Efficient Decision-based Black-box Adversarial Attacks on Face Recognition

Yinpeng Dong[1], Hang Su[1], Baoyuan Wu[2], Zhifeng Li[2], Wei Liu[2], Tong Zhang[3], Jun Zhu[1]*
[1] Dept. of Comp. Sci. and Tech., BNRist Center, State Key Lab for Intell. Tech. & Sys.,
[1] Institute for AI, THBI Lab, Tsinghua University, Beijing, 100084, China
[2] Tencent AI Lab [3] Hong Kong University of Science and Technology
dyp17@mails.tsinghua.edu.cn, suhangss@mail.tsinghua.edu.cn, wubaoyuan1987@gmail.com
michaelzfli@tencent.com, wl2223@columbia.edu, tongzhang@tongzhang-ml.org, dcszj@mail.tsinghua.edu.cn

Abstract

Face recognition has obtained remarkable progress in recent years due to the great improvement of deep convolutional neural networks (CNNs). However, deep CNNs are vulnerable to adversarial examples, which can cause fateful consequences in real-world face recognition applications with security-sensitive purposes. Adversarial attacks are widely studied as they can identify the vulnerability of the models before they are deployed. In this paper, we evaluate the robustness of state-of-the-art face recognition models in the decision-based black-box attack setting, where the attackers have no access to the model parameters and gradients, but can only acquire hard-label predictions by sending queries to the target model. This attack setting is more practical in real-world face recognition systems. To improve the efficiency of previous methods, we propose an evolutionary attack algorithm, which can model the local geometry of the search directions and reduce the dimension of the search space. Extensive experiments demonstrate the effectiveness of the proposed method that induces a minimum perturbation to an input face image with fewer queries. We also apply the proposed method to attack a real-world face recognition system successfully.

1. Introduction

Recent progress in deep convolutional neural networks (CNNs) [26, 29, 11] has led to substantial performance improvements in a broad range of computer vision tasks. Face recognition, as one of the most important computer vision tasks, has been greatly facilitated by deep CNNs [31, 28, 23, 33, 16, 32, 5]. There are usually two sub-tasks in face recognition: face verification and face identification [12, 15]. The former distinguishes whether a pair of face images represent the same identity, while the latter classifies an image to an identity. The state-of-the-art face recognition models realize these two tasks by using deep CNNs to extract face features that have minimum intra-class variance and maximum inter-class variance. Due to the excellent performance of these models, face recognition has been widely used for identity authentication in enormous applications, such as finance/payment, public access, criminal identification, *etc*.

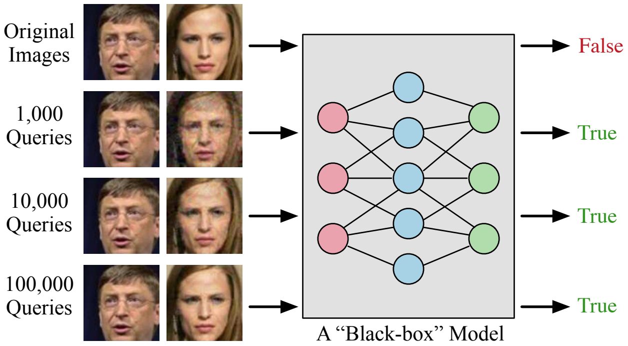

Figure 1. Demonstration of the decision-based black-box attack setting. Given a black-box model, the attackers use queries to generate adversarial examples with minimum perturbations.

Despite the great success in various applications, deep CNNs are known to be vulnerable to adversarial examples [30, 9, 19, 6]. These maliciously generated adversarial examples are often indistinguishable from legitimate ones for human observers by adding small perturbations. But they can make deep models produce incorrect predictions. The face recognition systems based on deep CNNs have also been shown their vulnerability against such adversarial examples. For instance, adversarial perturbations can be made to the eyeglass that, when worn, allows attackers to evade being recognized or impersonate another individual [24, 25]. The insecurity of face recognition systems in real-world applications, especially those with sensitive purposes, can cause severe consequences and security issues.

To evaluate the robustness of face recognition systems in real-world applications, adversarial attacks can serve as an important surrogate, as they can identify the vulnerability of

*Corresponding author.

these systems [2] and help to improve the robustness [9, 18]. However, existing attack methods [24, 25] for face recognition are mainly based on the *white-box* scenario, where the attackers know the internal structure and parameters of the system being attacked. Accordingly, the attack objective function can be directly optimized by gradient-based methods. This setting is clearly impractical in real-world cases, when the attackers cannot get access to the model details. Instead, we focus on a more realistic and general *decision-based black-box* setting [1], where no model information is exposed except that the attackers can only query the target model and obtain corresponding hard-label predictions. The goal of attacks is to generate adversarial examples with minimum perturbations by limited queries. This attack scenario is much more challenging, since that the gradient cannot be directly computed and the predicted probability is not provided. On the other hand, it is much more realistic and important, because most of the real-world face recognition systems are black-box and only provide hard-label outputs. To the best of our knowledge, it is the first attempt to conduct adversarial attacks on face recognition in this setting.

Several methods [1, 14, 4] have been proposed to perform decision-based black-box attacks. However, they lack the efficiency in the sense that they usually require a tremendous number of queries to converge, or get a relatively large perturbation given a limited budget of queries. Therefore, we consider how to efficiently generate adversarial examples for decision-based black-box attacks by inducing a smaller perturbation to each sample with fewer queries.

To address the aforementioned issues, we propose an **evolutionary attack** method for query-efficient adversarial attacks in the decision-based black-box setting. Given the attack objective function, the proposed method is able to optimize it in the black-box manner through queries only. Our method can find better search directions by modeling their local geometry. It further improves the efficiency by reducing the dimension of the search space. We apply the proposed method to comprehensively study the robustness of several state-of-the-art face recognition models, including SphereFace [16], CosFace [32], and ArcFace [5], under the decision-based black-box scenario. Extensive experiments conducted on the most popular public-domain face recognition datasets such as Labeled Face in the Wild (LFW) [12] and MegaFace Challenge [15] demonstrate the effectiveness of the proposed method. We further apply our method to attack a real-world face recognition system to show its practical applicability. In summary, our major contributions are:

- We propose a novel evolutionary attack method under the decision-based black-box scenario, which can model the local geometry of the search directions and meanwhile reduce the dimension of the search space. The evolutionary attack method is generally applicable for any image recognition task, and significantly improves the efficiency over existing methods.
- We thoroughly evaluate the robustness of several state-of-the-art face recognition models by decision-based black-box attacks in various settings. We demonstrate the vulnerability of these face models in this setting.
- We show the practical applicability of the proposed method by successfully attacking a real-world face recognition system.

2. Related Work

Deep face recognition. DeepFace [31] and DeepID [28] treat face recognition as a multi-class classification problem and use deep CNNs to learn features supervised by the softmax loss. Triplet loss [23] and center loss [33] are proposed to increase the Euclidean margin in the feature space between classes. The angular softmax loss is proposed in SphereFace [16] to learn angularly discriminative features. CosFace [32] uses the large margin cosine loss to maximize cosine margin. The additive angular margin loss is proposed in ArcFace [5] to learn highly discriminative features.

Adversarial attacks on face recognition. Deep CNNs are highly vulnerable to adversarial examples [30, 9, 19]. Face recognition has also been shown the vulnerability against attacks. In [24], the perturbations are constrained to the eyeglass region and generated by gradient-based methods, which fool face recognition systems even in the physical world. The adversarial eyeglasses can also be produced by generative networks [25]. However, these methods rely on the white-box manipulations of face recognition models, which is unrealistic in real-world applications. Instead, we focus on evaluating the robustness of face recognition models in the decision-based black-box attack setting.

Black-box attacks. Black-box attacks can be divided into transfer-based, score-based and decision-based attacks. Transfer-based attacks generate adversarial examples for a white-box model and attack the black-box model based on the transferability [17, 6]. In score-based attacks, the predicted probability is given by the model. Several methods rely on approximated gradients to generate adversarial examples [3, 14]. In decision-based attacks, we can only obtain the hard-label predictions. The boundary attack method is based on random walk on the decision boundary [1]. The optimization-based method [4] formulates this problem as a continuous optimization problem and estimate the gradient for optimization. However, it needs to calculate the distance to the decision boundary along a direction by binary search. In [14], the predicted probability is estimated by hard-label predictions. Then, the natural evolution strategy (NES) is used to maximize the target class probability or minimize the true class probability. These methods generally require a large number of queries to generate an adversarial example with a minimum perturbation, or converge to a large perturbation with few queries .

3. Methodology

In this section, we first introduce the decision-based black-box attack setting against a face recognition model, and then detail the proposed evolutionary attack method.

3.1. Attack Setting

Let $f(\boldsymbol{x}) : \mathcal{X} \to \mathcal{Y}$ ($\mathcal{X} \subset \mathbb{R}^n$) denote the face recognition model that predicts a label for an input face image. For face verification, the model relies on another face image to identify whether the pair of images belong to the same identity, and outputs a binary label in $\mathcal{Y} = \{0, 1\}$. For face identification, the model $f(\boldsymbol{x})$ compares the input image $\boldsymbol{x}$ with a gallery set of face images, and then classifies $\boldsymbol{x}$ as a specific identity. So it can be viewed as a multi-class classification task, where $\mathcal{Y} = \{1, 2, ..., K\}$, with K being the number of identities. Although the face recognition model $f(\boldsymbol{x})$ uses an additional face image or a set of face images for recognizing $\boldsymbol{x}$, we do not explicitly describe the dependency of $f(\boldsymbol{x})$ on the compared images for simplicity.

Given a real face image $\boldsymbol{x}$, the goal of attacks is to generate an adversarial face image $\boldsymbol{x}^*$ in the vicinity of $\boldsymbol{x}$ but is misclassified by the model. It can be obtained by solving a constrained optimization problem

$$\min_{\boldsymbol{x}^*} \mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x}), \quad \text{s.t. } \mathcal{C}(f(\boldsymbol{x}^*)) = 1, \tag{1}$$

where $\mathcal{D}(\cdot, \cdot)$ is a distance metric, and $\mathcal{C}(\cdot)$ is an adversarial criterion that takes 1 if the attack requirement is satisfied and 0 otherwise. We use the L_2 distance as $\mathcal{D}$. The constrained problem in Eq. (1) can be equivalently reformulated as the following unconstrained optimization problem

$$\min_{\boldsymbol{x}^*} \mathcal{L}(\boldsymbol{x}^*) = \mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x}) + \delta\big(\mathcal{C}(f(\boldsymbol{x}^*)) = 1\big), \tag{2}$$

where $\delta(a) = 0$ if a is true, otherwise $\delta(a) = +\infty$. By optimizing Eq. (2), we can obtain an image $\boldsymbol{x}^*$ with a minimum perturbation, which is also adversarial according to the criterion. Note that in the above objective functions, $\mathcal{C}$ cannot be defined as a continuous criterion such as cross-entropy loss, since that the model $f(\boldsymbol{x})$ only gives discrete hard-label outputs in this problem. In particular, we specify $\mathcal{C}$ according to the following two types of attacks.

Dodging attack corresponds to generating an adversarial image that is recognized wrong or not recognized. Dodging attack could be used to protect personal privacy against excessive surveillance. For face verification, given a pair of face images belonging to the same identity, the attacker seeks to modify one image and make the model recognize them as not the same identity. So the criterion is $\mathcal{C}(f(\boldsymbol{x}^*)) = \mathbb{I}(f(\boldsymbol{x}^*) = 0)$, where $\mathbb{I}$ is the indicator function. For face identification, the attacker generates an adversarial face image with the purpose that it is recognized as any other identity. The criterion is $\mathcal{C}(f(\boldsymbol{x}^*)) = \mathbb{I}(f(\boldsymbol{x}^*) \neq y)$, where y is the true identity of the real image $\boldsymbol{x}$.

Impersonation attack works as seeking an adversarial image recognized as a specific identity, which could be used to evade the face authentication systems. For face verification, the attacker tries to find an adversarial image that is recognized as the same identity of another image, while the original images are not from the same identity. The criterion is $\mathcal{C}(f(\boldsymbol{x}^*)) = \mathbb{I}(f(\boldsymbol{x}^*) = 1)$. For face identification, the generated adversarial image needs to be classified as a specific identity y^*, so $\mathcal{C}(f(\boldsymbol{x}^*)) = \mathbb{I}(f(\boldsymbol{x}^*) = y^*)$.

3.2. Evolutionary Attack

Since we cannot get access to the configuration and parameters of $f(\boldsymbol{x})$ but can only send queries to probe the model, we resort to black-box optimization techniques to minimize the objective function in Eq. (2). Gradient estimation methods [20, 8, 7] approximate the gradient of the objective function by finite difference and update the solution by gradient descent, which are commonly used for score-based black-box attacks, when the predicted probability is given by the model [3, 14]. However, in the case of hard-label output, the attack objective function is discontinuous and the output is insensitive to small input perturbations. So the gradient estimation methods cannot be directly used. Some methods [4, 14] successfully reformulate the discontinuous optimization problem in Eq. (2) as some continuous optimization problems and use gradient estimation methods for optimization. But they need to calculate the distance of a point to the decision boundary or estimate the predicted probability by the hard-label outputs, which are less efficient as demonstrated in the experiments. Therefore, we consider how to directly optimize Eq. (2) efficiently.

In this paper, we propose a novel **evolutionary attack** method to solve the black-box optimization problem. Our method is based on a simple and efficient variant of covariance matrix adaptation evolution strategy (CMA-ES) [10], which is the (1+1)-CMA-ES [13]. In each update iteration of the (1+1)-CMA-ES, a new offspring (candidate solution) is generated from its parent (current solution) by adding a random noise, the objective of these two solutions are evaluated, and the better one is selected for the next iteration. This method is capable for solving black-box optimization problems. However, directly applying the (1+1)-CMA-ES to optimize Eq. (2) is inefficient due to the high dimension of $\boldsymbol{x}^*$. Considering the query limit in decision-based black-box attacks for face images, the original (1+1)-CMA-ES may be infeasible. To accelerate this algorithm, we design an appropriate distribution to sample the random noise in each iteration, which can model the local geometry of the search directions. We also propose several techniques to reduce the dimension of the search space by considering the special characteristics of this problem.

The overall evolutionary attack algorithm is outlined in Algorithm 1. Rather than the original n-dimensional input

Algorithm 1 The evolutionary attack algorithm

```
Input: The attack objective function $\mathcal{L}(\boldsymbol{x}^*)$; the original face image $\boldsymbol{x}$; the dimension $n \in \mathbb{N}_+$ of the input space ($\boldsymbol{x}^* \in \mathbb{R}^n$); the dimension $m \in \mathbb{N}_+$ of the search space; the number of coordinates $k \in \mathbb{N}_+$ for stochastic coordinate selection.
Input: The total number of queries $T$.
 1: Initialize $\mathbf{C} = \mathbf{I}_m$, $\boldsymbol{p}_c = \mathbf{0}$, $\sigma, \mu, c_c, c_{cov} \in \mathbb{R}_+$, $\tilde{\boldsymbol{x}}^* \in \mathbb{R}^n$;
 2: for $t = 1$ to $T$ do
 3:     Sample $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{C})$;
 4:     Select $k$ coordinates from $m$ with probability proportional to each diagonal element in $\mathbf{C}$;
 5:     Set the non-selected coordinates of $\boldsymbol{z}$ to 0;
 6:     Upscale $\boldsymbol{z}$ to $\mathbb{R}^n$ by bilinear interpolation and obtain $\tilde{\boldsymbol{z}}$;
 7:     $\tilde{\boldsymbol{z}} \leftarrow \tilde{\boldsymbol{z}} + \mu(\boldsymbol{x} - \tilde{\boldsymbol{x}}^*)$;
 8:     if $\mathcal{L}(\tilde{\boldsymbol{x}}^* + \tilde{\boldsymbol{z}}) < \mathcal{L}(\tilde{\boldsymbol{x}}^*)$ then
 9:         $\tilde{\boldsymbol{x}}^* \leftarrow \tilde{\boldsymbol{x}}^* + \tilde{\boldsymbol{z}}$;
10:         Update $\boldsymbol{p}_c$ and $\mathbf{C}$ by $\boldsymbol{z}$ according to Eq. (3) and Eq. (4);
11:     end if
12: end for
13: return $\tilde{\boldsymbol{x}}^*$.
```

space, we perform search in a lower dimensional space $\mathbb{R}^m$ with $m < n$. In each iteration, we first sample a random vector $\boldsymbol{z}$ from $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{C})$ such that $\boldsymbol{z} \in \mathbb{R}^m$, where $\mathbf{C}$ is a diagonal covariance matrix to model the local geometry of the search directions. We then select k coordinates randomly for search, according to the assumption that only a fraction of pixels are important for finding an adversarial image. We keep the value of the selected k coordinates of $\boldsymbol{z}$ by setting the others to 0. We upscale $\boldsymbol{z}$ to the input space by bilinear interpolation and get $\tilde{\boldsymbol{z}} \in \mathbb{R}^n$. We further add a bias to $\tilde{\boldsymbol{z}}$ to minimize the distance between the adversarial and original images. We finally test whether we get a better solution. If we indeed find a better solution, we jump to it and update the covariance matrix. In the following, we will give a detailed description of each step in the algorithm.

3.2.1 Initialization

In Algorithm 1, $\tilde{\boldsymbol{x}}^*$ should be initialized at first (in Step 1). If the initial $\tilde{\boldsymbol{x}}^*$ does not satisfy the adversarial criterion, $\mathcal{L}(\tilde{\boldsymbol{x}}^*)$ equals to $+\infty$. For subsequent iterations, adding a random vector can rarely make the search point adversarial due to that deep CNNs are generally robust to random noises [30], and thus the loss function will keep being $+\infty$. So we initialize $\tilde{\boldsymbol{x}}^*$ with a sample that already satisfies the adversarial criterion. The following updates will also keep $\tilde{\boldsymbol{x}}^*$ adversarial, and at the same time minimize the distance between $\tilde{\boldsymbol{x}}^*$ and $\boldsymbol{x}$. For dodging attack, the initial $\tilde{\boldsymbol{x}}^*$ can be simply set as a random vector. For impersonation attack, we use the target image as the initial point of $\tilde{\boldsymbol{x}}^*$.

3.2.2 Mean of Gaussian Distribution

We explain why we need to add a bias term to the random vector in Step 7. Assume now that the dimension of the search space is the same as that of the input space and we select all coordinates for search (*i.e.*, $k = m = n$). In each iteration, a random vector $\boldsymbol{z}$ is sampled from a Gaussian distribution. In general, the distribution should be unbiased (with zero mean) for better exploration in the search space. But in our problem, sampling the random vector from a zero mean Gaussian distribution will result in nearly zero probability of updates as $n \to \infty$, given by Theorem 1.

Theorem 1 *(Proof in Appendix A) Assume that the covariance matrix* $\mathbf{C}$ *is positive definite. Let* λ_{max} *and* $\lambda_{min}(> 0)$ *be the largest and smallest eigenvalues of* $\mathbf{C}$*, respectively. Then, we have*

$$P_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{C})}\big(\mathcal{L}(\tilde{\boldsymbol{x}}^* + \boldsymbol{z}) < \mathcal{L}(\tilde{\boldsymbol{x}}^*)\big) \leq \frac{4\lambda_{max}\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}\|^2}{\sigma^2\lambda_{min}^2 n^2}.$$

From Theorem 1, we need to draw $\mathcal{O}(n^2)$ samples from the zero mean Gaussian distribution for only one successful update, which is inefficient and costly when n is large. This happens because in high dimensional search space, a randomly drawn vector $\boldsymbol{z}$ is almost orthogonal to $\tilde{\boldsymbol{x}}^* - \boldsymbol{x}$, thus the distance $\mathcal{D}(\tilde{\boldsymbol{x}}^* + \boldsymbol{z}, \boldsymbol{x})$ will be rarely smaller than $\mathcal{D}(\tilde{\boldsymbol{x}}^*, \boldsymbol{x})$. To address this problem, the random vector $\boldsymbol{z}$ should be sampled from a biased distribution towards minimizing the distance of $\tilde{\boldsymbol{x}}^*$ from the original image $\boldsymbol{x}$. So we add a bias term $\mu(\boldsymbol{x} - \tilde{\boldsymbol{x}}^*)$ to $\tilde{\boldsymbol{z}}$ (the same as $\boldsymbol{z}$ when $k = m = n$) in Step 7, where μ is a critical hyper-parameter controlling the strength of going towards the original image $\boldsymbol{x}$. We will specify the update procedure of μ in Sec. 3.2.6.

3.2.3 Covariance Matrix Adaptation

The adaptation of covariance matrix $\mathbf{C}$ is suitable for solving non-separable optimization problems since it can model the local geometry of the search directions [10]. For example, an appropriately set covariance matrix can make the random vectors generated predominantly in the direction of narrow valleys. In learning all pair-wise dependencies between dimensions, the storage and computation complexity of the covariance matrix is at least $\mathcal{O}(m^2)$, which is unacceptable when m is large. For black-box adversarial attacks, the dimension of the search space is extremely large (*e.g.*, $m = 45 \times 45 \times 3$ in our experiments). Therefore, we relax the covariance matrix to be a diagonal matrix for efficient computation. Inspired by [22] which uses a diagonal covariance matrix for CMA-ES, we design an update rule for the adaptation of the diagonal covariance matrix $\mathbf{C}$ (in Step 10) after each successful trial as

$$\boldsymbol{p}_c = (1 - c_c)\boldsymbol{p}_c + \sqrt{c_c(2 - c_c)}\frac{\boldsymbol{z}}{\sigma}, \tag{3}$$

$$c_{ii} = (1 - c_{cov})c_{ii} + c_{cov}(\boldsymbol{p}_c)_i^2, \tag{4}$$

where $\boldsymbol{p}_c \in \mathbb{R}^m$ is called the evolution path as it stores the exponentially decayed successful search directions; for $i = 1, ..., m$, c_{ii} is the diagonal element of $\mathbf{C}$ and $(\boldsymbol{p}_c)_i$ is the i-th element of $\boldsymbol{p}_c$. c_c and c_{cov} are two hyper-parameters

Model		SphereFace [16]				CosFace [32]				ArcFace [5]			
Queries		1,000	5,000	10,000	100,000	1,000	5,000	10,000	100,000	1,000	5,000	10,000	100,000
Dodging	Boundary [1]	2.3e-2	9.3e-3	7.0e-4	1.9e-5	2.0e-2	7.5e-3	7.7e-4	1.6e-5	2.4e-2	1.6e-2	1.5e-3	2.3e-5
	Optimization [4]	1.2e-2	2.9e-3	1.3e-3	7.1e-5	1.1e-2	2.9e-3	1.3e-3	6.6e-5	1.5e-2	5.4e-3	2.6e-3	9.9e-5
	NES-LO [14]	1.4e-1	3.8e-2	2.4e-2	7.4e-3	1.4e-1	3.5e-2	2.0e-2	6.5e-3	1.4e-1	3.9e-2	2.3e-2	1.5e-2
	Evolutionary	**1.6e-3**	**8.9e-5**	**3.4e-5**	**1.3e-5**	**1.7e-3**	**9.1e-5**	**3.3e-5**	**1.1e-5**	**2.8e-3**	**1.5e-4**	**5.2e-5**	**1.6e-5**
Impersonation	Boundary [1]	1.5e-2	6.3e-3	5.7e-4	1.6e-5	1.1e-2	2.9e-3	2.8e-4	7.4e-6	2.0e-2	9.2e-3	1.2e-3	1.7e-5
	Optimization [4]	1.1e-2	3.3e-3	1.3e-3	6.1e-5	7.7e-3	1.9e-3	7.1e-4	2.8e-5	1.6e-2	7.0e-3	3.3e-3	7.7e-5
	NES-LO [14]	8.4e-2	2.6e-2	1.7e-2	5.5e-3	9.3e-2	2.0e-2	1.2e-2	3.1e-3	9.3e-2	3.0e-2	1.9e-2	8.1e-3
	Evolutionary	**1.2e-3**	**7.2e-5**	**2.9e-5**	**1.2e-5**	**6.5e-4**	**3.7e-5**	**1.5e-5**	**5.3e-6**	**2.3e-3**	**1.2e-4**	**3.9e-5**	**1.2e-5**

Table 1. The results on face verification. We report the average distortion (MSE) of the adversarial images generated by different methods for SphereFace, CosFace, and ArcFace given 1,000, 5,000, 10,000, and 100,000 queries, based on the LFW dataset.

of CMA. An intuitive explanation of this update is that the variance along the past successful directions should be enlarged for future search.

3.2.4 Stochastic Coordinate Selection

For adversarial attacks, the perturbations added to the images could be very sparse to fool deep CNNs [27], indicating that only a fraction of coordinates (pixels) are sufficient for finding the adversarial images. We can also accelerate the black-box optimization if we could identify the important coordinates. However, this is non-trivial in the decision-based black-box attack setting. Fortunately, our algorithm provides a natural way to find the useful coordinates for search since the elements in the diagonal covariance matrix $\mathbf{C}$ represent the preferred coordinates of the past successful trials, *i.e.*, larger c_{ii} indicates that searching along the i-th coordinate may induce a higher success rate based on the past experience. According to this, in each iteration we select k ($k \ll m$) coordinates to generate the random vector $\boldsymbol{z}$ with the probability of selecting the i-th coordinate being proportional to c_{ii} (in Step 4-5).

3.2.5 Dimensionality Reduction

It has been proved that the dimensionality reduction of the search space is useful for acceleration of black-box attacks [3]. Based on this, we sample the random vector $\boldsymbol{z}$ in a lower dimensional space $\mathbb{R}^m$ with $m < n$ (in Step 3). We then adopt an upscaling operator to project $\boldsymbol{z}$ to the original space $\mathbb{R}^n$ (in Step 6). Note that we do not change the dimension of an input image but only reduce the dimension of the search space. Specifically, we use the bilinear interpolation method as the upscaling operator.

3.2.6 Hyper-parameter Adjustment

There are also several hyper-parameters in the proposed algorithm, including σ, μ, c_c, and c_{cov}. We simply set $c_c = 0.01$ and $c_{cov} = 0.001$. σ is set as $0.01 \cdot \mathcal{D}(\tilde{\boldsymbol{x}}^*, \boldsymbol{x})$ based on the intuition that σ should shrink gradually when the distance from $\boldsymbol{x}$ decreases. μ is a critical hyper-parameter that needs to be tuned carefully. If μ is too large, the search point may probably violate the adversarial criterion and the success rate of updates is low. On the other hand, if μ is too small, we would make little progress towards minimizing the distance between $\tilde{\boldsymbol{x}}^*$ and $\boldsymbol{x}$ although the success rate is high. So we adopt the 1/5th success rule [21], which is a traditional method for hyper-parameter control in evolution strategies, to update μ as $\mu = \mu \cdot \exp(P_{\text{success}} - {}^1\!/_5)$, where P_{success} is the success rate of several past trials.

4. Experiments

In this section, we present the experimental results to demonstrate the effectiveness of the proposed evolutionary attack method. We comprehensively evaluate the robustness of several state-of-the-art face recognition models under the decision-based black-box attack scenario. We further apply the proposed method to attack a real-world face recognition system to demonstrate its practical applicability.

4.1. Experimental Settings

Target models. We study three state-of-the-art face recognition models, including SphereFace [16], CosFace [32] and ArcFace [5]. In testing, the feature representation for each image is first extracted by these models. Then, the cosine similarity between feature representations of different images are calculated. Finally, we use the thresholding strategy and nearest neighbor classifier for face verification and identification, respectively.

Datasets. We conduct experiments on the Labeled Face in the Wild (LFW) [12] and MegaFace [15] datasets. For face verification, in each dataset, we select 500 pairs of face images for dodging attack, in which each pair represent the same identity. And, we select another 500 pairs of face images for impersonation attack, in which the images of each pair are from different identities. For face identification, in each dataset, we select 500 images of 500 different identities to form a gallery set, and corresponding 500 images of the same identities to form a probe set. We perform dodging and impersonation attacks for images in the probe set. For impersonation attack, the target identity is chosen randomly. The input image size (*i.e.*, the dimension of the input space n) is $112 \times 112 \times 3$. All the selected images can be correctly recognized by the three face recognition models.

Compared methods. We compare the performance of the evolutionary attack method with all existing methods for

Model		SphereFace [16]				CosFace [32]				ArcFace [5]			
Queries		1,000	5,000	10,000	100,000	1,000	5,000	10,000	100,000	1,000	5,000	10,000	100,000
Dodging	Boundary [1]	2.4e-2	6.5e-3	4.7e-4	1.4e-5	2.0e-2	5.1e-3	5.4e-4	1.2e-5	3.1e-2	1.7e-2	1.6e-3	2.3e-5
	Optimization [4]	1.1e-2	2.1e-3	8.3e-4	4.6e-5	1.0e-2	2.0e-3	8.2e-4	4.0e-5	2.0e-2	6.1e-3	2.7e-3	9.8e-5
	NES-LO [14]	1.4e-1	4.0e-2	2.5e-2	5.5e-3	1.5e-1	3.6e-2	2.2e-2	4.7e-3	1.5e-1	4.5e-2	3.1e-2	1.3e-2
	Evolutionary	**1.3e-3**	**6.6e-5**	**2.5e-5**	**9.9e-6**	**1.2e-3**	**6.2e-5**	**2.3e-5**	**7.5e-6**	**3.2e-3**	**1.6e-4**	**5.4e-5**	**1.6e-5**
Impersonation	Boundary [1]	2.4e-2	1.1e-2	1.7e-3	3.6e-5	2.5e-2	8.9e-3	1.3e-3	2.3e-5	2.5e-2	1.3e-2	2.5e-3	3.8e-5
	Optimization [4]	1.9e-2	7.7e-3	3.7e-3	1.6e-4	1.9e-2	7.1e-3	3.3e-3	1.1e-4	2.0e-2	1.1e-2	6.0e-3	3.5e-4
	NES-LO [14]	7.9e-2	3.8e-2	2.8e-2	1.0e-2	8.8e-2	3.7e-2	2.7e-2	8.8e-3	8.8e-2	3.4e-2	2.3e-2	1.1e-2
	Evolutionary	**2.5e-3**	**1.6e-4**	**6.3e-5**	**2.3e-5**	**2.2e-3**	**1.3e-4**	**4.6e-5**	**1.5e-5**	**3.7e-3**	**2.5e-4**	**8.8e-5**	**2.6e-5**

Table 2. The results on face identification. We report the average distortion (MSE) of the adversarial images generated by different methods for SphereFace, CosFace, and ArcFace given 1,000, 5,000, 10,000, and 100,000 queries, based on the LFW dataset.

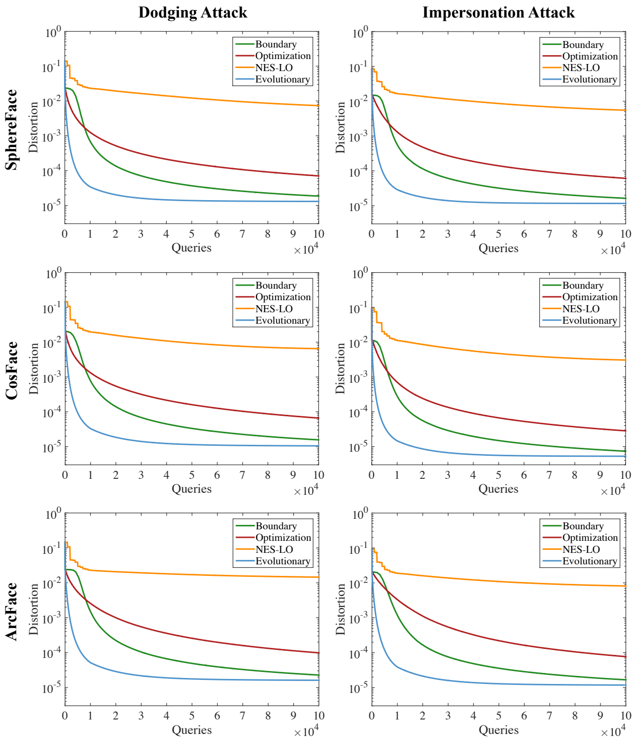

Figure 2. The results on face verification. We show the curves of the average distortion (MSE) of the adversarial images generated by different attack methods for SphereFace, CosFace, and ArcFace over the number of queries, based on the LFW dataset.

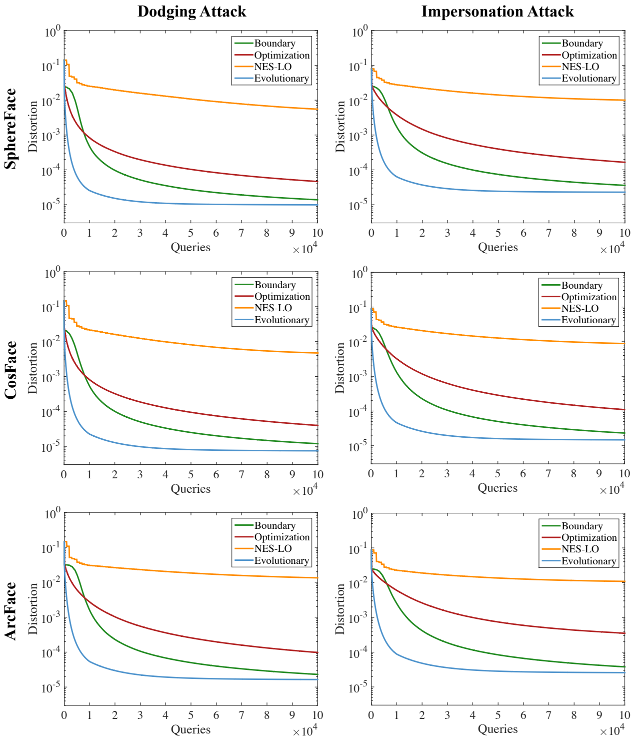

Figure 3. The results on face identification. We show the curves of the average distortion (MSE) of the adversarial images generated by different attack methods for SphereFace, CosFace, and ArcFace over the number of queries, based on the LFW dataset.

decision-based black-box attacks, including the boundary attack method [1], optimization-based method [4] and an extension of NES in the label-only setting (NES-LO) [14].

Evaluation metrics. For all methods, the generated adversarial examples are guaranteed to be adversarial. So we measure the distortion between the adversarial and original images by mean square error (MSE) to evaluate the performance of different methods[1]. We set a maximum number of queries to be 100,000 for each image across all experiments. Due to the space limitation, we leave the results on the MegaFace dataset in **Appendix B**. The results on both datasets are consistent. Our method is generally applicable beyond face recognition. We further present the results on the ImageNet dataset in **Appendix C**.

4.2. Experimental Results

We report the results on the LFW dataset in this section. We perform dodging attack and impersonation attack by Boundary, Optimization, NES-LO and the proposed Evolutionary method against SphereFace, CosFace, and ArcFace, respectively. For our method, we set the dimension of the search space as $m = 45 \times 45 \times 3$, and $k = m/20$ for stochastic coordinate selection. For other methods, we adopt the default settings. We calculate the distortion (MSE) of the adversarial images generated by each method averaged over the selected 500 images. And, the distortion curves over the number of queries for face verification are shown in Fig. 2, while those for face identification in Fig. 3. Besides,

[1] Images are normalized to $[0, 1]$.

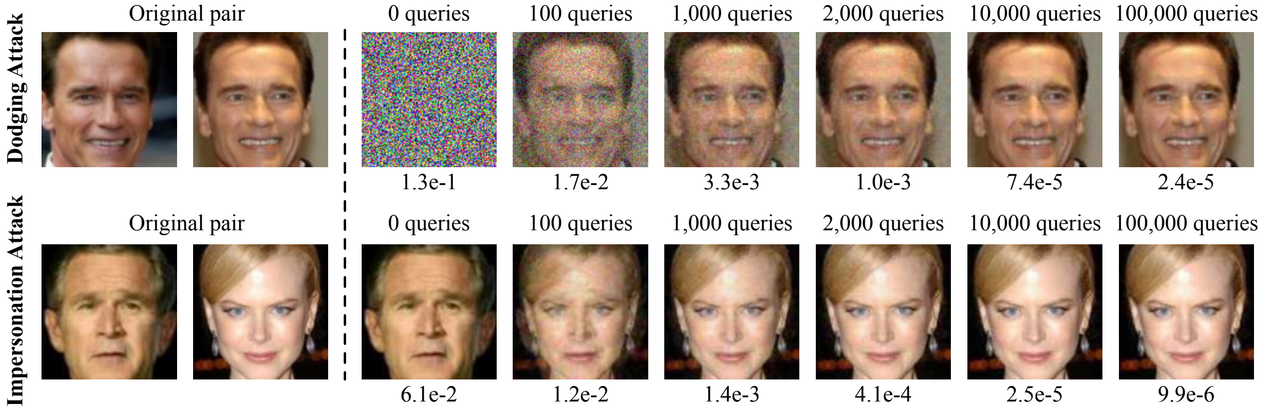

Figure 4. Examples of dodging and impersonation attacks on face verification for the ArcFace [5] model. The initial adversarial image is a random noise or the target image for each kind of attacks. The distortion between the adversarial image and the original image decreases gradually. We show the total number of queries and the mean square error until each point.

	SphereFace	CosFace	ArcFace
wo/ CMA, wo/ SCS	2.6e-4/1.9e-4	2.5e-4/9.2e-5	4.2e-4/2.6e-4
w/ CMA, wo/ SCS	2.4e-4/1.8e-4	2.3e-4/8.5e-5	3.8e-4/2.5e-4
w/ CMA, w/ SCS ($\mathbf{C}$)	**1.7e-4/1.3e-4**	**1.6e-4/6.4e-5**	**2.6e-4/1.7e-4**
w/ CMA, w/ SCS ($\mathbf{I}_n$)	2.0e-4/1.5e-4	1.9e-4/7.5e-5	3.0e-4/2.0e-4

Table 3. Comparisons of the evolutionary method with four settings: without CMA or SCS; with CMA, without SCS; with CMA and SCS where the selection probability is proportional to the elements in $\mathbf{C}$; with CMA and SCS where the selection probability is set equally. We report the average distortion (MSE) given 10,000 queries for dodging/impersonation attacks on face verification.

for 1,000, 5,000, 10,000, and 100,000 queries, we report the corresponding distortion values of different methods for face verification in Table 1, while those for face identification in Table 2. Two visual examples are also presented in Fig. 4 for dodging and impersonation attacks.

Above results demonstrate that our method converges much faster and achieves smaller distortions compared with other methods consistently across both tasks (*i.e.*, face verification and identification), both attack settings (*i.e.*, dodging and impersonation) and all face models. For example, as shown in Table 1 and 2, given 5,000 queries our method obtains the distortions which are about 30 times smaller than those generated by the second best method (*i.e.*, Optimization), which validates the effectiveness of the proposed method. From Fig. 4, it can be seen that 2,000 queries are sufficient to generate visually indistinguishable adversarial examples. For NES-LO, the hard-label predictions are first used to estimate the predicted probability (*e.g.*, 25 queries) and then it approximates the gradient by NES (*e.g.*, 40 trials). In consequence, this method requires more than 1,000 queries for only one update, which leads to the worst results.

It should be noted that the face recognition models are extremely vulnerable to adversarial examples. These models can be fooled in the black-box manner by adversarial examples with about only $1e^{-5}$ distortions, which are visually imperceptible for humans, as shown in Fig. 4.

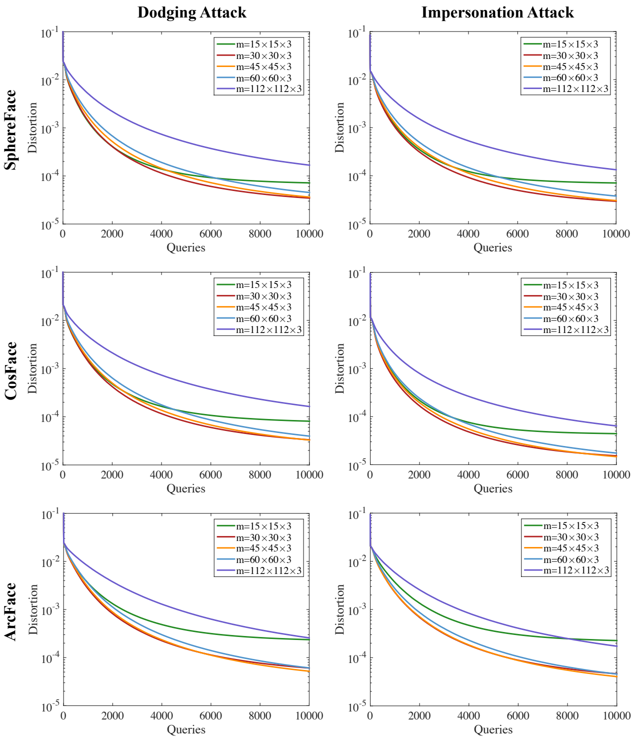

Figure 5. We show the curves of the average distortion (MSE) of the adversarial images generated by the evolutionary method with different dimensions of the search space over the number of queries. We perform dodging and impersonation attacks against SphereFace, CosFace, and ArcFace on face verification.

4.3. Ablation Study

We perform ablation study in this section to validate the effectiveness of each component in the proposed method. We conduct experiments based on face verification on the LFW dataset. In particular, we study the effects of covariance matrix adaptation, stochastic coordinate selection and dimensionality reduction respectively.

Covariance matrix adaptation (CMA). To examine the usefulness of CMA, we compare CMA with a baseline method that the covariance matrix is set to $\mathbf{I}_n$ without updating. We do not include stochastic coordinate selection or dimensionality reduction in this part for solely examining the effect of CMA. We show the results of the average distortion given 10,000 queries in the first two rows of Table 3. CMA improves the results over the baseline method.

Stochastic coordinate selection (SCS). We study two aspects of SCS. The first is whether SCS is useful. The second is whether we should select the coordinates with probability being proportional to the diagonal elements in the covariance matrix $\mathbf{C}$. We further perform experiments with SCS, where we compare the performance of SCS with the selection probability of each coordinate being proportional to each diagonal element in $\mathbf{C}$ or $\mathbf{I}_n$ (equal probability for each coordinate). By comparing the 2-4 rows of Table 3, it can be seen that SCS is beneficial for obtaining better results and sampling coordinates with probability proportional to c_{ii} is better than sampling with equal probability.

Dimensionality reduction. We finally study the influence of dimensionality reduction. We set the dimension m of the search space as $15\times15\times3$, $30\times30\times3$, $45\times45\times3$, $60\times60\times3$, and $112\times112\times3$. We perform dodging and impersonation attacks against SphereFace, CosFace, and ArcFace with each m, and compare the results in Fig. 5. It can be seen that the evolutionary method converges faster in a lower dimensional search space. However, if the dimension of the search space is too small (*e.g.*, $15\times15\times3$), the attack results in relatively large distortions. So we choose a medium dimension as $45\times45\times3$ in the above experiments.

4.4. Attacks on a Real-World Application

In this section, we apply the evolutionary attack method to the face verification API in Tencent AI Open Platform[2]. This face verification API allows users to upload two face images, and outputs a similarity score of them. We set the threshold to be 90, *i.e.*, if the similarity score is larger than 90, the two images are predicted to be the same identity; and if not, they are predicted to be different identities.

We choose 10 pairs of images from the LFW dataset to perform impersonation attack. The original two face images of each pair are from different identities. We generate a perturbation for one of them and make the API recognize the adversarial image to be the same identity as the other image. We set the maximum number of queries to be 10,000. We use the proposed evolutionary method to attack the face verification API and compare the results with Boundary [1] and Optimization [4]. We do not present the result of NES-LO [14], as it fails to generate an adversarial image within 10,000 queries. We show the average distortion between the adversarial and original images in Table 4. Our method still obtains a smaller distortion than other baseline methods.

Attack Method	Distortion (MSE)
Boundary [1]	1.63e-2
Optimization [4]	1.71e-2
Evolutionary	**2.54e-3**

Table 4. The results of impersonation attack on the real-world face verification API. We report the average distortion (MSE) of the selected 10 pairs of images by different attack methods.

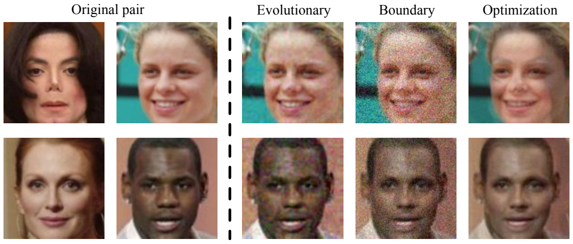

Figure 6. Examples of impersonation attack on the real-world face verification API. We show the original pairs of images as well as the adversarial images generated by each method.

We also show two examples in Fig. 6. It can be seen that the adversarial images generated by our method are more visually similar to the original images, while those generated by other methods have large distortions, making them distinguishable from the original images.

5. Conclusion

In this paper, we proposed an evolutionary attack algorithm to generate adversarial examples in the decision-based black-box setting. Our method improves the efficiency over the other methods by modeling the local geometry of the search directions and meanwhile reducing the dimension of the search space. We applied the proposed method to comprehensively study the robustness of several state-of-the-art face recognition models, and compared against the other methods. The extensive experiments consistently demonstrate the effectiveness of the proposed method. We showed that the existing face recognition models are extremely vulnerable to adversarial attacks in the black-box manner, which raises security concerns for developing more robust face recognition models. We finally attacked a real-world face recognition system by the proposed method, demonstrating its practical applicability.

Acknowledgements

Most of this work was done when Yinpeng Dong was intern at Tencent AI Lab, supported by the Tencent Rhino-Bird Elite Training Program. Yinpeng Dong, Hang Su, and Jun Zhu are supported by the National Key Research and Development Program of China (No. 2017YFA0700904), NSFC Projects (Nos. 61620106010, 61621136008, 61571261), Beijing NSF Project (No. L172037), Tiangong Institute for Intelligent Computing, NVIDIA NVAIL Program, DITD Program JCKY2017204B064, and the projects from Siemens and Intel.

[2] https://ai.qq.com/product/face.shtml#compare

References

[1] W. Brendel, J. Rauber, and M. Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*, 2018. 2, 5, 6, 8

[2] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017. 2

[3] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017. 2, 3, 5

[4] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018. 2, 3, 5, 6, 8

[5] J. Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018. 1, 2, 5, 6, 7

[6] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 1, 2

[7] A. D. Flaxman, A. T. Kalai, and H. B. Mcmahan. Online convex optimization in the bandit setting:gradient descent without a gradient. In *Sixteenth ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394, 2005. 3

[8] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. 3

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1, 2

[10] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001. 3, 4

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[12] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 1, 2, 5

[13] C. Igel, T. Suttorp, and N. Hansen. A computational efficient covariance matrix update and a (1+ 1)-cma for evolution strategies. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 453–460. ACM, 2006. 3

[14] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018. 2, 3, 5, 6, 8

[15] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016. 1, 2, 5

[16] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 1, 2, 5, 6

[17] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017. 2

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2

[19] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *CVPR*, 2017. 1, 2

[20] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. 3

[21] I. Rechenberg. Evolutionsstrategien. In *Simulationsmethoden in der Medizin und Biologie*, pages 83–114. Springer, 1978. 5

[22] R. Ros and N. Hansen. A simple modification in cma-es achieving linear time and space complexity. In *International Conference on Parallel Problem Solving from Nature*, pages 296–305. Springer, 2008. 4

[23] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 2

[24] 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 *ACM Sigsac Conference on Computer and Communications Security*, pages 1528–1540, 2016. 1, 2

[25] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. *arXiv preprint arXiv:1801.00349*, 2017. 1, 2

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[27] J. Su, D. V. Vargas, and S. Kouichi. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*, 2017. 5

[28] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014. 1, 2

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 1

[30] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1, 2, 4

[31] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1, 2

[32] 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 *CVPR*, 2018. 1, 2, 5, 6

[33] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 1, 2