Discrete Point-wise Attack Is Not Enough: Generalized Manifold Adversarial Attack for Face Recognition

Qian Li, Yuxiao Hu*, Ye Liu, Dongxiao Zhang, Xin Jin†, Yuntian Chen†
Eastern Institute for Advanced Study, Ningbo Zhejiang, China
{QianL01205, huyuxiao1205, liuye66a}@gmail.com
{dzhang, jinxin, ychen}@eias.ac.cn

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.01971

Abstract

Classical adversarial attacks for Face Recognition (FR) models typically generate discrete examples for target identity with a single state image. However, such paradigm of point-wise attack exhibits poor generalization against numerous unknown states of identity and can be easily defended. In this paper, by rethinking the inherent relationship between the face of target identity and its variants, we introduce a new pipeline of Generalized Manifold Adversarial Attack (GMAA)[1] to achieve a better attack performance by expanding the attack range. Specifically, this expansion lies on two aspects – GMAA not only expands the target to be attacked from one to many to encourage a good generalization ability for the generated adversarial examples, but it also expands the latter from discrete points to manifold by leveraging the domain knowledge that face expression change can be continuous, which enhances the attack effect as a data augmentation mechanism did. Moreover, we further design a dual supervision with local and global constraints as a minor contribution to improve the visual quality of the generated adversarial examples. We demonstrate the effectiveness of our method based on extensive experiments, and reveal that GMAA promises a semantic continuous adversarial space with a higher generalization ability and visual quality.

1. Introduction

Thanks to the rapid development of deep neural networks (DNNs), the face recognition (FR) networks [5, 27] has been applied to identity security identification systems in a variety of crucial applications, such as face unlocking and face payment on smart devices. However, it has been observed that DNNs are easily fooled by adversarial examples and offer incorrect assessments [9, 23], which has risks of unauthorized access to FR systems and stealing personal privacy through poisoned data. These 'well-designed' adversarial examples reveal the aspects of FR models that are vulnerable to be attacked, which makes the adversarial attack a meaningful work to provide reference for improving model robustness.

Following the point-wise paradigm, previous methods typically tend to attack a single target identity sample with discrete adversarial examples illustrated in Fig. 1. However, these methods are not strong enough both in target domain and adversarial domain. Concretely, for the target domain, attacking a single identity image has a poor generalization on those unseen faces (even if they belong to the same person) in realistic scenarios. For example, Fig. 2 shows these adversarial examples which were used to attack the *target* (a girl) have a disappointingly lower success rate of attacking the identity with other unseen states. We analyze that's because attacking a single image of target identity overfits the generation of adversarial examples to some fixed factors such as expression, makeup style, etc. In addition to the target domain, we naturally consider the weakness of adversarial domain in the existing methods. Most adversarial attack methods optimize a L_p bounded perturbation [24, 34] based on the gradient, which limits the problem to searching for discrete adversarial examples in a hypersphere of the clean sample and ignores the continuity of the generated adversarial domain. For example, the recently proposed methods based on makeup style transformation [16, 35, 37] focus on generating finite adversarial examples that correspond to discrete makeup references. Such methods all overemphasize on mining discrete adversarial examples within a limited scope and ignore the importance of the continuity in adversarial space. The weakness of current adversarial attack tasks motivates us, on the one hand, to explore how to generate adversarial examples that are more general to various target identity's states and, on the other hand, to upgrade adversarial domain from discrete points to continuous manifold for a stronger attack.

In this paper, we introduce a new paradigm dubbed Gen-

*Co-first author
†Corresponding author
[1] https://github.com/tokaka22/GMAA

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.01971

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:37:19 UTC from IEEE Xplore. Restrictions apply.

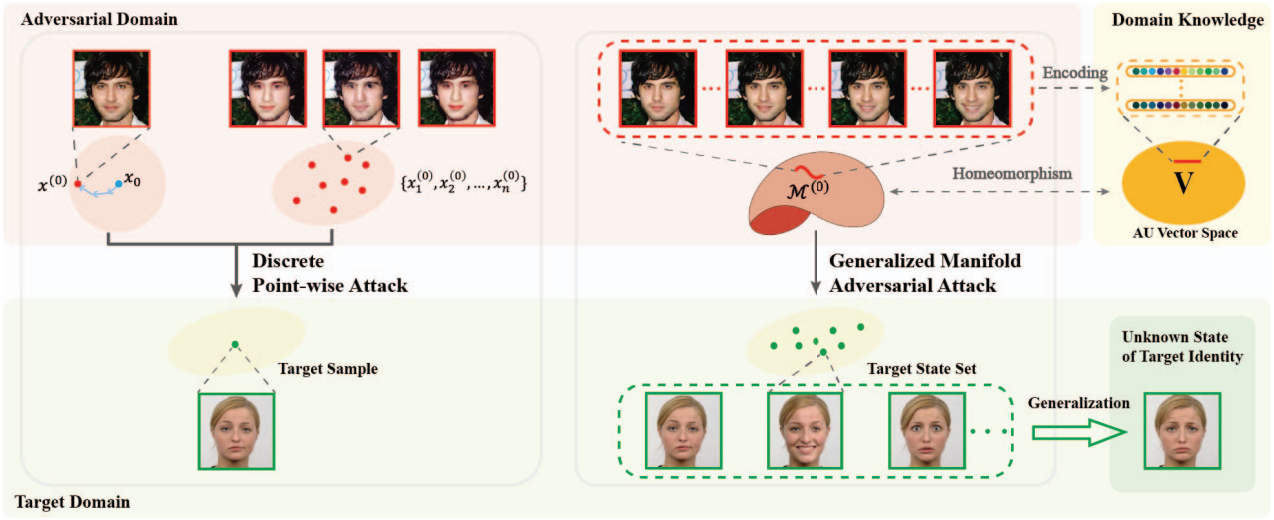

Figure 1. The core idea comparison. Discrete point-wise attack methods leverage a single state of the target identity during training and provide discrete adversarial examples. By attacking the target identity's state set and employing domain knowledge, our core idea aims to Generalized Manifold Adversarial Attack (GMAA), which promises a semantic continuous adversarial manifold with a higher generalization ability on the target identity with unknown state.

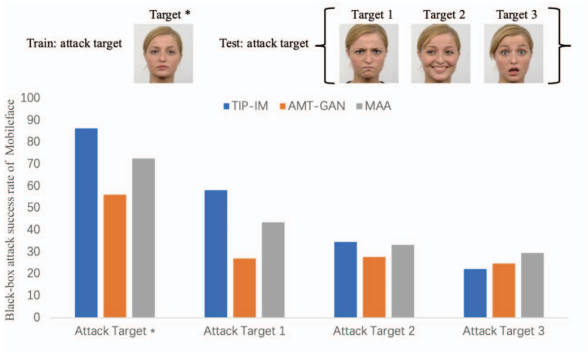

Figure 2. The black-box attack success rate on the Mobileface of attacking target * 1, 2 and 3 during the testing. The three methods are exclusively trained on target *.

eralized Manifold Adversarial Attack (GMAA) depicted in Fig. 1, which achieves a higher attack success rate by providing semantic continuous adversarial domains while expanding the target to be attacked from an instance to a group set. Specifically, we train adversarial examples to attack a target identity set rather than a single image, which increases the attack success rate on the target identity with different states. The expansion of target domain naturally prompts us to consider enhancing the adversarial domain. Inspired by the success of some recent *data and knowledge dual driven methods* [2, 18, 36], we explore a low-dimensional manifold near the sample according to the domain knowledge, which is a simple yet highly-efficient continuous embedding scheme and can be used to augment the data. For such manifold, the data in it share the visual identity same as the original sample and also lie in the decision boundaries of target identity. More specifically, we employ the Facial Action Coding System (FACS) [10] as prior domain knowledge (a kind of instantiation) to expand adversarial domain from discrete points to manifold. Through using FACS, the *expressions* can be encoded into a vector space, and by which the adversarial example generator could produce an manifold that is homogeneous with the expression vector space and possesses semantic continuity. In addition, in order to build an adversarial space with high visual quality, we employ four expression editors in GMAA pipeline to supervise the adversarial example generation in terms of global structure and local texture details. A transferability enhancement module is also introduced to drive the model to mine robust and transferable adversarial features. Extensive experiments have shown that these components work well on a wide range of baselines and black-box attack models.

Our contributions are summarized as follows.

- We first pinpoint that the popular adversarial attack methods generally face generalization difficulty caused by the limited point-wise attack mechanism.
- To enhance the performance of adversarial attack, a new paradigm of Generalized Manifold Adversarial Attack (GMAA) is proposed with an improved attack success rate and better generalization ability.
- GMAA considers the enhancement in terms of both target domain and adversarial domain. For the target domain, it expands the target to be attacked from one to many to encourage a good generalization. For the adversarial domain, the domain knowledge is embedded to strengthen the attack effect from discrete points to continuous manifold.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:37:19 UTC from IEEE Xplore. Restrictions apply.

- We instantiate GMAA in the face expression state space for a semantic continuous adversarial manifold and use it to attack a state set of the target identity. As a minor contribution, GMAA supervises the adversarial example generator w.r.t global structure and local details with the pre-trained expression editors for a high visual quality.

2. Related work

2.1. Adversarial Attacks on Face Recognition

Adversarial attacks can be divided into white-box attacks and black-box attacks. White-box attacks [11, 24, 32] must entail full information of the target neural networks, however, the parameters and architecture of a real-world face recognition system are typically hard to access. Hence, it is more practical to consider black-box adversarial attacks in face recognition scenarios, which demand high transferability for the unknown FR target models or commercial application programming interfaces (APIs). Query-based adversarial attacks [9, 13], a form of black-box attack, is inappropriate for realistic applications since it optimizes adversarial examples by repeatedly accessing the target model during training. Adding transferable adversarial perturbation is another effective black-box attack form [4, 7, 8, 34]. Unfortunately, the perturbation caused by this method always makes images become unrealistic and unnatural. Besides, limited by the way of pixel-to-pixel similarity computation, the model of this branch could only look for a limited number of discrete adversarial examples around the clean sample. The patch-based adversarial attack [20, 28, 33] severely degrades the visual quality of images, because the adversarial patch does not blend well with the clean background due to the abrupt shift in pixel values around the boundary. Numerous recent works attempt to instantiate adversarial perturbations with different makeup styles [16, 35, 37]. However, such makeup attacks typically result in an unnatural visual appearance due to gender constraints – female images have a higher attack success rate and visual quality than male ones. Regardless of the previous methods (e.g. [26] focuses on a target sample to obtain optimal feature-map interpolation, and [16] generates adversarial examples with various makeup styles to attack the target identity), they all tend to generate discrete adversarial examples for a single target identity sample and ignore the importance of continuity of adversarial space, which might be crucial to the attack performance.

2.2. Facial Expression Editing

As another challenging task in facial analysis, face expression editing is also related to our study, which aims at modifying facial expressions in a reasonable manner while preserving identity completeness. In recent years, generative adversarial networks (GANs) have achieved surprising advances in facial expression editing: GCGAN [29] uses the facial geometry as prior knowledge to guide the generation. ExprGAN [6] exploits a controller to adjust the intensity of face expression editing. StarGAN [3] introduces a cycle consistency loss to maintain the identity content invariant. However, these methods are all limited on the discrete expressions generation. GANimation [25] utilizes *Action Units* [10] to define an expression space and generate continuous-change facial images. EF-GAN [31] achieves progressive facial expression editing with a local-focused cascade GAN structure, and produces fewer artifacts and blurs in large-gap expression transformations. In this paper, we borrow the idea of *Action Units* as domain prior knowledge (from face expression editing) to re-define the adversarial attack task on a continuous manifold, to finally strengthen attack effect.

3. Generalized Manifold Adversarial Attack

3.1. Problem Definition

Adversarial face attack tasks can be separated into targeted attacks (i.e. impersonation attacks) and non-targeted attacks (i.e. dodging attacks). Targeted attacks force the generated adversarial examples to have a predetermined output towards the target FR model, while the non-targeted attacks mislead the target FR model to provide incorrect random classification for the adversarial examples. To capture the adversarial examples that can impersonate a specific target identity under face authentication systems, we mainly consider the targeted attack task.

Current methods always define the targeted attack task as an optimization problem, which can be formalized as

$$\begin{aligned} \min L_{adv} &= \min_{\boldsymbol{\theta}} Dist(\mathcal{C}(\boldsymbol{x}^*), \mathcal{C}(\boldsymbol{x}')), \\ \boldsymbol{x}' &= G(\boldsymbol{x}; \boldsymbol{\theta}). \end{aligned} \tag{1}$$

where $\boldsymbol{x}^*$ is the pre-specified target image belonging to the sample space $\Omega \subset \mathbb{R}^{3\times H\times W}$, $Dist(\cdot)$ represents a metric of difference, $\mathcal{C}$ represents the feature extractor of FR neural networks, and G maps the clean sample $\boldsymbol{x} \in \Omega$ to the adversarial version $\boldsymbol{x}'$ with the parameter $\boldsymbol{\theta}$.

In this paper, we re-define the problem from a broader standpoint. To obtain highly generalized adversarial examples that are more threatening to the target identity with unknown state, the adversarial version $\boldsymbol{x}'$ attacks the state set $\mathcal{S}$ of the target identity during training. In order to capture an adversarial manifold $\mathcal{M}$ instead of discrete adversarial examples, we aim to construct a distribution on the $\mathcal{M}$. The new task can be formalized as Eq. 2.

$$\begin{aligned} \min L_{adv} &= \min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}_i^* \sim \mathcal{S}, \boldsymbol{x}' \sim \mathcal{M}} Dist(\mathcal{C}(\boldsymbol{x}_i^*), \mathcal{C}(\boldsymbol{x}')), \\ \mathcal{M} &= G(\boldsymbol{x}; \boldsymbol{\theta}). \end{aligned} \tag{2}$$

In addition to the adversarial attack tasks, we force G to map $\boldsymbol{x}$ to $\boldsymbol{x}'$ according to the given expression AU vector. The Facial Action Coding System (FACS) is applied as

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:37:19 UTC from IEEE Xplore. Restrictions apply.

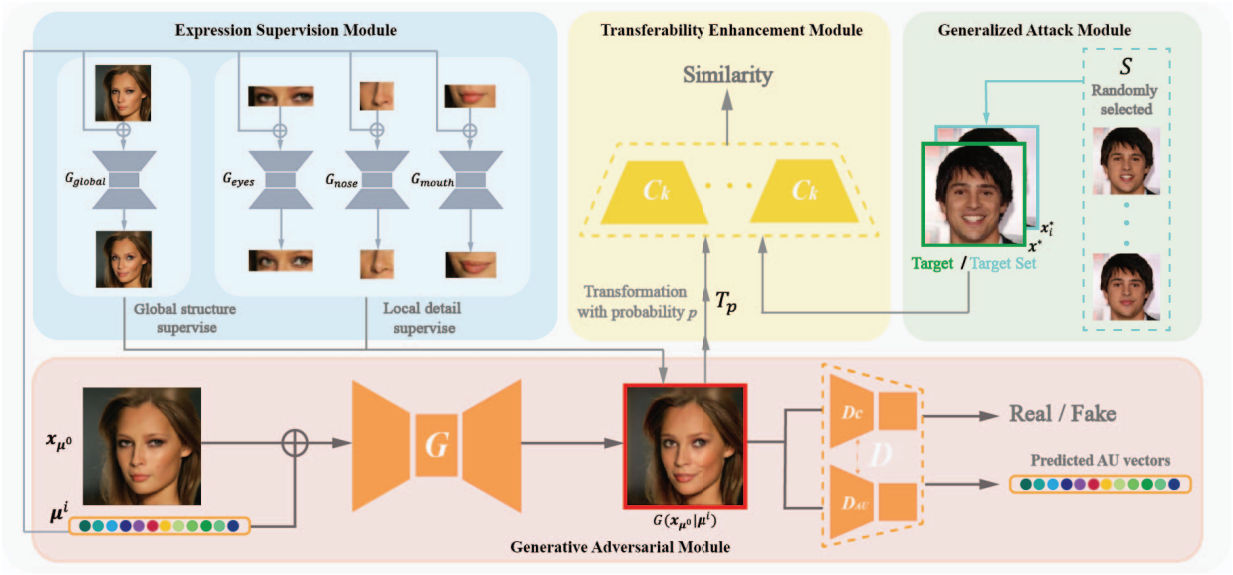

Figure 3. Overview of GMAA. A face image and a random AU vector are sent to the generative adversarial module as inputs. Meanwhile, the expression supervision module generates supervisory signals based on the AU vector and the input image (or detail patches cropped by the landmarks). In the transferability enhancement module, aiming at white-box FR models, the transformed output attacks the target identity's state set, which is provided by the generalized attack module.

prior domain knowledge to establish a continuous expression state space. Specifically, every expression is encoded by an n-element expression vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$ which corresponds to N facial action units. Each $\mu_n \in \boldsymbol{\mu}$ represents the magnitude of muscle activity in the n-th region of the face, which indicates that the AU vector is a continuous embedding scheme for different expressions. Thus, for the input image $\boldsymbol{x}_{\boldsymbol{\mu}^0} \in \mathbb{R}^{3\times H\times W}$ with the expression encoded by $\boldsymbol{\mu}^0 \in \mathbb{R}^N$ and the given expression encoded by $\boldsymbol{\mu}^i \in \mathbb{R}^N$, G is a binary mapping $G : (\boldsymbol{x}_{\boldsymbol{\mu}^0}, \boldsymbol{\mu}^i) \rightarrow \boldsymbol{x}'_{\boldsymbol{\mu}^i}$, where $\boldsymbol{x}'_{\boldsymbol{\mu}^i}$ has the same visual label with $\boldsymbol{x}_{\boldsymbol{\mu}^0}$ and wearing the expression encoded by $\boldsymbol{\mu}^i$.

3.2. Generalized Manifold Adversarial Attack

We establish the adversarial examples' distribution on a manifold via WGAN-GP [12]. GMAA comprises a generative adversarial module, an expression supervision module, a transferability enhancement module, and a generalized attack module. In contrast, Manifold Adversarial Attack (MAA) omits the generalized attack module, as it merely extends the adversarial domain from a point to a manifold. The structure of our proposed method is depicted in Fig. 3.

Generative adversarial module. The generative adversarial module (red box in Fig. 3) includes a generator G, a discriminator D_c and an AU predictor D_{AU}, where D_c and D_{AU} both lie in D and share partial parameters.

As inputs, the generator G receives a clean sample $\boldsymbol{x}_{\boldsymbol{\mu}^0}$ and a given expression AU label $\boldsymbol{\mu}^i$, which aims to produce adversarial examples wearing the expression matching to the supplied AU label and maintain the same visual identity with $\boldsymbol{x}_{\boldsymbol{\mu}^0}$. The discriminator D_c learns to distinguish real images from generated images. Meanwhile, generated images deceive the discriminator D_c to force the outputs of the generator G match the real distribution. Our G and D_c are trained using WGAN-GP [12], and the critic loss function we employ is

$$\begin{aligned} L^D_{critic} =& \lambda_c(1 - D_c(\boldsymbol{x}_{\boldsymbol{\mu}^0}))^2 + \lambda_c(D_c(G(\boldsymbol{x}_{\boldsymbol{\mu}^0}|\boldsymbol{\mu}^i))^2 \\ &+ \lambda_{gp}(\|\nabla_{\widetilde{\boldsymbol{x}}} D_c(\widetilde{\boldsymbol{x}})\|_2 - 1)^2, \end{aligned} \tag{3}$$

$$L^G_{critic} = \lambda_c(1 - D_c(G(\boldsymbol{x}_{\boldsymbol{\mu}^0}|\boldsymbol{\mu}^i)))^2, \tag{4}$$

where $\widetilde{\boldsymbol{x}}$ is the random interpolation distribution between the real distribution and the generated images' distribution. To ensure that the generated image match the provided expression code $\boldsymbol{\mu}^i$, we employ AU regression loss to establish the consistency of the generated expression with $\boldsymbol{\mu}^i$. Specifically, the AU predictor D_{AU} learns the AU coding rules by real images and their AU labels (can be obtained by the open source framework Openface [1]), and the G reduces the AU error between the generated expression and $\boldsymbol{\mu}^i$ to satisfy the given expression. The loss function can be formulated as:

$$\begin{aligned} L^D_{AU} &= \lambda_{AU}\|D_{AU}(\boldsymbol{x}_{\boldsymbol{\mu}^0}) - \boldsymbol{\mu}^0\|_2^2, \\ L^G_{AU} &= \lambda_{AU}\|D_{AU}(G(\boldsymbol{x}_{\boldsymbol{\mu}^0}|\boldsymbol{\mu}^i)) - \boldsymbol{\mu}^i\|_2^2. \end{aligned} \tag{5}$$

Expression supervision module. The expression supervision module (blue box in Fig. 3) protects the visual identity of adversarial examples and guides G in expression editing by generating global and local facial supervisory signals. The global branch focuses on structural features of the face, whereas the local branch protects important facial details and reduces artifacts and blurs caused by the global branch.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:37:19 UTC from IEEE Xplore. Restrictions apply.

Specifically, a global editor and three local editors are pre-trained to provide supervisory signals. For the local editors, we crop the eyes, nose, and mouth pixel patches based on face landmarks first. Then the input image and three detail patches are fed into the corresponding generator G_{global} and G_j ($j \in J$ ={eyes, nose, mouth}) with the input AU vector $\boldsymbol{\mu}$, respectively. Each generator has the network structure similar to [25], which provides the color response $M_c^{\boldsymbol{\mu}}$ and the attention response $M_a^{\boldsymbol{\mu}}$. $M_c^{\boldsymbol{\mu}}$ and $M_a^{\boldsymbol{\mu}}$ force networks to pay attention to the expression change region and protect the remainder regions from disturbance. The ultimate supervisory signals can be obtained as follows, where $\otimes$ denotes element-wise multiplication.

$$G_j(\boldsymbol{x}_{in}|\boldsymbol{\mu}) = M_a^{\boldsymbol{\mu}} \otimes M_c^{\boldsymbol{\mu}} + (1 - M_a^{\boldsymbol{\mu}}) \otimes \boldsymbol{x}_{in}. \quad (6)$$

The global editor focuses more on large scale features, such as shape and position of the five senses, and tends to produce artifacts and blurs in the detail region, while the local editors concentrate on significant local features and provide finer details. Therefore, we deploy the global editor to supervise the structural information of adversarial examples and the local editors to supervise local specifics. The loss of expression supervision module can be expressed as

$$\begin{aligned} L_{exp}^G &= \lambda_g \mathrm{SSIM}[G(\boldsymbol{x}_{\boldsymbol{\mu}^0}|\boldsymbol{\mu}^i), G_{global}(\boldsymbol{x}_{\boldsymbol{\mu}^0}|\boldsymbol{\mu}^i)] \\ &+ \lambda_l \sum_{j \in J} \mathrm{MSE}[\mathrm{Crop}_j(G(\boldsymbol{x}_{\boldsymbol{\mu}^0}|\boldsymbol{\mu}^i)), G_j(\mathrm{Crop}_j(\boldsymbol{x}_{\boldsymbol{\mu}^0})|\boldsymbol{\mu}^i)], \end{aligned} \quad (7)$$

where Crop_j denotes the crop operation of local region j according to the face landmarks.

Transferability enhancement module. To improve the transferability of adversarial examples and the black-box attack success rate, we introduce the transferability enhancement module (yellow box in Fig. 3) from [16]. The generated adversarial examples are transformed with probability p by the function T_p (resize with padding or add noise), and then attack K pre-trained high-precision FR models $\{C_k\}_{k=1,...K}$, which perform as white-box models during the training of G. The adversarial attack loss function is formulated as

$$L_{adv}^G = \frac{\lambda_{adv}}{K} \sum_{k=1}^{K} [1 - \cos(\mathcal{C}_k(\boldsymbol{x}^*), \mathcal{C}_k(T_p(G(\boldsymbol{x}_{\boldsymbol{\mu}^0}|\boldsymbol{\mu}^i))))] \quad (8)$$

Generalized attack module. The generalized attack module (green box in Fig. 3) intends to raise the attack success rate on the unseen face belonging to the target identity, which can be introduced into other adversarial attack approaches. The adversarial loss L_{adv}^G in GMAA can be further expressed as

$$\mathbb{E}_{\boldsymbol{x}_i^* \sim \mathcal{S}} \frac{\lambda_{adv}}{K} \sum_{k=1}^{K} [1 - \cos(\mathcal{C}_k(\boldsymbol{x}_i^*), \mathcal{C}_k(T_p(G(\boldsymbol{x}_{\boldsymbol{\mu}^0}|\boldsymbol{\mu}^i))))], \quad (9)$$

where $\boldsymbol{x}_i^*$ and $\mathcal{S}$ are defined in 2. However, numerous face recognition datasets and realistic scenarios do not fit this module since an identity only contain a few or single state image. Fortunately, since our method can accomplish both expression editing and adversarial attack, when we remove the loss associated with adversarial attack, we can obtain an expression editor, G_{exp}, by eliminating the adversarial effect L_{adv}^G, which can generate the expression state set $\mathcal{S}$ by different AU vectors.

Total loss function. Let $\boldsymbol{X}$ denotes the dataset, and V is the AU vector space. In particular, the given AU vector $\boldsymbol{\mu}^i$ is sampled randomly from V to train the generator G to learn the distribution of adversarial expression manifold. For the generator G, we have the loss function as follows,

$$L^G = \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{\mu}^0} \sim \boldsymbol{X}, \boldsymbol{\mu}^i \sim V}(L_{critic}^G + L_{AU}^G + L_{exp}^G + L_{adv}^G). \quad (10)$$

As for D, the total loss fuction can be obtained as follows,

$$L^D = \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{\mu}^0} \sim \boldsymbol{X}, \boldsymbol{\mu}^i \sim V}(L_{critic}^D + L_{AU}^D). \quad (11)$$

3.3. Continuity of the adversarial space

In this subsection, we illustrate more precisely the continuity of the adversarial space and provide a proof that our method establishes a continuous adversarial manifold.

Firstly, the definition of continuous adversarial space is shown in Def. 1.

Definition 1. *Let $\boldsymbol{x}_0 \in \mathbb{R}^{3\times H\times W}$, then $\mathcal{M}^0 = G(\boldsymbol{x}_0; \boldsymbol{\theta})$ is a continuous adversarial space if and only if*
(1) $\mathcal{M}^0$ is a subspace of $\mathbb{R}^{3\times H\times W}$.
(2) $\forall \boldsymbol{x}_i^0 \in \mathcal{M}$, $\boldsymbol{x}_i^0$ is an adversarial version of $\boldsymbol{x}_0$.

Then, we can prove that $\mathcal{M}^0$ is a continuous manifold.

Theorem 1. *$\mathcal{M}^0$ generated by G_0 is a continuous adversarial manifold, where $G_0 : V \to \mathcal{M}$ is a map when fixed the input $\boldsymbol{x}_0$ in G.*

Proof. (1) If $\mathcal{M}^0$ homogeneous with the AU vector space V, it is obviously that $\mathcal{M}^0$ is a subspace of $\mathbb{R}^{3\times H\times W}$.

(1.1) $\forall\ \boldsymbol{\mu},\ \boldsymbol{\nu} \in V$, if $G_0(\boldsymbol{\mu}) = G_0(\boldsymbol{\nu})$, then $D_{AU}(G_0(\boldsymbol{\mu})) = D_{AU}(G_0(\boldsymbol{\nu}))$, we have $\boldsymbol{\mu} = \boldsymbol{\nu}$ and $G_0 : V \to \mathcal{M}^0$ is a single shot. Besides, $\forall \boldsymbol{x}^0 \in \mathcal{M}^0$, $\exists \boldsymbol{\mu} = D_{AU}(\boldsymbol{x}^0) \in V$ s.t. $G_0(D_{AU}(\boldsymbol{x}^0)) = \boldsymbol{x}^0$, then G_0 is a surjection. Thus, G_0 is a bijection.

(1.2) $\forall \boldsymbol{x}_1^0,\ \boldsymbol{x}_2^0 \in \mathcal{M}^0$, we define $d(\boldsymbol{x}_1^0, \boldsymbol{x}_2^0) = \|D_{AU}(\boldsymbol{x}_1^0) - D_{AU}(\boldsymbol{x}_2^0)\|_2$ as the metric between $\boldsymbol{x}_1^0$ and $\boldsymbol{x}_2^0$ in $\mathcal{M}^0$. Since $(V, \|\cdot\|_2)$ is a metric space, and $D_{AU}(\boldsymbol{x}_i^0) \in V$, $\forall \boldsymbol{x}_i^0 \in \mathcal{M}^0$, we prove that d is a metric on $\mathcal{M}^0$.

(Positivity) $d(\boldsymbol{x}_1^0, \boldsymbol{x}_2^0) \geq 0$, and if $d(\boldsymbol{x}_1^0, \boldsymbol{x}_2^0) = 0$, according to the definition of d we have $D_{AU}(\boldsymbol{x}_1^0) = D_{AU}(\boldsymbol{x}_2^0)$. Since G_0 is a bijection, we have $\boldsymbol{x}_1^0 = G_0(D_{AU}(\boldsymbol{x}_1^0)) = G_0(D_{AU}(\boldsymbol{x}_2^0)) = \boldsymbol{x}_2^0$.

(Symmetry) $d(\boldsymbol{x}_1^0, \boldsymbol{x}_2^0) = \|D_{AU}(\boldsymbol{x}_1^0) - D_{AU}(\boldsymbol{x}_2^0)\|_2 = \|D_{AU}(\boldsymbol{x}_2^0) - D_{AU}(\boldsymbol{x}_1^0)\|_2 = d(\boldsymbol{x}_2^0, \boldsymbol{x}_1^0)$.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:37:19 UTC from IEEE Xplore. Restrictions apply.

Table 1. Black-box attack success rate

	CelebA-HQ				LFW			
	IRSE50	IR152	Facenet	Mobileface	IRSE50	IR152	Facenet	Mobileface
Clean	3.68	3.08	1.31	8.43	3.20	0.06	0.04	5.00
PGD [23]	24.20	13.37	5.86	28.72	31.30	10.20	7.40	33.50
MI-FGSM [7]	38.90	20.76	9.25	40.48	38.20	14.20	7.60	39.40
SemanticAdv [26]	26.53	10.24	7.80	55.32	33.60	10.40	8.80	37.40
TIP-IM [34]	44.20	16.09	14.46	**65.36**	32.80	15.20	13.00	**79.00**
AMT-GAN [16]	51.06	15.63	11.63	33.27	40.72	25.23	13.89	35.67
MAA	**60.40**	**29.43**	**18.91**	56.13	**55.80**	**29.20**	**18.00**	60.80

Table 2. The unbolded numbers are the black-box ASR of attacking the test target *, 1, 2, 3, and all the models are trained on target *. The bolded numbers are the results of the models that train on the state set. For example, if we train the MAA to attack the state set without target 1, the adversarial examples attack target 1 with an 11.43% success rate on Facenet during the testing period.

	Target*		Target 1		Target 2		Target 3	
	Facenet	Mobileface	Facenet	Mobileface	Facenet	Mobileface	Facenet	Mobileface
TIP-IM [34] / **G-TIP-IM**	17.68	86.33	4.54 / **7.62**	58.03 / **70.93**	10.75 / **20.42**	34.42 / **49.20**	11.93 / **19.41**	22.21 / **42.43**
AMT-GAN [16] / **G-AMT-GAN**	16.12	55.95	8.22 / **13.23**	26.99 / **47.14**	9.78 / **17.12**	27.67 / **43.93**	10.91 / **16.16**	24.69 / **42.37**
MAA / **GMAA**	25.22	72.62	11.43 / **17.84**	43.44 / **67.50**	13.30 / **21.71**	33.08 / **41.24**	12.64 / **19.15**	29.56 / **47.21**

(Triangle inequality) $\forall \boldsymbol{x}_3^0 \in \mathcal{M}^0$, we have $d(\boldsymbol{x}_1^0, \boldsymbol{x}_3^0) + d(\boldsymbol{x}_3^0, \boldsymbol{x}_2^0) = \|D_{AU}(\boldsymbol{x}_1^0) - D_{AU}(\boldsymbol{x}_3^0)\|_2 + \|D_{AU}(\boldsymbol{x}_3^0) - D_{AU}(\boldsymbol{x}_2^0)\|_2 \geq \|D_{AU}(\boldsymbol{x}_1^0) - D_{AU}(\boldsymbol{x}_2^0)\|_2 = d(\boldsymbol{x}_1^0, \boldsymbol{x}_2^0)$.

Thus, $(\mathcal{M}^0, d)$ is a metric space. Then we need to prove that G_0 is a continuous mapping. $\forall \boldsymbol{\mu}, \boldsymbol{\nu} \in V$, $\forall \epsilon > 0$, let $\delta = \epsilon$, when $\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2 < \delta$, we have $d(G_0(\boldsymbol{\mu}) - G_0(\boldsymbol{\nu})) = \|D_{AU}(G_0(\boldsymbol{\mu})) - D_{AU}(G_0(\boldsymbol{\nu}))\|_2 = \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2 < \delta = \epsilon$. We get the continuity of G_0.

(1.3) It is obviously that $D_{AU}|_{\mathcal{M}^0} : \mathcal{M}^0 \to V$ is the inverse mapping of G_0. $\forall \boldsymbol{x}_1^0, \boldsymbol{x}_2^0 \in \mathcal{M}^0, \epsilon > 0$, let $\delta = \epsilon$, when $d(\boldsymbol{x}_1^0, \boldsymbol{x}_2^0) = \|D_{AU}(\boldsymbol{x}_1^0) - D_{AU}(\boldsymbol{x}_2^0)\|_2 < \delta$, we have $\|D_{AU}|_{\mathcal{M}^0}(\boldsymbol{x}_1^0) - D_{AU}|_{\mathcal{M}^0}(\boldsymbol{x}_2^0)\|_2 = \|D_{AU}(\boldsymbol{x}_1^0) - D_{AU}(\boldsymbol{x}_2^0)\|_2 < \delta = \epsilon$. We get the continuity of the inverse map of G_0.

In conclusion, G_0 is the homeomorphism of V to the manifold $\mathcal{M}^0$, i.e. $\mathcal{M}^0$ homogeneous with the AU vector space V. Since AU vector space is a finite dimensional vector space, $\mathcal{M}^0$ is a subspace of $\mathbb{R}^{m \times n \times l}$.

(2) By the loss function 8, 7 and the back propagation, $\boldsymbol{x}_i^0$ is influenced by the 8. It is obviously that $\boldsymbol{x}_i^0 \in \mathcal{M}^0$ is an adversarial example when the model is well-trained.

Thus, the manifold $\mathcal{M}^0$ generated by G_0 is a continuous adversarial manifold. □

Remark 1. *Since the $\mathcal{M}^0$ generated by G_0 is a continuous adversarial manifold when fixed the $\boldsymbol{x}_0$, then we can assert over the sample space Ω, the adversarial examples space generated by G constitutes an adversarial fiber bundle.*

Secondly, the definition of semantic continuous adversarial space is shown in Def. 2.

Definition 2. *$\mathcal{M}^0$ generated by $\boldsymbol{x}_0 \in \mathbb{R}^{3 \times H \times W}$ is a semantic continuous adversarial space if and only if*
(1) $\mathcal{M}^0$ is a continuous adversarial space.
(2) $\forall \boldsymbol{x}_1^0, \boldsymbol{x}_2^0 \in \mathcal{M}^0$, if $\boldsymbol{x}_1^0$ is close to $\boldsymbol{x}_2^0$ on the $\mathcal{M}^0$, then $\boldsymbol{x}_1^0$ and $\boldsymbol{x}_2^0$ satisfy the semantic consistency.

We can state that $\mathcal{M}^0$ is a semantic continuous manifold.

Theorem 2. *$\mathcal{M}^0$ generated by G_0 is a semantic continuous adversarial manifold, where $G_0 : V \to \mathcal{M}$ is a map when fixed the input $\boldsymbol{x}_0$ in G.*

Proof. Since we have proved that D_{AD} is a continuous mapping, we have $D_{AU}(\boldsymbol{x}_1^0)$ is close to $D_{AU}(\boldsymbol{x}_2^0)$ when $\boldsymbol{x}_1^0$ is close to $\boldsymbol{x}_2^0$, which means the AU vectors of $\boldsymbol{x}_1^0$ and $\boldsymbol{x}_2^0$ are very close. Thus, the semantic information of $\boldsymbol{x}_1^0$ and $\boldsymbol{x}_2^0$ is close. □

4. Experiments

4.1. Experimental setting

Implementation details. We set λ_c, λ_{gp}, λ_{AU}, λ_g, λ_l, λ_{adv} to be 1, 10, 250, 20, 20, 25, respectively. Our method is trained by an Adam optimizer with the learning rate 0.0001 and the exponential decay rates set to be $(\beta_1, \beta_2) = (0.5, 0.99)$. We evaluate the black-box attack performance of the models utilizing the *attack success rate*(ASR) at FAR@0.01 and the confidence scores returned by commercial APIs.

Dataset. We train the model on two public datasets: 1) CelebA-HQ [19] is a high-quality face dataset, which contains 30,000 face images. 2) LFW [17] is a challenging dataset that collects 13,233 images with complex environmental factors and is a common dataset for face recognition tasks. We remove the images whose AU confidence is below 95% as extracted by Openface [1], then randomly select 10% of each dataset as the test set and the remaining images as the training set. Four pairs of images from CelebA [22] with the same identity are used as attack targets for training and testing, respectively, since CelebA [22] contains multiple images of one identity. Besides, the real state set in subsection 4.3 are obtained from the RaFD dataset [21].

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:37:19 UTC from IEEE Xplore. Restrictions apply.

Table 3. The effect of the image quantity in state set on improving the generalizability of the adversarial example. The values n_i in $n_1/n_2/n_3$ represents each expression state using i images.

	Target 1		Target 2		Target 3	
	Facenet	Mobileface	Facenet	Mobileface	Facenet	Mobileface
G-TIP-IM [34]	7.6 / 7.1 / 7.2	70.9 / 70.3 / 70.1	20.4 / 21.5 / 21.4	49.2 / 49.2 / 49.2	19.4 / 19.6 / 19.4	39.4 / 39.6 / 39.5
G-AMT-GAN [16]	13.2 / 14.4 / 13.9	47.1 / 44.8 / 45.3	17.1 / 16.9 / 18.9	43.9 / 45.4 / 43.8	16.2 / 15.4 / 15.7	42.4 / 44.6 / 45.4
GMAA	17.8 / 18.2 / 17.7	67.5 / 68.5 / 69.3	21.7 / 21.3 / 19.6	41.2 / 42.6 / 40.3	19.2 / 19.9 / 21.3	47.2 / 45.6 / 44.7

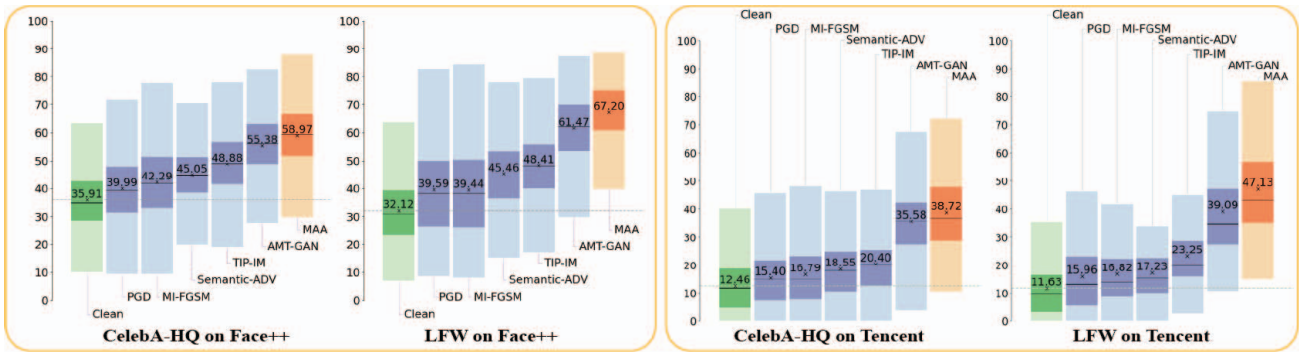

Figure 4. The confidence scores returned by Face++ and Tencent. The dashed line represents the average confidence level of clean samples.

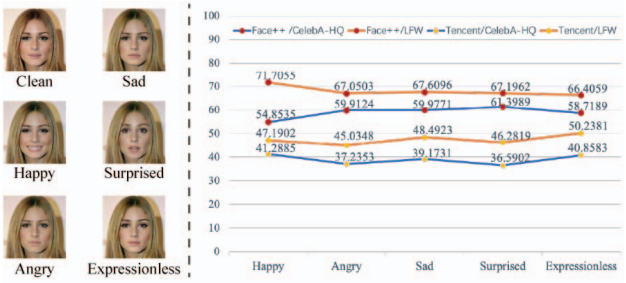

Figure 5. The left side of the image depicts the five expression states, while the right side of the image depicts the influence of varied AU on the attack performance of Face++ and Tencent.

Table 4. This table shows the black-box ASR results of attacking the test target, which is the same person as the train image highlighted by a green square in Fig. 3. The unbolded numbers represent the results of training on the single target image that is shown in Fig. 3, while the bolded numbers are the results of training on the generated state set.

	Facenet	Mobileface
TIP-IM [34] / **G-TIP-IM**	5.80 / **9.50**	17.20 / **23.5**
AMT-GAN [16] / **G-AMT-GAN**	4.04 / **8.27**	9.82 / **12.45**
MAA / **GMAA**	6.60 / **10.60**	13.50 / **21.60**

Competitors. We compare our approach to the baselines PGD [23], MI-FGSM [7], SemanticAdv [26], TIP-IM [34] and AMT-GAN [16]. Since our work belongs to the branch of GAN based unrestricted adversarial attack [30], which is budget-free. Similar to [30], we compare our method MAA/GMAA to both restricted and unrestricted adversarial methods w.r.t attack performance and visual naturalness. All restricted methods are setted to the size of perturbation $\epsilon = 12$. And all baselines are equipped with the transferability enhancement module for a fair comparison.

Target models. Following [16], we choose IR152 [14], IRSE50 [15], Facenet [27] and Mobileface [5] as the attacked FR models, with three of them serving as white-box models during training and the remaining as the black-box model for testing.

4.2. Comparison Study

This subsection shows the comparison results of the MAA method and competitors in terms of attack performance and visual quality.

Comparison of attack performance on commercial API. We evaluate the performance of each method against the commercial APIs Face++[2] and Tencent[3]. Fig. 4 exhibits the average confidence score of Face++ and Tencent between the adversarial example and the test image of the target identity. Our method MAA achieves the highest score, outperforming all competitors on both datasets, as shown in Fig. 4. Furthermore, in Fig. 5, we display the confidence scores for five typical expressions (happy, angry, sad, surprised, and expressionless) to demonstrate that different AU vectors have little effect on attack performance and are more influenced by the dataset.

Comparison of black-box attack success rate. Since our model can give us a semantic continuous adversarial manifold and we want to make sure the comparison is fair, we randomly sample the AU vector to get test adversarial examples for calculating the black-box ASR. Tab. 1 shows the

[2] https://www.faceplusplus.com/

[3] https://cloud.tencent.com/document/product/867/44987

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:37:19 UTC from IEEE Xplore. Restrictions apply.

Figure 6. The images with green frames are the clean samples, while the images with blue frames are the results of TIP-IM [34]. In the case of the AMT-GAN [16], we chose 8 makeup styles at random, and the visualization results are shown as images with an orange frame. The images highlighted by red frames are the results of MAA, which are generated by a set of AU vectors. Please refer to the supplementary material for more high-definition magnified visualization results.

black-box ASR of each method under four FR models and two datasets. Obviously, our method has good performance for black-box attacks, i.e., it has strong transferability.

Comparison of visual quality. We choose TIP-IM and AMT-GAN, two recent approaches with high black-box ASR, as benchmarks for our assessment of visual quality. Fig. 6 shows the adversarial examples of each method, and the target image is shown in Fig. 3 highlighted by a green square. In particular, to demonstrate that our method can generate semantically continuous adversarial examples, Fig. 6 displays the adversarial examples generated by MAA that continuously transform on four expressions (expressionless, disgusted, happy and surprised in succession). Note that although only 20 adversarial examples are presented in Fig. 6, our method can generate an infinite number of adversarial examples by continuously interpolating between AU vectors since MAA establishes a correspondence with the AU vector space. Moreover, our method has a natural visual quality and is gender-insensitive.

4.3. Attack state set

Attack real state set. To avoid serendipity, two FR models, three different test targets, and three adversarial attack methods were employed to assess the effectiveness of attacking state set on enhancing the adversarial examples' generalizability. Particularly, one of the FR models is Facenet with high accuracy, and the other is Mobileface with a lightweight network. Targets *, 1, 2 and 3 are shown in Fig. 2. The state set consists of several common expression states (angry, contemptuous, disgusted, fearful, happy, sad and surprised), and each correlates to an image. By comparing the results in Tab. 2, we can summarize that the model trained to attack target * generalizes poorly to test targets 1, 2, 3, whereas adversarial examples generated on the state set generalize better to the test target, even though the test target is not used to train the model.

We try to add more images in the state set to further improve the performance. Tab. 3 shows the ASR results that the target state set contains each expression state corresponding to 1/ 2/ 3 image(s), respectively. We demonstrate that the generalizability of the adversarial examples can be effectively strengthened as long as the state set contains a single image of each state, and additional images have minimal impact on the results.

Attack generated state set. For face image datasets that do not contain rich states, we generate the target state set by the pre-trained generator G_{exp} with AU vectors of common expressions. Tab. 4 shows the comparison results of training on a single target and a generated state set, demonstrating that training the model on the target state set enhances the generalization ability of adversarial examples.

5. Future Work

Considering the ubiquitous applicability of expressions in the facial adversarial attacks, the expression state space was chosen to implement GMAA in this paper. We employ the FACS as the prior domain knowledge to implement GMAA, while the paradigm GMAA can be broadly generalized by integrating other domain information, such as illumination, posture, etc. By selecting different state spaces, our work can be generalized to other adversarial attacks with more general image categories.

6. Conclusion

In this paper, we provide a novel paradigm GMAA that broadens both target domain and adversarial domain to enhance the performance of adversarial attack. For the target domain, GMAA optimizes generalization to the target identity by attacking the state set instead of a single image. Additionally, GMAA leverages the domain knowledge to expand adversarial domain from discrete points to semantic continuous manifold. Numerous comparative experiments have verified that GMAA has a better attack performance and a more natural visual quality than other competitors. Moreover, the generalized attack module can be extended to a wide of applications.

7. Acknowledgement

This work was supported and partially funded by the National Natural Science Foundation of China (Grant No. 62106116). This work was also supported in part by ZJNSFC under Grant LQ23F010008. We also would like to thank Han Fang of NUS for the valuable discussion.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:37:19 UTC from IEEE Xplore. Restrictions apply.

References

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018. 4, 6

[2] Yuntian Chen and Dongxiao Zhang. Integration of knowledge and data in machine learning. *arXiv preprint arXiv:2202.10337*, 2022. 2

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 3

[4] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020. 3

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1, 7

[6] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. *CoRR*, abs/1709.03842, 2017. 3

[7] 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*, pages 9185–9193, 2018. 3, 6, 7

[8] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 3

[9] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019. 1, 3

[10] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 2, 3

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[12] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 4

[13] Ying Guo, Xingxing Wei, Guoqiu Wang, and Bo Zhang. Meaningful adversarial stickers for face recognition in physical world. *CoRR*, abs/2104.06728, 2021. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 7

[16] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15014–15023, 2022. 1, 3, 5, 6, 7, 8

[17] 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 *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 6

[18] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. 2

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[20] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826. IEEE, 2021. 3

[21] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010. 6

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 6

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 6, 7

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 3

[25] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 3, 5

[26] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *European Conference on Computer Vision*, pages 19–37. Springer, 2020. 3, 6, 7

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:37:19 UTC from IEEE Xplore. Restrictions apply.

[27] 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, 7

[28] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 3

[29] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 627–635, 2018. 3

[30] Yang Song and et al. Constructing unrestricted adversarial examples with generative models. *NIPS*, 2018. 7

[31] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5021–5030, 2020. 3

[32] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 3

[33] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11845–11854, 2021. 3

[34] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3897–3907, 2021. 1, 3, 6, 7, 8

[35] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. *arXiv preprint arXiv:2105.03162*, 2021. 1, 3

[36] Su Hang. Zhang Bo, Zhu Jun. Toward the third generation of artificial intelligence (in chinese). *Scientia Sinica Informationis*, 50(9):1281–1302, 2020. 2

[37] Zheng-An Zhu, Yun-Zhong Lu, and Chen-Kuo Chiang. Generating adversarial examples by makeup attacks on face recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2516–2520. IEEE, 2019. 1, 3

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:37:19 UTC from IEEE Xplore. Restrictions apply.