. Tang, K. Fawaz, S. Jha, and S. Banerjee, "Face-off: Adversarial face obfuscation." in *2021 Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 2, 2021, pp. 369–390.

[8] V. Cherepanova, M. Goldblum, H. Foley, S. Duan, J. P. Dickerson, G. Taylor, and T. Goldstein, "Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition," *CoRR*, vol. abs/2101.07922, 2021. [Online]. Available: https://arxiv.org/abs/2101.07922

[9] I. Evtimov, P. Sturmfels, and T. Kohno, "Foggysight: A scheme for facial lookup privacy," in *2021 Proceedings on Privacy Enhancing Technologies*. PoPETs, 2021.

[10] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[13] E. Radiya-Dixit and F. Tramèr, "Data poisoning won't save you from facial recognition," in *ICML Workshop on Adversarial Machine Learning (AdvML)*, 2021. [Online]. Available: https://arxiv.org/abs/2106.14851

[14] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a.html

[15] V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson, "Fairness through robustness: Investigating robustness disparity in deep learning," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 466–477. [Online]. Available: https://doi.org/10.1145/3442188.3445910

[16] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 11 492–11 501. [Online]. Available: https://proceedings.mlr.press/v139/xu21b.html

[17] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and Simile Classifiers for Face Verification," in *IEEE International Conference on Computer Vision (ICCV)*, Oct 2009.

[18] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 1735–1742.

[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

[20] A. Holmes, "533 million facebook users' phone numbers and personal data have been leaked online," *Business Insider Tech*, 2021.

[21] D. Harwell, "This facial recognition website can turn anyone into a cop — or a stalker," May 2021. [Online]. Available: https://www.washingtonpost.com/technology/2021/05/14/pimeyes-facial-recognition-search-secrecy/

[22] G. L. Goodwin, "Facial recognition technology: Federal law enforcement agencies should have better awareness of systems used by employees," GAO Report GAO-21-105309, Washington, DC: USA, 2021 [Online]. [Online]. Available: https://www.gao.gov/products/gao-21-105309

[23] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=Sys6GJqxl

[24] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[25] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 1528–1540.

[26] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, S. Capkun and F. Roesner, Eds. USENIX Association, 2020, pp. 1589–1604. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/shan

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[28] L. van der Maaten and G. Hinton, "Viualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.

[29] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15. MIT Press, 2003. [Online]. Available: https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf

[30] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2668–2677. [Online]. Available: http://proceedings.mlr.press/v80/kim18d.html

[31] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[32] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 214–226. [Online]. Available: https://doi.org/10.1145/2090236.2090255

[33] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, pp. 3315–3323, 2016.

[34] B. L. WELCH, "The Generalization of 'Student's' Problem When Several Different Population Variances Are Involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 01 1947. [Online]. Available: https://doi.org/10.1093/biomet/34.1-2.28

[35] R. A. Alexander and D. M. Govern, "A new and simpler approximation for anova under variance heterogeneity," *Journal of Educational Statistics*, vol. 19, no. 2, pp. 91–101, 1994. [Online]. Available: http://www.jstor.org/stable/1165140

[36] Face++, "Face++," https://www.faceplusplus.com.

[37] "Facial Recognition | Microsoft Azure," Jun. 2022, [Online; accessed 5. Jun. 2022]. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/face

[38] "Detecting and analyzing faces - Amazon Rekognition," Jun. 2022, [Online; accessed 5. Jun. 2022]. [Online]. Available: https://docs.aws.amazon.com/rekognition/latest/dg/faces.html

[39] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[40] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[41] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[42] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Mądry, "Adversarially robust generalization requires more data," *arXiv preprint arXiv:1804.11285*, 2018.

[43] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1929–1938. [Online]. Available: https://proceedings.mlr.press/v80/hashimoto18a.html

[44] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 1310–1320. [Online]. Available: https://proceedings.mlr.press/v97/cohen19c.html

[45] J. D'Orazio, S. Jarrett, A. Amaro-Ortiz, and T. Scott, "Uv radiation and the skin," *International journal of molecular sciences*, vol. 14, no. 6, pp. 12 222–12 248, Jun 2013. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23749111

[46] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[47] P. L. Bartlett, D. J. Foster, and M. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6241–6250.

[48] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring generalization in deep learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5949–5958.

[49] K. Scaman and A. Virmaux, "Lipschitz regularity of deep neural networks: Analysis and efficient estimation," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 3839–3848.

[50] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel, "Evaluating the robustness of neural networks: An extreme value theory approach," in *International Conference on Learning Representations (ICLR)*, may 2018.

[51] U. V. Luxburg and O. Bousquet, "Distance-based classification with lipschitz functions," in *J. Mach. Learn. Res.*, 2003.

[52] H. Zhang, P. Zhang, and C.-J. Hsieh, "Recurjac: An efficient recursive algorithm for bounding jacobian matrix of neural networks and its applications," in *AAAI Conference on Artificial Intelligence (AAAI), arXiv preprint arXiv:1810.11783*, dec 2019.

[53] T.-W. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, D. Boning, and I. S. D. A. Daniel, "Towards fast computation of certified robustness for relu networks," in *International Conference on Machine Learning (ICML)*, july 2018.

[54] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan, "Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning," 2020.

[55] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1548–1558.

[56] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, p. 64–73, Jan. 2016. [Online]. Available: https://doi.org/10.1145/2812802

[57] C. Song and V. Shmatikov, "Overlearning reveals sensitive attributes," *arXiv preprint arXiv:1905.11742*, 2019.

[58] V. Cherepanova, S. Reich, S. Dooley, H. Souri, M. Goldblum, and T. Goldstein, "A deep dive into dataset imbalance and bias in face identification," *arXiv preprint arXiv:2203.08235*, 2022.

[59] A. Rajabi, R. B. Bobba, M. Rosulek, C. V. Wright, and W. chi Feng, "On the (im)practicality of adversarial perturbation for image privacy," *Proceedings on Privacy Enhancing Technologies*, vol. 2021,

no. 1, pp. 85–106, 2021. [Online]. Available: https://doi.org/10.2478/popets-2021-0006

[60] S. Qin, "Bias and fairness of evasion attacks in image perturbation," Master's thesis, Central Washington University, 2021.

[61] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020, pp. 23–27.

[62] F. von Luschan, *Beiträge zur Völkerkunde der deutschen Schutzgebiete*. D. Reimer, 1897.

A Mathematics for the Analytical Model

To solve the optimization problem posed in eq. (9), we examine the likelihood function. Let $p_{\mathcal{D}_g}$ now denote the PDF of the image of distribution $\mathcal{D}_g$ as it appears in the 1-D PCA embedding space. We assume $\gamma \leq 1$. We aim to find the strength of the perturbation necessary to push an example $\boldsymbol{x}$ in a $\frac{p_{\mathcal{D}_a}[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta})]}{p_{\mathcal{D}_b}[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta})]}$ exceeds one, where:

$$1 \leq \left(\frac{p_{\mathcal{D}_a}\left[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta}) \mid \boldsymbol{\delta} \propto \frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2} \right]}{p_{\mathcal{D}_b}\left[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta}) \mid \boldsymbol{\delta} \propto \frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2} \right]} \right) \tag{11}$$

$$= \left((2\pi)^{-1/2} (\mathrm{f}(\boldsymbol{q}_1))^{-1} \right)$$
$$\times \exp\left\{ -\frac{1}{2} \left[\mathrm{f}\left(\boldsymbol{x} + \eta \frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2} - \boldsymbol{\mu}_a \right) \right]^2 (\mathrm{f}(\boldsymbol{q}_1))^{-2} \right\}$$
$$\left[\left((2\gamma\pi)^{-1/2} (\mathrm{f}(\boldsymbol{q}_1))^{-1} \right) \right.$$
$$\left. \times \exp\left\{ -\frac{\gamma}{2} \cdot \left[\mathrm{f}\left(\boldsymbol{x} + \eta \frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2} - \boldsymbol{\mu}_b \right) \right]^2 (\mathrm{f}(\boldsymbol{q}_1))^{-2} \right\} \right]^{-1} \tag{12}$$

$$\leq \exp\left\{ -\frac{1}{2} \left[\mathrm{f}\left(\boldsymbol{x} + \eta \frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2} \right) + \mathrm{f}(\boldsymbol{\mu}_b) \right]^2 \right\}$$
$$\times \exp\left\{ -\frac{\gamma}{2} \cdot \left[\mathrm{f}\left(\boldsymbol{x} + \eta \frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2} \right) - \mathrm{f}(\boldsymbol{\mu}_b) \right]^2 \right\}^{-1} \tag{13}$$

We now solve the following the following inequality for η

$$1 \leq \exp\left\{ -\frac{1}{2} \left[\mathrm{f}\left(\boldsymbol{x} + \eta \frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2} \right) + \mathrm{f}(\boldsymbol{\mu}_b) \right]^2 \right\}$$
$$\times \exp\left\{ -\frac{\gamma}{2} \cdot \left[\mathrm{f}\left(\boldsymbol{x} + \eta \frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2} \right) - \mathrm{f}(\boldsymbol{\mu}_b) \right]^2 \right\}^{-1} \tag{14}$$

After some algebra, we arrive at the following interval solution for η.

$$\eta > \frac{2b(1-\gamma)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}} + a\gamma + a + \mathrm{f}(\boldsymbol{\mu}_b)(1-\gamma)}{b(\gamma-1)} \tag{15}$$

AND

$$\eta < \frac{2b(\gamma-1)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}} + a\gamma + a + \mathrm{f}(\boldsymbol{\mu}_b)(1-\gamma)}{b(\gamma-1)} \tag{16}$$

Where, for notational compactness, we have:

$$a = \mathrm{f}(\boldsymbol{x}) \tag{17}$$
$$b = \mathrm{f}\left(\frac{(\boldsymbol{\mu}_b - \boldsymbol{x})}{\|\boldsymbol{\mu}_b - \boldsymbol{x}\|_2} \right) \tag{18}$$

Since the right-side of inequality (14) upper bounds the right-side of inequality (11), we know that any solution for inequality (14) is also a solution for inequality (11).

Since inequality (16) is a bound on ε which provides a guarantee on when eq. (9) is infeasible. Hence we conclude that eq. (9) may be feasible only when

$$\varepsilon \geq \max\left\{ 0, \frac{2b(\gamma-1)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}} + a\gamma + a + \mathrm{f}(\boldsymbol{\mu}_b)(1-\gamma)}{b(\gamma-1)} \right\} \tag{19}$$

B Additional Experiments

This section contains additional experiments furthering our understanding of face recognition/obfuscation fairness issues.

B.1 Face Obfuscation and Demographics

While adversarial examples cause a misclassification on a particular identity, these adversarial examples do not necessarily cause their demographic attributes to change. To validate this claim, we generate untargeted adversarial examples on the LFW dataset and classify these perturbed images using a classifier that predicts the race attribute from a face [61]. Of each identity in the dataset, only 8% of the identities' race attribute change after adversarial perturbations are added. We visualize this in fig. 15, where the embeddings of the natural examples and adversarial examples are plotted according to their ground truth demographics using t-SNE. Comparing the plots within figs. 15a and 15b, a heavy overlap is observed between the embeddings of both natural and adversarial examples of the same race and the same sex. These results further addressed confirm our intuition from section 4.2; an obfuscated face is unlikely to depart it's demographic group within the embedding space. This is especially true when the groups are clustered in the embedding space.
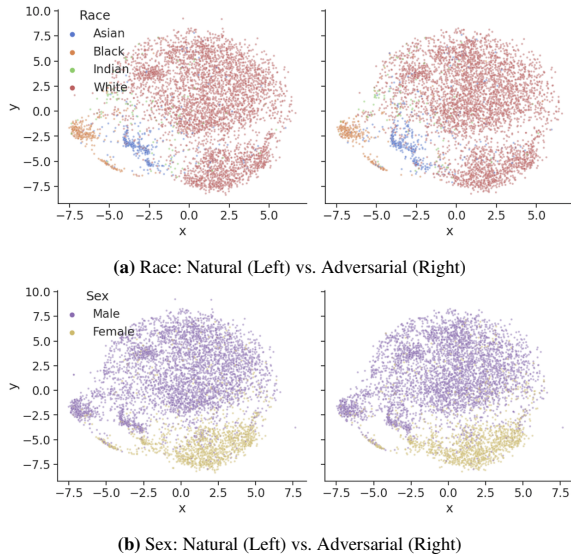
(a) Race: Natural (Left) vs. Adversarial (Right)

(b) Sex: Natural (Left) vs. Adversarial (Right)

Figure 15: t-SNE of the natural and adversarial embeddings. Note that the clusters do not change, identities are still rooted within their own demographic.

I	II	III	IV	V	VI
0xF4F2F5	0xFAF0EF	0xFFF9E2	0xE4C567	0x8F573C	0x2D2024
0xEDECEA	0xF3EBE6	0xF1E8C4	0xE2C26A	0x7A4C2C	0x14152A
0xFAF9F7	0xF4F1EB	0xF0E3AE	0xE0C27C	0x642D0E	
0xFDFBE7	0xFBFCF4	0xE1D394	0xDFB978	0x652C1A	
0xFDF6E7	0xFCF8EE	0xF2E398	0xC8A664	0x602D1B	
0xFEF7E6	0xFEF6E2	0xECD7A0	0xBD9862	0x562E24	
	0xFFF9E2	0xECDA86	0x9D6B41	0x3E1A0D	

Table 3: This chart depicts colors in the Fitzpatrick scale [45], which is derived from Von Luschan's chromatic scale [62]. Colors are in hexadecimal.

B.2 Black-Box Obfuscation Disparities

To understand whether the performance disparities between demographics (as discussed in section 4.2) manifest in commercial face recognition systems, we tested the success of the perturbed faces against three commercial face recognition systems. These Face Recognition APIs were Face++, Microsoft Azure Face API, and Amazon AWS Rekognition. In fig. 7, we observe large differences in obfuscation success for each race demographic. This performance disparity can be attributed to the larger perturbation sizes associated with Asian and White identities (fig. 5). Black identities were observed to have smaller perturbations, and higher local Lipschitz constants (fig. 16), suggesting that decreased obfuscation success can stem from disparities in the robustness of each demographic.

B.3 Targeted vs. Untargeted Obfuscation

Figure 17 portrays the untargeted obfuscation success rates for the pre-trained Facenet model. Untargeted success is also provided for the black-box setting with the OpenFace model (fig. 18). In fig. 14, the targeted success rates are provided for the Race-Balanced and Sex-Balanced Facenet models. We observe similar trends as discussed in sections 4.2 and 4.3.
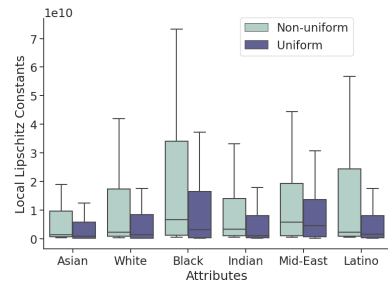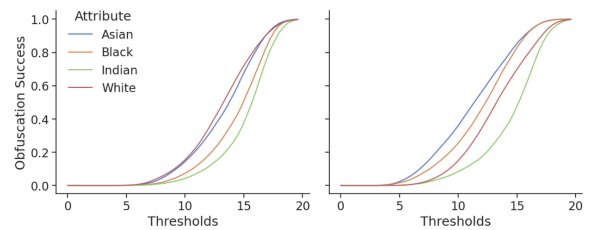

Figure 16: Upper bounds on local Lipschitz constants estimated using Fast-Lin and RecurJac. Models trained on non-uniform demographic distributions suffer from higher instability. The maximum constants for each demographic is 2-3 times larger in the non-uniform model than in the uniform model.
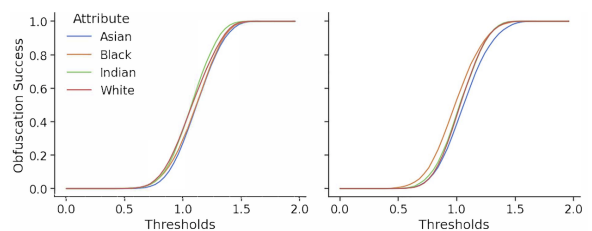

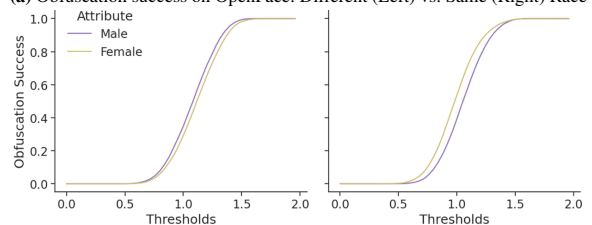**(a)** Obfuscation success on FaceNet: Different (Left) vs. Same (Right) Race
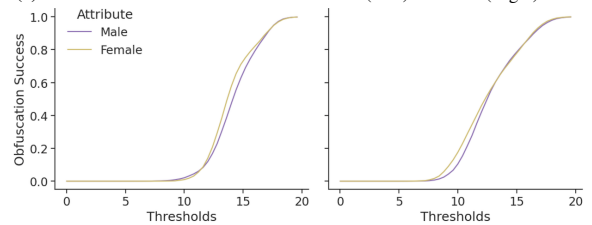
(b) Obfuscation success on FaceNet: Different (Left) vs. Same (Right) Sex

Figure 17: Untargeted obfuscation success evaluated on the FaceNet metric embedding network in a white-box setting.

(a) Obfuscation success on OpenFace: Different (Left) vs. Same (Right) Race

(b) Obfuscation success on OpenFace: Different (Left) vs. Same (Right) Sex

Figure 18: Untargeted obfuscation success evaluated on the OpenFace metric embedding network in a black-box setting.