Transferable Adversarial Face Attack with Text Controlled Attribute

Wenyun Li[1,2], **Zheng Zhang**[1,2] *, **Xiangyuan Lan**[2,3] *, **Dongmei Jiang**[2]

[1]Harbin Institute of Technology, Shenzhen, China
[2]Pengcheng Laboratory , Shenzhen, China
[3]Pazhou Laboratory (Huangpu), Guangzhou, China
liwy@pcl.ac.cn, darrenzz219@gmail.com, lanxy@pcl.ac.cn, jiangdm@pcl.ac.cn

Abstract

Traditional adversarial attacks typically produce adversarial examples under norm-constrained conditions, whereas unrestricted adversarial examples are free-form with semantically meaningful perturbations. Current unrestricted adversarial impersonation attacks exhibit limited control over adversarial face attributes and often suffer from low transferability. In this paper, we propose a novel Text Controlled Attribute Attack (TCA^2) to generate photorealistic adversarial impersonation faces guided by natural language. Specifically, the category-level personal softmax vector is employed to precisely guide the impersonation attacks. Additionally, we propose both data and model augmentation strategies to achieve transferable attacks on unknown target models. Finally, a generative model, *i.e*, Style-GAN, is utilized to synthesize impersonated faces with desired attributes. Extensive experiments on two high-resolution face recognition datasets validate that our TCA^2 method can generate natural text-guided adversarial impersonation faces with high transferability. We also evaluate our method on real-world face recognition systems, *i.e*, Face++ and Aliyun, further demonstrating the practical potential of our approach.

Extended version — https://arxiv.org/abs/2412.11735

Introduction

Recent studies have shown that deep learning-based face recognition (FR) model systems are vulnerable to adversarial examples (Vakhshiteh, Nickabadi, and Ramachandra 2021; Dong et al. 2019; Zhang et al. 2022). Adding deliberately designed but imperceptible noise to a clean image can fool even state-of-the-art commercial FR models (Ali et al. 2021). This vulnerability poses a direct threat to socially critical applications such as customs inspection and mobile device face identification. Consequently, the security community has increasingly focused on studying adversarial examples to improve the robustness and generalization ability of existing FR systems.

Early well-studied works focus on norm-constrained attacks, where the adversarial image lies within an ϵ-neighborhood of a real sample using the L_p distance metric to evaluate the strength of the adversarial example (Szegedy et al. 2013; Wang et al. 2021). Common values for p include 0, 2, and ∞. With a sufficiently small ϵ, the adversarial image is quasi-indistinguishable from the natural sample. Such norm-based attacks have demonstrated outstanding adversarial performance against face recognition (FR) systems. However, there are limitations: 1) Despite being designed to be indistinguishable, norm-based attacks may still contain visible perturbations, making them detectable by human eyes or specially designed detectors (Massoli et al. 2021); 2) Numerous adversarial defense methods have been introduced to counter norm-based attacks, leading to a relatively low attack success rate against real-world FR models (Madry et al. 2017).

Recently proposed unrestricted adversarial attack (UAA) methods generate adversarial images with more stealthy and semantically meaningful perturbations compared to noise-based adversarial attacks. Some unrestricted adversarial attacks hide the adversarial perturbations as decorative accessories like glasses (Sharif et al. 2016) or hat (Komkov and Petiushko 2021) to improve stealthiness. Notably, makeup-based adversarial attacks (Yin et al. 2021) generate perturbations as natural makeup. Moreover, (Li et al. 2021) generate more natural face images than previous methods with the help of pre-trained GANs. Adv-Attribute (Jia et al. 2022) automatically injects a pre-defined pattern from a target image to complete an adversarial edit. Adv-Diffusion (Liu et al. 2024) extracts latent codes from both source and target images, then exploits a diffusion model to generate adversarial faces. Although these UAA methods demonstrate improved stealthiness, they have limited ability to change attributes and primarily edit a pre-defined set of semantic information from the target image. We argue that it is essential to rapidly generate adversarial images with specific attributes, such as skin color, expression, or hairstyle, guided by attribute text. Such adversarial attacks can help security researchers expose vulnerabilities in existing FR models due to changes in facial attributes that are likely to occur.

Notwithstanding their effectiveness in attacking face recognition (FR) systems, these adversarial attacks have significant limitations. Although some previous works achieve relative transferability in black-box scenarios, they still struggle to attack FR models in real-world scenarios. Specifically, works such as Adv-Attribute (Jia et al. 2022) and

*Corresponding authors.

Figure 1: The visualization of source face, other different adversarial face and target face. The second image is from GAN inversion. Some representations of norm-based adversarial examples(Madry et al. 2017) and unrestricted adversarial examples(Komkov and Petiushko 2021) are shown. TCA^2 generates the 5^{th} image.

Adv-Diffusion (Liu et al. 2024) tend to generate impersonated faces optimized for a specific model, leading to overfitting to the source model. This limitation drives us to develop a more transferable attack that can generalize well to real-world FR systems.

To address the aforementioned shortcomings, this paper proposes a Text Controlled Attribute Attack (TCA^2) to generate adversarial perturbations guided by text prompts as shown in Figure 1. By feeding the targeted image into the FR model, a discriminative category-level softmax vector is produced to guide the impersonation attack. To enhance black-box transferability, we employ both data and model augmentation strategies. For data augmentation, we adopt simple random resizing and padding. For model augmentation, we apply a meta-learning paradigm to simulate white-box and black-box FR environments, further improving transferability. The framework of our TCA^2 is shown in Figure 2.

Our contributions can be summarized as follows:

1. We propose a novel Text Controlled Attribute Attack (TCA^2) to generate semantically meaningful perturbations guided by text prompts. Significantly, unlike existing unrestricted adversarial attacks (UAA), our TCA^2 offers rich face attribute editing capabilities under text guidance;
2. Both data and model augmentation techniques are employed to generate adversarial images that are more transferable to unknown black-box face recognition (FR) models;
3. Extensive experiments validate the superior effectiveness and transferability of our method compared to other state-of-the-art attack techniques on two high-resolution datasets.

Related Work

Norm-based Adversarial Examples

Many adversarial attack algorithms (Ryu, Park, and Choi 2021) have demonstrated that deep learning face recognition (FR) models are vulnerable to adversarial samples. Traditional adversarial examples against FR focus on norm-constrained conditions. For a given FR model $\mathcal{F}(x) : \mathcal{X} \rightarrow \mathbb{R}^d$ and a face image $x \in \mathbb{R}^n$, the adversarial image $\hat{x} \in \mathbb{R}^n$ satisfies the condition $|x - \hat{x}|\, p < \epsilon$ and $\mathcal{F}(x) \neq \mathcal{F}(\hat{x})$. Common values for p are 0, 2, and ∞, and ϵ is a sufficiently small value to ensure the perturbation is imperceptible. Adversarial attacks against FR models can be categorized as impersonation (targeted) attacks and dodging (untargeted) attacks based on whether their goal is to make the FR classify the adversarial face image as a specified $\hat{y}$ or any $\hat{y} \neq y$. Representative norm-based methods include the Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014), which uses a first-order approximation of the function for faster adversarial example generation, and Projected Gradient Descent (PGD) (Madry et al. 2017), an iterative variant of FGSM that provides a strong first-order attack through multiple steps of gradient ascent. Carlini and Wagner (C&W) (Carlini and Wagner 2017) proposed stronger optimization-based attacks for L_0, L_2, and $L\infty$ via improved objective functions. AdvGAN (Xiao et al. 2018) proposes a GAN network to efficiently generate adversarial examples. These methods can easily fool victim neural networks. However, norm-based adversarial examples can still be detected by humans or adversarial detectors (Massoli et al. 2021). Consequently, several defense mechanisms against such attacks have been proposed, such as adversarial training (Madry et al. 2017).

Unrestricted Adversarial Attack

Traditional adversarial perturbations are constrained by norm bounds, whereas unrestricted adversarial attacks (UAA) are not subject to such limitations. These attacks have been extensively studied in image classification tasks (Sharif et al. 2016; Brown et al. 2017; Karmon, Zoran, and Goldberg 2018). UAA generates adversarial images with semantically meaningful perturbations compared to noise-based adversarial attacks. Some UAAs have been proposed by generating adversarial wearable accessories like glasses (Sharif et al. 2016) or hat(Komkov and Petiushko 2021) to fool the FR model. However, such colorful patches are easily noticeable, leading to poor stealthiness. Makeup-based UAA (Yin et al. 2021; Guetta et al. 2021) are developed against FR models by generating perturbations in the form of makeup, but these generated makeups still appear unnatural to humans. More recently, some attacks (Li et al. 2021) have been introduced using pre-trained generative models. These works demonstrate an excellent ability to generate adversarial examples containing fewer artifacts compared to previous

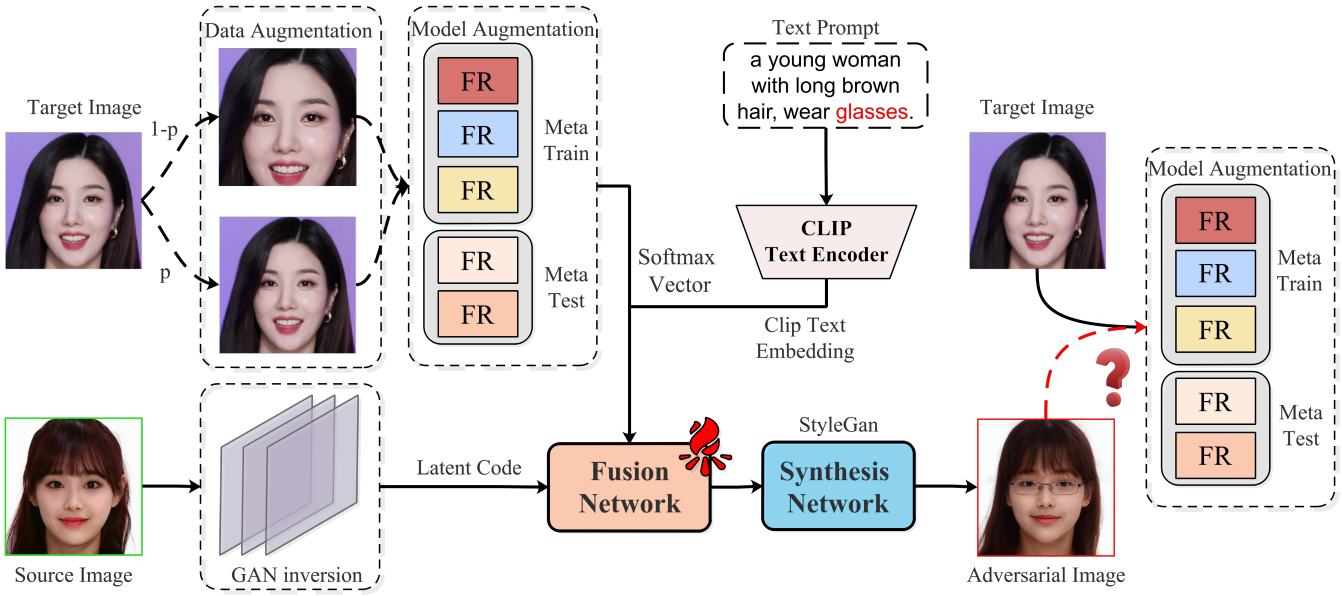

Figure 2: The overall framework of our proposed Text Controlled Attribute Attack (TCA^2). Our proposed method involves searching for adversarial faces on the generative StyleGAN manifold by optimizing the parameters of a Fusion Network. The generated photo-realistic adversarial faces can deceive state-of-the-art face recognition (FR) systems under the guidance of text prompts. To effectively represent the semantics of an impersonated person, a softmax vector is employed to perform a targeted attack against the FR system. The FR models in the framework are randomly selected from either the meta-train set or the meta-test set.

UAA.

Transferable Attack

Adversarial transferability refers to the ability of adversarial examples, generated on a white-box model (the source model), to successfully deceive a black-box model (the target model). Black-box adversarial attacks, which follow the most common practice, are more feasible in real-world scenarios and hold greater research significance compared to white-box attacks. While most adversarial attacks are designed for white-box models (Madry et al. 2017), they exhibit poor transferability when applied directly to black-box models. To address this limitation, various black-box attack strategies have been proposed, including query-based and transfer attacks. Query-based attacks (Ilyas et al. 2018) estimate gradients through a large number of queries to the target model. However, these methods require an extensive number of queries, making them susceptible to detection by the target system. In contrast, transfer attacks are more efficient. Although some previous works have achieved relatively high transferability in black-box settings, they often overfit to the source model. In our work, we employ both data and model augmentation techniques to enhance the transferability of our TCA^2 method in black-box scenarios.

Method

Problem Formulation

The objective of adversarial face attacks is to deceive the target face recognition (FR) model using adversarial perturbations. Specifically, an impersonation attack seeks to cause the FR model to misclassify a face as another specific identity by introducing subtle perturbations. Most prior studies have learned these perturbations under norm constraints to ensure stealthiness. In our work, we relax these strict constraints to explore unconstrained adversarial attacks (UAA).

Let $x_s \in \mathcal{X} \subset \mathbb{R}^n$ denote the given source face image, and let $x_t \in \mathcal{X} \subset \mathbb{R}^n$ denote the target face image to be impersonated. Let $\mathcal{F}(x) : \mathcal{X} \to \mathbb{R}^d$ be a face recognition model that extracts the normalized facial feature representation for identification. The optimization process of the impersonation face attack can be expressed as follows:

$$\underset{\hat{x}_s}{argmin} \mathtt{S}(\mathcal{F}(\hat{x}_s), \mathcal{F}(x_t)) \tag{1}$$

$$\underset{\hat{x}_s}{argmin} \mathcal{D}(\hat{x}_s, x_s) \tag{2}$$

where $\mathtt{S}(\cdot)$ denotes to the identity similarity, $\mathcal{D}$ represents the perceptual distance used in unrestricted adversarial attacks. We adopt the most common perceptual network LPIPS(Zhang et al. 2018) to measure the difference between the clean face image x_s and its corresponding adversarial image $\hat{x}_s$. A natural language text prompt t is adopted to control the generation of $\hat{x}_s$ according to the intent of adversary. The objective of our approach is to generate a high-quality adversarial face, denoted as $\hat{x}_s$, which closely resembles its original image except for the attribute controlled by the text prompt t. Additionally, $\hat{x}_s$ is designed to effectively mislead the black-box face recognition system, causing it to

misidentify $\hat{x}_s$ as x_t.

Preliminaries

A latent code in the style space of StyleGAN (Karras et al. 2020) can be projected into a specific image. Following the approach of (Li et al. 2021), we manipulate in the latent space of StyleGAN directly. Let G_L denote the generator network with L layers in StyleGAN. The random noise z is sampled from a uniform distribution Z and then transformed into a style vector ω via a nonlinear mapping network f. The intermediate latent code ω consists of L copies, i.e., $\omega = [\omega_1, \omega_2, \cdots, \omega_L] \in \mathbb{R}^{L\times 512}$. Each ω_m within ω represents the latent code input to the L_m layer of G_{L_m}. This ω_m is projected into the L_m layers and controls the m^{th} level of style in the synthesized image, where $m \in 1, 2, \cdots, L$. The corresponding attribute at the m^{th} level varies with changes in the value of ω_m. It is important to note that ω_m at different depths influences generated attributes to varying degrees: shallow layers control coarse attributes, middle layers control intermediate attributes, and deep layers control fine attributes. This impact is further illustrated in the Supplementary Materials.

In addition to the latent code ω, a noise term η, also sampled from the uniform distribution Z, is introduced to control the stochastic variations of the generated image at each layer. The noise term η typically affects uncorrelated attributes, such as the fine details of hair strands in a generated face. Since ω is entangled with semantically meaningful attributes, this work aims to control ω with a text prompt t to generate the desired adversarial image capable of fooling the target FR.

Text Controlled Attribute Attack

Previous works (Jia et al. 2022) have introduced semantically meaningful perturbations to create transferable adversarial examples against face recognition (FR) systems by injecting specific styles or patterns from a target image. However, these methods face two significant limitations: 1) **Inability to control the adversarial attributes.** While a real attribute vector corresponding to specific styles, such as smiling or glasses, is provided to control the generation details, the process automatically injects a pre-defined pattern from the target image. This means that the attacker cannot control the type of injected attribute, nor can they introduce a pattern outside the predefined attribute candidates. This limitation severely restricts the adversary's ability to generate the desired adversarial face. 2) **Low adversarial transferablity.** The semantically meaningful adversarial perturbations (Jia et al. 2022; Qiu et al. 2020; Liu et al. 2024) are optimized based on a single target FR model. As a result, the generated adversarial examples are highly coupled with the white-box FR model, which significantly reduces their transferability when used to attack black-box FR models with different architectures and parameters. Our work addresses these two challenges by focusing on enhancing both the control over adversarial attributes and the transferability of the generated adversarial faces.

Text-controlled Adversarial Face Generation Our text-controlled adversarial face generator (illustrated in Fig. 2) leverages the robust joint multimodal representation capabilities of the vision-language pretrained model, specifically CLIP (Radford et al. 2021). Given a text prompt t, the CLIP textual encoder $CLIP_t$ projects it into a shared embedding space as $E_t = CLIP_t(t)$, where E_t represents the textual embedding of the prompt t. For a clean face image x_s, a StyleGAN inverter network $Inv(\cdot)$ converts it into the corresponding style latent code, denoted as $\omega_s = Inv(x_s)$. To increase diversity, we apply random resizing and padding operations as data augmentation to the target face image. The augmented target face is then fed into FR to generate a softmax vector v, which guides the generation of the adversarial face image. Subsequently, the textual embedding E_t, latent code ω_s, and target face representation v are concatenated and fused. This process can be formalized as follows:

$$\omega_s^* = M_{\Theta_M}([\omega_s, E_t, v]) \tag{3}$$

where $\mathcal{M}_{\Theta_M}$ is a Multi-Level Fusion Network with learnable parameter Θ_M. Then the adversarial image is generated by $\hat{x}_s = G_L(\omega_s^*)$.

Additionally, we aim to align the adversarial face image with a controlling natural language prompt. For an image $\hat{x}_s$ guided by the text prompt t, our goal is for the image $\hat{x}_s$ to exhibit the attributes described by the text prompt t. Specifically, we use CLIP to bridge the gap between the text prompt t and the image $\hat{x}_s$. The textual guidance loss is defined as follows:

$$\mathcal{L}_{guide} = CLIP(\hat{x}_s, t) \tag{4}$$

where $CLIP(\cdot,\cdot)$ represents a pre-trained vision-language model. Additionally, apart from the attribute specified in the text prompt t, we aim to preserve all other attributes in the adversarial face image. This is similar to norm-based attacks, where the goal is to minimize the pixel-level differences between the clean image and the adversarial one. We apply the same principle to maintain minimal perceptual changes. The perception preservation loss is defined as follows:

$$\mathcal{L}_{perc} = D(\hat{x}_s, x_s) \tag{5}$$

where D represents a perceptual network pretrained using LPIPS. Our ultimate goal is for the adversarial face image $\hat{x}_s$ to effectively deceive the face recognition model $\mathcal{F}$. Specifically, in the context of an impersonation attack, we aim for the similarity scores between $\hat{x}_s$ and the target image x_t to be higher than those of other pairs. In our approach, we use the cosine similarity loss as our adversarial impersonation loss, defined as follows:

$$\mathcal{L}_{adv} = cos(\mathcal{F}(\hat{x}_s), \mathcal{F}(x_t)) \tag{6}$$

where $cos(\cdot)$ is the cosine similarity function. Finally, combining the three loss functions, we have

$$\mathcal{L}_{impe} = \lambda_{guide}\mathcal{L}_{guide} + \lambda_{perc}\mathcal{L}_{perc} + \mathcal{L}_{adv} \tag{7}$$

where λ_{guide} and λ_{perc} are hyperparameters that balance the contributions of the respective loss terms. Here, $\mathcal{L}_{adv}$ represents the adversarial objective as defined in Eq.6. Meanwhile, $\mathcal{L}_{guide}$ and $\mathcal{L}_{perc}$ correspond to the text-guided control and the preservation of other attributes in Eq.4 and Eq.5 respectively.

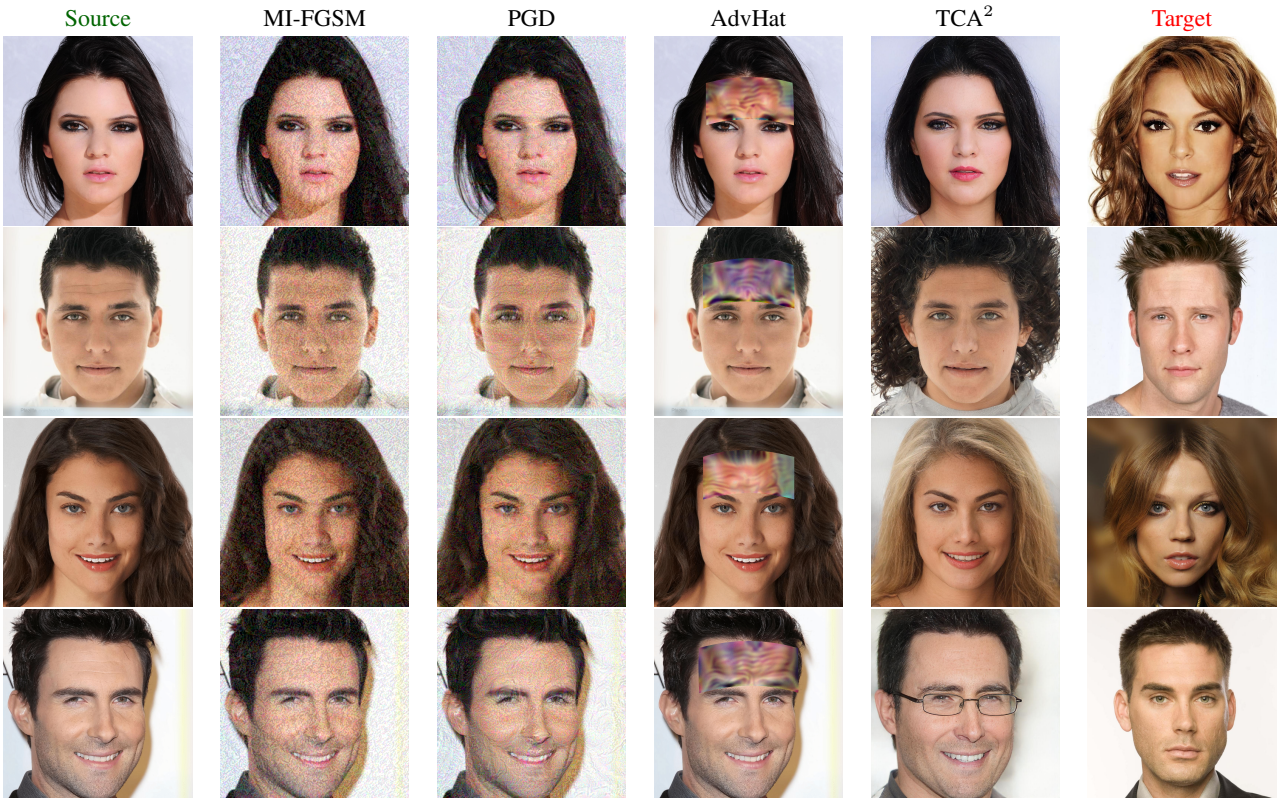

Figure 3: The visualization of source face, other different adversarial face and target face. Each image row is source, MI-FGSM, PGD, AdvHat, TCA^2, and target face, respectively.

Enhance the Transferablity of Target Adversarial Face Previous impersonation attacks(Jia et al. 2022; Liu et al. 2024) generate adversarial face images by optimizing a specific white-box surrogate face recognition (FR) model. This approach inevitably leads to overfitting to the white-box model, resulting in poor attack performance when targeting a black-box FR model. To address this issue, we adopt both data and model augmentation techniques to enhance the transferability of our TCA^2 method: 1) **Data Augmentation**. Previous works (Xie et al. 2019; Xiong et al. 2022) have demonstrated that data augmentation strategies help prevent adversarial examples from overfitting to specific data patterns. We employ this strategy to improve the generalization of the target identity. Specifically, we apply an input transformation $T(\cdot)$ with a stochastic dropout component, following the approach used in DIM (Xie et al. 2019). In our TCA^2, this involves applying random resizing and padding operations on the target face x_t; 2) **Meta Learning**. To mitigate overfitting to a specific surrogate model, some works (Liu et al. 2017; Gubri et al. 2022) have proposed enhancing the white-box model to improve generalization. Notably, model ensemble methods (Liu et al. 2017) simply aggregate losses from multiple models. However, obtaining multiple models can be challenging, and ensemble methods still tend to overfit to the combined white-box models. Inspired by recent adversarial research (Fang et al. 2022), we employ meta-learning to simulate both white-box and black-box environments. Specifically, given a total of $T+1$ FR models, we randomly select T models for the meta-train set and 1 model for the meta-test set. In each iteration, the meta-train and meta-test sets are reshuffled from the $T+1$ FR models. We first evaluate Eq.6 on the meta-test set, then jointly optimize the current loss using the ensemble losses from both the meta-train and meta-test sets. The meta-learning details of our TCA^2 approach are provided in the Supplementary Materials.

Experiments

Experiment Settings

Implementation details In our experiments, we utilize StyleGAN2, pretrained on the FFHQ face dataset, as our generative model. For adversarial text guidance, we employ the CLIP model, which is pretrained on the WIT dataset. For GAN inversion, we adopt the BDInvert method (Kang, Kim, and Cho 2021). We collect 18 text prompts representing diverse facial styles for the text guidance (details provided in the Supplementary Material). For training optimization, we use the Adam optimizer with β_1 set to 0.9, β_2 set to 0.999, and a learning rate of 0.01. The training process is run for 50 epochs. We set the values of λ_{guide} and λ_{perc} to 0.5 and 0.05, respectively. All experiments are conducted using PyTorch on a V100 GPU with 32 GB of memory.

Dataset We conducted experiments using two publicly available facial datasets: (1) the CelebA-Identity dataset (Na, Ji, and Kim 2022), which is a subset of the CelebA-HQ dataset (Huang et al. 2018) comprising 307 identities. Each identity includes at least 15 facial images, totaling 5,478 images, all at a resolution of 1024×1024 pixels. (2) The KID-F dataset[1], also known as the K-pop Idol Dataset - Female (KID-F), consists of approximately 6,000 high-quality facial images of Korean female idols. For our experiments, we selected about 2,000 images representing 100 identities from the KID-F dataset. We randomly chose 1,000 images from different identities as source images from both datasets. Additionally, five images were selected as target facial images in each dataset.

Attacked Threaten Model To validate our proposed impersonation attack against face recognition, we trained models on the two aforementioned datasets. Specifically, well-pretrained MobileFace (Chinaev, Chigorin, and Laptev 2018), IRSE50 (Hu, Shen, and Sun 2018), IR152 (Deng et al. 2019), and FaceNet (Schroff, Kalenichenko, and Philbin 2015) were fine-tuned on the CelebA-Identity and KID-F datasets. All FR models aligned the input face images via MTCNN (Zhang et al. 2016) during the preprocessing step.

Evaluation Metrics Following (Deb, Zhang, and Jain 2020), we used the *attack success rate* (ASR) to evaluate our proposed TCA^2. The ASR is defined as the proportion of adversarial faces that are misclassified by the face recognition model. The ASR for an impersonation attack is formulated as follows:

$$ASR = \frac{\sum_i^N 1_\tau \left(\cos\left[\mathcal{F}\left(\hat{x}_s^i\right), \mathcal{F}\left(x_t^i\right)\right] > \tau\right)}{N} \times 100\% \quad (8)$$

where 1_τ denotes the indicator function, $\hat{x}_s$ and x_t represent the generated adversarial face and the target face, respectively. Threshold τ is set as 0.01 FAR (False Acceptance Rate), and N is the number of images. ASR measures the proportion of source-target pairs whose similarity scores exceed τ out of all source-target pairs. More details can be found in the Supplementary Material.

Additionally, we report the PSNR, SSIM (Wang et al. 2004), and FID (Heusel et al. 2017) scores to evaluate the imperceptibility of TCA^2. Higher PSNR and SSIM scores indicate greater similarity to the original images, while a lower FID score suggests more realistic images.

Baseline methods We compare our TCA^2 approach with recent noise-based and unrestricted adversarial attacks against face recognition. The noise-based methods include MI-FGSM (Dong et al. 2018), and PGD (Madry et al. 2017). Unrestricted adversarial attacks include Adv-Makeup (Yin et al. 2021), Adv-Hat (Komkov and Petiushko 2021), Adv-Attribute (Jia et al. 2022), (Li et al. 2021), Latent-HSJA (Na, Ji, and Kim 2022), and Adv-Diffusion (Liu et al. 2024). Unlike traditional norm-based methods, which strictly ensure that perturbations do not exceed a set boundary, unrestricted adversarial attacks do not provide a strict guarantee that the perturbations will stay within the attribute bounds. All experimental settings closely follow those described in the original papers. Additional information is provided in the Supplementary Materials.

[1]https://github.com/PCEO-AI-CLUB/KID-F

Experimental Results

Baseline comparison In this section, we present the experimental results of TCA^2 on both datasets against four different pretrained face recognition models in black-box attack scenarios. To ensure a fair comparison, all TCA^2 results are averaged over five text style prompts.

We evaluate the black-box attack performance of TCA^2 on four face recognition models, namely MobileFace (Chinaev, Chigorin, and Laptev 2018), IRSE50 (Hu, Shen, and Sun 2018), IR152 (Deng et al. 2019), and FaceNet (Schroff, Kalenichenko, and Philbin 2015). Notably, we adopt a leave-one-out strategy, where three face recognition (FR) models are treated as available white-box models to train our TCA^2 framework, with the remaining model used as the target black-box model. Tab.1 reports the ASR results of TCA^2 in impersonation attacks against the target models on the CelebA-Identity and KID-F datasets. In most cases, TCA^2 achieves a higher ASR than other traditional norm-based adversarial attacks (Dong et al. 2018; Madry et al. 2017) and unrestricted adversarial attacks (Yin et al. 2021; Komkov and Petiushko 2021; Li et al. 2021; Jia et al. 2022; Liu et al. 2024).

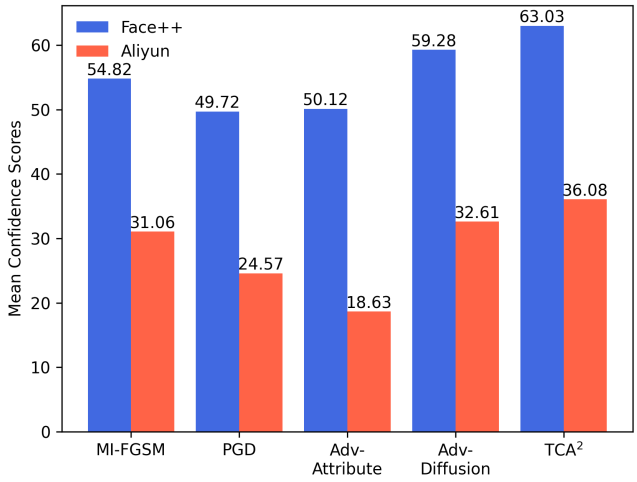

Figure 4: Mean confidence scores returned from commercial APIs, *i.e*, Face++ and Aliyun.

Image Quality Assessment In addition to the effectiveness of adversarial examples, their concealment is a crucial quality. Ideally, an adversarial image should be indistinguishable from a non-adversarial image. The FID scores of TCA^2 on two datasets, reported in Supplementary Materials, assess the naturalness of the generated images. Leveraging the capabilities of StyleGAN, TCA^2 is able to produce high-quality, photorealistic adversarial face images. Additionally, the PSNR and SSIM results are presented in Supplementary

Dataset	CelebA-Identity				KID-F			
Target Model	MobileFace	IRSE50	IR152	FaceNet	MobileFace	IRSE50	IR152	FaceNet
Clean	12.68	3.80	1.09	2.65	2.70	3.66	0.69	4.17
Inverted	13.57	2.74	0.66	3.32	1.81	2.52	1.45	6.66
MI-FGSM$_{\text{(CVPR'18)}}$	59.89	70.37	39.00	34.90	56.15	65.58	37.74	32.47
PGD$_{\text{(ICLR'18)}}$	50.10	61.73	45.26	38.85	43.05	62.80	40.01	33.33
Adv-Makeup	15.59	54.48	36.37	32.00	17.64	50.03	40.97	32.38
Adv-Hat$_{\text{(ICPR'21)}}$	8.40	7.23	2.74	5.27	5.99	7.77	10.09	6.44
Adv-Attribute$_{\text{(NeurIPS'21)}}$	45.59	56.81	38.67	30.81	42.59	53.77	39.90	34.81
(Li et al. 2021)$_{\text{(CVPR'21)}}$	28.60	52.74	30.73	33.78	28.63	47.04	28.70	33.46
Latent-HSJA$_{\text{(ECCV'22)}}$	14.40	37.15	14.77	12.64	15.99	33.51	16.71	15.82
Adv-Diffusion$_{\text{(AAAI'24)}}$	72.27	**80.08**	52.88	36.12	65.55	**83.79**	52.01	35.99
TCA^2	**73.10**	78.42	**53.72**	**42.26**	**65.57**	82.31	**53.51**	**39.64**

Table 1: Attack success rate (ASR %) of impersonation attack against the face recognition task on the CelebA-Identity and KID-F datasets. We choose four FR models (*i.e.*, MobileFace, IRSE50, IR152 and FaceNet) to evaluate the attack methods.

Original

w/o text guidance

w text guide

Figure 5: The visualizations of the text prompt's style on the visual quality of the output images. These images successfully deceive the facial recognition (FR) model. The text prompt used was "*A female face with open mouth*".

Original

w/o restriction

w text restriction

Figure 6: The visualizations of impact of perception preservation on the visual quality of the output images. The same text prompt, '*A female face with red lipstick*,' is applied to both images.

Materials. Visualizations of the adversarial faces generated by different attack methods are shown in Figure 3.

Attacks on commercial model via APIs To further validate the effectiveness and transferability of TCA^2 in real-world scenarios, we evaluated our algorithm using two well-known commercial face verification APIs: Face++ and Aliyun. The experimental results on the CelebA-Identity dataset are presented in Figure 4. As shown in Figure 4, TCA^2 outperforms the state-of-the-art method, Adv-Diffusion, on both commercial models.

Ablation Study In this section, we will report some ablation results to evaluate the contributions of our TCA^2 components.

- **Style text prompt:** As illustrated in Figure 5, without the guidance of a text prompt, the generated image tends to introduce globally visible perturbations aimed at optimizing the adversarial objective. These global perturbations have two main consequences: first, they can result in the generated image appearing unnatural; second, they limit the ability to offer users the option to select a desired style attribute compared to the clean image.
- **Perception preservation:** To evaluate the impact of perception preservation in TCA^2, we removed all perception preservation constraints. The visualization is presented in Figure 6. Our perception preservation constraints guide the adversarial optimization to explore the vicinity of the original latent code.

Additionally, we analyze the impact of other variables, specifically the hyperparameters λ_{guide} and λ_{perc}. Furthermore, we perform an ablation study to assess the impact on transferability. These results are provided in the Supplementary Materials.

Conclusion

In this paper, we proposed a novel approach that leverages natural language to guide the style latent code of StyleGAN2 in generating photorealistic face images capable of conducting impersonation attacks against face recognition systems. Additionally, TCA^2 demonstrates superior attack generalization across different face recognition models, making the generated adversarial images highly transferable to unknown black-box systems. Extensive experiments show that the faces generated using our method not only embody the desired attributes specified in the controlled text but also successfully deceive state-of-the-art face recognition systems, including commercial APIs, with a high success rate. However, defense mechanisms against unrestricted adversarial attacks remain underexplored. In future work, we plan to investigate more generalized defense strategies to enhance the robustness of face recognition systems against both norm-based and unrestricted adversarial attacks.

Acknowledgments

This research was partially supported by the Shenzhen Science and Technology Program (Grant No. RCYX20221008092852077), the National Natural Science Foundation of China (Grant Nos. 62372132 and 62402252), Pengcheng Laboratory Project (Grant No. PCL2024Y02).

References

Ali, W.; Tian, W.; Din, S. U.; Iradukunda, D.; and Khan, A. A. 2021. Classical and modern face recognition approaches: a complete review. *Multimedia Tools and Applications*, 80: 4825–4880.

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *IEEE SP*, 39–57. Ieee.

Chinaev, N.; Chigorin, A.; and Laptev, I. 2018. MobileFace: 3D Face Reconstruction with Efficient CNN Regression. arXiv:1809.08809.

Deb, D.; Zhang, J.; and Jain, A. K. 2020. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10. IEEE.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*, 4685–4694.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting Adversarial Attacks With Momentum. In *CVPR*, 9185–9193.

Dong, Y.; Su, H.; Wu, B.; Li, Z.; Liu, W.; Zhang, T.; and Zhu, J. 2019. Efficient decision-based black-box adversarial attacks on face recognition. In *CVPR*, 7714–7722.

Fang, S.; Li, J.; Lin, X.; and Ji, R. 2022. Learning to learn transferable attack. In *AAAI*, volume 36, 571–579.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gubri, M.; Cordy, M.; Papadakis, M.; Traon, Y. L.; and Sen, K. 2022. Lgv: Boosting adversarial example transferability from large geometric vicinity. In *ECCV*, 603–618. Springer.

Guetta, N.; Shabtai, A.; Singh, I.; Momiyama, S.; and Elovici, Y. 2021. Dodging attack using carefully crafted natural makeup. *arXiv preprint arXiv:2109.06467*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *CVPR*, 7132–7141.

Huang, H.; He, R.; Sun, Z.; Tan, T.; et al. 2018. Introvae: Introspective variational autoencoders for photographic image synthesis. *NeurIPS*, 31.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-box adversarial attacks with limited queries and information. In *ICML*, 2137–2146. PMLR.

Jia, S.; Yin, B.; Yao, T.; Ding, S.; Shen, C.; Yang, X.; and Ma, C. 2022. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. *NeurIPS*, 35: 34136–34147.

Kang, K.; Kim, S.; and Cho, S. 2021. Gan inversion for out-of-range images with geometric transformations. In *ICCV*, 13941–13949.

Karmon, D.; Zoran, D.; and Goldberg, Y. 2018. Lavan: Localized and visible adversarial noise. In *ICML*, 2507–2515. PMLR.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *CVPR*, 8110–8119.

Komkov, S.; and Petiushko, A. 2021. Advhat: Real-world adversarial attack on arcface face id system. In *2020 International Conference on Pattern Recognition (ICPR)*, 819–826. IEEE.

Li, D.; Wang, W.; Fan, H.; and Dong, J. 2021. Exploring adversarial fake images on face manifold. In *CVPR*, 5789–5798.

Liu, D.; Wang, X.; Peng, C.; Wang, N.; Hu, R.; and Gao, X. 2024. Adv-diffusion: imperceptible adversarial face identity attack via latent diffusion model. In *AAAI*, volume 38, 3585–3593.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2017. Delving into Transferable Adversarial Examples and Black-box Attacks. In *ICLR*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Massoli, F. V.; Carrara, F.; Amato, G.; and Falchi, F. 2021. Detection of face recognition adversarial attacks. *Computer Vision and Image Understanding*, 202: 103103.

Na, D.; Ji, S.; and Kim, J. 2022. Unrestricted Black-Box Adversarial Attack Using GAN with Limited Queries. In *ECCV*, 467–482. Springer.

Qiu, H.; Xiao, C.; Yang, L.; Yan, X.; Lee, H.; and Li, B. 2020. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *ECCV*, 19–37. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Ryu, G.; Park, H.; and Choi, D. 2021. Adversarial attacks by attaching noise markers on the face against deep face recognition. *Journal of Information Security and Applications*, 60: 102874.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Sharif, M.; Bhagavatula, S.; Bauer, L.; and Reiter, M. K. 2016. 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*, 1528–1540.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Vakhshiteh, F.; Nickabadi, A.; and Ramachandra, R. 2021. Adversarial attacks against face recognition: A comprehensive study. *IEEE Access*, 9: 92735–92756.

Wang, X.; Zhang, Z.; Wu, B.; Shen, F.; and Lu, G. 2021. Prototype-Supervised Adversarial Network for Targeted Attack of Deep Hashing. In *CVPR*, 16357–16366.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4): 600–612.

Xiao, C.; Li, B.; Zhu, J.-Y.; He, W.; Liu, M.; and Song, D. 2018. Generating adversarial examples with adversarial networks. In *IJCAI*, 3905–3911.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2730–2739.

Xiong, Y.; Lin, J.; Zhang, M.; Hopcroft, J. E.; and He, K. 2022. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *CVPR*, 14983–14992.

Yin, B.; Wang, W.; Yao, T.; Guo, J.; Kong, Z.; Ding, S.; Li, J.; and Liu, C. 2021. Adv-makeup: A new imperceptible and transferable attack on face recognition. *arXiv preprint arXiv:2105.03162*.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.*, 23(10): 1499–1503.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.

Zhang, Z.; Wang, X.; Lu, G.; Shen, F.; and Zhu, L. 2022. Targeted Attack of Deep Hashing Via Prototype-Supervised Adversarial Networks. *IEEE Trans. Multim.*, 24: 3392–3404.