Fairness Properties of Face Recognition and Obfuscation Systems

Harrison Rosenberg
University of Wisconsin–Madison
hrosenberg@ece.wisc.edu

Brian Tang
University of Wisconsin–Madison
bjtang2@wisc.edu

Kassem Fawaz
University of Wisconsin – Madison
kfawaz@wisc.edu

Somesh Jha
University of Wisconsin – Madison
jha@cs.wisc.edu

Abstract

The proliferation of automated face recognition in various commercial and government sectors has caused significant privacy concerns for individuals. A recent, popular approach to address these privacy concerns is to employ evasion attacks against the metric embedding networks powering face recognition systems. Face obfuscation systems generate imperceptible perturbations, when added to an image, cause the face recognition system to misidentify the user. The key to these approaches is the generation of perturbations using a pre-trained metric embedding network followed by their application to an online system, whose model might be proprietary. This dependence of face obfuscation on metric embedding networks, which are known to be unfair in the context of face recognition, surfaces the question of demographic fairness – *are there demographic disparities in the performance of face obfuscation systems?* To address this question, we perform an analytical and empirical exploration of the performance of recent face obfuscation systems that rely on deep embedding networks. We find that metric embedding networks are demographically aware; they cluster faces in the embedding space based on their demographic attributes. We observe that this effect carries through to face obfuscation systems: faces belonging to minority groups incur reduced utility compared to those from majority groups. For example, the disparity in average obfuscation success rate on the online Face++ API can reach up to 20 percentage points. We present an intuitive analytical model to provide insights 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 media coverage of data breaches, violation of privacy laws, and the adoption of face recognition by law enforcement entities have shed light on the significant security and privacy implications of face recognition systems. To mitigate the growing privacy concerns, face obfuscation systems have been proposed to hide the identity of users. Several of these systems, such as Face-Off [7], LowKey [8], and Foggysight [9], leverage the 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 end-user applications: perturbations are often visually acceptable to the user; several features of social media applications, such as face-augmenting filters, do not suffer; and the obfuscation mechanism runs locally, without access to target face recognition systems. The last feature is enabled by the transferability property of evasion attacks, whereby perturbations are effective against different models running the same or similar task.

These face obfuscation systems suffer major shortcomings, including the recently identified ability of face recognition systems to adapt and use perturbed faces to improve its performance [13] – perturbed faces can be re-identified in the future. In this work, we uncover another shortcoming of such systems: *there are demographic disparities in the performance of face obfuscation systems that are based on evasion attacks.* 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?
- How do training set demographics affect the behavior of face recognition and obfuscation?

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] 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 empirical and analytical characterization yields the following insights about the 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 that face recognition systems are better at differentiating individuals in different demographic groups than differentiating individuals within the same demographic. Without access explicit access to demographic information, we find that 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], a proxy for targeted obfuscation, and LowKey [8], a proxy for untargeted obfuscation. For minority groups, compared to the white-box setting, we find that stronger perturbations are necessary to successfully obfuscate a face in the black-box setting, such as the Face++ face recognition API. 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.

How do training set demographics affect the behavior of face recognition and obfuscation? We train and evaluate metric embedding networks trained on two subsets of VGGFace2: one dataset is balanced in the race demographic and another is balanced in the sex demographic. Metric embedding networks provide more uniform performance for the demographics upon which their training data was balanced. Though performance is more uniform when training data are demographically balanced, performance disparities remain.

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), to formalize intuition regarding the behavior of face recognition and obfuscation when conditioning on the demographic group. Our model reveals two insights that help explain our empirical observations: the efficacy of the embedding function depends on the balance in sampling from different groups and an obfuscated face is more likely to belong to its original demographic group when embeddings are clustered.

2 Background

To understand the context of face obfuscation systems, it is necessary to discuss some of the notation, background, and tools used with neural networks and face recognition. We use the following notation throughout this paper.

[1] We plan to make our codebase and results publicly available.

2.1 Notation and Terminology

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 in our setting, $\boldsymbol{x}$ refers to a cropped face, d to the number of pixels in the cropped face multiplied by three (referring to the RGB color channels), and $\mathcal{Y}$ to the set of identities.

Throughout this paper, scalars are denoted by lowercase standard-typeface letters, vectors are denoted by boldface lowercase letters, sets are denoted by capital letters and matrices are denoted by boldface capital letters. Given a vector $\boldsymbol{z}$, z_j denotes the j^{th} entry in vector $\boldsymbol{z}$. Given a matrix $\boldsymbol{A}$, $\boldsymbol{A}_{ij}$ denotes the entry of matrix $\boldsymbol{A}$ at row i, column j. If a matrix is itself indexed by a subscript, such as $\boldsymbol{A}_b$, the example at row i, column j is denoted $(\boldsymbol{A}_b)_{ij}$. Probability distributions are denoted with calligraphic capital letters. We will use $\mathcal{D}$ to represent the distribution from which examples in 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 $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Sometimes we abuse notation such that metric ρ is given only one argument. When this is the case, the second argument to the metric ρ is implicitly a size conforming zero vector.

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 [16]. 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.

2.2 Face Recognition System

Face recognition systems are predominantly based on *metric embedding networks*. A metric embedding network, denoted by $\mathrm{f}_k : \mathbb{R}^d \to \mathbb{R}^k$, is a neural network which takes an RGB face image $\boldsymbol{x}$ as input, and returns a k dimensional embedding. We 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 images belonging to the same identity have lower pairwise distance in the embedding space. Likewise, images belonging to different identities are intended to have high pairwise distance in the embedding space. To achieve such behavior, metric embedding networks typically trained with one of two classes of loss functions: contrastive loss [17] and triplet loss [3]. The functionality of a face recognition system is depicted in fig. 1.

Upon obtaining an embedding output by a metric embedding network, one can perform tasks such as clustering, matching, or classification. Often, distance metrics such as ℓ_2 norm or cosine similarity to determine the distance between embeddings. 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 $\|\boldsymbol{x} - \boldsymbol{x}\|_2 \leq \tau$.

To measure how well a metric embedding network f performs this matching task, we use TPR_z which is a parametrized notion of true positive rate. The match threshold τ is chosen such that it satisfies a false acceptance rate upper-bounded by z.

$$\mathrm{TPR}_z(S, S') \triangleq \sum_{(\boldsymbol{x}_i, y_i, \boldsymbol{x}'_i, y') \in \{S, S' | y = y'\}}^{m} \frac{\mathbb{1}\left[\|\mathrm{f}(\boldsymbol{x}) - \mathrm{f}(\boldsymbol{x}')\|_2 < \tau\right]}{|\{S, S' \mid y = y'\}|} \tag{1}$$

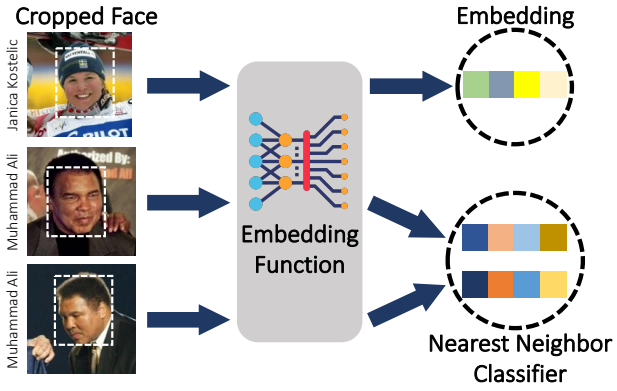


Figure 1: For well-trained embedding functions, embeddings of images belonging to the same identity will have smaller pairwise distances than the pairwise distances between embeddings of different identities. Note that a *metric embedding network* refers to an "embedding function" with a DNN architecture.

where threshold τ is defined as

$$\begin{aligned} \max \quad & \tau \\ \text{s.t.} \quad & z > \sum_{(\boldsymbol{x},y,\boldsymbol{x}',y') \in \{S,S' | y \neq y'\}} \frac{\mathbb{1}\left[\|\mathrm{f}(\boldsymbol{x}) - \mathrm{f}(\boldsymbol{x}')\|_2 < \tau\right]}{|\{S, S' \mid y \neq y'\}|} \end{aligned} \tag{2}$$

and the right side of the inequality is the false acceptance rate.

2.3 Principal Component Analysis

Neural networks are notoriously difficult to analyze. Perhaps the best known, theoretical exploration of neural network performance was done by Arora et al [18]. The authors primarily consider networks with one hidden layer. A tractable analysis of more complicated architectures remains an open problem. As metric embedding networks used in face recognition systems nearly always have more than one hidden layer, and so we conclude there exists no suitable explanatory, tractable analysis of metric embedding network performance. Hence, we use, as a proxy for metric embedding networks, the Principal Component Analysis (PCA).

PCA forms the backbone for our analytical model (section 4.3). PCA is a well-known linear dimensionality reduction algorithm [19]. By finding a new basis for the dataset, PCA can uncover hidden or otherwise non-obvious patterns therein. In particular, PCA constructs an ordered orthonormal basis for a given mean-centered dataset. The orthonormal basis vectors are arranged in decreasing order of their variance contribution to the original dataset.

2.4 t-SNE

The t-Distributed Stochastic Neighbor Embedding (t-SNE) [20] is another dimensionality reduction technique which is useful in visualizing high dimensional embedding spaces. t-SNE is often used in visualizing image datasets or deep neural network embeddings in two or three dimensions. It is a variant of Stochastic Neighbor Embeddings [21] which avoids crowding data points and can capture the implicit structure of data. In this paper, we use t-SNE to visualize the embeddings of faces as in fig. 2; such plots aid our discussion of the embedding space geometry.

2.5 TCAV

TCAV [22] is a tool used to evaluate the interpretability of a deep neural network. Using user-specified high-level concepts, such as patterns or colors, decision boundaries are created by training linear classifiers on a neural network's activations for a given concept's examples. Concept Activation Vectors (CAVs) are extracted from the vector orthogonal to this decision boundary, and a statistical significance test is performed on these CAVs and the model's gradients. Though TCAV has been traditionally used to map learned concepts in classification networks, we extend its functionality to metric embedding networks so that we may better understand the contributing factors in a deep neural network's predictions.

2.6 Relevant Fairness Definitions

There is no single definition of fairness, mathematical or otherwise [23]. 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 the success rates of face obfuscation systems as measured by perturbation success rate. When these success rates are equal, they satisfy the fairness constraint known as statistical parity [24]. We also compare, by demographic, the strength of such perturbations necessary to yield a successful face obfuscation. For targeted evasion attacks, we study the strength of such perturbations in 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 5.3, we aim to determine if training metric embedding networks on balanced datasets yields metric embedding networks which are approximately equal in their parametrized True Positive Rate TPR_z. An optimization constraint explicitly requiring equal true positive rates between groups is the notion of fairness known as equalized odds [25].

2.7 Adversarial Machine Learning

Goodfellow et al. [10] discovered an interesting property of deep neural networks: they are vulnerable to adversarial examples. Small structured perturbations, imperceptible to the human eye, may cause the network to misclassify a given input sample. This branch of machine learning research resulted in the formulation of many attack algorithms, the most common of which are evasion attacks, including Fast Gradient Sign Method (FGSM) [11], Projected Gradient Descent (PGD) [11], and Carlini-Wagner (CW) [12] attacks. The bulk of study in adversarial machine learning focuses on attacks that algorithmically generate ℓ_p norm-bounded perturbations by performing a noisy gradient-based optimization procedure. We further discuss adversarial machine learning in the context of face recognition in the next section.

3 Face Obfuscation

Researchers have demonstrated that systems which leverage principles from adversarial machine learning can provide users with privacy utility in the presence of face recognition systems. These privacy benefits are encapsulated in systems known as face obfuscation systems. In this section, we present the notation and discuss the work relevant to such face obfuscation systems.

3.1 Threat Model

In the face obfuscation setting, an end user considers the machine learning provider to be the main threat. They wish to apply (imperceptible) perturbations to their faces, prior to uploading them, so that the presence

of faces are correctly recognized, but the predicted identity of faces is incorrect. Such protection might be beneficial in the event of social media data leaks [26], preventing cyber-stalking [27], protecting data from web scrapers, hiding online activity from big government entities [28], and more.

State-of-the-art designs for face obfuscation systems leverage evasion attacks. Such attacks were described in section 2.7. Unlike other attacks, evasion attacks assume the user is not able to poison the face recognition system. An important property of adversarial examples, especially for vision tasks such as face recognition, is their ability to transfer across models. This property allows users to generate adversarial examples without requiring access to an online face recognition API and without directly querying said model. Instead, face obfuscation systems leverage this properties by querying surrogate models to generate the perturbed faces. In a *white-box* setting, examples are generated with the intention of attacking the surrogate model. In the *black-box* setting, examples are generated on the surrogate model and applied to a black-box model, by leveraging the transferability property.

3.2 Face Obfuscation Systems

The earliest work in this vein of face obfuscation research explores physical adversarial examples. For example, Sharif et al. [29] present physically realizable glasses that allow a human user to impersonate a target individual and evade face detection systems. Recently, there has been a shift in focus towards preserving user privacy online from malicious web scrappers, data breaches, and intrusive service providers. Thus, researchers have focused more on digital face obfuscation systems, which are more suited for social media and online applications. Examples of such systems include FAWKES [30], 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 with an attempt to hide the user's identity. These systems are the main focus of this paper.

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

3.3 Evasion Attacks on Face Recognition

Evasion attacks may be divided into targeted and untargeted varieties. To define the two attacks, we first describe the embedding centroid. Given a dataset S, denote by $\boldsymbol{c}_{\mathrm{f},y}$ the embedding centroid for identity y as computed on embedding function f is:

$$\boldsymbol{c}_{\mathrm{f},y} \triangleq \frac{1}{|\{(\boldsymbol{x}', y') \in S \mid y' = y\}|} \sum_{(\boldsymbol{x}', y') \in S} \mathrm{f}(\boldsymbol{x}') \cdot \mathbb{1}\left[y' = y\right] \tag{3}$$

Untargeted Attacks: Recall from section 2.1 that a metric ρ given only one explicit argument is assumed to have a size-conforming zero vector as its second argument. Given an embedding function $\mathrm{f} : \mathcal{X} \to \mathcal{Y}$, unlabeled example $\boldsymbol{x}$, and metrics ρ_1, ρ_2 an *untargeted attack* may be formulated as:

$$\begin{gathered} \min_{\boldsymbol{\delta} | \boldsymbol{x}+\boldsymbol{\delta} \in \mathcal{X}} \rho_1(\boldsymbol{\delta}) \\ \text{s.t.} \arg\min_{y' \in \mathcal{Y}} \rho_2(\boldsymbol{c}_{\mathrm{f},y'}, \mathrm{f}(\boldsymbol{x})) \neq \arg\min_{y' \in \mathcal{Y}} \rho_2(\boldsymbol{c}_{\mathrm{f},y'}, \mathrm{f}(\boldsymbol{x} + \boldsymbol{\delta})) \end{gathered} \tag{4}$$

The problem with the formulation above is that the constraint is non-convex. To make the attack implementable in practice, the constraints must be relaxed. For a labeled example $(\boldsymbol{x}, y)$, the optimization

objective which yields an untargeted perturbation can be written as:

$$\underset{\boldsymbol{\delta}|\boldsymbol{x}'+\boldsymbol{\delta}\in\mathcal{X}}{\arg\max}\ \rho_2(\mathrm{f}(\boldsymbol{x}'+\boldsymbol{\delta}), \boldsymbol{c}_{\mathrm{f},y})\ \text{ s.t. }\ \rho_1(\boldsymbol{\delta}) \leq \epsilon \tag{5}$$

In this paper, we use the LowKey attack [8] to instantiate eq. (5). In this case, PGD is use to solve the optimization problem, ρ_1 is the Learned Perceptual Image Patch Similarity (LPIPS) metric [31], 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 target label y', embedding function $\mathrm{f} : \mathcal{X} \to \mathcal{Y}$, unlabeled example $\boldsymbol{x}$, and metrics ρ_1, ρ_2 a *targeted attack* may be formulated as follows:

$$\min_{\boldsymbol{\delta}|\boldsymbol{x}+\boldsymbol{\delta}\in\mathcal{X}} \rho_1(\boldsymbol{\delta}) \text{ s.t. } y' = \underset{y\in\mathcal{Y}}{\arg\min}\, \rho_2(\boldsymbol{c}_{\mathrm{f},y}, \mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta})) \tag{6}$$

Similar to the untargeted case, attacks on face recognition systems relax the above constraints to arrive at a convex optimization formulation. Here, we state one such relaxation from Face-Off [7], which we use in the rest of this paper. First, we define the multi-class hinge loss. The multi-class hinge loss enables a convex relaxation of the constraints in the targeted attack eq. (6). Given a perturbed example $\boldsymbol{x}+\boldsymbol{\delta}$, target label y', and positive real number κ, the multi-class hinge loss is denoted by $\mathrm{G}_\kappa(\boldsymbol{x}+\boldsymbol{\delta}, y')$ where:

$$\begin{aligned}\mathrm{G}_\kappa(\boldsymbol{x}+\boldsymbol{\delta}, y') \triangleq \max\Big\{0, \kappa + \rho_2(\boldsymbol{x}+\boldsymbol{\delta}, \boldsymbol{c}_{\mathrm{f},y'}) \\ - \max_{y\neq y'} \rho_2(\boldsymbol{x}+\boldsymbol{\delta}, \boldsymbol{c}_{\mathrm{f},y})\Big\}\end{aligned} \tag{7}$$

For a labeled example $(\boldsymbol{x}, y)$, and target label $y' \neq y$, the CW attack can minimize the following optimization objective:

$$\underset{\boldsymbol{\delta}|\boldsymbol{x}+\boldsymbol{\delta}\in\mathcal{X}}{\arg\min}\ \rho_1(\boldsymbol{\delta}) \text{ s.t. } \rho_1(\boldsymbol{\delta}) \leq \epsilon \text{ and } \mathrm{G}_\kappa(\boldsymbol{x}+\boldsymbol{\delta}, y') \leq 0 \tag{8}$$

4 Characterization of Face Recognition and Obfuscation

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. Our objective is to study if face obfuscation techniques equally benefit different demographic groups. The experiments focus on answering three questions:

1. ***Bias in Face Recognition*** **What is the baseline bias present in face recognition systems?** What are the initial biases in already-existing face recognition systems? If face recognition systems are biased, where are these biases learned within the network? Our experiments concur with existing literature: when conditioning by demographic, there is indeed a disparity in performance of face recognition systems. Furthermore, 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 that face recognition systems are less robust to perturbations applied to faces from minority demographic groups; the strength of the perturbation necessary to change the identity is smaller than for samples from majority demographic groups. Performance disparities are evident between demographics in both the targeted and untargeted cases, and this gap widens for source-target pairs of the same demographic group. Consequently, obfuscated faces tend to remain classified as a member of the same demographic group.

3. ***Dataset Balancing*: How does the balance of training set demographics impact the utility of face obfuscation?** We observe that FaceNet models trained on balanced subsets of VGGFace2 reduce the performance gap of face obfuscation systems on the demographic groups upon which the dataset was balanced. We demonstrate this for two training sets: a training dataset balanced on sex and another balanced on race.

4.1 Experimental Setup

Datasets and Models: We utilize the LFW and VGGFace2 datasets in the rest of this evaluation; these datasets are described in section 4.2. In the white-box setting, we generate perturbed faces using the FaceNet model. In the black-box setting, we test for the transferability of the perturbed faces to the Face++ face recognition API [32] and to a pre-trained OpenFace model [33]. The OpenFace model is of architecture similar to FaceNet in that it is an Inception ResNet network; it is trained on $96 \times 96 \times 3$ images and is tuned to output 128-dimension embeddings. Trained on both the CASIA-WebFace [34] and FaceScrub [35] datasets, OpenFace achieves a 0.973 area under the curve (AUC) on the LFW dataset.
Attacks: We perform our evaluation using untargeted and targeted variants of face obfsucation systems, as described in section 3. Utilizing the Face-off face obfuscation system [7] and the FaceNet model [3], we generate adversarial examples for a subset of the LFW [5] dataset. For sections 5.1 and 5.2 we use 5 as our margin value (κ in eq. (8)) and in section 5.3 we use 1 as the margin value. We generated untargeted attacks using the Lowkey [8] attack as described in section 3.3. The ensemble used in the LowKey attack includes two pre-trained ArcFace models [4] and two pre-trained Cosface models [36]. We report the obfuscation success rate as an indicator of the perturbation's effectiveness. In the targeted case, it measures the proportion of perturbed faces which match their intended targets. In the untargeted case, it measures the proportion of perturbed faces that evade their source identity.
Scenarios: We consider six demographic attributes, four for race (Black, White, Indian, Asian) and two for sex (female, male). We sample 50 identities per attribute from the LFW dataset[2]. We refer to these images as the source images, and we use them as inputs to the untargeted attack. For the targeted attacks, we create two scenarios:

- **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 sex and 2450 source-target pairs of the same race.
- **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 sex, 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.

4.2 Face Recognition Setup

Our evaluations focus on two widely-used datasets. For tasks which require multiple images per identity such as training models and running TCAV, we use the VGGFace2 dataset [6]. Though this dataset contains 8631 identities with an average of 362.6 images per identity, it is demographically imbalanced. The demographic split is approximately 60% male and 40% female; and 73% White, 9% Black, 6% Asian, 4% Indian, 4% Middle Eastern, and 4% Latino. During inference time we use the Labeled Faces in the Wild (LFW)

[2]The LFW demographic attributes were annotated using attribute classifiers described by Kumar et al. [16].

dataset [5]. For our evaluation in section 5.1, we utilize a FaceNet metric embedding network [3] pre-trained on the VGGFace2 dataset [6], which achieves $\mathrm{TPR}_{0.001}$ of .9965 on the LFW dataset [5]. This Inception ResNet v1 network is trained on $160 \times 160 \times 3$ images and outputs 128-dimensional embeddings. In the next section, we will explore how a neural network trained on an imbalanced dataset performs differently for faces of each demographic group.

4.3 An analytical model for face recognition

To study how the training data distribution affects fairness properties of face obfuscation, we examine a hierarchical mixture of Gaussians. This model is inspired by the hierarchical nature of popular face recognition datasets. We use a k-component PCA as an embedding function: a PCA embedding function which only utilizes projections onto the leading k principal components to perform its dimensionality reduction. Such analysis is common in the space of machine learning. 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 dataset, there are two mutually exclusive groups: group a and group b. Sometimes a placeholder g is used to represent a group $\mathsf{g} \in \{\mathsf{a}, \mathsf{b}\}$. The mean vector for population groups a and b are denoted by $\boldsymbol{\mu}_{\mathsf{a}} \in \mathbb{R}^d$ and $\boldsymbol{\mu}_{\mathsf{b}} \in \mathbb{R}^d$, respectively. Moreover, $\boldsymbol{\mu}_{\mathsf{a}} = -\boldsymbol{\mu}_{\mathsf{b}}$, $\|\boldsymbol{\mu}_{\mathsf{a}}\|_2 = 1$, and $\|\boldsymbol{\mu}_{\mathsf{b}}\|_2 = 1$.

The i^{th} identity in group g is denoted by $\boldsymbol{\nu}_{\mathsf{g},i} \in \mathbb{R}^d$. The j^{th} image representing identity $\boldsymbol{\nu}_{\mathsf{g},i}$ is denoted by $\boldsymbol{x}_{\mathsf{g},i,j} \in \mathbb{R}^d$. $\boldsymbol{\Sigma}_{\mathsf{a}} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{\Sigma}_{\mathsf{b}} \in \mathbb{R}^{d \times d}$ are diagonal covariance matrices. Furthermore, $\boldsymbol{\Sigma}_{\mathsf{a}} = \gamma \boldsymbol{\Sigma}_{\mathsf{b}}$ where γ is a positive, real-valued scalar. As we will see, when relating this analytical model to experiments on existing face recognition datasets γ captures numerical imbalances in population groups.

An identity $\boldsymbol{\nu}_{\mathsf{g},i}$ is drawn from the identity distribution $\mathcal{N}(\boldsymbol{\mu}_{\mathsf{g}}, \boldsymbol{\Sigma}_{\mathsf{g}})$. The identity distribution may be thought of as a hyperprior on images. An image $\boldsymbol{x}_{\mathsf{g},i,j}$ is drawn from $\mathcal{N}(\boldsymbol{\nu}_{\mathsf{g},i}, \beta \boldsymbol{I})$ where β is a positive, real-valued number. For each identity $\boldsymbol{\nu}_{\mathsf{g},i}$, exactly m images are drawn from $\mathcal{N}(\boldsymbol{\nu}_{\mathsf{g},i}, \beta \boldsymbol{I})$. Lastly, we denote by $\mathcal{D}_{\mathsf{g}}$ the distribution of images in group g.

The synthetic data distribution is a linear combination of image distributions for group a ($\mathcal{D}_{\mathsf{a}}$) and group b ($\mathcal{D}_{\mathsf{b}}$). Let $\alpha_{\mathsf{a}} \in [0, 1]$ denote the proportion of examples drawn from group a. $\alpha_{\mathsf{b}} \in [0, 1]$ is defined similarly. Since the population groups are mutually exclusive, we know $\alpha_{\mathsf{a}} + \alpha_{\mathsf{b}} = 1$. The synthetic data distribution, we denote by $\mathcal{D}$ is a linear combination, that is $\mathcal{D} = \alpha_{\mathsf{a}} \mathcal{D}_{\mathsf{a}} + \alpha_{\mathsf{b}} \mathcal{D}_{\mathsf{b}}$. S denotes a sampling of images from $\mathcal{D}$.

Though it is certainly true that our hierarchical Gaussian training data distribution and accompanying PCA embedding are simpler than datasets used in face recognition and face recognition systems, respectively, we believe that this simplicity allows us to show how the fairness issues we observe in face obfuscation systems directly relate to first principles in machine learning. The deep relationship between our analytical model and the experiments will become more concrete in the experiments section (section 5).

5 Experiments on Face Obfuscation Systems

These experiments are designed to 1) provide a baseline on the level of unfairness in already existing face recognition systems, 2) showcase the impact of face obfuscation on existing face recognition systems, and 3) study the utility of dataset balancing on fairness implications of face obfuscation.

5.1 Error Disparity Among Groups

Each experiment is designed to elucidate fairness and privacy issues in existing face recognition systems. We highlight 4 results: 1) Tables 5 and 6 show the discrepancy in classification accuracy amongst existing face recognition systems; 2) Figure 3 demonstrates neural networks learn skin tones in early layers; 3)

	Male	Female	White	Asian	Black	Indian
$\mathrm{TPR}_{0.001}$	.9618	.8516	.9594	.8536	.9242	1.000
AUC	.9994	.9977	.9996	.9981	.9995	1.000
N	10000	5000	10000	2500	1240	20

Table 1: Same demographic matching performance on LFW. Both accuracy and validation accuracy are lower compared to Table 2. $\mathrm{TPR}_{0.001}$: Parametrized True Positive Rate with False Acceptance Rate of 0.001, N: number of pairs.

	Male	Female	White	Asian	Black	Indian
$\mathrm{TPR}_{0.001}$	.9678	.9436	.9732	.9448	.9742	1.000
AUC	.9995	.9988	.9993	.9998	.9999	1.000
N	10000	5000	10000	2500	1240	20

Table 2: Any demographic matching performance on LFW.

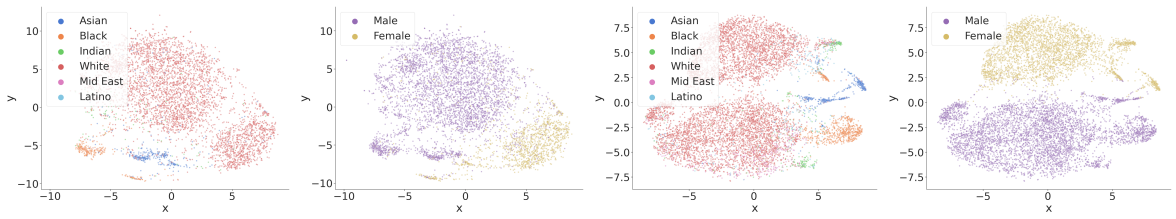
Figure 2 provides a visual depiction of the demographic-awareness of metric embedding networks though they are provided no explicit demographic information; 4) Figure 4 shows the stability of metric embedding networks as measured by local Lipschitz constants.

Face Recognition System Empirical Performance: Tables 1 and 2 demonstrate how training deep neural networks on an imbalanced dataset can result in performance discrepancies. When evaluating this model using parametrized True Positive Rate $\mathrm{TPR}_{0.001}$, the pairs of identities selected with only the same race or sex (table 1) perform worse when compared to pairs of identities selected without any such demographic restriction (table 2). This indicates it is easier to get a false accept when the source and target face are in the same demographic group. Indeed these discrepancies are statistically significant [3]: The difference in $\mathrm{TPR}_{0.001}$ between sexes is statistically significant as is the difference in $\mathrm{TPR}_{0.001}$ between races. For the same demographic matching, utilizing the Alexander-Govern test (a multi-sample generalization of Welch's t-Test) [37], the p-values are 4.26×10^{-46} and 6.71×10^{-24}, for sex and race respectively. For any demographic matching of sex and race, utilizing the Alexander-Govern test, the p-values are 4.11×10^{-6} and 0.000163, respectively. Formal statements of the null hypothesis its statistical test are deferred to the appendix.

What the Metric Embedding Network Learns To understand *what* the metric embedding network architecture learns, we plot the resulting structure of an embedding of FaceNet using t-SNE [20] in fig. 2. Interestingly, clustering by demographic appears in the embedding space despite the network never having explicit access to demographic attributes.

Where the Metric Embedding Network Learns We utilize TCAV [22] to investigate if intermediate layers of the network learn to distinguish demographic attributes. As it was originally defined for classification networks, the TCAV framework is retrofitted to metric embedding networks. In particular, 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. For the Facenet metric embedding network, the layers examined include early and intermediate activations towards the end of each inception ResNet block. Our evaluation involves 9 identities for each of the following four races: Asian, Black, Indian, White. In total, we run

[3] We take $p < 0.05$ to indicate statistical significance



(a) LFW - Race (b) LFW - Sex (c) VGGFace2 - Race (d) VGGFace2 - Sex

Figure 2: t-SNE [20] of the embedding spaces generated using the LFW [5] and VGGFace2 [6] datasets. Embeddings of identities are colored by race and sex. Distinct clusters exist for each demographic group, some clusters are more spread out than others.

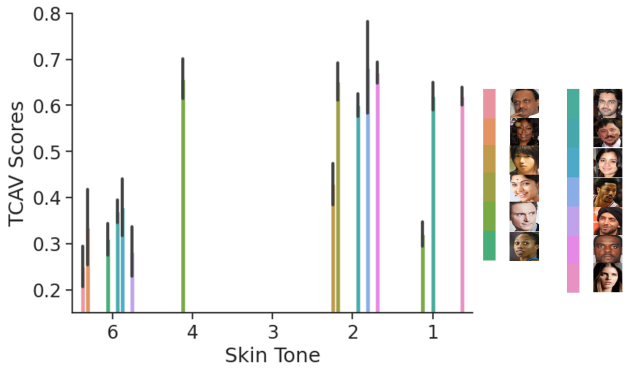


Figure 3: TCAV [22] scores for skin tones. The x-axis represents a range from darker skin (left) to lighter skin (right). TCAV was performed on intermediate layers early in the network. The network learns these concepts for other layers.

TCAV on 36 identities, with each identity containing 100 images of a person's face.

The concept we study is skin tone. Of the six skin tones on the Fitzpatrick scale [38], we select and annotate five skin tones ranging from pale white to dark brown. Each skin tone concept contains 75 images cropped from our subset of identities, and each facial region concept contains 250 cropped images. For the FaceNet metric embedding network, examined layers include early and intermediate activations towards the end of each Inception ResNet block. In fig. 3, we see high utilization of the skin tone concepts by earlier layers in the neural network. This suggests metric embedding networks learn to differentiate between skin tones in early layers.

Stability Properties of Face Recognition: By examining estimates of the local Lipschitz constants, we investigate the stability of face recognition 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 less generalizable [39–43]. We use the RecurJac [44] and Fast-Lin [45] bound algorithms to estimate the local Lipschitz constant for small neural networks trained on datasets with balanced and imbalanced distributions of demographic groups. Note that due to computational considerations, larger metric embedding networks, such as FaceNet, are not amenable to RecurJac and similar estimators. As such, we train two small VGG16 [46] classification networks on balanced and imbalanced datasets of 20 identities with the same number of face images (5 identities per race) from the VGGFace2 dataset. The imbalanced dataset mimics the distribution of the VGGFace2 dataset where White faces have more representation in the number of face images.

Figure 4 shows the distributions of the upper bounds on the estimated local Lipschitz constants for the imbalanced and balanced classifiers; this distribution is plotted for the training set. We observe larger upper

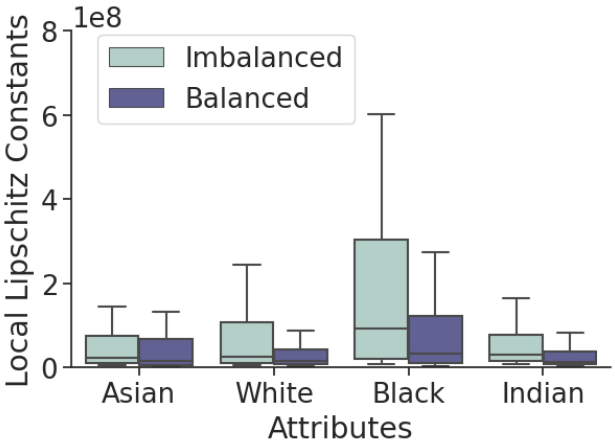


Figure 4: Upper bounds on local Lipschitz constants estimated using Fast-Lin and RecurJac. Models trained on imbalanced demographic distributions suffer from higher instability. The maximum constants for each demographic is 2-3 times larger in the imbalanced model than in the balanced model.

bounds for models utilizing demographically imbalanced training data. 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.

Connection to Analytical Model: In the context of machine learning, an ideal embedding function $\mathrm{f}_k : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is one which preserves distances within a dataset. While this property is important for face recognition, it is not the only consequential property of embedding functions. In the context of fair face recognition, we concern ourselves with how well the embedding represents a particular group from the lens of that group. Understanding such behavior for metric embedding networks is intractable given the state of current literature in neural network analysis, so we turn to our analytical model and PCA for intuition. To aid our analysis, we define the relative projection distance to capture how well the k component PCA embedding function represents a particular group, from the lens of that group:

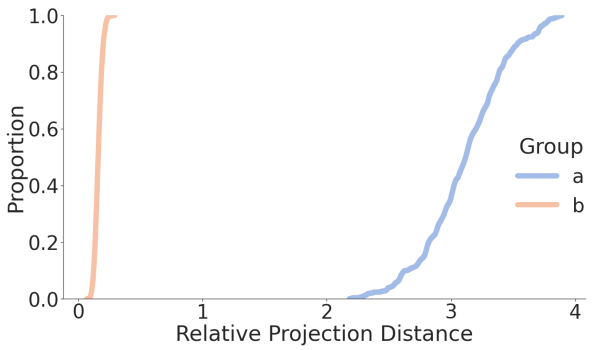
Definition 5.1. Let S be a sample of images drawn from the overall synthetic distribution $\mathcal{D}$. The relative projection distance of a point $\boldsymbol{x}$, a member of group g, with respect to the leading k principal components of a dataset S is denoted by $\rho_{\mathrm{rp},S,\mathrm{g},k} : \mathcal{X} \rightarrow \mathbb{R}^+$. More precisely:

$$\rho_{\mathrm{rp},S,\mathrm{g},k}(\boldsymbol{x}) \triangleq \frac{\left\| \left(\boldsymbol{x} - \sum_{i=1}^{k} \left\{ \frac{\boldsymbol{q}_i^\mathsf{T} \boldsymbol{x}}{\|\boldsymbol{q}_i\|_2 \|\boldsymbol{x}\|_2} \boldsymbol{q}_i \right\} \right) \right\|_2}{\sum_{j=1}^{d} \left(\boldsymbol{\Sigma}_\mathrm{g}\right)_{jj}}, \tag{9}$$

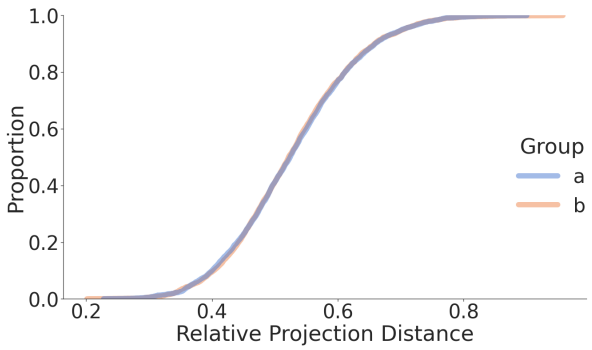
where the eigen-decomposition of the covariance of overall synthetic data distribution $\mathcal{D}$ is $\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\mathsf{T}$. Furthermore, $\boldsymbol{Q}$ may be decomposed as $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_d]^\mathsf{T}$.

Let us explain the relative projection distance: the numerator is the norm of the portion of sample $\boldsymbol{x}$ which is not captured by f_k. The denominator represents a group specific normalization factor representing the overall variance within all identity vectors $\{\boldsymbol{\nu}_{\mathsf{a},1}, \boldsymbol{\nu}_{\mathsf{a},2}, \ldots, \boldsymbol{\nu}_{\mathsf{b},1}, \boldsymbol{\nu}_{\mathsf{b},2}, \ldots, \}$. It is this normalization factor which allows us to show how error can be measured in the context of group g.

Given our interest in studying the impact of the frequency of each group in a training set on the efficacy of a learned embedding network, we focus on how the relative projection distance relates to local Lipschitz constant for faces in each group (a and b): Figure 5 portrays the groupwise distribution of relative projection distances. The case in which the local Lipschitz constants for each group are equal is captured by fig. 5b. Figure 5a captures the setting in which the local Lipschitz constants for each group are non-equal. This



(a) ECDF of relative projection distance is plotted for groups a and b with $\gamma = \frac{1}{100}$. The number of examples in group a is 250 whereas the number of examples in group b is 4750.



(b) ECDF of relative projection distance is plotted for groups a and b with $\gamma = 1$. The number of examples in each group is 2500.

Figure 5: The empirical cumulative distribution function (ECDF) of relative projection distance is plotted for $\gamma = \frac{1}{100}$ and $\gamma = 1$.

discrepancy in Lipschitz constant manifests itself as γ tending away from 1. In particular, each subfigure has 5000 total identities with $m = 50$ images sampled per identity. Figure 5a is constructed such that $\mathbf{\Sigma}_\mathsf{a} = \frac{1}{100}\mathbf{\Sigma}_\mathsf{b}$ with 250 examples in group a and 4750 examples in b. Figure 5b is constructed such that $\mathbf{\Sigma}_\mathsf{a} = \mathbf{\Sigma}_\mathsf{b}$ and the number of examples in each group is 2500. Furthermore, observe how in fig. 5a, the distribution of relative projection distances for group a is shifted right and more dispersed than the distribution of relative projection distances for group b. We formalize this intuition with the following proposition:

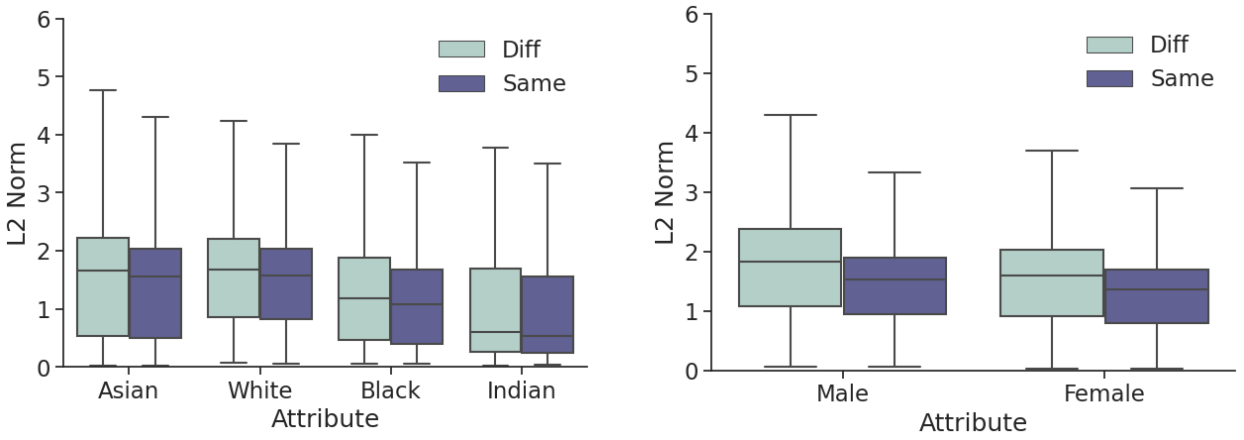
Proposition 5.2. For fixed $\boldsymbol{\mu}_\mathsf{a}$ fixed α_a, and fixed $\mathbf{\Sigma}_\mathsf{a}$, as γ approaches 0, the relative projection distance of examples in group a (defined in eq. (9)) increases.

Relating relative projection distance to the experiments on face datasets: minority groups tend to have larger local Lipschitz constants (fig. 4), meaning their distribution of relative projection distances is more dispersed than the distribution of relative projection distances for a more frequent population group. Further evidence is provided in inequality (12) and the associated proof. There we provide a relationship between Lipschitz constants and data dispersion when metric embedding networks are applied to Gaussian Mixtures. We conclude that greater dispersion in relative projection distances contributes to the performance disparities incurred by minority groups in face recognition.

5.2 Effectiveness of Obfuscation

In our attack scenarios, we are interested in understanding how the strength of generated perturbations differs with each demographic group and how the geometry of the embedding space induced by the metric embedding network impacts fairness properties of face obfuscation. Our experiments examine the distributions of ℓ_2 norms for each perturbation conditioned on demographic attributes, the obfuscation success in terms of both untargeted and targeted obfuscation success rates, and the success metrics in the black-box setting.

Face Obfuscation Induced Perturbation Norms Using the adversarial examples generated from the targeted attack, we examine the distribution of perturbation norms conditioned by demographic attribute. The perturbation norms appear smaller for same demographic pairs than for different demographic pairs in fig. 6. A formal analysis confirms this observation: Utilizing Welch's t-Test with non-equal variance [47] and the



(a) ℓ_2 Norm: Race

(b) ℓ_2 Norm: Sex

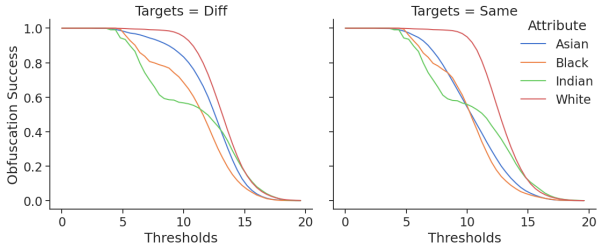
Figure 6: The distribution of adversarial perturbation sizes generated using the CW [48] attack. Along with the disparity among demographics, the perturbation sizes tend to be smaller when the target identity is a member of the same demographic group as the source identity.

Alexander-Govern test [37], the differences in perturbation strength between demographic groups are statistically significant with p-values of 7.65×10^{-33}, 3.34×10^{-28}, 2.50×10^{-291}, and 3.12×10^{-143} when the matching targets are in the same sex demographic group, different sex demographic group, same race demographic group, and different race demographic group, respectively. Furthermore, utilizing Welch's t-Test with non-equal variance, 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. Formal statements of null hypotheses and statistical tests are deferred to the appendix. Users may wish to select a target identity of the same race or sex to optimize face obfuscation system utility, but 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) users who perturb within the same demographic may leak demographic information to an adversary.

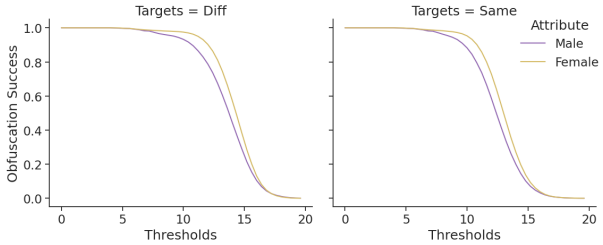
Success of Face Obfuscation: In the case of targeted adversarial examples, we also observe their success to be dependent on the demographic attribute. In fig. 7, faces with the White race attribute exhibit a higher obfuscation success rate. When performing the black-box evaluation on the OpenFace model, we observe a similar trend. Examples generated on faces from the majority group still 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.

We test the success of the perturbed faces against the Face++ face recognition API. In fig. 8, we present the distribution of the obfuscation success rate in untargeted examples. For race, we observe higher success rates for White and Asian faces, which agrees with the conclusions drawn from fig. 4 and fig. 6. Meanwhile, Indian and Black faces suffer from less robustness in the network, decreased perturbation sizes, and lower obfuscation success rate. Similarly, for the sex attribute, we observe a slightly higher success rate for the male faces.

Connection to Analytical Model: From our experiments, we have observed there is a discrepancy, when conditioning by demographic, in the strength of perturbations necessary to successfully obfuscate an image. As was the case in the previous section, the state of existing neural network literature does not provide the necessary tools to analyze face obfuscation behavior. Consequently, we turn to our analytical model for a

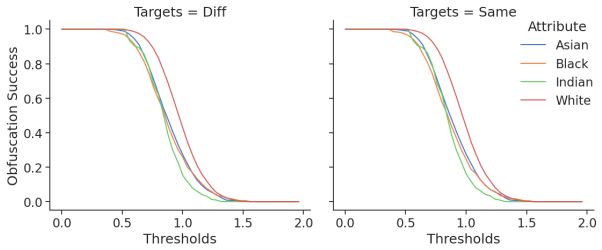


(a) Targeted: Different Race (Left) vs. Same Race (Right)

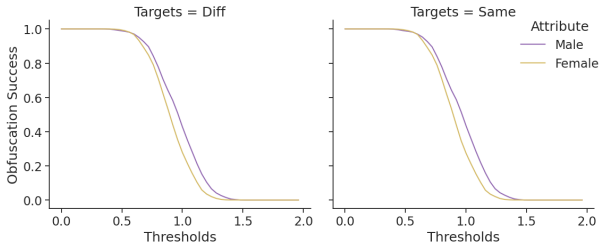


(b) Targeted: Different Sex (Left) vs. Same Sex (Right)

Figure 7: Targeted obfuscation success evaluated on the FaceNet metric embedding network in a white-box setting.



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



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

Figure 8: Targeted obfuscation success evaluated on OpenFace metric embedding network in a black-box setting.

probabalistic interpretation of face obfuscation:

Let us begin by defining some notation for this analysis. Denote by $p_{\mathcal{Q}}$ the probability density function for probability distribution $\mathcal{Q}$. Given an example $\boldsymbol{x}$, define the maximum likelihood estimator for population group as $\psi(\boldsymbol{x}) = \arg\max_{\mathsf{g}\in\{\mathsf{a},\mathsf{b}\}} p_{\mathcal{D}_\mathsf{g}}(\boldsymbol{x})$.

Given an example image $\boldsymbol{x}$ in group g, we study how difficult it is to construct a perturbation $\boldsymbol{\delta}$ such that $\psi(\boldsymbol{x}+\boldsymbol{\delta}) \neq \mathsf{g}$. The synthetic data distribution described in this section is a natural medium with which to quantify the necessary strength of such a perturbation $\boldsymbol{\delta}$:

For the sake of simplicity, consider the 1-PCA embedding function $\mathrm{f}_1 : \mathbb{R}^d \to \mathbb{R}$. Let sample $\boldsymbol{x}$ drawn from the overall synthetic data distribution $\mathcal{D}$ and without loss of generality assume this $\boldsymbol{x}$ is a member of group a. We will assume group a is the minority group and so we assume that $\gamma \leq 1$, $p_{\mathcal{D}_\mathsf{a}}[\boldsymbol{\mu}_\mathsf{a}] > p_{\mathcal{D}_\mathsf{b}}[\boldsymbol{\mu}_\mathsf{a}]$, and $p_{\mathcal{D}_\mathsf{b}}[\boldsymbol{\mu}_\mathsf{b}] > p_{\mathcal{D}_\mathsf{a}}[\boldsymbol{\mu}_\mathsf{b}]$. Given an adversarial perturbation $\boldsymbol{\delta}$, we assume that perturbation $\boldsymbol{\delta}$ is norm-bounded and in the direction $\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}$ as it is known to be an adversarial direction. That is, we assume $\|\boldsymbol{\delta}\|_2 \leq \epsilon$ where ϵ is a non-negative real number. To capture how strong this perturbation $\boldsymbol{\delta}$ can be such that $\psi(\boldsymbol{x}+\boldsymbol{\delta})$ still evaluates to group a, we quantify the values of ϵ for which the following optimization problem is infeasible, thereby guaranteeing $\psi(\boldsymbol{x}+\boldsymbol{\delta}) = \mathsf{a}$:

$$\begin{aligned} \min_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_2\leq\epsilon} \quad & \|\boldsymbol{\delta}\|_2 \\ \text{s.t.} \quad & \mathop{\mathbb{P}}_{\boldsymbol{x}+\boldsymbol{\delta}\sim\mathcal{D}_\mathsf{b}}[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta})] > \mathop{\mathbb{P}}_{\boldsymbol{x}\sim\mathcal{D}_\mathsf{a}}[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta})] \\ \boldsymbol{\delta} \quad & = \eta \cdot \frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2} \text{ where } \eta \in \mathbb{R} \end{aligned} \tag{10}$$

For notational compactness, we denote:

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

This optimization objective in eq. (10) guaranteed to be infeasible when:

$$\epsilon < \max\left\{0, \frac{2b(\gamma-1)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}} + a\gamma + a + \mathrm{f}(\boldsymbol{\mu}_{\mathsf{b}})(1-\gamma)}{b(\gamma-1)}\right\} \tag{11}$$

This result is further detailed in appendix B.3.

Therefore we conclude that $\psi(\boldsymbol{x}+\boldsymbol{\delta}) = \mathsf{a}$ when inequality (5.2) holds. Furthermore, note that the bound in inequality (11) is not tight, as it becomes more loose as γ approaches 0. Within section 5.2, we notice that for a fixed $\boldsymbol{x}$, $\boldsymbol{\Sigma}_{\mathsf{a}}$, and $\boldsymbol{\mu}_{\mathsf{a}}$ for which $\psi(\boldsymbol{x}) = \mathsf{a}$, as γ approaches 0, the set of perturbations $\boldsymbol{\delta}$ for which our estimator predicts the group to be within the minority group, decreases in size.

Relating to experiments in this section on existing face recognition datasets, it is the numerical imbalance in demographic groups which affects the strength of the perturbation necessary to perturb an image in one demographic such that it is classified as a member of another demographic. Firstly, the numerical imbalance in groups affects the local Lipschitz constants for each demographic groups; this manifests itself as a discrepancy in groupwise covariance and is captured by γ as discussed previously. Furthermore, given the numerical imbalances of demographics in training sets, it is likely that the embedding of a face within a minority demographic is near the majority demographic than it is for the embedding of a face within a majority demographic to be near the embeddings within a minority demographic. Hence, it tends to be easier for a member of a minority demographic to be obfuscated into the majority demographic than vice versa. This analysis agrees with the perturbation norms shown in fig. 6.

5.3 Balancing the Dataset

We train three metric embedding networks of the FaceNet architecture. The first is trained on all of the VGGFace2 [6] dataset's training split, a total of 8631 identities. The second dataset, henceforth referred to as Sex-Balanced VGGFace2, 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. We remove examples from VGGFace2 so we do not introduce any artifacts data augmentation. We use similar reasoning to construct a dataset consisting of 307 identities for each of the race groups. We call this dataset, containing 1842 total identities, Race-Balanced VGGFace2. Race-Balanced VGGFace2 and Sex-Balanced VGGFace2 are the training sets used to obtain Race-Balanced FaceNet and Sex-Balanced FaceNet, respectively.

Balanced Training Empirical Performance: Tables 3 and 4 depict the performance of FaceNet metric embedding networks trained on balanced data. Compared to the models trained on the original VGGFace2 dataset (table 5), the models trained on balanced data exhibit performance that is more uniform, as measured by the parametrized true positive rate $\mathrm{TPR}_{0.2}$, across the demographics upon which training data was balanced. There does appear to be a difference in performance depending on whether examples are matched against only the same demographic group (table 3) or whether examples are matched against examples in all demographic groups (table 4). Even though the models are trained on balanced data, performance disparities on the demographics upon which the training data are balanced are still statistically significant: utilizing Welch's t-Test with unequal variance, the p-values for the sex demographic are 0.0002 and 7.47×10^{-5} when matching on the same and different demographics, respectively. Utilizing the Alexander-Govern test, the p-values for the race demographic are 0.000952 and 1.37×10^{-56} when matching on the same and

	Sex-balanced		Race-balanced			
	Male	**Female**	**White**	**Asian**	**Black**	**Indian**
$\text{TPR}_{0.2}$	.5884	.5184	.5918	.6480	.6145	.8000
AUC	.7802	.7382	.7845	.8049	.7786	.9200
N	10000	5000	10000	2500	1240	20

Table 3: Same demographic matching performance on LFW. The performance, when conditioning on race, is evaluated on the model trained with race-balanced data. The performance, when conditioning on sex, is evaluated on the model trained with sex-balanced data.

	Sex-balanced		Race-balanced			
	Male	**Female**	**White**	**Asian**	**Black**	**Indian**
$\text{TPR}_{0.2}$	.6320	.6568	.6224	.8304	.6629	.8000
AUC	.8016	.8143	.7984	.8858	.8215	.8450
N	10000	5000	10000	2500	1240	20

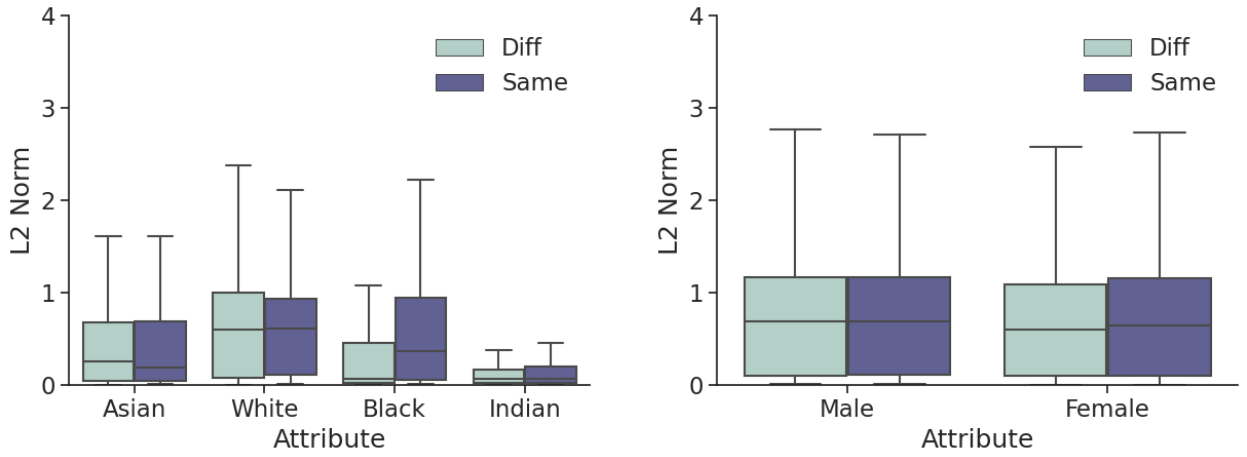
Table 4: Any demographic matching performance on LFW. The setting is that of table 3

different demographics, respectively. Given the statistical significance of these performance differences, we conclude that a balanced training set is not the only factor which mitigates performance disparities in face recognition systems.

Perturbation Norms Induced By Face Obfuscation on Models Trained with Balanced Data: To evaluate the demographic disparities in face obfuscation performance, we plot the strength of perturbations generated by Face-Off on the sex-balanced and race-balanced FaceNets. Though the models are trained on balanced data, statistically significant differences still exist between demographics: utilizing Welch's t-Test with unequal variance, the p-values for the sex demographic are 7.40×10^{-8} and 0.0157857 when matching on the same and different demographics, respectively. Utilizing the Alexander-Govern test, the p-values for the race demographic are both $< 10^{-150}$ when matching on the same and different demographics. For a model trained on the Sex-Balanced VGGFace2, there is less statistical significance in performance disparity when compared to the pretrained FaceNet standard model. On the other hand, balancing on race yields statistically significant performance disparities. We attribute this increased statistical significance associated with a race balanced training set to: 1) a the greatly reduced size of Race-Balanced VGGFace2 compared to the original, unmodified VGGFace2 dataset and 2) empirical results suggest metric embedding networks learn to distinguish skin tones (see discussion regarding fig. 3), not the race attribute itself. Additional results regarding fig. 9 are deferred to the appendix.

Success of Face Obfuscation on Models With Balanced Training Data: Figure 10 reports the obfuscation success rate for the targeted attacks; these attacks are generated in the white-box setting. As evident from the plots, there is a discrepancy in the obfuscation success rate between the female and male attributes. In the same sex scenario (fig. 10a-right), the obfuscation success rate of male faces outperforms that of female faces, which is opposite to the different sex scenario (fig. 10a-left). After balancing the dataset, the gaps between the male and female users diminish (fig. 10b).

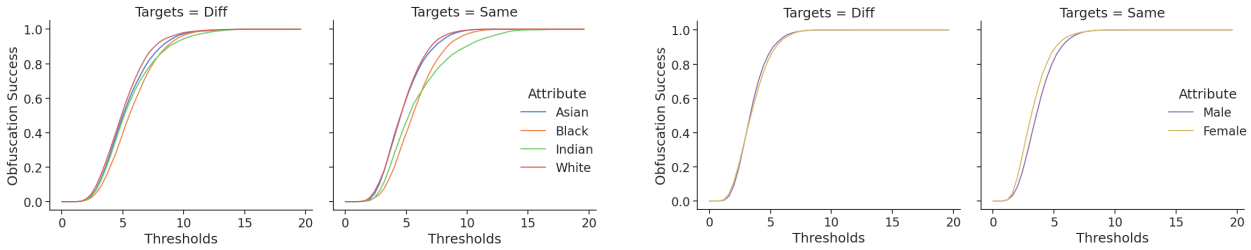
Visualizing Embeddings from Models Trained on Balanced Data: To visualize what metric embedding networks trained on balanced data learn, we plot the resulting structure of their embeddings using t-SNE in fig. 11. Compared to the t-SNE plots for pretrained models (fig. 2), demographic groups appear to be less distinct within the embedding space. Though demographic groups are less distinct, clustering behavior is



(a) Race Balanced Training Data

(b) Sex Balanced Training Data

Figure 9: The distribution of adversarial perturbation sizes generated using the CW [48] attack on the Race-Balanced and Sex-Balanced models.



(a) Obfuscation success rates on the Race-Balanced FaceNet: Different (Left) vs. Same (Right) Race

(b) Obfuscation success rates on the Sex-Balanced FaceNet: Different (Left) vs. Same (Right) Sex

Figure 10: Untargeted obfuscation success on the Race-Balanced and Sex-Balanced models in a white-box setting.

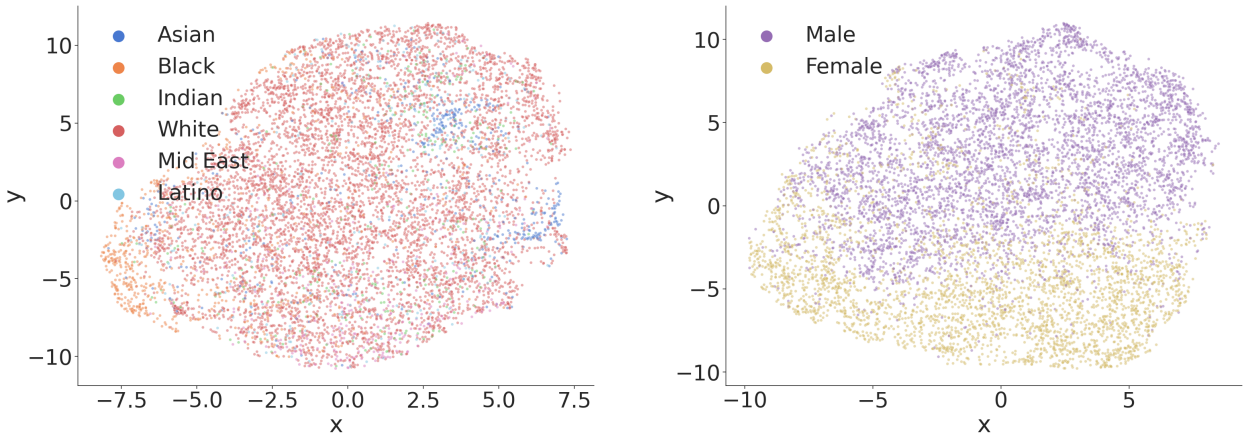
still apparent, which indicates that demographic awareness persists in the face of dataset balancing.

6 Discussion

In this section, we discuss the main insights about the performance of obfuscation systems and their demographic disparities. We draw these insights from the analytical model of section 4.3 and evaluations of section 5.

6.1 Potential Remedies

We have discussed how disparities in the demographic distributions of the training set can significantly impact face obfuscation performance. However, this is not an easily solvable problem. Finding a diverse large-scale dataset to train face recognition models has proven difficult. Datasets, such as FairFace [49] and IBM's Diversity in Faces [50], attempt to address this problem. But, as seen in our experiments, balancing demographics in a dataset is not straightforward. For example, a commonly used proxy for skin-tone is race, but the two are not perfectly correlated. Additionally, users may be interested in combinations of different



(a) Race Balanced - Race

(b) Sex Balanced - Sex

Figure 11: t-SNE [20] of the embedding spaces generated using the VGGFace2 [6] datasets. Figures 11a and 11b use as embedding functions Race-Balanced FaceNet and Sex-Balanced FaceNet, respectively. Embeddings of identities are colored by race and sex. Distinct clusters are less evident for each demographic group.

demographic attributes, which are difficult to balance in practice. Some work has sought to address this: Serna et al. propose Sensitive Loss, a "discrimination-aware" Triplet Loss derived tuning procedure for pre-trained models [51]. This tuning procedure involves adding a layer to the network, then tweaking this layer using only identities drawn from the same demographic group.

Demographic imbalance in the training set does not account for all possible demographic disparities in face obfuscation. Imbalances in the evaluation set can affect the privacy utility of a model trained on a demographically balanced dataset. Consider the pathological example in which a social media user has $m \gg 1$ connections. Of the m connections, $m-1$ of them are demographic group a and the last connection is in demographic group b. The $m-1$ individuals in group a can easily hide their identities amongst each other as their embeddings are relatively clustered, yet as our experiments have shown, the one connection in group b will likely have trouble hiding his identity because his embedding is likely to be distant from the embeddings of individuals in group a. This provides additional evidence for the work of Kulynych et al [52], who demonstrate membership inference attacks [53] may be more successful against minority groups.

6.2 Threats to Validity

We discuss some of the limitations of our analytical and experimental approaches. Our intention for section 4.3 is to create 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. For example, the construction of PCA means there is no difference in local Lipschitz constants when conditioned on demographic groups. Second, our results regarding local Lipschitz constants is only an approximation, as the non-linear activation functions of large-scale neural networks are computationally complex. The evaluated networks are not of typical metric embedding network architecture, but are instead small classification models. Estimating local Lipschitz constant on large embedding networks is infeasible given the state of current estimators.

Another threat to validity is the limited number of publicly available face recognition datasets are scarce, and datasets with labeled demographic attributes are even harder to come by. Such publicly available datasets are rather small, on the order of thousands of identities. As such, our results may not general-

ize to datasets orders of magnitude larger than currently available face recognition datasets.

7 Related Work

Researchers have recently started looking into the intersection of face recognition and machine learning fairness. Buolamwini and Gebru study commercially available face recognition datasets and classifiers [14]. Their findings indicate three prominent commercial face classifiers, across three benchmark datasets, have disparate performance across demographic groups. Follow-up research has attempted to address these demographic disparities through balanced datasets [49, 54]. Such datasets improve the generalization performance but do not address demographic disparities completely. These findings are consistent with our results in section 5, where we observe that a demographically balanced dataset may still contain biases towards specific attributes such as skin tone.

In general, demographic disparities seem to contribute to the phenomenon of overlearning, a term coined by Song and Shmatikov [55]. Overlearning refers to the phenomenon in which models implicitly learn to recognize sensitive patterns not part of the original learning objective. Indeed, metric embedding networks trained on faces do overlearn. They learn not just a dimensionality reduction on faces, but a demographically aware dimensionality reduction on such faces.

More related to our research, Nanda et al. discuss the relationship between fairness and robustness of face recognition [15]. They show that faces from minority groups require smaller adversarial perturbations to misclassify compared to those for majority groups; this discrepancy is a factor of the underlying datasets, models, and hyperparameter tuning. Our work differs along three dimensions. First, we study the robustness of face recognition in the context of the metric embedding networks – the real-world setting for such systems. Second, we present an empirical characterization of face obfuscation systems that build on evasion attacks; our characterization extends beyond the perturbation size to the success rate in white-box and black-box settings. Further, we study the real-world implications of demographic disparities of face obfuscation systems, which prevent a user from choosing targets outside their demographic group. Third, we present an analytical model of an embedding function to reason about the robustness of face recognition and obfuscation systems. Finally, a Master's thesis by Qin [56] evaluates the performance discrepancies in the FAWKES [30] conditioned on skin tones. FAWKES is a data poisoning-based technique for face obfuscation. Using a well-established image similarity scoring function known as DSSIM, Qin finds differences in perturbation visibility for certain demographics. In comparison, our work characterizes why these discrepancies exist in the adversarial setting and how they impact the privacy utility of all face obfuscation systems.

8 Conclusion

Face recognition systems have seen increased usage in online and physical 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, but also decrease the performance of face obfuscation among underrepresented demographic groups. Imbalances in training set demographics are indicative of the root of these privacy utility issue. To mitigate the effects of this incidental privacy leak, we must not only develop loss functions for training fair metric embedding networks, but also develop techniques to characterize if such privacy leaks will occur. Further, we must identify optimization procedures which mitigate such privacy leaks.

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] 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.

[17] 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.

[18] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," 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. 322–332. [Online]. Available: https://proceedings.mlr.press/v97/arora19a.html

[19] K. P. F.R.S., "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901. [Online]. Available: https://doi.org/10.1080/14786440109462720

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

[21] 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

[22] 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

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

[24] 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

[25] 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.

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

[27] 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/

[28] 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

[29] 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.

[30] 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

[31] 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.

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

[33] 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.

[34] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[35] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 343–347.

[36] 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.

[37] 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

[38] 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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[47] 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

[48] 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.

[49] 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.

[50] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith, "Diversity in faces," 2019.

[51] 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.

[52] B. Kulynych, M. Yaghini, G. Cherubin, M. Veale, and C. Troncoso, "Disparate vulnerability to membership inference attacks," *arXiv e-prints*, pp. arXiv–1906, 2019.

[53] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[54] 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

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

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

[57] C. Dan, Y. Wei, and P. Ravikumar, "Sharp statistical guaratees for adversarially robust gaussian classification," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2345–2355.

[58] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *In 7th International Conference on Learning Representations (ICLR)*, 2019.

[59] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[60] 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.

A Metric Embedding Networks with a Gaussian Mixture Model

To further motivate the connection between robustness of an embedding and its Lipschitz constant, we consider a Gaussian mixture model. These models have been considered in the theoretical analysis of robustness in the classification setting [57,58]. illustrates this Gaussian mixture model setting. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the Gaussian distribution in $\mathbb{R}^d$ with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}$ a $d \times d$ positive-definite matrix. We will consider Gaussian distributions of the form $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}_d)$.

Let $\mathcal{X} \times \mathcal{Y}$ (where $\mathcal{Y} = \{-1, 1\}$) be generated from a distribution $\mathcal{D}$ as follows: $y \in \mathcal{Y}$ is equally probable with probability $\frac{1}{2}$ and given y, generate z according to $\mathcal{N}(y\boldsymbol{\mu}, \boldsymbol{I}_d)$. Furthermore, assume that ρ_θ is L-Lipschitz, i.e., for all z and z' in $\mathcal{X}$:

$$\rho_\theta(\mathrm{z}, \mathrm{z'}) \leq \quad L\rho(\mathrm{z}, \mathrm{z'})$$

for choice of metric ρ.

We have the following concentration of measure result from Theorem 5.2.2 [59].

Theorem A.1. (Gaussian concentration) Consider a random vector $\mathrm{z} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$ and a Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$. Then

$$\| f(\mathrm{z'}) - \mathbb{E}\left[f(\mathrm{z})\right] \|_{\psi_2} \quad \leq \quad C \parallel f \parallel_{Lip},$$

where $\| f \|_{Lip}$ is the Lipschitz constant of f, and $\| \cdot \|_{\psi_2}$ is the sub-Gaussian metric.

Let $\mathcal{S}^d$ be a d-dimensional unit-sphere. Consider a metric embedding network model $f_\theta : \mathbb{R}^d \to \mathcal{S}^d$, and let ρ_θ be the associated distance metric. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_{-1}$ be the for labels 1 and -1 respectively. Consider the two functions defined as follows: $f_1(\mathrm{z}) = \rho_\theta(\boldsymbol{x}_1, \mathrm{z})$ and $f_{-1}(\mathrm{z}) = \rho_\theta(\boldsymbol{x}_{-1}, \mathrm{z})$ (the functions correspond to the distances from the two anchors).

$$\begin{aligned} \omega_1 &= \mathop{\mathbb{E}}_{\mathrm{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}_d)} \left[f_1(\mathrm{z})\right] \\ \omega_{-1} &= \mathop{\mathbb{E}}_{\mathrm{z} \sim \mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{I}_d)} \left[f_{-1}(\mathrm{z})\right] \end{aligned}$$

We first show that f_1 is L-Lipschitz if f_θ is L-Lipschitz. Take $\mathrm{z}, \mathrm{z'} \in \mathcal{X}$,

$$\begin{aligned} |f_1(\mathrm{z}) - f_1(\mathrm{z'})| &= |\rho_\theta(\boldsymbol{x}_1, \mathrm{z}) - \rho_\theta(\boldsymbol{x}_1, \mathrm{z'})| \\ &\leq \rho_\theta(\mathrm{z}, \mathrm{z'}) \\ &\leq L\rho(\mathrm{z}, \mathrm{z'}) \end{aligned}$$

As a result, $\| f_1 \|_{Lip} = \| f_\theta \|_{Lip}$. Intuitively, if the Lipschitz constant of f_1 is lower, the points drawn from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}_d)$ get closer to ω_1. In other words, as the Lipschitz constant of embedding gets smaller, the "point clouds" corresponding to the two Gaussian distributions in the mixture get farther apart, because they are concentrated more around their means.

Next we formalize this intuition. Let $E\left[\mathrm{z}, \boldsymbol{x}_1, \boldsymbol{x}_{-1}\right]$ represent the event that z is closer to $\boldsymbol{x}_{-1}$ than $\boldsymbol{x}_1$. We prove the following:

$$P_{\mathrm{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}_d)}(1_{\mathbb{E}[\mathrm{z}, \boldsymbol{x}_1, \boldsymbol{x}_{-1}]}) \leq 2 \exp\left(-\frac{C' z^2}{\| f_1 \|_{Lip}}\right) \tag{12}$$

In the equation given above, $C' > 0$ is a positive constant, and z is given by the following expression:

$$\frac{\rho_\theta(\boldsymbol{x}_{-1}, \boldsymbol{x}_1)}{2} - \omega_1$$

Notice that $P_{\mathrm{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)}(1_{\mathbb{E}[\mathrm{z},\boldsymbol{x}_1,\boldsymbol{x}_{-1}]})$ represents the probability that a point drawn from $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)$ is closer to $\boldsymbol{x}_{-1}$ than $\boldsymbol{x}_1$, and hence represents an "undesirable event". Also note that the upper bound goes down as the Lipschitz constant $\parallel f_1 \parallel_{Lip}$ goes down, and thus confirming our intuition. Next we prove Equation 12.

Let X be a sub-Gaussian random variable, then the following equation is well-known:

$$P(\mid X \mid \geq t) \quad \leq \quad 2\exp\left(\frac{-ct^2}{\parallel X \parallel_{\psi_2}^2}\right) \tag{13}$$

To prove the Equation 12, we use the following sequence of inequalities (let $q = P_{\mathrm{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)}(1_{\mathbb{E}[\mathrm{z},\boldsymbol{x}_1,\boldsymbol{x}_{-1}]})$)

$$\begin{aligned} q &\leq P_{\mathrm{z}\sim\mathcal{N}(\boldsymbol{\mu},I_d)}\left(f_1(\mathrm{z}) \geq \frac{\rho_\theta(\boldsymbol{x}_{-1},\boldsymbol{x}_1)}{2}\right) \\ &\leq P_{\mathrm{z}\sim\mathcal{N}(\boldsymbol{\mu},I_d)}\left(\mid f_1(\mathrm{z}) - \omega_1 \mid \geq \frac{\rho_\theta(\boldsymbol{x}_{-1},\boldsymbol{x}_1)}{2} - \omega_1\right) \\ &\leq 2\exp\left(-\frac{C'z^2}{\parallel f_1 \parallel_{Lip}}\right) \end{aligned}$$

The first step follows from the following observation: if $f_1(\mathrm{z})$ is less than $\frac{\rho_\theta(\boldsymbol{x}_{-1},\boldsymbol{x}_1)}{2}$ then z is closer to $\boldsymbol{x}_1$ than $\boldsymbol{x}_{-1}$. The next two steps use Theorem A.1 and Equation 13.

B Supplemetal Mathematics for the Analytical Model

In this section we perform provide additional results for our analytical model as described in **??**

B.1 Important Lemmas and Identities

Lemma B.1 (Law of Total Variance). Let X and Y be random variables in the same probability space. Furthermore, let the variance of both X and Y be finite, then

$$\mathrm{Var}(Y) = \mathbb{E}\left[\mathrm{Var}\left(Y \mid X\right)\right] + \mathrm{Var}\left(\mathbb{E}\left[Y \mid X\right]\right) \tag{14}$$

Lemma B.2 (Law of Total Covariance). Let X, Y and Z be random variables in the same probability space. Furthermore, let the covariance of X and Y be finite, then

$$\begin{aligned} \mathrm{cov}(X,Y) = \mathbb{E}\left[\mathrm{cov}(X,Y \mid Z)\right] \\ + \mathrm{cov}\left(\mathbb{E}\left[X \mid Z\right], \mathbb{E}\left[Y \mid Z\right]\right) \end{aligned} \tag{15}$$

B.2 Decomposing the Covariance

Let us compute the diagonal components of the overall covariance matrix $\boldsymbol{\Sigma}$ of synthetic data distribution $\mathcal{D}$ in terms of $\boldsymbol{\Sigma}_{\mathsf{a}}$, $\boldsymbol{\Sigma}_{\mathsf{b}}$ the covariance matrices for identity distributions $\mathcal{D}_{\mathsf{a}}$ and $\mathcal{D}_{\mathsf{b}}$, respectively:

$$\begin{aligned} (\boldsymbol{\Sigma})_{jj} &= \alpha_{\mathsf{a}}(\boldsymbol{\Sigma}_{\mathsf{a}})_{jj} + (1-\alpha_{\mathsf{a}})(\boldsymbol{\Sigma}_{\mathsf{b}})_{jj} \\ &\quad + \alpha_{\mathsf{a}}\left((\boldsymbol{\mu}_{\mathsf{a}})_j - (\boldsymbol{\mu}_{\mathsf{ab}})_{jj}\right)^2 \\ &\quad + (1-\alpha_{\mathsf{a}})\left((\boldsymbol{\mu}_{\mathsf{b}})_j - (\boldsymbol{\mu}_{\mathsf{ab}})_{jj}\right)^2 \end{aligned} \tag{16}$$

$$\begin{aligned} &= \alpha_{\mathsf{a}}(\boldsymbol{\Sigma}_{\mathsf{a}})_{jj} + (1-\alpha_{\mathsf{a}})(\boldsymbol{\Sigma}_{\mathsf{b}})_{jj} \\ &\quad + \alpha_{\mathsf{a}}(1-\alpha_{\mathsf{a}})((\boldsymbol{\mu}_{\mathsf{a}})_j - (\boldsymbol{\mu}_{\mathsf{b}})_j)^2 \end{aligned} \tag{17}$$

Now we compute the off-diagonal components of the overall covariance matrix $\mathbf{\Sigma}$ in terms of $\mathbf{\Sigma}_\mathsf{a}$ and $\mathbf{\Sigma}_\mathsf{b}$:

$$
\begin{aligned}
(\mathbf{\Sigma})_{jk} &= \alpha_\mathsf{a} \left(\mathbf{\Sigma}_\mathsf{a}\right)_{jk} + (1-\alpha_\mathsf{a}) \left(\mathbf{\Sigma}_\mathsf{b}\right)_{jk} \\
&\quad + \alpha_\mathsf{a} \Big((\boldsymbol{\mu}_\mathsf{a})_j - (\alpha_\mathsf{a}(\boldsymbol{\mu}_\mathsf{a})_j + (1-\alpha_\mathsf{a})(\boldsymbol{\mu}_\mathsf{b})_j) \Big) \\
&\quad \times \Big((\boldsymbol{\mu}_\mathsf{a})_k - (\alpha_\mathsf{a}(\boldsymbol{\mu}_\mathsf{a})_k + (1-\alpha_\mathsf{a})(\boldsymbol{\mu}_\mathsf{b})_k) \Big) \\
&\quad + (1-\alpha_\mathsf{a}) \Big((\boldsymbol{\mu}_\mathsf{b})_j - (\alpha_\mathsf{a}(\boldsymbol{\mu}_\mathsf{a})_j + (1-\alpha_\mathsf{a})(\boldsymbol{\mu}_\mathsf{b})_j) \Big) \\
&\quad \times \Big((\boldsymbol{\mu}_\mathsf{b})_k - (\alpha_\mathsf{a}(\boldsymbol{\mu}_\mathsf{a})_k + (1-\alpha_\mathsf{a})(\boldsymbol{\mu}_\mathsf{b})_k) \Big) \qquad (18) \\
&= \alpha_\mathsf{a} (\mathbf{\Sigma}_\mathsf{a})_{jk} + (1-\alpha_\mathsf{a}) (\mathbf{\Sigma}_\mathsf{b})_{jk} \\
&\quad + \alpha_\mathsf{a} \Big((\boldsymbol{\mu}_\mathsf{a})_j - (\alpha_\mathsf{a}(\boldsymbol{\mu}_\mathsf{a})_j + (1-\alpha_\mathsf{a})(-\boldsymbol{\mu}_\mathsf{a})_j) \Big) \\
&\quad \times \Big((\boldsymbol{\mu}_\mathsf{a})_k - (\alpha_\mathsf{a}(\boldsymbol{\mu}_\mathsf{a})_k + (1-\alpha_\mathsf{a})(-\boldsymbol{\mu}_\mathsf{a})_k) \Big) \\
&\quad + (1-\alpha_\mathsf{a}) \Big((-\boldsymbol{\mu}_\mathsf{a})_j - (\alpha_\mathsf{a}(\boldsymbol{\mu}_\mathsf{a})_j + (1-\alpha_\mathsf{a})(-\boldsymbol{\mu}_\mathsf{a})_j) \Big) \\
&\quad \times \Big((-\boldsymbol{\mu}_\mathsf{a})_k - (\alpha_\mathsf{a}(\boldsymbol{\mu}_\mathsf{a})_k + (1-\alpha_\mathsf{a})(-\boldsymbol{\mu}_\mathsf{a})_k) \Big) \qquad (19)
\end{aligned}
$$

Putting it all together, we arrive at the following decompostion of the synthetic data distribtion covariance matrix:

$$
(\mathbf{\Sigma}) = \alpha_\mathsf{a}(\mathbf{\Sigma}_\mathsf{a}) + (1-\alpha_\mathsf{a})(\mathbf{\Sigma}_\mathsf{b}) + 4\alpha_\mathsf{a}(1-\alpha_\mathsf{a})\boldsymbol{\mu}_\mathsf{a}\boldsymbol{\mu}_\mathsf{a}^\mathsf{T} \qquad (20)
$$

B.3 Adversarial Attacks

To solve the optimization problem posed in eq. (10), we examine the likelihood function. Let $p_{\mathcal{D}_\mathsf{g}}$ now denote the PDF of the image of distribution $\mathcal{D}_\mathsf{g}$ 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 $\frac{p_{\mathcal{D}_\mathsf{a}}[\mathsf{f}(\boldsymbol{x}+\boldsymbol{\delta})]}{p_{\mathcal{D}_\mathsf{b}}[\mathsf{f}(\boldsymbol{x}+\boldsymbol{\delta})]}$ exceeds one, where:

$$1 \leq \left(\frac{p_{\mathcal{D}_\mathsf{a}}\left[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta}) \mid \boldsymbol{\delta} \propto \frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right]}{p_{\mathcal{D}_\mathsf{b}}\left[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta}) \mid \boldsymbol{\delta} \propto \frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right]} \right) \tag{21}$$

$$\begin{aligned} &= \left((2\pi)^{-1/2}(\mathrm{f}(\boldsymbol{q}_1))^{-1}\right) \\ &\quad \times \exp\left\{-\frac{1}{2}\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}-\boldsymbol{\mu}_\mathsf{a}\right)\right]^2(\mathrm{f}(\boldsymbol{q}_1))^{-2}\right\} \\ &\quad \left[\left((2\gamma\pi)^{-1/2}(\mathrm{f}(\boldsymbol{q}_1))^{-1}\right)\right. \\ &\quad \left.\times \exp\left\{-\frac{\gamma}{2}\cdot\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}-\boldsymbol{\mu}_\mathsf{b}\right)\right]^2(\mathrm{f}(\boldsymbol{q}_1))^{-2}\right\}\right]^{-1} \end{aligned} \tag{22}$$

$$\begin{aligned} &\leq \exp\left\{-\frac{1}{2}\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right)+\mathrm{f}\left(\boldsymbol{\mu}_\mathsf{b}\right)\right]^2\right\} \\ &\quad \times \exp\left\{-\frac{\gamma}{2}\cdot\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right)-\mathrm{f}\left(\boldsymbol{\mu}_\mathsf{b}\right)\right]^2\right\}\Bigg]^{-1} \end{aligned} \tag{23}$$

We now solve the following the following inequality for η

$$\begin{aligned} 1 \leq &\exp\left\{-\frac{1}{2}\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right)+\mathrm{f}\left(\boldsymbol{\mu}_\mathsf{b}\right)\right]^2\right\} \\ &\times \exp\left\{-\frac{\gamma}{2}\cdot\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right)-\mathrm{f}\left(\boldsymbol{\mu}_\mathsf{b}\right)\right]^2\right\}\Bigg]^{-1} \end{aligned} \tag{24}$$

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

$$\eta > \frac{2b(1-\gamma)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}}+a\gamma+a+\mathrm{f}(\boldsymbol{\mu}_\mathsf{b})(1-\gamma)}{b(\gamma-1)} \tag{25}$$

AND

$$\eta < \frac{2b(\gamma-1)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}}+a\gamma+a+\mathrm{f}(\boldsymbol{\mu}_\mathsf{b})(1-\gamma)}{b(\gamma-1)} \tag{26}$$

Where, for notational compactness, we denoted the following:

$$a = \mathrm{f}(\boldsymbol{x}) \tag{27}$$

$$b = \mathrm{f}\left(\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right) \tag{28}$$

Since the right-side of inequality (24) upper bounds the right-side of inequality (21), we know that any solution for inequality (24) is also a solution for inequality (21).

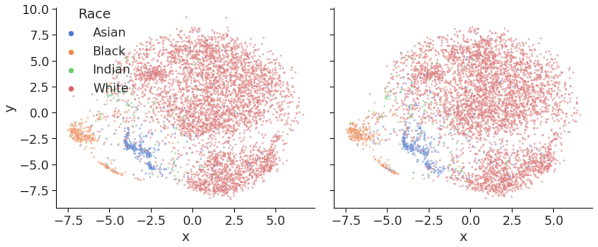
Since inequality (26) is a bound on ϵ which provides a guarantee on when eq. (10) is infeasible. Hence we conclude that eq. (10) may be feasible only when

$$\epsilon \geq \max \left\{ 0, \frac{2b(\gamma-1)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}} + a\gamma + a + \mathrm{f}(\boldsymbol{\mu}_\mathsf{b})(1-\gamma)}{b(\gamma-1)} \right\} \tag{29}$$

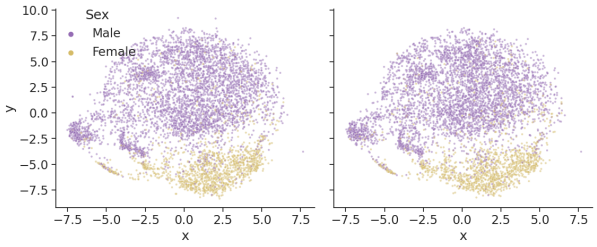
C Additional Experiments

C.1 Face Obfuscation's Impact on Demographic

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 [60]. Of each identity in the dataset, only 8% of the identities' race attribute change after adversarial perturbations are added. We visualize this in fig. 12, where the embeddings of the natural examples and adversarial examples are plotted according to their demographics using t-SNE. Comparing the plots within figs. 12a and 12b, a heavy overlap is observed between the embeddings of the natural and adversarial examples of the same race and sex. These results further addressed confirm our intuition from section 5.2; an adversarial perturbation does not push away the sample from its demographic group, when the groups are clustered.



(a) Race: Natural (Left) vs. Adversarial (Right)



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

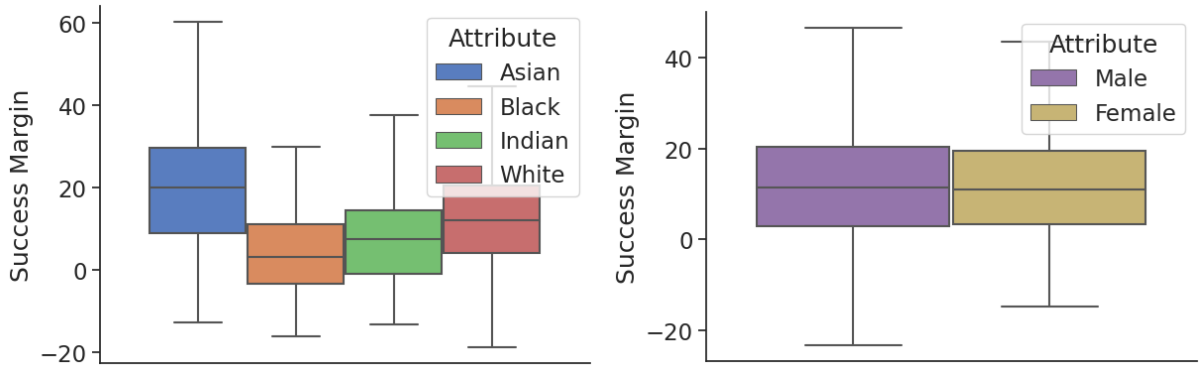
Figure 12: t-SNE [20] of the embedding spaces generated using both natural and untargeted adversarial LFW [5] examples. Embeddings of identities are colored by race and sex. Note that the clusters do not change: embeddings for adversarial images are still within the source demographic.

C.2 Obfuscation Disparities in Black-Box Settings

To understand whether the performance disparities between demographics (as discussed in section 5.2) manifest in commercial face recognition systems, we tested the success of the perturbed faces against the Face++ face recognition API. In fig. 13. We observe large differences in obfuscation success rates dependent on the race demographic. This performance disparity can be attributed to the larger perturbation sizes associated with Asian and White identities (fig. 6). These demographics were observed to have higher local Lipschitz constants (fig. 4), suggesting real-world consequences stemming from disparities in the robustness of each demographic.

C.3 Targeted and Untargeted Obfuscation Success Rates

The results depicted in fig. 14 portray the untargeted obfuscation success rates for the pre-trained Facenet model. These untargeted obfuscation success rates are also provided for the black-box setting with the

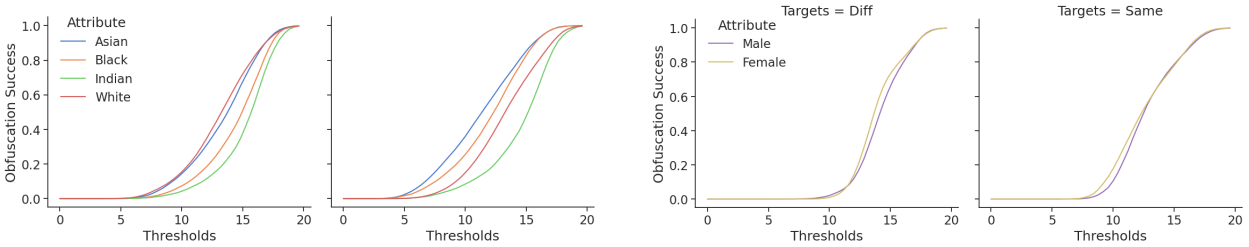


(a) Untargeted Face++: Race

(b) Untargeted Face++: Sex

Figure 13: Untargeted adversarial examples generated using LowKey [8] and evaluated on Face++.

OpenFace model (fig. 15). In fig. 16, the targeted success rates are provided for the Race-Balanced and Sex-Balanced Facenet models. We observe similar trends as discussed in sections 5.2 and 5.3.



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

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

Figure 14: Untargeted obfuscation success evaluated on the FaceNet metric embedding network in a white-box setting.

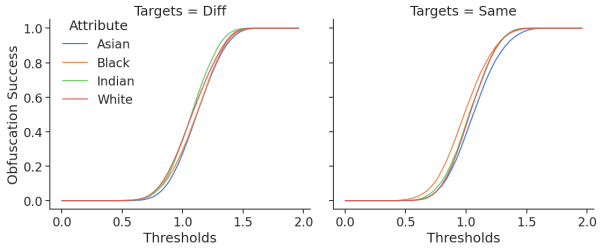
C.4 Imbalanced Facenet

A third dataset containing the entirety of VGGFace2 was used to train another Facenet model, Imbalanced Facenet. The t-SNE visualizations for the imbalanced Facenet models from section 5.3 demonstrate a clustering behavior with fewer overlapping data points than in fig. 11. The performance discrepancies and clusters resulting from imbalances in the dataset are also impacted by the training scheme. Note that the overlearned pre-trained softmax Facenet model in fig. 2 produces much tighter clusters in the embedding space than the Imbalanced Facenet trained using triplet loss.

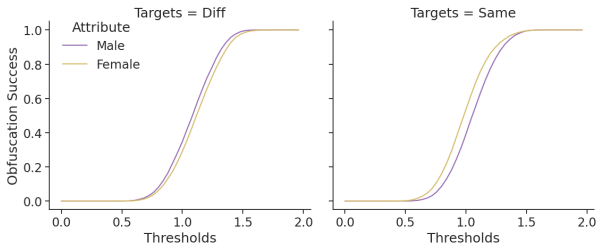
The following tables depict the performance of the Imbalanced Facenet model trained utilizing the same hyperparameters as those used to train the balanced models.

D Statistical Significance Tests

The statistical significance of each result for the experiments performed in section 5 has been verified with the following statistical tests.

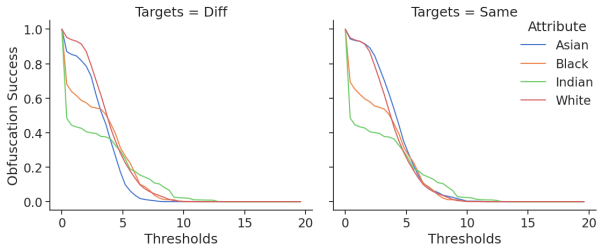


(a) Obfuscation success rates on OpenFace: Different (Left) vs. Same (Right) Race

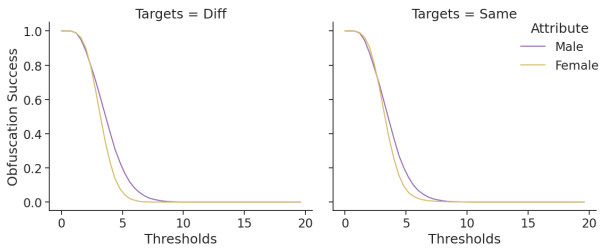


(b) Obfuscation success rates on OpenFace: Different (Left) vs. Same (Right) Sex

Figure 15: Untargeted obfuscation success evaluated on the OpenFace metric embedding network in a black-box setting.



(a) Obfuscation success rates on the Race-Balanced FaceNet: Different (Left) vs. Same (Right) Race



(b) Obfuscation success rates on the Sex-Balanced FaceNet: Different (Left) vs. Same (Right) Sex

Figure 16: Targeted obfuscation success evaluated on the Race-Balanced and Sex-Balanced FaceNets in a white-box setting.

Null Hypothesis D.1. The accuracy of examples in each demographic group is the same. (Statistical Test regarding tables 1 and 2)

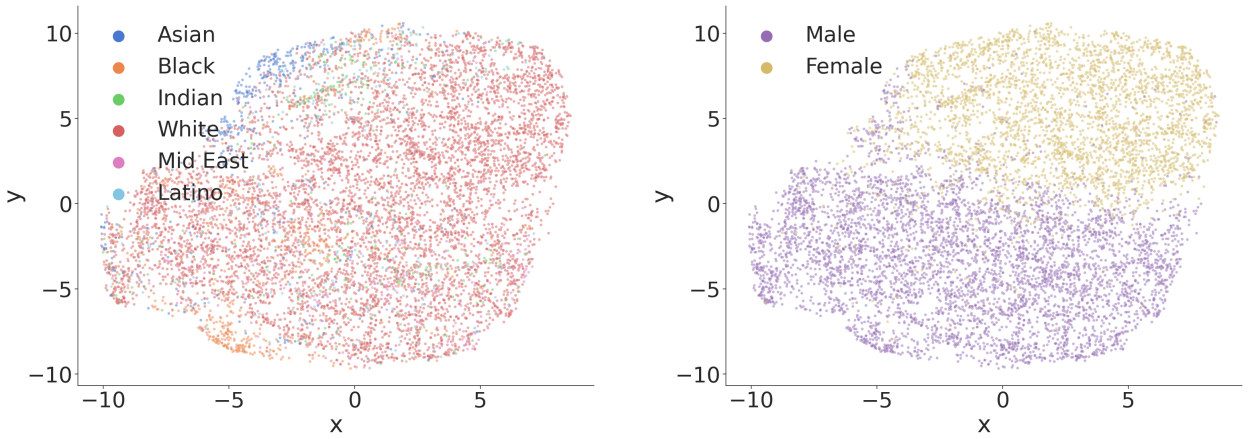
Statistical Test. Utilizing the Alexander-Govern test [37], we conclude that the differences in accuracy between sexes is statistically significant. For the data in table 1, the difference in accuracy between sexes is statistically significant as is the difference in accuracy between races. Utilizing the Alexander-Govern test [37], the pvalues are 5.89×10^{-5} and 0.000995, respectively. For the data in table 2, the difference in accuracy between sexes is statistically significant as is the difference in accuracy between races. Utilizing the Alexander-Govern test [37], the pvalues are 0.24 and 0.00270, respectively.

Null Hypothesis D.2. For a given demographic the mean ℓ_2 norm for a different demographic target is the same as the mean ℓ_2 norm for a same demographic target. (Statistical test regarding fig. 6)

Statistical Test. Utilizing Welch's t-Test with non-equal variance [47], Table 7 compares the ℓ_2 norm of perturbations which have a member of the same demographic group as a target identity, with perturbations that have a member of the different demographic target identity. We see that, for source faces within the Male, Female, Asian and White demographic groups, there is a statistically significant difference in the distributions of perturbations which target the same demographic group and perturbations targeting a different demographic group. For source faces within the Black and Indian demographic groups, the discrepancy is not statistically significant.

Null Hypothesis D.3. The mean ℓ_2 norm perturbation is the same across all population groups. (Statistical test regarding fig. 6)

Statistical Test. Utilizing Welch's t-Test with non-equal variance [47] and the Alexander-Govern test (a multi-sample generalization of Welch's t-test) [37], Table 8 compares the ℓ_2 norm of perturbations across the



(a) Imbalanced - Race

(b) Imbalanced - Sex

Figure 17: t-SNE [20] of the embedding spaces generated using, as the embedding function, the metric embedding network trained on the entirety of the VGGFace2 [6] dataset. Embeddings of identities are colored by race and sex.

	Male	**Female**	**White**	**Asian**	**Black**	**Indian**
$\mathrm{TPR}_{0.2}$	.5616	.5136	.6108	.6112	.5645	.8000
AUC	.7669	.7307	.7951	.7952	.7538	.8400
N	10000	5000	10000	2500	1240	20

Table 5: Same demographic matching performance on LFW and imbalanced.

sexes and across a selection of racial groups. There is a statistically significant difference in the distributions of perturbation strength when conditioning by demographic group.

Null Hypothesis D.4. When a metric embedding network is trained on balanced data, the parametrized true positive rate $\mathrm{TPR}_{0.2}$ is equal for the demographic groups upon which the training dataset was balanced. (Statistical test regarding tables 3 and 4)

Statistical Test. Utilizing Welch's t-Test with non-equal variance [47] and the Alexander-Govern test (a multi-sample generalization of Welch's t-test) [37], Table 9 shows the statistical significance of the parametrized true positive rate $\mathrm{TPR}_{0.2}$ among population groups. Though the training data are balanced, we conclude that the differences in parametrized true positive rate $\mathrm{TPR}_{0.2}$ are still statistically significant.

Null Hypothesis D.5. For a given demographic the mean ℓ_2 norm for a different demographic target is the same as the mean ℓ_2 norm for a same demographic target. This test depicts the difference in the Race-Balanced and Sex-Balanced Facenets. (Statistical test regarding fig. 9)

Statistical Test. Utilizing Welch's t-Test with non-equal variance [47], Table 10 compares the ℓ_2 norm of perturbations which have a member of the same demographic group as a target identity, with perturbations that have a member of the different demographic target identity. For the Race-Balanced FaceNet, only Male, Female and Black demographic groups have a statistically significant difference in the distributions of perturbations which target the same demographic group and perturbations targeting a different demographic group. For the Sex-Balanced FaceNet, only for the female demographic is there is a statistically significant difference in the distributions of perturbations which target the same demographic group and perturbations targeting a different demographic group. For source faces within the Black and Indian demographic groups,

	Male	**Female**	**White**	**Asian**	**Black**	**Indian**
$\mathrm{TPR}_{0.2}$	.6166	.6936	.6404	.7232	.6516	.7000
AUC	.7966	.8314	.8055	.8532	.8084	.8300
N	10000	5000	10000	2500	1240	20

Table 6: Any demographic matching performance on LFW and imbalanced.

Demographic	p**-value**
Male	6.61×10^{-24}
Female	2.00×10^{-26}
Asian	0.00110
Black	0.241
White	0.00180
Indian	0.0859

Table 7: Statistical significance of perturbations targeted at examples in the same demographic group vs. targeted at examples in a different demographic group. p-values were obtained by Welch's t-test with non-equal variance.

the discrepancy is not statistically significant.

Null Hypothesis D.6. The mean ℓ_2 norm perturbation is the same across all population groups. This test depicts the difference in the Race-Balanced and Sex-Balanced Facenets. (Statistical test regarding fig. 9)

Statistical Test. Utilizing Welch's t-Test with non-equal variance [47] and the Alexander-Govern test (a multi-sample generalization of Welch's t-test) [37], Table 11 compares the ℓ_2 norm of perturbations across the sexes and across a selection of racial groups. There is a statistically significant difference in the distributions of perturbation strength when conditioning by demographic group.

Demographic	p-value
Target: Same Sex	7.94×10^{-5}
Target: Different Sex	2.24×10^{-38}
Target: Same Race	3.03×10^{-170}
Target: Different Race	1.53×10^{-71}

Table 8: Statistical significance of differences, between demographic groups, of perturbations. p-values were obtained by Welch's t-test with non-equal variance and the Alexander-Govern test.

Target Demographic	Race Balanced Training Data p-value	Sex Balanced Training Data p-value
Targets: Same Sex	9.63×10^{-9}	0.000288
Targets: All Data	0.0340	7.47×10^{-5}
Targets: Same Race	0.000952	7.47×10^{-12}
Targets: All Data	1.32×10^{-56}	1.10×10^{-34}

Table 9: Statistical significance of differences, between demographic groups, of $\mathrm{TPR}_{0.2}$. p-values were obtained by Welch's t-test with non-equal variance. and the Alexander-Govern test.

Demographic	Race Balanced Training Data p-value	Sex Balanced Training Data p-value
Male	0.0014576	0.8492978
Female	4.06×10^{-27}	1.03×10^{-12}
Asian	0.1214498	0.8310723
Black	0.0184528	0.3342724
White	0.8952117	0.8796809
Indian	0.8990931	0.3579986

Table 10: Statistical significance of perturbations targeted at examples in the same demographic group vs. targeted at examples in a different demographic group (Race-Balanced and Sex-Balanced). p-values were obtained by Welch's t-test with non-equal variance.

Demographic	Race Balanced Training Data p-value	Sex Balanced Training Data p-value
Target: Same Sex	9.92×10^{-23}	7.40×10^{-8}
Target: Different Sex	1.21×10^{-83}	0.0157857
Target: Same Race	$< 10^{-150}$	$< 10^{-150}$
Target: Different Race	$< 10^{-150}$	2.01×10^{-196}

Table 11: Statistical significance of differences, between demographic groups, of perturbations (Race-Balanced and Sex-Balanced). p-values were obtained by Welch's t-test with non-equal variance.