Towards Face Encryption by Generating Adversarial Identity Masks

Xiao Yang[1], Yinpeng Dong[1,3], Tianyu Pang[1], Hang Su[1,2]*, Jun Zhu[1,2,3], Yuefeng Chen[4], Hui Xue[4]

[1] Dept. of Comp. Sci. and Tech., Institute for AI, BNRist Center, Tsinghua-Bosch Joint ML Center
[1] THBI Lab, Tsinghua University, Beijing, 100084, China
[2] Pazhou Lab, Guangzhou, 510330, China   [3] RealAI   [4] Alibaba Group

{yangxiao19, dyp17, pty17}@mails.tsinghua.edu.cn
{suhangss, dcszj}@mail.tsinghua.edu.cn   {yuefeng.chenyf, hui.xueh}@alibaba-inc.com

Abstract

As billions of personal data being shared through social media and network, the data privacy and security have drawn an increasing attention. Several attempts have been made to alleviate the leakage of identity information from face photos, with the aid of, e.g., image obfuscation techniques. However, most of the present results are either perceptually unsatisfactory or ineffective against face recognition systems. Our goal in this paper is to develop a technique that can encrypt the personal photos such that they can protect users from unauthorized face recognition systems but remain visually identical to the original version for human beings. To achieve this, we propose a targeted identity-protection iterative method (TIP-IM) to generate adversarial identity masks which can be overlaid on facial images, such that the original identities can be concealed without sacrificing the visual quality. Extensive experiments demonstrate that TIP-IM provides 95%+ protection success rate against various state-of-the-art face recognition models under practical test scenarios. Besides, we also show the practical and effective applicability of our method on a commercial API service.

1. Introduction

The blooming development of social media and network has brought a huge amount of personal data (e.g., photos) shared publicly. With the growing ubiquity of deep neural networks, these techniques dramatically improve the capabilities for the face recognition systems to deal with personal data [6, 26, 37, 46], but as a byproduct, also increase the potential risks for privacy leakage of personal information. For example, an unauthorized third party may scrabble and identify the shared photos on social media (e.g., Twitter, Facebook, LinkedIn, *etc.*) without the permission of
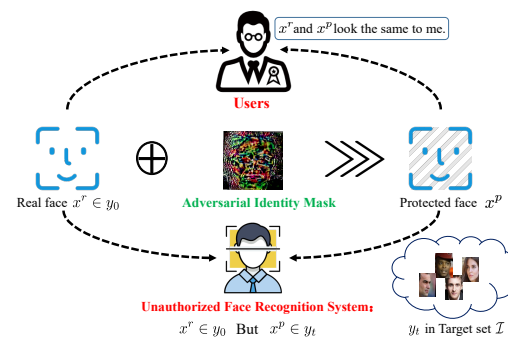
---
*Corresponding author

Figure 1. An illustrative example of targeted identity protection. When users share a photo x^r on social media (e.g., Twitter, Facebook, *etc.*), unauthorized applications could scrabble this identity y_0 based on face recognition systems, resulting in the privacy leakage of personal information. Thus we provide an effective identity mask tool to generate a protected image x^p, which can conceal the corresponding identity by misleading the malicious systems to predict it as a wrong target identity y_t in an authorized or virtual target set, which can be provided by the service providers.

their owners, resulting in cybercasing [23]. Therefore, it is imperative to provide users an effective way to protect their private information from being unconsciously identified and exposed by the excessive unauthorized systems, without affecting users' experience.

The past years have witnessed the progress for face encryption in both the security and computer vision communities. Among the existing techniques, obfuscation-based methods are widely studied. Conventional obfuscation techniques [48], such as blurring, pixelation, darkening, and occlusion, are maybe either perceptually satisfactory or effective against recognition systems [28, 31, 35]. The recent advance in generative adversarial networks (GANs) [15] provides an appealing way to generate more realistic images for obfuscation [14, 43, 44, 49, 25]. However, the resultant obfuscated images have significantly different visual appearances compared with the original images due to the exaggeration and suppression of some discriminative features, and occasionally generate unnatural output images with un-

desirable artifacts [44].

Recent researches have found that adversarial examples can evade the recognition of a FR system [52, 45, 16, 40] by overlaying adversarial perturbations on the original images [1]. It becomes an appealing way to apply an adversarial perturbation to conceal one's identity, even under a more strict constraint of impersonating some authorized or generated face images when available (e.g., given by the social media services). It provides a possible solution to specify the output, which may avoid an invasion of privacy to other persons if the resultant image is recognized as an arbitrary identity [1]. It should nevertheless be noted that although the adversarial perturbations generated by the existing methods (e.g., PGD [22] and MIM [8]) have a small intensity change (e.g., 12 or 16 for each pixel in $[0, 255]$), they may still sacrifice the visual quality for human perception due to the artifacts as illustrated in Fig. 2, and similar observation is also elaborately presented in [54, 38] that ℓ_p-norm adversarial perturbations can not fit human perception well. Moreover, the current adversarial attacks are mainly dependent on either the *white-box* control of the target system [40, 32] or the tremendous number of *model queries* [10], which are impractical in real-world scenarios (e.g., unauthorized face recognition systems on social media) for identity protection.

In this paper, we involve some valuable considerations from a general user's perspective and propose to alleviate the identity leakage of personal photos in real-world social media. We focus on face identification in particular, a typical sub-task in face recognition, the goal of which is to identify a real face image in an unknown gallery identity set (see Sec. 3), since it can be adopted by unauthorized applications for recognizing the identity information of users. As stated in Fig. 1, *face encryption* is to block the ability of automatic inference on malicious applications, making them predict a wrong authorized or virtual target by the service providers. In general, little is known about the face recognition system and no direct query access is possible. Therefore, we need to generate adversarial masks against a surrogate known model with the purpose of deceiving a *black-box* face recognition system. Moreover, we try to not affect the user experience when users share the protected photos on social media, and simultaneously conceal their identities from unauthorized recognition systems. Thus, the protected images should also be visually natural from the corresponding original ones, otherwise it may introduce undesirable artifacts as a result.

To address the aforementioned challenges, we propose a **targeted identity-protection iterative method** (TIP-IM) for face encryption against black-box face recognition systems. The proposed method generates adversarial identity masks that are both transferable and imperceptible. A good

transferability implies that a model can effectively deceive other black-box face recognition systems, meanwhile the imperceptibility means that a photo manipulated by an adversarial identity mask is visually natural for the human observers. Specifically, to ensure the generated images are not arbitrarily misclassified as other identities, we randomly choose a set of face images from a dataset collected from the internet as the specified targets in our experiments[2]. Our method obtains superior performance against white-box and black-box face systems with multiple target identities via a novel iterative optimization algorithm.

Extensive experiments under practical and challenging open-set [3] test scenarios [26] demonstrate that our algorithm provides 95+% protection success rate against white-box face systems, and outperforms previous methods by a margin even against various state-of-the-art algorithms. Besides, we also demonstrate its effectiveness in a real-world experiment by considering a commercial API service. Our main contributions are summarized as

- We involve some valuable considerations to protect privacy against unauthorized identification systems from the user's perspective, including targeted protection, natural outputs, black-box face systems, and unknown gallery set.

- We propose a targeted identity-protection iterative method (TIP-IM) to generate an adversarial identity mask, in which we consider multi-target sets and introduce a novel optimization mechanism to guarantee effectiveness under various scenarios.

2. Related Work

In this section, we review related work on face encryption. Typical information encryption [27, 41, 24] requires to encode message to protect information from exposed by unauthorized parties, whereas face encryption aims to protect users' facial information from being unconsciously identified and exposed by the unauthorized AI recognition systems. We provide a comprehensive comparison between the previous methods and ours in Tab. 1.

Obfuscation-based methods. Several works have been developed to protect private identity information in personal photos against face or person recognition systems. Earlier works [48, 34] study the performance of these systems under various simple image obfuscation methods, such as blurring, pixelation, darkening, occlusion, *etc*. These methods have been shown to be ineffective against the current

---

[1]The practical FR system will obtain the similarity rankings from the candidate image library, which could be mistaken for someone else.

[2]We choose face images from the Internet only for experimental illustration, which simulates the authorized or generated target set of face images provided by the service providers.

[3]The testing identities are disjoint from the training sets, which is regarded as the opposite of close-set classification task.

	Evasion [48]	PP-GAN [49]	Inpainting [43]	Replacement [44]	Eyeglasses [40]	Evolutionary [10]	LOTS [36]	GAMAN [32]	Ours
Unknown gallery set	No	No	No	No	No	**Yes**	No	No	**Yes**
Target identity	No	No	No	No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Black-box model	**Yes**	**Yes**	**Yes**	**Yes**	No	*Yes (Queries)*	No	No	**Yes**
Natural outputs	No	**Yes**	**Yes**	**Yes**	No	No	**Yes**	No	**Yes**
Same faces	*Partially*	*Partially*	No	No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 1. A comparison among different methods w.r.t the unknown gallery set, targeted misclassification of the output faces, black-box face models, natural outputs, and whether the output faces are recognized as the same identities as the original ones for human observers.

recognition systems [31, 28, 25], since they can adapt to the obfuscation patterns. More sophisticated techniques have been proposed thereafter. For example, generative adversarial networks (GANs) [15] provide a useful way to synthesize realistic images on the data distribution for image obfuscation [49]. In [43], the obfuscated images are generated by head in-painting conditioned on the detected face landmarks. However, these image obfuscation methods often change the visual appearances of face images and even lead to unnatural outputs, limiting their utility for users.

Adversarial methods. Deep neural networks are susceptible to adversarial examples [45, 16, 7, 33], so are the face recognition models [40, 10, 51]. Fawkes [39] fools unauthorized facial recognition models by introducing adversarial examples into training data. A recent work [32] proposes to craft protected images from a game theory perspective. However, our work is different from their previous works in three aspects. First, we focus on the unknown face systems without changing training data [39], while [32] assumes the white-box access to the target model. Second, we consider the open-set face identification protocol with an unknown gallery set rather than a closed-set classification scenario. Ours can provide better protection success rate against unknown recognition systems on more practical open-set scenarios. Third, we have the ability to control the naturalness of the protected images under the ℓ_p norm.

Differential privacy. As one of the popular definitions of privacy, differential privacy (DP) [11, 12] has been introduced in the context of machine learning and data statistics, which requires that the returned information about an underlying dataset is robust to any change of one individual, thus protecting the privacy of entities. Along this routine, many promising DP techniques [12, 29] and practical applications [3, 2] on DP have been developed. While DP withholds the existence of entities in a dataset, in this paper we focus on concealing the identity of a single one by exploiting the vulnerability of neural networks [45].

3. Adversarial Identity Mask

Let $f(\boldsymbol{x}) : \mathcal{X} \rightarrow \mathbb{R}^d$ denote a face recognition model that extracts a fixed length feature representation in $\mathbb{R}^d$ for an input face image $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^n$. Given the metric $\mathcal{D}_f(\boldsymbol{x}_1, \boldsymbol{x}_2) = \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|_2^2$ that measures the feature distance between two face images, face recognition compares the distance between a probe image and a gallery set of face images $\mathcal{G} = \{\boldsymbol{x}_1^g, ..., \boldsymbol{x}_m^g\}$, and returns the identity whose face image has the nearest feature distance with the probe image.

In this paper, we involve some valuable considerations from the user's perspective, to protect user's photos against an illegal face recognition systems, as illustrated in Fig. 1. Specifically, to conceal the true identity y of a user's image $\boldsymbol{x}^r$, we aim to generate a protected image $\boldsymbol{x}^p$ by adding an adversarial identity mask $\boldsymbol{m}^a$ to $\boldsymbol{x}^r$ which can be denoted by $\boldsymbol{x}^p = \boldsymbol{x}^r + \boldsymbol{m}^a$ to make the face recognition system predict $\boldsymbol{x}^p$ as a different authorized identity or virtual identity corresponding to a generated image. Rather than specifying a single target identity for generating the protected image, we choose an identity set $\mathcal{I} = \{y_1, ..., y_k\}$, i.e., we allow the face recognition system to recognize the protected image as an arbitrary one of the target identities in $\mathcal{I}$ rather than a single one, which makes identity protection easier to achieve due to the relaxed constraints.

Formally, let $\mathcal{G}_y = \{\boldsymbol{x} | \boldsymbol{x} \in \mathcal{G}, \mathcal{O}(\boldsymbol{x}) = y\}$ denote a subset of $\mathcal{G}$ containing all face images belonging to the true identity y of $\boldsymbol{x}^r$, with $\mathcal{O}$ being an oracle to give the ground-truth identity labels, and $\mathcal{G}_\mathcal{I} = \bigcup_{1 \le i \le k} \mathcal{G}_{y_i}$ denote the face images belonging to the target identities of $\mathcal{I}$ in the gallery set $\mathcal{G}$. To conceal the identity of $\boldsymbol{x}^r$, the protected image $\boldsymbol{x}^p$ should satisfy the constraint as

$$\exists \boldsymbol{x}^t \in \mathcal{G}_\mathcal{I}, \forall \boldsymbol{x} \in \mathcal{G}_y : \mathcal{D}_f(\boldsymbol{x}^p, \boldsymbol{x}) > \mathcal{D}_f(\boldsymbol{x}^p, \boldsymbol{x}^t). \quad (1)$$

It ensures that the feature distance between the generated protected image $\boldsymbol{x}^p$ and a target identity's image $\boldsymbol{x}^t$ in $\mathcal{G}_\mathcal{I}$ is smaller than that between $\boldsymbol{x}^p$ and any image belonging to the true identity y in $\mathcal{G}_y$.

We involve more practical considerations from a general user's perspective than the previous studied setting, in the following three aspects.

Naturalness. To make the protected image indistinguishable from the corresponding original one, a common practice is to restrict the ℓ_p ($p = 2, \infty$, *etc.*) norm between the protected and original examples, as $\|\boldsymbol{m}^a\|_p \le \epsilon$. However, the perturbation under the ℓ_p norm can not naturally fit human perception well [54, 38], as also illustrated in Fig. 2. Therefore, we require that the protected image should look natural besides the constraint of the ℓ_p norm bound, to make it constrained on the data manifold of real images [42], thus achieving imperceptible for human eyes. We use an objective function to promote the naturalness of the protected image, which will be specified in the following section.

PGD

MIM

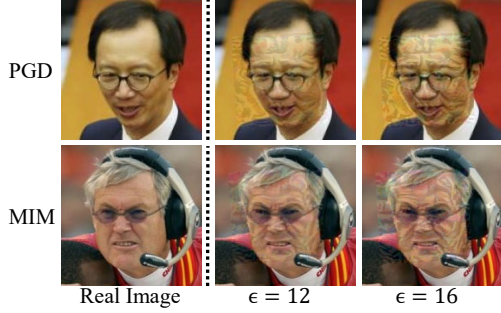Real Image $\epsilon = 12$ $\epsilon = 16$

Figure 2. Illustration of different perturbations under the l_∞ norm. More examples are presented in Appendix D.

Unawareness of gallery set. For a real-world face recognition system, we have no knowledge of its gallery set $\mathcal{G}$, meaning that we are not able to solve Eq. (1) directly, while previous works assume the availability of the gallery set or a closed-set protocol (i.e., no gallery set). To address this issue, we use substitute face images for optimization. In particular, we collect an image set $\hat{\mathcal{G}}_\mathcal{I}$ containing face images that belong to the target identities of $\mathcal{I}$ as a surrogate for $\mathcal{G}_\mathcal{I}$; and use $\{\boldsymbol{x}^r\}$ directly instead of $\mathcal{G}_y$. The rationality of using substitute images is that face representations of one identity are similar, and thus the representation of a protected image optimized to be similar to the substitutes can also be close to images belonging to the same target identity in the gallery set.

Unknown face systems. In practice, we are also unaware of the face recognition model, include its architecture, parameters, and gradients. Previous methods rely on the white-box access to the target model, which are impractical in real-world scenarios for identity protection. Thus we adopt a surrogate white-box model against which the protected images are generated, with the purpose of improving the transferability of adversarial masks against unknown face systems.

In summary, our considerations are designed to simulate the real-world scenarios with minimum assumptions of the target face recognition system, which is also more challenging than previously studied settings.

4. Methodology

To achieve the above requirements, we propose a **targeted identity-protection iterative method** (TIP-IM) to generate protected images in this section.

4.1. Problem Formulation

To generate a protected image $\boldsymbol{x}^p$ that is both effective for obfuscation against face recognition systems and visually natural for human eyes, we formalize the objective of targeted privacy-protection function as

$$\min_{\boldsymbol{x}^t, \boldsymbol{x}^p} \mathcal{L}_{iden}(\boldsymbol{x}^t, \boldsymbol{x}^p) = \mathcal{D}_f(\boldsymbol{x}^p, \boldsymbol{x}^t) - \mathcal{D}_f(\boldsymbol{x}^p, \boldsymbol{x}^r)$$
$$\text{s.t. } \|\boldsymbol{x}^p - \boldsymbol{x}^r\|_p \leq \epsilon, \mathcal{L}_{nat}(\boldsymbol{x}^p) \leq \eta, \quad (2)$$

where $\boldsymbol{x}^t \in \hat{\mathcal{G}}_\mathcal{I}$, and $\mathcal{L}_{iden}$ is a relative identification loss that enables the generated $\boldsymbol{x}^p$ to increase the distance gap between a targeted image $\boldsymbol{x}^t$ and the original image $\boldsymbol{x}^r$ in the feature space. $\mathcal{L}_{nat} \leq \eta$ is a constraint condition that makes $\boldsymbol{x}^p$ look natural. We also restrict the ℓ_p norm of the perturbation to be smaller than a constant ϵ such that the visual appearance does not change significantly. Note that for the unawareness of gallery set, we use the substitute face images $\hat{\mathcal{G}}_\mathcal{I}$ in our objective (2); for the unknown model, we generate a protected image $\boldsymbol{x}^p$ against a surrogate white-box model with the purpose of fooling the black-box model based on the transferability. Thus the requirements in the proposed targeted identity-protection function can be fulfilled by solving Eq. (2).

Although the perturbation is somewhat small due to the ℓ_p norm constraint in Eq. (2), it can be still perceptible and not natural for human eyes, as shown in Fig. 2 . Therefore, we add a loss $\mathcal{L}_{nat}$ into our objective to explicitly encourage the naturalness of the generated protected image. In this paper, we adopt the *maximum mean discrepancy* (MMD) [4] as $\mathcal{L}_{nat}$, because it is an effective non-parametric and differentiable metric capable of comparing two data distributions and evaluating the imperceptibility of the generated images. In our case, given two sets of data $\mathbf{X}^p = \{\boldsymbol{x}_1^p, ..., \boldsymbol{x}_N^p\}$ and $\mathbf{X}^r = \{\boldsymbol{x}_1^r, ..., \boldsymbol{x}_N^r\}$ comprised of N generated data and N real images, MMD calculates their discrepancy by

$$\text{MMD}(\mathbf{X}^p, \mathbf{X}^r) = \big|\big|\frac{1}{N}\sum_{i=1}^N \phi(\boldsymbol{x}_i^p) - \frac{1}{N}\sum_{j=1}^N \phi(\boldsymbol{x}_j^r)\big|\big|_\mathcal{H}^2, \quad (3)$$

where $\phi(\cdot)$ maps the data to a reproducing kernel Hilbert space (RKHS) [4]. We adopt the same $\phi(\cdot)$ as in [4]. By minimizing MMD between the samples $\mathbf{X}^p$ from the generated distribution and the samples $\mathbf{X}^r$ from the real data distribution, we can constrain $\mathbf{X}^p$ to lie on the manifold of real data distribution, meaning the protected images in $\mathbf{X}^p$ will be as natural as real examples.

Since MMD is a differentiable metric and defined on the batches of images, we thus integrate MMD into Eq. (2) and rewrite our objective with a batch-based formulation[4] as

$$\min_{\mathbf{X}^p} \mathcal{L}(\mathbf{X}^p) = \frac{1}{N}\sum_{i=1}^N \mathcal{L}_{iden}(\boldsymbol{x}_i^t, \boldsymbol{x}_i^p) + \gamma \cdot \text{MMD}(\mathbf{X}^p, \mathbf{X}^r),$$
$$\text{s.t. } \|\boldsymbol{x}_i^p - \boldsymbol{x}_i^r\|_p \leq \epsilon, \quad (4)$$

where $\boldsymbol{x}_i^t \in \hat{\mathcal{G}}_\mathcal{I}$ and γ is a hyperparameter to balance these two losses.

4.2. Targeted Identity-Protection Iterative Method

Given the overall loss $\mathcal{L}(\mathbf{X}^p)$ in Eq. (4), we can therefore generate the batch of protected images $\mathbf{X}^p$ by minimizing

---

[4]In case of there is only one single image or a small number of images, we can augment the images with multiple transformations to make up a large batch, with the results shown in Appendix B.

$\mathcal{L}(\mathbf{X}^p)$. Given the definitions of $\mathcal{L}_{iden}$ and MMD in Eq. (2) and Eq. (3), $\mathcal{L}(\mathbf{X}^p)$ is a differentiable function w.r.t. $\mathbf{X}^p$, and thus we can iteratively apply fast gradient method [22] multiple times with a small step size α to generate protected images by minimizing the loss $\mathcal{L}(\mathbf{X}^p)$. In particular, we optimize $\mathbf{X}^p$ via

$$\mathbf{X}_{t+1}^p = \Pi_{\{\mathbf{X}^r, \ell_p, \epsilon\}}\big(\mathbf{X}_t^p - \alpha \cdot \text{Normalize}(\nabla_\mathbf{X} \mathcal{L}(\mathbf{X}_t^p))\big), \quad (5)$$

where $\mathbf{X}_t^p$ is the batch of protected images at the t-th iteration, Π is the projection function that projects the protected images onto the ℓ_p norm bound, and $\text{Normalize}(\cdot)$ is used to normalize the gradient (e.g., a sign function under the ℓ_∞ norm bound or the ℓ_2 normalization under the ℓ_2 norm bound). We perform the iterative process for a total number of T iterations and get the final protected images as $\mathbf{X}_T^p$. To prevent protected images from falling into local minima and improve their transferability for other black-box face recognition models, we incorporate the momentum technique [8] into the iterative process.

4.3. Search Optimal x^t via Greedy Insertion

When there is only one target face image in $\hat{\mathcal{G}}_\mathcal{I}$, we do not need to consider how to select x^t to effectively incorporate into Eq. (4). When the target set $\hat{\mathcal{G}}_\mathcal{I}$ contains multiple target images, it offers more potential optimization directions to get better performance. Therefore, we develop an optimization algorithm to search for the optimal target while generating protected images as Eq. (5). Specifically, for the iterative procedure in Eq. (5) with T iterations, we select a representative target for each protected image in $\hat{\mathcal{G}}_\mathcal{I}$ at each iteration for updates, which belongs to a subset selection problem.

Definition 1. *Let S_t denote the set of the selected targets from $\hat{\mathcal{G}}_\mathcal{I}$ at each iteration until the t-th iteration. Let F denote a set mapping function that outputs a gain value (larger is better) in $\mathbb{R}$ for a set. For $x^t \in \hat{\mathcal{G}}_\mathcal{I}$, we define $\Delta(x^t|S_t) = F(S_t \cup \{x^t\}) - F(S_t)$ be the marginal gain of F at S_t given x^t.*

Formally, as the iteration gets increasing in the iteration loop, if the marginal gain decreases monotonically, then F will belong to the family of submodular functions [55]. For a submodular problem, a greedy algorithm can be used to find an approximate solution, and it has been shown that submodularity will have a $(1 - 1/e)$-approximation [30] for monotinely submodular functions. Although our iterative identity-protection method is not guaranteed to be strictly submodular, the solution based on greedy insertion still plays an obvious role even if submodularity is not strictly decreased [55, 13], which evaluate the theoretical results and justify that the greedy algorithm has a performance guarantee for the maximization of approximate submodularity. Therefore, we adopt the greedy insertion solution

---

Algorithm 1: Search Optim. via Greedy Insertion

Input: The privacy-protection objective function $\mathcal{L}_{iden}$ from Eq. (2); a real face x^r and a multi-identity face images $\hat{\mathcal{G}}_\mathcal{I}$; a feature representation function f; a gain function G.

Input: The protected image x^p generated before the current iteration.

Output: The best target image x^{t^*} in $\hat{\mathcal{G}}_\mathcal{I}$.

1   $g_{best} \leftarrow 0$; $x^{t^*} \leftarrow$ None;
2   **for** x^t in $\hat{\mathcal{G}}_\mathcal{I}$ **do**
3   Get the loss $\mathcal{L}_{iden}(x^t, x^p)$ via Eq. (2);
4   Compute the gradient $\nabla_x \mathcal{L}_{iden}(x^t, x^p)$;
5   Generate candidate protected image $\hat{x}^p = \Pi_{\{x^r, \ell_p, \epsilon\}}(x^p - \alpha \cdot \text{Normalize}(\nabla_x \mathcal{L}_{iden}(x^t, x^p)))$;
6   Calculate $g = \text{G}(\hat{x}^p)$;
7   **if** $g > g_{best}$ **then**
8         $g_{best} \leftarrow g$; $x^{t^*} \leftarrow x^t$;
9   **end**
10   **end**

---

as an approximately optimal solution for our multi-target problem.

By analyzing the above setting, we perform the approximate submodular optimization by greedy insertion algorithm, which calculates the gain of every object from the target set at each iteration and integrates the object with the largest gain into current subset S_t by Definition 1 as

$$S_{t+1} = S_t \cup \{\arg\max_{x^t \in \hat{\mathcal{G}}_\mathcal{I}} \Delta(x^t|S_t)\}. \quad (6)$$

To achieve this, we need to define the above set mapping function F. In particular, we specify F as first generating a protected image x^p given the targets in S_t for t iterations via Eq. (5), and then using a function G to compute a gain value by Definition 1. An appropriate gain function G should choose examples that are effective for minimizing $\mathcal{L}_{iden}(x^t, x^p)$ at each iteration. It is noted that G must also have positive value and a larger value indicates better performance. Based on this analysis, we design a feature-based similarity gain function as

$$\text{G}(x^p) = \log\big(1 + \max_{x^t \in \hat{\mathcal{G}}_\mathcal{I}} \exp(\mathcal{D}_f(x^p, x^r) - \mathcal{D}_f(x^p, x^t))\big), \quad (7)$$

where the algorithm tends to select a target closer to the real image in the feature space at each iteration. The algorithm is summarized in Algorithm 1.

5. Experiments

In this section, we conduct extensive experiments in the aspect of identity protection to demonstrate the effective-

Model	Backbone	Loss	Parameters (M)
FaceNet [37]	InceptionResNetV1	Triplet	27.91
SphereFace [26]	Sphere20	A-Softmax	28.08
CosFace [46]	Sphere20	LMCL	22.67
ArcFace [6]	IR-SE50	Arcface	43.80
MobileFace [5]	MobileFaceNet	Softmax	1.20
ResNet50 [18]	ResNet50	Softmax	40.29

Table 2. Chosen target models that lie in various settings, including different architectures and training objectives.

ness of the proposed method. We thoroughly evaluate different properties of our method based on various state-of-the-art face recognition models.[5]

5.1. Experimental Settings

Datasets. The experiments are constructed on the Labeled Face in the Wild (LFW) [19] and MegFace [21] datasets. We involve some additional considerations to draw near realistic testing scenarios: 1) **practical gallery set**: we first select 500 different identities as the protected identities. Meanwhile, we randomly select an image from each identity as the probe image (total 500 images), and the *other* images (not selecting one template) for each identity are assembled to form a gallery set because the gallery of face encryption in social media includes multiple images per identity (more difficult yet practical for simultaneously concealing multiple images per identity); 2) **target identities**: we randomly select another 10 identities as $\mathcal{I}$ from a dataset in the Internet named MS-Celeb-1M [17]. We select one image for each of these target identities to form $\hat{\mathcal{G}}_\mathcal{I}$ and the remaining images are integrated into the gallery set, which can ensure the unawareness of gallery set that target images in the optimized phase are different from ones in the testing; 3) **additional identities:** we add additional 500 identities to the gallery set, which accords with a realistic test scenario. Thus we construct two challenging yet practical data scenarios (over 1k identities and total 10K images).

Target models. We select models with diverse backbones and training losses to fully demonstrate the ability to protect user privacy in Tab. 2. In experiments, we first use MTCNN [53] to detect faces in the image, then align the images and crop them to 112×112, meaning that the identity masks are executed only in the face area. Only one model is used as known model to generate the identity masks, and test the protection performance in other unknown models.

Compared Methods. We investigate many adversarial face encryption methods [40, 32], which essentially depend on single-target adversarial attack method [22]. Advanced MIM [8] introduces the momentum into iterative process [22] to improve the black-box transferability, and DIM and TIM [50, 9] aim to achieve better transferability by input or gradient diversity. Note that TIM only focus on evading defense models and experimentally also achieve worse performance than MIM and DIM. Thus MIM and

---

[5]Code at https://github.com/ShawnXYang/TIP-IM.

DIM are regarded as more effective single-target black-box algorithms as comparison. As original DIM only support single-target attack in the iterative optimization, we thus incorporate a multi-target version for DIM via a dynamic assignment from same target set in the inner minimization, named MT-DIM. Besides, we study the influence of other multi-target optimization methods. We denote an additional gain function based on Eq. (7) as $\mathrm{G}_1(\boldsymbol{x}) = \log\left(1 + \sum_{\boldsymbol{x}^t \in \hat{\mathcal{G}}_\mathcal{I}} \exp(\mathcal{D}_f(\boldsymbol{x}, \boldsymbol{x}^r) - \mathcal{D}_f(\boldsymbol{x}, \boldsymbol{x}^t))\right)$ which is named *Center-Opt*. Center-Opt promotes protected images to be updated towards the mean center of target identities in the feature space, which is similarly adopted in [36]. Note that single-target methods calculate optimal result as final report by attempting a target from the same target set. We set the number of iterations as $N = 50$, the learning rate $\alpha = 1.5$ and the size of perturbation $\epsilon = 12$ under the ℓ_∞ norm bound, which are identical for all the experiments.

Evaluation Metrics. To comprehensively evaluate the **protection success rate**, we report Rank-N targeted identity success rate named *Rank-N-T* and untargeted identity success rate named *Rank-N-UT* (higher is better), which are consistent with the evaluation of face recognition [6, 46]. Specifically, given a probe image $\boldsymbol{x}$ and a gallery set $\mathcal{G}$ with at least one image of the same identity with $\boldsymbol{x}$, meanwhile $\mathcal{G}$ has images of target identities. The face recognition algorithm ranks the distance $\mathcal{D}_f$ for all images in the gallery to $\boldsymbol{x}$. Rank-N-T means that *at least one* of the top N images belongs to the target identity, whereas Rank-N-UT needs to satisfy that top N images do not have the same identity as $\boldsymbol{x}$. In this paper, we report Rank-1-T / Rank-1-UT and Rank-5-T / Rank-5-UT. Note that Rank-1-T / Rank-1-UT (Accuracy / Misclassification) is the most common evaluation metric in prior works, whereas Rank-5-T / Rank-5-UT can provide a comprehensive understanding since it is not sure whether the image will reappear in the top-K candidates. All methods including single-target methods adopt the same target identities and evaluation criterion for a fair comparison.

To test the **imperceptibility** of the generated protected images, we adopt the standard quantitative measures—PSNR (dB) and structural similarity (SSIM) [47], as well as MMD in the face area. For SSIM and PSNR, a larger value means better image quality, whereas a smaller MMD value indicates superior performance.

5.2. Effectiveness of Black-box Face Encryption

We first generate protected images against ArcFace, MobileFace, and ResNet50 respectively, by the proposed TIP-IM. We then feed the generated protected images to all face models for testing the performance in Tab. 3 and Tab. 4. Our algorithm achieves nearly two times of the success rates than previous state-of-the-art method MT-DIM in terms of Rank-1-T and Rand-5-T, and outperforms other methods by a large margin, whereas SSIM values among compared

	Method	ArcFace		MobileFace		ResNet50		SphereFace		FaceNet		CosFace	
		R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T
ArcFace	MIM [8]	94.0*	96.9*	14.3	45.8	8.2	32.4	3.1	14.5	3.1	17.9	1.7	10.1
	DIM [50]	94.8*	97.6*	16.8	48.0	10.8	34.8	4.2	15.6	4.4	19.0	2.6	11.0
	MT-DIM [50]	34.8*	68.2*	18.8	53.6	15.8	46.0	3.8	18.4	9.6	33.6	2.0	11.2
	Center-Opt	59.4*	84.6*	36.8	66.0	28.8	57.6	6.6	21.4	11.8	35.4	3.8	13.0
	TIP-IM	**97.2***	**98.8***	**69.8**	**90.6**	**56.0**	**80.6**	**13.2**	**32.0**	**32.8**	**56.2**	**11.4**	**31.0**
MobileFace	MIM [8]	8.1	27.9	96.1*	98.3*	26.7	61.5	4.5	18.0	3.7	19.4	0.3	4.1
	DIM [50]	9.4	28.8	96.6*	98.4*	28.8	63.2	6.2	19.2	4.8	20.6	1.2	5.2
	MT-DIM [50]	10.6	30.2	40.2*	73.2*	18.6	49.4	6.8	22.2	9.8	27.0	2.0	11.0
	Center-Opt	14.8	41.8	53.0*	83.4*	21.8	53.6	5.8	25.6	12.2	29.6	3.6	12.4
	TIP-IM	**44.0**	**68.2**	**96.6***	**99.2***	**62.8**	**85.8**	**12.6**	**29.8**	**28.8**	**46.2**	**12.4**	**31.2**
ResNet50	MIM [8]	5.2	24.7	24.6	56.5	30.1*	64.8*	8.1	23.4	4.9	20.7	0.9	5.6
	DIM [50]	7.0	26.4	26.6	57.2	31.4*	65.0*	9.2	24.2	6.8	22.8	2.0	6.6
	MT-DIM [50]	14.2	37.6	22.4	52.6	31.4*	65.0*	5.2	16.8	9.4	29.2	1.8	8.8
	Center-Opt	13.2	37.14	26.8	58.6	41.6*	73.4*	6.8	21.0	9.2	27.2	2.4	12.0
	TIP-IM	**34.2**	**56.8**	**62.4**	**83.4**	**95.6***	**98.2***	**11.4**	**25.6**	**23.2**	**40.0**	**10.8**	**26.2**

Table 3. Rank-1-T and Rank-5-T (%) of black-box identity protection against different models on LFW. * indicates white-box results.

	Attack	ArcFace		MobileFace		ResNet50		SphereFace		FaceNet		CosFace	
		R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U
ArcFace	DIM [50]	95.8*	91.8*	67.6	58.2	58.2	48.0	79.6	68.2	67.4	53.6	74.2	62.8
	MT-DIM [50]	96.0*	93.6*	73.6	65.2	64.8	54.2	82.8	73.0	73.0	60.0	74.4	63.8
	TIP-IM	**97.4***	**96.4***	**79.4**	**68.8**	**70.4**	**56.8**	**85.2**	**76.6**	**73.4**	**63.4**	**84.0**	**73.8**
MobileFace	DIM [50]	60.2	45.4	96.4*	92.0*	72.2	59.0	80.0	69.6	68.4	53.6	75.6	62.2
	MT-DIM [50]	66.4	52.8	95.4*	95.4*	77.6	68.4	83.2	73.2	74.6	58.6	76.0	62.8
	TIP-IM	**68.6**	**58.8**	**96.6***	**94.8***	**81.4**	**71.2**	**84.4**	**74.0**	**77.6**	**60.4**	**79.4**	**68.2**
ResNet50	DIM [50]	77.6	50.4	80.6	72.2	95.4*	91.6*	80.4	65.8	69.0	53.2	64.2	61.8
	MT-DIM [50]	79.0	52.4	84.8	75.2	94.4*	93.8*	82.2	72.4	77.8	60.8	77.0	65.8
	TIP-IM	**83.6**	**59.6**	**87.0**	**81.8**	**96.8***	**94.6***	**85.8**	**75.0**	**79.4**	**65.4**	**83.6**	**73.0**

Table 4. Rank-1-UT and Rank-5-UT (%) of black-box identity protection against different models on LFW. * indicates white-box attacks.

	Metric	$\gamma = 0.0$	$\gamma = 1.0$	$\gamma = 2.0$	$\gamma = 3.0$
ArcFace	PSNR($\uparrow$)	25.26	25.59	26.08	27.63
	SSIM($\uparrow$)	0.6520	0.6690	0.6986	0.7817
	MMD($\downarrow$)	0.7567	0.7562	0.7554	0.7518
MobileFace	PSNR($\uparrow$)	25.24	25.18	25.72	27.19
	SSIM($\uparrow$)	0.6490	0.6523	0.6828	0.7533
	MMD($\downarrow$)	0.7567	0.7568	0.7559	0.7525
ResNet50	PSNR($\uparrow$)	25.14	25.26	25.74	27.21
	SSIM($\uparrow$)	0.6507	0.6595	0.6897	0.760
	MMD($\downarrow$)	0.7570	0.7567	0.7558	0.7525

Table 5. The average PSNR (db), SSIM, and MMD of the protected images generated by TIP-IM with different γ.

Figure 3. Comparison of SSIM for different methods.

methods are very similar in Fig. 3. MT-DIM obtains more acceptable performance than single method DIM, indicating that multi-target setting yields a better black-box transferability. It can be also observed that different multi-target methods will influence the performance, and proposed TIP-IM defined in Eq. (7) achieves better performance than Center-Opt, . We also report the results of Rank-1-UT and Rank-5-UT in Tab. 4, which can still maintain the best performance with an average accuracy of Rank-1-UT over 80% for black-box models. As a whole, TIP-IM provides more promising multi-target optimization direction, making generated protected images more effective for black-box models. Note that the protected images generated by ArcFace have excellent transferability to the other black-box models. Thus we will have priority to select ArcFace or ensemble mechanism [8] as the substitute model for better performance in practical application. Due to the space limitation, we leave the results of MegFace in Appendix A.

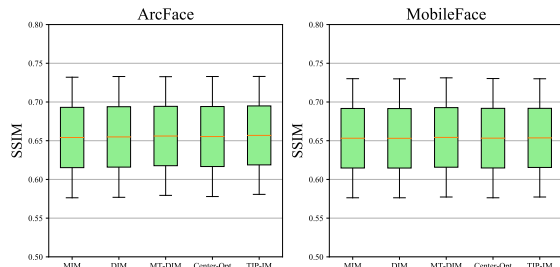**Comparison experiments about target images.** We test the performance of different numbers of targets in Appendix C. We experimentally find 10 target identities in this paper is enough, which implies that small increases in the number of targets can obtain impressive performance in spite of taking slight timing cost. We specify some *generated* images from StyleGAN [20] as target images. The results show that our algorithm still has excellent black-box performance of identity protection. In practical applications, we can arbitrarily specify the available and authorized target identity set or generated face images, and our algorithm is applicable to any target set.

5.3. Naturalness

To examine whether our algorithm is able to control the naturalness of protected images in the process of generating the samples, we perform experiments with different coefficient γ. Tab. 5 shows the evaluation results of different face recognition models (including ArcFace, MobileFace, and ResNet50) w.r.t three different metrics—PSNR, SSIM, and MMD. As γ increases, the visual quality of the generated images is getting better based on different metrics, which is also consistent with the example in Fig. 4. Therefore, conditioned on different coefficient γ, we can control the degree
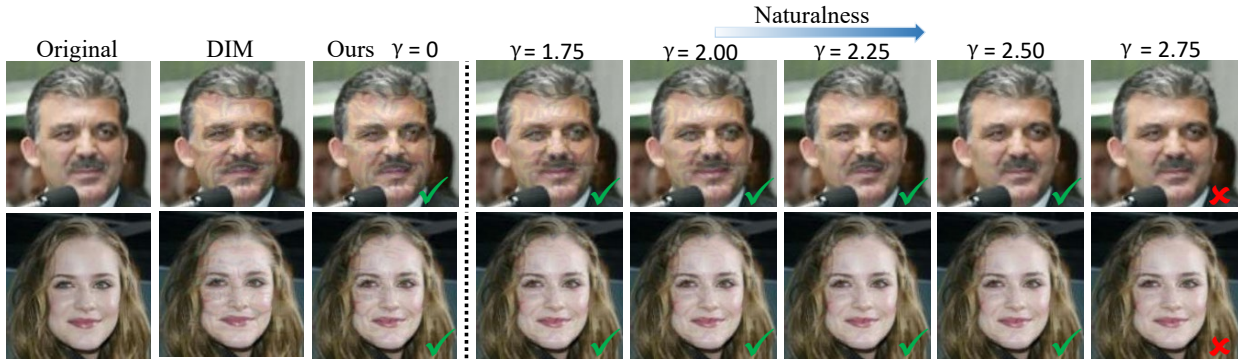
Figure 4. Experiments on how different γ affects the performance. Green hook refers to successful targeted identity protection while red hook refers to failure, which also implies a trade-off on effectiveness and naturalness. Best view when zoom in.
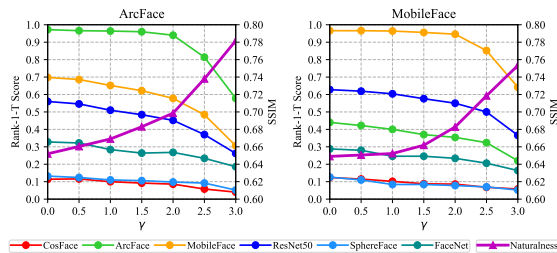

Figure 5. Rank-1-T score and SSIM of protected images generated by TIP-IM with different γ against different models.

of the generated protected images. Apart from quantitative measures, we also performed naturalness manipulation for different γ dynamically in Fig. 4. The image looks more natural as the γ increases, whereas to a certain extent identity protection tends to fail. We also perform a more general evaluation on all given recognition models in Fig. 5. As γ increases, SSIM values perform a general downward trend for Rank-1-T accuracy, meaning that appropriate γ is crucial for transferability and naturalness.

Practicability. Face encryption focuses on generating effective and natural adversarial identity masks, which cannot be realized by most previous adversarial attacks w.r.t. the effectiveness (in Tab. 3) and naturalness (in Fig. 3 and Fig. 4). In practical applications, users can adopt proposed TIP-IM to adjust γ to control stronger obfuscation performance (effectiveness) or visual quality (naturalness).

5.4. Effectiveness on a Real-World Application

In this section, we apply our proposed TIP-IM to test the identity protection performance on a commercial face search API available at Tencent AI Open Platform[6]. The working mechanism and training data are completely unknown for us. To simulate the privacy data scenario, we use the same gallery set described above. We choose 20 probe faces from above probe set to execute face search based on similarity ranking in this platform. All 20 probe faces can be identified at Rank1. Then we generate corresponding protected images from probe faces to execute face search.

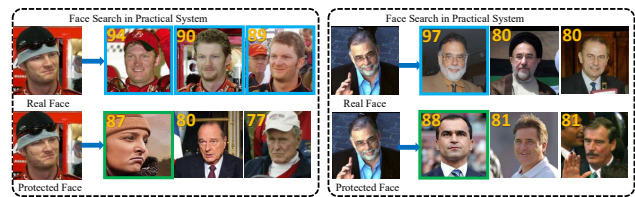[6] https://ai.qq.com/product/face.shtml

Figure 6. Examples of face encryption on the real-world face recognition API. We separately use real and protected faces by TIP-IM as probes to do face search and show top three results by similarity. Blue boxes represent the faces with same identities as probe faces and green boxes imply the faces belonging to targeted identities. Similarity scores with probe face are marked in yellow.

For return rankings there exists 6 target identities in rank1 and 16 in rank5. Note that those faces with the same identity also show a decreasing similarity in different degrees, which also illustrates the effectiveness for black-box face system, and two examples are shown in Fig. 6.

6. Conclusion

In this paper, we studied the problem of identity protection by simulating realistic identification systems in the social media. Extensive experiments show that proposed TIP-IM method enables users to protect their private information from being exposed by the unauthorized identification systems while not affecting the user experience in social media.

Acknowledgements

References

[1] 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

[2] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 273–282, 2007. 3

[3] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, 2005. 3

[4] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 4

[5] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobile-facenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018. 6

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1, 6

[7] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness. *arXiv preprint arXiv:1912.11852*, 2019. 3

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5, 6, 7

[9] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[10] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[11] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008. 3

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006. 3

[13] Uriel Feige, Vahab S Mirrokni, and Jan Vondrák. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011. 5

[14] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9378–9387, 2019. 1

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 3

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 2, 3

[17] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016. 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, pages 630–645. Springer, 2016. 6

[19] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Technical report*, 2007. 6

[20] 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*, pages 8110–8119, 2020. 7

[21] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016. 6

[22] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshops*, 2017. 2, 5, 6

[23] Martha Larson, Zhuoran Liu, SFB Brugman, and Zhengyu Zhao. Pixel privacy. increasing image appeal while blocking automatic inference of sensitive scene information. 2018. 1

[24] Fengjun Li, Bo Luo, and Peng Liu. Secure information aggregation for smart grids using homomorphic encryption. In *2010 first IEEE international conference on smart grid communications*, pages 327–332. IEEE, 2010. 2

[25] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 3

[26] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 1, 2, 6

[27] Haiping Lu, Karl Martin, Francis Bui, Konstantinos N Plataniotis, and Dimitris Hatzinakos. Face recognition with biometric encryption for privacy-enhancing self-exclusion. In

2009 16th International Conference on Digital Signal Processing, pages 1–8. IEEE, 2009. 2

[28] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016. 1, 3

[29] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007. 3

[30] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978. 5

[31] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016. 1, 3

[32] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1491–1500. IEEE, 2017. 2, 3, 6

[33] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. Boosting adversarial training with hypersphere embedding. *arXiv preprint arXiv:2002.08619*, 2020. 3

[34] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, 47:131–151, 2016. 2

[35] José Marconi Rodrigues, William Puech, Peter Meuel, Jean-Claude Bajard, and Marc Chaumont. Face protection by fast selective encryption in a video. 2006. 1

[36] Andras Rozsa, Manuel Günther, and Terranee E Boult. Lots about attacking deep features. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 168–176. IEEE, 2017. 3, 6

[37] 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*, pages 815–823, 2015. 1, 6

[38] Ayon Sen, Xiaojin Zhu, Liam Marshall, and Robert Nowak. Should adversarial attacks use pixel p-norm? *arXiv preprint arXiv:1906.02439*, 2019. 2, 3

[39] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 1589–1604, 2020. 3

[40] 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 *ACM Sigsac Conference on Computer and Communications Security*, pages 1528–1540, 2016. 2, 3, 6

[41] Gurpreet Singh. A study of encryption algorithms (rsa, des, 3des and aes) for information security. *International Journal of Computer Applications*, 67(19), 2013. 2

[42] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6976–6987, 2019. 3

[43] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018. 1, 3

[44] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 553–569, 2018. 1, 2, 3

[45] 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. 2, 3

[46] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 1, 6

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[48] Michael J Wilber, Vitaly Shmatikov, and Serge Belongie. Can we still avoid automatic face detection? In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016. 1, 2, 3

[49] Yifan Wu, Fan Yang, and Haibin Ling. Privacy-protective-gan for face de-identification. *arXiv preprint arXiv:1806.08906*, 2018. 1, 3

[50] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6, 7

[51] Xiao Yang, Fangyun Wei, Hongyang Zhang, and Jun Zhu. Design and interpretation of universal adversarial patches in face detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 174–191. Springer, 2020. 3

[52] 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. 2

[53] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 6

[54] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 3

[55] Yuxun Zhou and Costas J Spanos. Causal meets submodular: Subset selection with directed information. In *Advances In*

Neural Information Processing Systems, pages 2649–2657, 2016. 5