Towards Characterizing the Semantic Robustness of Face Recognition

Juan C. Pérez[1,2], Motasem Alfarra[1], Ali Thabet[3], Pablo Arbeláez[2], Bernard Ghanem[1]
[1]King Abdullah University of Science and Technology (KAUST),
[2]Center for Research and Formation in Artificial Intelligence, Universidad de los Andes,
[3]Reality Labs

Abstract

Deep Neural Networks (DNNs) lack robustness against imperceptible perturbations to their input. Face Recognition Models (FRMs) based on DNNs inherit this vulnerability. We propose a methodology for assessing and characterizing the robustness of FRMs against semantic perturbations to their input. Our methodology causes FRMs to malfunction by designing adversarial attacks that search for identity-preserving modifications to faces. In particular, given a face, our attacks find identity-preserving variants of the face such that an FRM fails to recognize the images belonging to the same identity. We model these identity-preserving semantic modifications via direction- and magnitude-constrained perturbations in the latent space of StyleGAN. We further propose to characterize the semantic robustness of an FRM by statistically describing the perturbations that induce the FRM to malfunction. Finally, we combine our methodology with a certification technique, thus providing (i) theoretical guarantees on the performance of an FRM, and (ii) a formal description of how an FRM may model the notion of face identity.

1. Introduction

Deep Neural Networks (DNNs) have achieved impressive performance across fields such as computer vision [27], natural language processing [43], and reinforcement learning [44]. Despite their remarkable success, DNNs are particularly vulnerable against imperceptible perturbations to their input, known as adversarial attacks [25, 59]. The unexpected vulnerability of DNNs against adversarial attacks highlights our narrow understanding of these models and their limitations [22, 69].

This vulnerability further poses potentially negative ramifications in the real-world. Specifically, the deployment of DNNs for security-critical applications may be hampered, since "why" or "how" these systems fail is largely unknown. A case of utmost importance in security-critical applications is that of Face Recognition Models (FRMs). These systems have been the central subject of large amounts of research and engineering [36], and their use is widespread in everyday life, ranging from unlocking phones or personal computers to entering buildings or passing through airport security. Thus, understanding FRMs and their failure modes can constrain how and when to trust FRMs in the real-world. More importantly, interpreting FRMs can provide guides towards a more responsible and ethical use.

The pervasive vulnerability of DNNs against adversarial attacks calls for a unified methodology to study the robustness of FRMs in realistic settings. Specifically, we argue for studying the *semantic* robustness of FRMs, concurring with other works [4, 31], which account for semantic considerations in robustness settings. Towards this objective, some works studied adversarial perturbations to attack [20, 32, 65] and diagnose [26, 50] FRMs.

Other works criticized the physical and/or semantic realism of traditional adversarial perturbations, and developed sophisticated frameworks to introduce physical [41, 52, 53] or semantic [32, 48] considerations. Despite such progress in exploring the vulnerability of FRMs against perturbations, there is still no consensus regarding a methodology for studying the semantic robustness of FRMs.

In this work, we propose and deploy a methodology for systematically assessing and characterizing the semantic robustness of Face Recognition Models. Our methodology achieves this objective by modeling identity-preserving semantic modifications via constrained perturbations in the latent space of Generative Adversarial Networks (GANs) [24], specifically the popular StyleGAN [34]. Please refer to Figure 1 for a visual guide through our methodology. Under this model of identity-preserving modifications, our methodology then connects such modifications with the domain of adversarial robustness [11, 59] to study the semantic robustness of FRMs.

Our methodology models identity-preserving modifications of semantic attributes by introducing constrained perturbations in the latent space of StyleGAN [34]. In particular, we leverage InterFaceGAN [54, 55], a recent method for interpreting the latent space of StyleGAN for synthetic face generation. Identity-preserving perturbations are con-

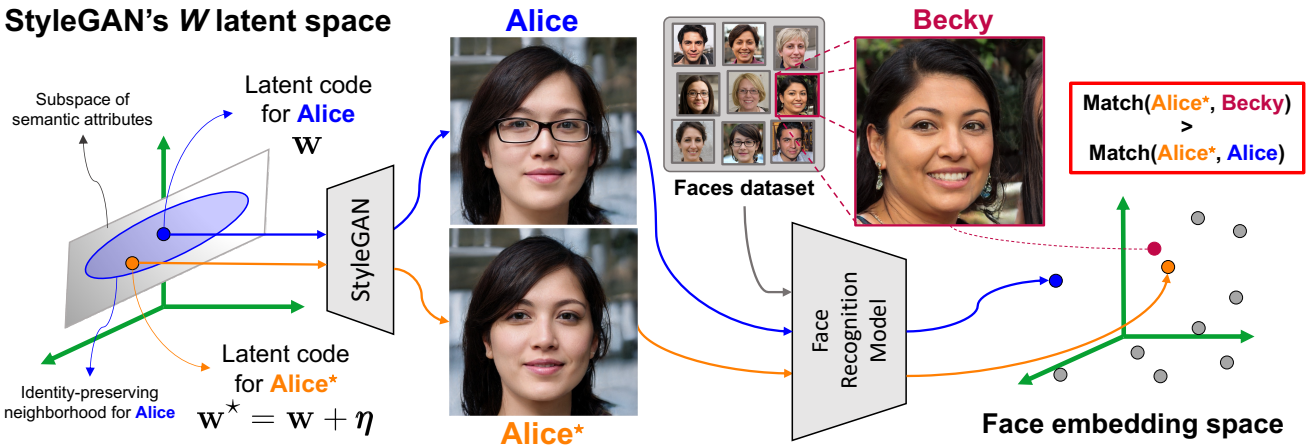

Figure 1. **Searching for identity-preserving modifications via StyleGAN's latent space.** Given **Alice**'s latent code ($\mathbf{w}$) and a subspace of semantic attributes, we draw an identity-preserving neighborhood around **Alice**. We search for **Alice***, a variant of **Alice**'s face, whose latent code $\mathbf{w}^\star = \mathbf{w} + \boldsymbol{\eta}$ lies in this neighborhood. We generate the images corresponding to both $\mathbf{w}$ and $\mathbf{w}^\star$ via StyleGAN and find that, despite remarkable similarities between the faces, a Face Recognition Model's embedding space may suggest to match **Alice*** with **Becky** rather than with **Alice**. Best viewed in color.

strained both in direction and magnitude: only the subspace spanned by certain attributes is allowed, and different attributes can be perturbed to different extents. We adapt adversarial attacks to this model of identity-preserving modifications, and then search for semantic adversarial examples for FRMs by employing constrained- and minimum-perturbation adversarial attacks [12, 42]. We then characterize the semantic robustness of an individual FRM through a statistical procedure that describes the adversarial examples that fool the FRM. Finally, we show how our methodology can leverage an approach for certified robustness. Certifying an FRM provides us with *(i)* theoretical guarantees on the FRM's performance and *(ii)* insights into how the FRM may model the notion of face identity, as delivered by a formal description of the extent to which a face's attributes can vary while the FRM's output remains constant.

Contributions. Our contributions are three-fold. **(1)** We propose a methodology for studying the robustness of Face Recognition Models (FRMs) against semantic perturbations. For that purpose, we extend widely-used paradigms of adversarial attacks to our methodology to search for semantic adversarial examples. **(2)** We propose a procedure for characterizing the semantic robustness of FRMs by statistically describing the semantic adversarial examples we find. **(3)** We show how our methodology can be combined with certification techniques, granting formal guarantees on the performance of an FRM against semantic perturbations and insights regarding how the FRM models identity.

2. Related Work

Adversarial attacks. Previous works [25, 59] showed that adversarial examples, *i.e.* images modified by small maliciously-crafted additive perturbations, could deteriorate the impressive recognition performance of DNNs. This observation led to research on designing procedures, or "attacks", to find adversarial examples for DNNs. Attacks can be dichotomously categorized into two paradigms [19] according to how the underlying optimization problem accounts for the perturbation's magnitude: either as a constraint [42], known as *constrained-perturbation attacks*, or as the objective itself [45], known as *minimum-perturbation attacks*. In this work, we find semantic adversarial examples for FRMs by adapting adversarial attacks from both paradigms to our methodology and searching in the latent space of StyleGAN. For constrained-perturbation attacks, we adopt Projected Gradient Descent (PGD) attacks [42], while for minimum-perturbation attacks, we adopt Fast Adaptive Boundary (FAB) attacks [12]. Moreover, we characterize the semantic robustness of a target FRM by proposing a statistical procedure to describe the adversarial examples found by each attack in terms of semantic attributes.

Certified robustness. Adversarial attacks can be used to empirically assess the robustness of DNNs [9, 10, 13]. However, an attack's inability to find adversarial examples for a DNN does not imply the nonexistence of adversarial examples for this DNN [2, 9]. To address this shortcoming, a line of works studied "certifiable robustness" [38, 39, 62]. This field studies models that are provably robust against additive input perturbations of restricted magnitude, thus guaranteeing the nonexistence of adversarial examples at such magnitude. Randomized smoothing [11] is one such approach and one of the main certification frameworks that scales to large DNNs and datasets. In this work, we extend randomized smoothing to combine it with our methodology. By certifying FRMs against semantic perturbations, we provide performance guarantees and insights into how FRMs recognize faces and, thus, model the notion of identity.

Adversarial examples for Face Recognition Models. Face Recognition Models (FRMs) are computer vision models, whose objective is recognizing human faces. Modern FRMs leverage DNNs to achieve impressive performance [16, 17, 51]. The discovery of adversarial examples led to a stream of works attacking FRMs. Some works perturbed the FRM's input in pixel space [20, 26, 66], while others proposed sophisticated attacks [14, 15, 58, 65] that accounted for physical [4, 41, 52, 53] and semantic [31, 32, 48] considerations in attacking FRMs in the real-world. These works showcased the vulnerability of FRMs against adversarial examples, both in pixel space and in more semantically-inclined spaces. Sharing spirit with our work, Song *et al.* [58] trained a class-conditional GAN and conducted attacks in its latent space. Similarly, Qiu *et al.* [48] interpolated in the latent space of an image-conditional GAN to search for semantic adversarial examples. Joshi *et al.* [31] optimized over a Fader [37] network's latent space to fool facial attribute classifiers. Ruiz *et al.* [50] searched for adversarial examples in a simulator's parametric space to detect weaknesses in FRMs. Most recently, Li *et al.* [40] fooled deepfake-detection by searching StyleGAN's latent space for adversarial examples. While earlier works address FRMs' vulnerability against semantic perturbations, a standard assessment of semantic robustness is still missing. Our work fills this gap in the literature, proposing a methodology to assess and characterize an FRM's semantic robustness by searching for identity-preserving examples that fool the FRM. We search for such examples by modeling semantic (and interpretable) manipulations of facial attributes via direction- and magnitude-constrained perturbations in StyleGAN's latent space.

GANs and interpretation methods. The advent of GANs [24] propelled works on generating images of remarkable visual quality [7, 33]. The impressive perceptual quality achieved by GANs [34, 35] suggested that the representations learnt by these models inherently captured concepts of our visual world. This observation stimulated research on interpreting the internal features learnt by GANs [3] and the GANs' latent space [64]. Recent works showed that this latent space not only encodes semantic concepts, but that such concepts can also be discovered [30, 56, 60, 61] and "controlled" [54, 55]. Our methodology leverages identity-preserving modifications by *(i)* building upon StyleGAN's capacity for generating human faces, and *(ii)* controlling facial attributes in StyleGAN's latent space via InterFaceGAN [54, 55].

3. Semantic Adversarial Attacks

Adversarial attacks usually fool a recognition model by imperceptibly modifying the pixels of an input image with an additive perturbation. These attacks find such perturbation by searching for incorrectly-classified images within a set of imperceptible perturbations. This set is often defined in pixel space as an ℓ_p-ball with a small radius ϵ, aiming at preserving the image's semantics. Thus, these attacks leave both the image *and* its semantics mostly unchanged. While analyzing these perturbations is of interest, here we aim for a more practical class of perturbations that could fool FRMs in the real-world. Thus, in this work, we aim to assess the robustness of FRMs against semantic perturbations.

3.1. Problem Formulation

Let $f : \mathcal{I} \to \mathcal{P}(\mathcal{Y})$ be an FRM that maps image $I \in \mathcal{I}$ into the probability simplex over the set of identities $\mathcal{Y}$. Given an image I of identity y, an attack aims at constructing $I^\star$, a perturbed version of I, considering two goals: *(i)* image similarity, *i.e.* the distance between the two images $d_{\mathcal{I}}(I, I^\star)$ is small for some notion of $d_{\mathcal{I}}$, and *(ii)* fooling the FRM, *i.e.* $I^\star$ is *not* recognized as y such that $\arg\max_i f^i(I^\star) \neq y$. These two goals may be misaligned, affecting the attack's formulation via constrained optimization. In particular, formulations differ in whether the goal of similarity is used as a constraint—and so the fooling goal is the objective—or vice versa. These two alternatives give rise to the paradigms of *constrained-perturbation* and *minimum-perturbation* attacks, respectively [19].

In this work, we find identity-preserving modifications by proposing attacks from both paradigms that model image similarity via distances in StyleGAN's latent space.

3.2. Identity-preserving Modifications

A StyleGAN model $G : \mathcal{W} \to \mathcal{I}$ generates images by mapping from latent space to image space. We consider a latent code $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^d$, which produces image $I = G(\mathbf{w})$. We can generate $I^\star$, a perturbed variant of I, by injecting a perturbation $\boldsymbol{\eta} \in \mathbb{R}^d$ on $\mathbf{w}$, that is $I^\star = G(\mathbf{w}^\star) = G(\mathbf{w} + \boldsymbol{\eta})$. However, we are not interested in introducing *any* perturbation, but rather perturbations that produce identity-preserving modifications on I.

We remark two observations for these modifications: *(i)* InterFaceGAN [55] finds directions along which latent codes can be modified to inject semantically-viable modifications, *e.g.* smile or pose directions, and *(ii)* constrained modifications along these directions should not modify the image's identity. Hence, we model identity-preserving modifications on I by constraining $\boldsymbol{\eta}$'s direction and magnitude. We next describe how we model each constraint.

Direction constraints. InterFaceGAN provides a set of N directions $\{\mathbf{v}_i\}_{i=1}^N$ in StyleGAN's latent space. Each unit-norm vector $\mathbf{v}_i \in \mathbb{R}^d$ specifies a direction along which a semantic face attribute changes. If these vectors are stacked into matrix $V \in \mathbb{R}^{N \times d}$, then constraining $\boldsymbol{\eta}$'s direction amounts to constraining $\boldsymbol{\eta}$ to lie in the subspace spanned by V's rows. We enforce this constraint by substituting $\boldsymbol{\eta} = V^\top \boldsymbol{\delta}$. The substitution accomplishes our goal while

Figure 2. **Identity-preserving modifications.** The row column is the original image, and the other rows are random variants within the respective identity-preserving neighborhood. Notice simultaneous changes in pose and smile, while eyeglasses change color or appear/disappear.

changing the attack's search space from $\mathbb{R}^d \ni \boldsymbol{\eta}$ to $\mathbb{R}^N \ni \boldsymbol{\delta}$. This change in search space benefits the attack's efficiency, since most likely $N \ll d = 512$. *In practice,* we derive V by drawing upon the $N = 5$ interpretable directions provided by InterFaceGAN. Thus, we build matrix V from the directions corresponding to attributes: "Pose", "Age", "Gender", "Smile" and "Eyeglasses".

Magnitude constraints. Given how we enforce the direction constraints, we constrain $\boldsymbol{\eta}$'s magnitude by constraining $\boldsymbol{\delta}$'s magnitude. While most works in robustness constrain with an ℓ_p norm, we argue this scheme is ill-suited for our purposes, since the scale in which semantic attributes vary may be incomparable across attributes. We thus introduce a symmetric and Positive-Definite (PD) matrix $M \in \mathbb{R}^{N \times N}$ to induce "comparability" across attributes. Given this matrix, we model $\boldsymbol{\delta}$'s magnitude as the norm induced by M. Formally, we constrain $\sqrt{\boldsymbol{\delta}^\top M\, \boldsymbol{\delta}} = \|\boldsymbol{\delta}\|_{M,2} \leq 1$[1] and so, the $\boldsymbol{\eta}$'s magnitude is controlled solely by M. *In practice,* we define M by noting each entry of $\boldsymbol{\delta} \in \mathbb{R}^N$ is associated with one direction from $\{\mathbf{v}_i\}_{i=1}^N$, in turn corresponding to a semantic attribute. Defining M is thus linked with the *maximum* allowable perturbation along each individual $\mathbf{v}_i$. Let the scalar ϵ_i define the maximum perturbation allowed along $\mathbf{v}_i$, then we have the condition $|\boldsymbol{\delta}_i| \leq \epsilon_i$. However, this condition still leaves M's definition ill-posed. We resolve this ambiguity by requiring M to enclose the minimum volume possible. With this requirement, we find that M must be the diagonal matrix $M = \text{diag}(\epsilon_1^{-2}, \ldots, \epsilon_N^{-2})$. We leave the details of this derivation to the **Appendix**.

[1]This formulation still allows bounding $\|\boldsymbol{\delta}\|_{M,2}$ by any $\epsilon > 0$, as common in adversarial robustness, by redefining M as $M := 1/\epsilon^2\, M$.

Summary: With the direction and magnitude constraints, we define the set of identity-preserving modifications as $\mathcal{S}(V, M) = \{V^\top \boldsymbol{\delta} \;:\; \|\boldsymbol{\delta}\|_{M,2} \leq 1\}$. We show examples of these modifications in Figure 2.

3.3. Constrained-perturbation Attacks

Based on our formulation of identity-preserving modifications, we outline a constrained-perturbation attack under our framework. In particular, for the composition $F(\mathbf{w}) = f(G(\mathbf{w})) : \mathcal{W} \to \mathcal{P}(\mathcal{Y})$, an attack constructs an identity-preserving modification $\boldsymbol{\delta}$ that fools the FRM f by solving:

$$\max_{\boldsymbol{\delta}} \; \mathcal{L}\left(F(\mathbf{w} + V^\top \boldsymbol{\delta}),\, y\right) \quad \text{s.t.} \quad \|\boldsymbol{\delta}\|_{M,2} \leq 1,$$

where $\mathcal{L}$ is a suitable loss function between probability distributions. This problem can be tackled with Projected Gradient Descent (PGD) [42], whose steps take the form:

$$\boldsymbol{\delta}^{k+1} = \prod_{\|\boldsymbol{\delta}\|_{M,2} \leq 1} \left(\boldsymbol{\delta}^k + \alpha\, \nabla_{\boldsymbol{\delta}} \mathcal{L}\left(F(\mathbf{w} + V^\top \boldsymbol{\delta}),\, y\right)\Big|_{\boldsymbol{\delta} = \boldsymbol{\delta}^k}\right),$$

where α is the step size and $\prod$ is the projection operator. While this formulation is similar to the classical PGD, we highlight a key difference: the set onto which updates are projected, that is $\|\boldsymbol{\delta}\|_{M,2} \leq 1$, is no longer an isotropic ℓ_p-ball, but rather an ellipsoid. Hence, we derive next an efficient projection procedure on ellipsoids, which is critical for the computational tractability of our iterative attacks.

Projecting to an ellipsoid. Formally, projecting a point $\boldsymbol{\delta}$ to the region defined by $\|\boldsymbol{\delta}\|_{M,2} \leq 1$ is defined as solving

$$\arg\min_{\boldsymbol{\delta}^\star} \frac{1}{2}\, \|\boldsymbol{\delta} - \boldsymbol{\delta}^\star\|_2^2, \qquad \text{s.t.} \;\; \boldsymbol{\delta}^{\star\top} M\, \boldsymbol{\delta}^\star \leq 1. \tag{1}$$

If $\boldsymbol{\delta}$ is inside the ellipsoid, then $\boldsymbol{\delta}^\star = \boldsymbol{\delta}$. Otherwise, we need to solve a variant of Problem (1), where the inequality constraint is replaced by an equality, *i.e.* search for $\boldsymbol{\delta}^\star$ on the ellipsoid's surface. Problem (1) is convex in $\boldsymbol{\delta}$ since M is positive definite, so we find $\boldsymbol{\delta}^\star$ with the Lagrangian:

$$L(\boldsymbol{\delta}^\star, \lambda) = \frac{1}{2}\, \|\boldsymbol{\delta}^\star - \boldsymbol{\delta}\|_2^2 + \lambda\, \left(\boldsymbol{\delta}^{\star\top} M\, \boldsymbol{\delta}^{\star\top} - 1\right).$$

Deriving the KKT conditions yields:

$$(\boldsymbol{I} + \lambda^\star\, M)\, \boldsymbol{\delta}^\star = \boldsymbol{\delta}^\star, \tag{2}$$

where $\boldsymbol{I}$ is the identity and $\lambda^\star \in \mathbb{R}$ is the root of the function

$$h(\lambda) = \boldsymbol{\delta}^\top\, (\boldsymbol{I} + \lambda\, M)^{-1}\, M\, (\boldsymbol{I} + \lambda\, M)^{-1}\, \boldsymbol{\delta} - 1.$$

Thus, to find $\boldsymbol{\delta}^\star$, we efficiently find $\lambda^\star$ via the bisection method, substitute into Eq. (2), and solve the linear system.

In practice, we define M as a diagonal matrix (Section 3.2). This structure implies that h can be evaluated without matrix multiplications nor inversions, and that Eq. (2) is a diagonal system that can be efficiently solved. Thus, our projection step is an inexpensive procedure that makes our attacks computationally tractable.

3.4. Minimum-perturbation Attacks

Analogous to constrained-perturbation attacks, we also outline a minimum-perturbation attack under our framework. In this paradigm, the attack aims to find the perturbation with the smallest magnitude that fools the FRM. Thus, based on our formulation of identity-preserving modifications, an attack that minimally modifies identity seeks to solve the following optimization problem:

$$\min_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_{M,2} \quad \text{s.t.} \quad \arg\max_i F^i(\mathbf{w} + V^\top \boldsymbol{\delta}) \neq y. \tag{3}$$

We adopt the state-of-the-art FAB attack [12] to solve Problem (3), as detailed in the **Appendix**.

3.5. Interpreting Adversarial Examples

Once we attack and find adversarial perturbations, we are interested in interpreting them. Each perturbation $\boldsymbol{\delta} \in \mathbb{R}^N$ has an associated energy $\|\boldsymbol{\delta}\|_{M,2}$, and entry $\boldsymbol{\delta}_i$ is related to modifying the i^{th} attribute. Hence, we discover trends in how an FRM weighs attributes to recognize faces by finding trends in how $\boldsymbol{\delta}$'s energy is distributed among the attributes.

We thus propose to describe these trends via a ranking of the energy spent by $\boldsymbol{\delta}$ on modifying each attribute. Therefore, we first compose a candidate ranking by collecting "votes" from the $\boldsymbol{\delta}$s that were found, and then validate the ranking by conducting statistical tests.

Composing a candidate ranking. The quantities being ranked must consider *(i)* the likely anisotropy of the attribute space and *(ii)* the energy of each perturbation found. Thus, we consider the *normalized* entries $\hat{\boldsymbol{\delta}}_i = \boldsymbol{\delta}_i^2 / (\epsilon_i^2 \, \|\boldsymbol{\delta}\|_{M,2})$. Based on these entries, each $\boldsymbol{\delta}$ casts weighed votes, which we sort to find a "winner" attribute. Each time a winner is found, we append the attribute to the ranking, and so we complete the ranking by iterating $(N-1)$ times. We evaluate for significant differences among the remaining attributes with Friedman's test before deciding each winner.

Validating the ranking. Once we have a candidate ranking, we validate it with a statistical test. In particular, we model a ranking of N attributes as $(N-1)$ pair-wise comparisons of adjacent items in the ranking. Thus, for each such pair of items we run a Wilcoxon signed-rank test. Hence, for each ranking, we obtain $(N-1)$ p-values testing for the local validity of the candidate ranking we propose.

3.6. Certifying Against Semantic Perturbations

We also outline a certified robustness approach under our framework. We consider composition F from Section 3.3 and adopt a certification formulation based on randomized smoothing. In particular, we specialize the definition of domain-smoothed classifiers [1, 47] to anisotropically-smooth [21] semantic directions defined by matrix V.

Definition 1. *Given a classifier $F(\mathbf{w}) : \mathcal{W} \to \mathcal{P}(\mathcal{Y})$, we define a semantically-smoothed classifier as:*

$$g(\mathbf{w}, \mathbf{p}) = \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,\Sigma)} \left[F\left(\mathbf{w} + V^T(\mathbf{p} + \boldsymbol{\epsilon})\right) \right].$$

In a nutshell, g's prediction for the image generated from latent code $\mathbf{w}$ is the expected value of F's predictions for semantic variants of the image, where such variants originate from perturbing $\mathbf{w}$. Moreover, $\mathbf{p}$ represents a canonical semantic perturbation of the original image. The following proposition shows that our smooth classifier g is certifiably robust against semantic perturbations along the directions defined by V. We leave the proof for the **Appendix**.

Proposition 1. *Let g assign class c_A for the input pair* $(\mathbf{w}, \mathbf{p})$, i.e. $\arg\max_c g^c(\mathbf{w}, \mathbf{p}) = c_A$ *with:*

$$p_A = g^{c_A}(\mathbf{w}, \mathbf{p}) \quad \textit{and} \quad p_B = \max_{c \neq c_A} g^c(\mathbf{w}, \mathbf{p})$$

then $\arg\max_c \; g^c(\mathbf{w}, \mathbf{p} + \boldsymbol{\delta}) = c_A \;\; \forall \, \boldsymbol{\delta}$ *such that:*

$$\sqrt{\boldsymbol{\delta}^T \Sigma^{-1} \boldsymbol{\delta}} \leq \frac{1}{2} \left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B) \right). \tag{4}$$

Here, Σ is the Gaussian covariance matrix and Φ is the Gaussian CDF. Proposition 1 guarantees the smooth classifier's prediction will be constant for all perturbations within the ellipsoid defined by Eq. (6). Note our result is not constrained to directions in a GAN's latent space: the smooth classifier is certifiable w.r.t. directions characterized by any matrix V. When $\Sigma = \sigma \boldsymbol{I}$, Eq. (6) reduces to isotropic certification as introduced in Randomized Smoothing [11]; consequently, other choices of Σ yield anisotropic certification.

4. Experiments

In this section, we assess and characterize the semantic robustness of off-the-shelf FRMs with our methodology. We first study robustness under a constrained-perturbation attack, *i.e.* PGD. Then, we study robustness under a minimum-perturbation attack, *i.e.* FAB. Finally, we run isotropic and anisotropic certifications on the FRMs.

4.1. Experimental details

FRMs. We target three renowned off-the-shelf FRMs: *(i)* ArcFace [17], *(ii)* a FaceNet [51] model trained on CASIA-Webface [67] that we refer to as "FaceNet^C", and *(iii)* a FaceNet model trained on VGGFace2 [8] that we refer to as "FaceNet^V". All models were retrieved from the public implementations *InsightFace* and *facenet-pytorch*.

Attributes' budget. Table 1 reports the budgets we assign to each attribute, *i.e.* the ϵ_i defining the maximum extent to which latent codes can be perturbed in the direction of each attribute without changing the identity. We establish these values by qualitatively and extensively exploring

Table 1. **Budget per attribute.** We report the budget assigned for each attribute, *i.e.* the maximum extent to which a latent code is allowed to vary in each direction while preserving identity.

ϵ_i for attribute:				
Pose	Age	Gender	Smile	Eyeglasses
0.5	0.5	0.2	0.8	0.5

StyleGAN's output. In particular, we set ϵ_i values that allowed StyleGAN to generate high-quality faces which, arguably, belong to the original identity. Figure 2 shows examples of faces following these ϵ_i attribute budgets.

Attacks. Unless stated otherwise, we always experiment with a StyleGAN-generated dataset of `100k` identities (of comparable size to FFHQ [34]), from which we extract `5k` identities to attack. We consider one image per identity. *PGD.* We use PGD with 10 iterations and 10 restarts. *FAB.* This attack has un-targeted and targeted versions. FAB's un-targeted version is impractical, since its computational cost scales with the number of identities in the dataset. Thus, in practice, we use FAB's *targeted* version, and refer to it simply as "FAB". We use FAB with 10 iterations, 10 restarts, and 10 target classes. We ablate PGD's and FAB's hyper-parameters in the **Appendix**.

Certification. Randomized Smoothing (RS) uses Monte Carlo sampling and a statistical test on the predicted class probability. We use 100 and $10,000$ samples to determine c_A and p_a, respectively, and a significance of $\alpha = 10^{-3}$ for the statistical test. Due to the computational cost of RS, we follow common practice [11] and certify 500 identities.

4.2. Attacks with PGD

Attacking each FRM with PGD reveals the model's semantic robust accuracy, *i.e.* the accuracy achieved by the model when under semantic attacks. We find the following robust accuracies: 84.9 for ArcFace, 76.9 for FaceNet^C, and 71.0 for FaceNet^V. That is, PGD attacks suggest ArcFace is more robust than FaceNet^C, which is, in turn, more robust than FaceNet^V. We show some of the adversarial examples that fooled ArcFace in Figure 3. We note that subtle changes in smiling, pose and, most notably, eyeglasses, cause the FRM to malfunction. Next, we take a closer look at the adversarial examples found by PGD by conducting the statistical procedure described in Section 3.5.

Interpreting adversarial PGD examples. We analyze how PGD spends its budget when constructing adversarial examples. Since PGD is a constrained-perturbation attack, we argue that the relative energy spent on modifying an attribute is related to *"the FRM's disproportionate sensitivity to modifications on such attribute"*. We characterize each FRM's semantic robustness by applying the procedure described in Section 3.5 on the semantic adversarial examples found by PGD, and report the ranking we obtain[2] in Table 2. We make two main observations about the extrema of the rankings, which hold for all FRMs: *(i)* the "Eyeglasses" attribute leads the ranking in 1$^{\text{st}}$ position, and *(ii)* the "Smile" and "Gender" attributes take the last two positions (4$^{\text{th}}$ and 5$^{\text{th}}$). Next, we discuss these observations.

Table 2. **Ranking of PGD's per-attribute energy spent.** The rankings suggest each FRM's disproportionate sensitivity against modifications to an attribute (relative to other attributes). We denote statistically-significant comparisons with "$>^\star$", and the rest with "$\geq$" (significance of 0.01).

Method	Ranking									
	1$^{\text{st}}$		2$^{\text{nd}}$		3$^{\text{rd}}$		4$^{\text{th}}$		5$^{\text{th}}$
ArcFace	E	$>^\star$	P	$>^\star$	A	$>^\star$	S	$>^\star$	G
FaceNet^C	E	$>^\star$	A	$>^\star$	P	$\geq$	G	$\geq$	S
FaceNet^V	E	$>^\star$	A	$>^\star$	P	$\geq$	G	$\geq$	S

Eyeglasses (E), Pose (P), Age (A), Smile (S), Gender (G)

First, we find of high interest that statistical validation can suggest how the presence/absence of eyeglasses is a strong cue on which FRMs rely, somewhat disproportionately, to recognize faces. This finding can be related to previous works [22, 23] that observe how DNNs learn "shortcuts" to solve tasks, thus hindering generalization. Moreover, we note that reliance on eyeglasses is not strange to the human visual system: humans also have difficulty recognizing people when glasses are added/removed. Additionally, our methodology's computation of the position in which eyeglasses rank may prove useful to improve the robustness of FRMs against addition and removal of eyeglasses.

Second, we observe that the smile and gender attributes fall last in the ranking. Thus, compared to other attributes, *neither* smile nor gender are attributes to which FRMs are disproportionately sensitive. That is, under the attack's constrained budget, modifying either smile or gender is largely ineffective: altering either *such that* the FRM is fooled would require an expense that exceeds the budget that was given to PGD. This observation can be read as a pleasant finding: we do not find evidence that FRMs can be fooled by *constrained* changes in smile nor gender. Lastly, we leave more detailed discussion with a brute-force approach for characterizing semantic robustness to the **Appendix**.

Robustness *vs.* dataset size. An FRM's chances of confusing individuals varies as the dataset size changes. We thus experiment with this factor and vary the number of identities in the dataset from `5k` to `1M` and conduct PGD attacks on the same `5k` identities as before. Figure 4a reports the robust accuracies for each dataset size we considered. As expected, the robust accuracies of all FRMs drop rapidly as the number of identities increases. Specifically, performances drop from around 85% when there are `5k` identities to around 70% when there are `1M` identities. Our experi-

[2] We leave implementation details to the **Appendix**.

Figure 3. **PGD attacks on Face Recognition Models (FRMs).** Given face **A** of an identity, we attack an FRM (ArcFace) to find **A***, an identity-preserving modified version, such that the FRM matches **A*** with **B** rather than with **A**.

ments show that an FRM's semantic robustness largely depends on the number of identities it is required to recognize. Hence, depending on the deployment setting, semantic robustness concerns may vary from negligible to problematic.

Attacking more identities. For computational feasibility, we considered a sample of `5k` out of the `100k` identities for our attacks. Here, we test whether this set of identities is a representative sample of the population. We thus fix the `100k` identities in the dataset and vary the amount of samples we attack from `1k` to `20k` and report the results in Figure 4b. We observe that there is virtually no variation in the semantic robustness of any FRM. These results suggest that our design choice of experimenting with `5k` samples provides a reasonable sample of the population for assessing the semantic adversarial robustness of FRMs.

Perturbation budget. In previous experiments, we searched for adversarial examples within the set of identity-preserving modifications by constraining $\|\boldsymbol{\delta}\|_{M,2} \leq \epsilon = 1$. Since our analysis relied on an empirical estimate of the identity-preserving region (*i.e.* M), this region might not be the tightest. Thus, we test how FRMs behave when this constraint is relaxed/tightened by varying ϵ from $1/4$ to 8. We report results in Figure 4c. As expected, the robustness of all FRMs drops rapidly when the semantic perturbation budget increases: ArcFace: $98.3 \rightarrow 4.1$, FaceNet^C: $95.4 \rightarrow 12.6$, and FaceNet^V: $93.3 \rightarrow 7.3$. It is worthwhile to note that allowing semantic perturbation budgets of $\epsilon > 1$ could lead to changing the generated face's identity.

4.3. FAB attack

We also assess each FRM's semantic robustness with FAB attacks. FAB searches over the subspace of semantic attributes, however, FAB does not guarantee that the adversarial examples it finds fall in the identity-preserving neighborhood. That is, while FAB may successfully find adversarial examples for *all* the instances it attacks, a human observer may no longer judge the discovered examples as belonging to the same identity.

We run FAB on each FRM, and find semantic adversarial examples for all the `5k` images we attack. The latent code $\mathbf{w}^\star = \mathbf{w} + V^\top \boldsymbol{\delta}$ of each adversarial example found has a perturbation budget $\|\boldsymbol{\delta}\|_{M,2}$. FAB finds few adversarial examples with $\|\boldsymbol{\delta}\|_{M,2} \leq 1$, that is, within the identity-preserving neighborhood; in particular: 3 for ArcFace, 10 for FaceNet^C and 13 for FaceNet^V. Given the uncertainty on M's tightness (due to its empirical estimation), and following common practice in robustness [19], We plot accuracy *vs.* perturbation budget curves for all FRMs in Figure 4d. Adversarial robustness is judged by how rapidly each curve drops as the perturbation budget increases. Thus, FAB's assessment suggests ArcFace is more robust than FaceNet^C, which is more robust than FaceNet^V, agreeing with PGD's ranking. We leave the interpretation of FAB's adversarial examples (via the procedure from Section 3.5 and a brute-force approach) to the **Appendix**.

4.4. FRM Certification

Isotropic certification. Following the methodology introduced in Section 3.6, we certify all FRMs with covariance $\Sigma = \sigma^2 \boldsymbol{I}$, and set $\sigma \in \{0.1, 0.25, 0.5, 0.75, 1\}$. In this setup, the certified region in Proposition 1 is a ball with radius $\|\boldsymbol{\delta}\|_2 \leq \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right) := R$, as derived in [1]. We denote this quantity as the *certified radius*.

Anisotropic certification. Following our consideration of anisotropic regions for preserving identity, we explore anisotropic certification by drawing upon recent work [21] that extends RS to anisotropic settings. Thus, we require a sensible candidate for an anisotropic Σ in Proposition 1, encoding *a priori* knowledge on the subspace of semantic attributes. Hence, we set $\Sigma = M^{-1}$, where M is the matrix encoding the magnitude constraints in our approach. The rationale behind this choice is that the Mahalanobis distance to the distribution $\mathcal{N}(0, M^{-1})$ draws precisely the ellipsoid described by M. For the experiments, we consider $\Sigma = \sigma M^{-1}$ and set $\sigma \in \{0.25, 0.5, 0.75, 1, 2, 2.5\}$. Since anisotropic regions lack a notion of radius, we follow [21]

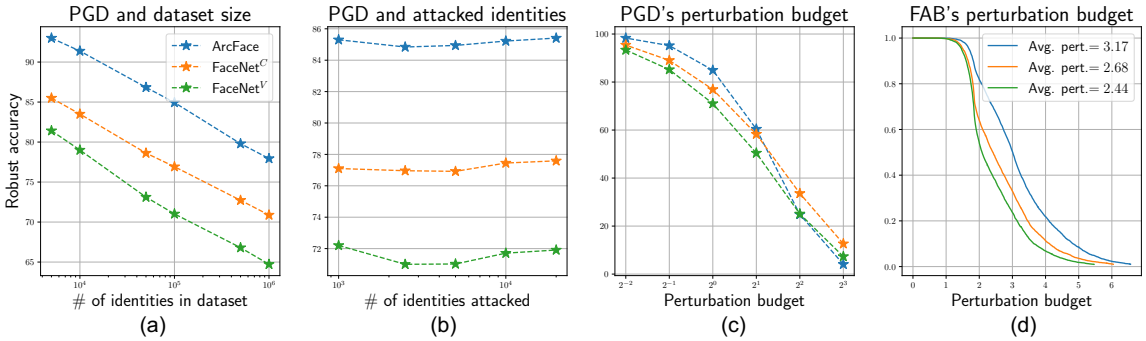

Figure 4. **Assessing semantic robustness via attacks.** We use PGD and FAB to assess the semantic robustness of three Face Recognition Models. For PGD, we report how robustness varies with the number of (a) identities in the dataset and (b) identities attacked. For PGD (c) and FAB (d), we show how robustness changes w.r.t. perturbation budget.

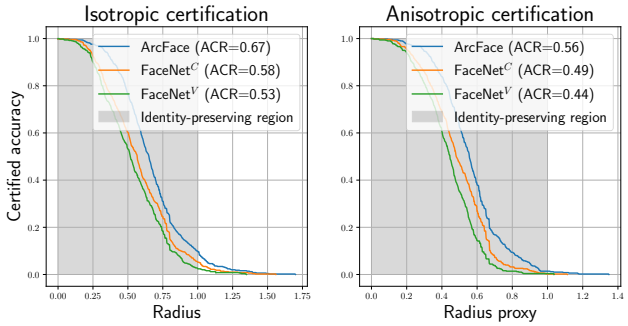

Figure 5. **Certifying Face Recognition Models (FRMs) via Randomized Smoothing.** Envelope curves of all FRMs both for isotropic (left) and anisotropic (right) certification.

and compute a *radius proxy*: the radius of a ball whose volume is equivalent to that of the certified region.

Results. We compute the best certificates for each FRM across all σ values, and report certified accuracy curves in Figure 5 for isotropic (left) and anisotropic (right) certification. Each point (x, y) in a curve implies that percentage $y\%$ of the dataset is both predicted correctly and has a certified radius of at least x. Moreover, we adopt common practice [68] and report the Average Certified Radius (ACR) for each FRM. We draw the following observations: *(i)* The certified accuracy within the identity-preserving region is remarkably low. That is, while all FRMs displayed substantial robustness against our attacks, certification demonstrates these models *can* be fooled by stronger attacks. Hence, we find FRMs are also extremely vulnerable to simple semantic perturbations. We argue this vulnerability is expected, as regular DNN training is not designed to resist against adversarial attacks. *(ii)* The ACRs under the anisotropic setting are smaller than those under the isotropic one. This can be a result of a sub-optimal choice of the matrix Σ.

5. Conclusions

We propose a methodology for assessing and characterizing the semantic robustness of Face Recognition Models (FRMs). Our methodology induces malfunction in FRMs by conducting direction- and magnitude-constrained search in StyleGAN's latent space, such that faces are modified but their identity is preserved. Under this framework, we attack FRMs, find adversarial examples, and then characterize the semantic robustness of FRMs by statistically describing the examples that lead them to fail. Finally, we demonstrate how our methodology can leverage a certification technique, allowing us to construct a formal description of what an FRM may conceive as a face's identity.

6. Limitations

The main focus of our study is the semantic robustness of a standalone FRM. However, in practice, we are unable to directly study the FRM, as we introduce a StyleGAN *before* the FRM. We model semantic directions in StyleGAN's latent space via InterFaceGAN. Thus, the conclusions we reach are limited by the weaknesses of StyleGAN and InterFaceGAN. Specifically, we underscore the following weaknesses: *(i)* there are no guarantees for StyleGAN's output, while impressive, to be clean of artifacts, *(ii)* StyleGAN's training data is presumably biased, thus affecting the diversity of generated faces, and *(iii)* the semantic directions found by InterFaceGAN still display some entanglement.

Further limitations relate to the types of attacks considered and the usage of GANs. In particular, while other types of attacks exist [5, 28], we focus exclusively on adversarial attacks, in which a target DNN is fooled via input manipulation. Moreover, while recently diffusion models have shown remarkable capacity for generating photo-realistic imagery [18, 29, 49], our work leveraged GANs exclusively, mainly due to their well-studied latent space and low computational burden, when compared to diffusion models.

References

[1] Motasem Alfarra, Adel Bibi, Naeemullah Khan, Philip H. S. Torr, and Bernard Ghanem. Deformrs: Certifying input deformations with randomized smoothing. *CoRR*, abs/2107.00996, 2021. 5, 7, 13

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 2

[3] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[4] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and D. A. Forsyth. Unrestricted adversarial examples via semantic manipulation. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 3

[5] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1467–1474, 2012. 8

[6] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 12

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[8] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 2018. 5

[9] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 2

[10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 2017. 2

[11] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019. 1, 2, 5, 6

[12] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pages 2196–2205. PMLR, 2020. 2, 5

[13] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020. 2

[14] Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser Nasrabadi. Fast geometrically-perturbed adversarial faces. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 3

[15] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020. 3

[16] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 5

[18] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 8

[19] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 7

[20] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[21] Francisco Eiras, Motasem Alfarra, M. Pawan Kumar, Philip H. S. Torr, Puneet K. Dokania, Bernard Ghanem, and Adel Bibi. ANCER: anisotropic certification via sample-wise volume maximization. *CoRR*, abs/2107.04570, 2021. 5, 7, 13

[22] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. 1, 6

[23] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems (NeurIPS)*, 2014. 1, 3

[25] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 2

[26] Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1, 3

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the*

IEEE International Conference on Computer Vision (ICCV), 2015. 1

[28] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, June 2021. 8

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 8

[30] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Advances in neural information processing systems (NeurIPS)*, 2020. 3

[31] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1, 3

[32] Kazuya Kakizaki and Kosuke Yoshida. Adversarial image translation: Unrestricted adversarial examples in face recognition systems. *AAAI Workshop on Artificial Intelligence Safety*, 2020. 1, 3

[33] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[34] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3, 6

[35] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[36] Yassin Kortli, Maher Jridi, Ayman Al Falou, and Mohamed Atri. Face recognition systems: A survey. *Sensors*, 2020. 1

[37] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[38] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, 2019. 2

[39] B Li, C Chen, W Wang, and L Carin. Second-order adversarial attack and certifiable robustness. *arXiv preprint arXiv: 1809.03113*, 2018. 2

[40] Dongze Li, Wei Wang, Hongxing Fan, and Jing Dong. Exploring adversarial fake images on face manifold. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[41] Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 3

[42] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 4

[43] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. 1

[44] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. 2013. 1

[45] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 2

[46] Nima Moshtagh et al. Minimum volume enclosing ellipsoid. *Convex optimization*, 2005. 12

[47] Gabriel Pérez, Juan C. Pérez, Motasem Alfarra, Silvio Giancola, and Bernard Ghanem. 3deformrs: Certifying spatial deformations on point clouds. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15148–15158. IEEE Computer Society, 2022. 5

[48] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 3

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 8

[50] Nataniel Ruiz, Adam Kortylewski, Weichao Qiu, Cihang Xie, Sarah Adel Bargal, Alan Yuille, and Stan Sclaroff. Simulated adversarial testing of face recognition models. *arXiv preprint arXiv:2106.04569*, 2021. 1, 3

[51] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015. 3, 5

[52] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael 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*, 2016. 1, 3

[53] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 2019. 1, 3

[54] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[55] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1, 3

[56] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[57] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006. 18

[58] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[59] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 1, 2

[60] Christos Tzelepis, Georgios Tzimiropoulos, and Ioannis Patras. WarpedGANSpace: Finding non-linear rbf paths in GAN latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[61] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning (ICML)*, 2020. 3

[62] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, 2018. 2

[63] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020. 12

[64] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision (IJCV)*, 2020. 3

[65] Lu Yang, Qing Song, and Yingqi Wu. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. *Multimedia Tools and Applications*, 2021. 1, 3

[66] Xiao Yang, Dingcheng Yang, Yinpeng Dong, Wenjian Yu, Hang Su, and Jun Zhu. Delving into the adversarial robustness on face recognition. *arXiv preprint arXiv:2007.04118*, 2020. 3

[67] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 5

[68] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. *International Conference on Learning Representations (ICLR)*, 2020. 8

[69] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021. 1