Adv-Diffusion: Imperceptible Adversarial Face Identity Attack via Latent Diffusion Model

Decheng Liu[1,2]*, Xijun Wang[3]*, Chunlei Peng[1,2]†, Nannan Wang[4], Ruimin Hu[5], Xinbo Gao[6]
[1]School of Cyber Engineering, Xidian University, Xi'an, China
[2]Key Laboratory of Artificial Intelligence, Ministry of Education, Shanghai, China
[3]School of Artifical Intelligence, Xidian University, Xi'an, China
[4]School of Telecommunications Engineering, Xidian University, Xi'an, China
[5]Hangzhou Institute of Technology, Xidian University, Xi'an, China
[6]Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China
dchliu@xidian.edu.cn, xdu.xjwang@gmail.com, clpeng@xidian.edu.cn, nnwang@xidian.edu.cn,
hrm1964@163.com, gaoxb@cqupt.edu.cn

Abstract

Adversarial attacks involve adding perturbations to the source image to cause misclassification by the target model, which demonstrates the potential of attacking face recognition models. Existing adversarial face image generation methods still can't achieve satisfactory performance because of low transferability and high detectability. In this paper, we propose a unified framework Adv-Diffusion that can generate imperceptible adversarial identity perturbations in the latent space but not the raw pixel space, which utilizes strong inpainting capabilities of the latent diffusion model to generate realistic adversarial images. Specifically, we propose the identity-sensitive conditioned diffusion generative model to generate semantic perturbations in the surroundings. The designed adaptive strength-based adversarial perturbation algorithm can ensure both attack transferability and stealthiness. Extensive qualitative and quantitative experiments on the public FFHQ and CelebA-HQ datasets prove the proposed method achieves superior performance compared with the state-of-the-art methods without an extra generative model training process. The source code is available at https://github.com/kopper-xdu/Adv-Diffusion.

Introduction

With the development of deep learning technology, face recognition systems have been applied in more and more real-world scenarios, which also brings increasing security risks of biometrics in the meantime. Recent works find that these deep neural network-based models are vulnerable to adversarial perturbations added in the original clean images (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2018). These well-designed perturbations added images, also called *adversarial examples*, demonstrate the strong potential of attacking against existing state-of-the-art face recognition models (Deb, Zhang, and Jain 2020; Sharif et al. 2016), even evaluating black-box attacking scenarios (Jia et al. 2022; Komkov and Petiushko 2021). Thus, exploring craft adversarial examples in face recognition is vital and significant in the field of biometrics and economic security.

Existing adversarial attacks on face recognition methods are roughly grouped into three categories: gradient-based methods, patch-based methods and stealthy-based methods. The goal of adversarial face examples is to successfully attack target models with strong transferability, and generate high-quality adversarial images with inconspicuous perturbations. Gradient-based adversarial examples methods are early explorations to add L_p bounded perturbations directly to source images, which could bring in arbitrary predictions when testing (Madry et al. 2018; Dong et al. 2018). By controlling the perturbation boundary, pixel-level attacking clues seem imperceptible and make it more feasible to invade deployed face recognition systems. However, some related studies have found that gradient-based methods are vulnerable to variant lighting conditions (Xiao et al. 2018) and perform poorly in black-box attacking evaluations (Xiao et al. 2021). Patch-based methods (Sharif et al. 2016; Komkov and Petiushko 2021) aim to generate adversarial local face patches to protect identity information in physical-world scenarios. These existing synthesized craft adversarial patches usually have specific color and texture patterns, which are more easily distinguished and fail to be stealthy. Formerly mentioned kinds of adversarial face examples methods are generating suitable perturbations in raw pixel space, ignoring the specific properties of face images inherently. (Qiu et al. 2020) first explored learning adversarial face examples with semantic appearance, which could help to make these images visually imperceptible. (Hu et al. 2022) leveraged the makeup transfer generative network to generate better visual-quality adversarial faces. Recently, (Jia et al. 2022) has focused on synthesizing adversarial faces with edited attributes from reference faces to improve the stealthiness of attacking information. However, noting that not any reference face images would provide suitable and precise semantic information (e.g. attribute text (Qiu et al. 2020), semantic clues (Jia et al. 2022)) as guidance, and *it is essential to learn adversarial semantic ap-*

---

*These authors contributed equally.

†Corresponding author.

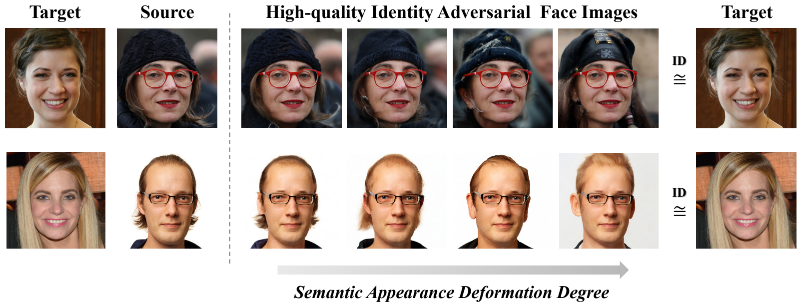

Figure 1: Illustrations of our proposed imperceptible adversarial face identity attack task, which can generate high-quality faces with target identities and similar source appearance information.

pearance automatically with high attack capabilities.

To address these issues, we propose a novel imperceptible adversarial face identity attack via latent diffusion model (Adv-Diffusion), which can generate adversarial perturbation in the latent space but not the raw pixel space and result in less perceptive perturbations surrounding the identity-sensitive regions as shown in Figure 1. Specifically, we leverage the latent diffusion model to construct the latent space for adversarial semantic perturbations. Because the diffusion model could offer excellent inductive bias for spatial data, and the learned latent space is perceptually equivalent to the raw pixel space (Rombach et al. 2022). To improve the key stealthiness performance, we design the identity-agnostic conditioned diffusion generative module. The pre-trained parsing model (Luo, Wang, and Tang 2012; Liu et al. 2020) is utilized to disentangle the face identity-sensitive regions (e.g. eyes, nose, mouth, etc.) and the face identity-agnostic region (e.g. hairstyle, decoration, background, etc.) by masking operation from a cognitive psychology perspective. Moreover, the identity-sensitive regions are regarded as the condition of the latent diffusion model when restoring adversarial face images. Additionally, the semantical adversarial perturbations are learned to add to the latent embedding through the adversarial attacking against face recognition. Noting that the proposed method is without any training data or complex deep network architectures, which is benefitted from the specific properties of the latent diffusion model.

The main contributions of our paper are summarized as follows:

- We propose a novel unified adversarial face image generation pipeline attacking face recognition in the latent space but not the raw pixel space, which could automatically learn effective adversarial semantic appearance with high attack capabilities and low perceptibility.
- We further propose the identity-sensitive conditioned diffusion generative module to guarantee the most adversarial appearance surrounding the identity-sensitive region, and the designed adaptive strength-based semantical adversarial perturbation is designed to ensure attack transferability and stealthiness.
- Experimental results on the public FFHQ and CelebA-HQ datasets illustrate the superior performance of the proposed Adv-Diffusion compared with the state-of-the-art adversarial face image generation methods. Meanwhile, the proposed method is without any training data or complex deep network architectures, which makes it easy to be deployed in real-world scenarios.

Related Work

Adversarial Attack on Face Recognition Model

Adversarial attacks on face recognition model is a growing field of research, and many studies have been conducted to evaluate the robustness of advanced face recognition models against adversarial attacks. In this paragraph, we review the representative attack methods on the face recognition model in three categories: gradient-based, patch-based and stealthy-based methods.

Gradient-based method. Gradient-based methods are one of the most widely used attack techniques in deep learning models. And several gradient-based methods have been proposed for adversarial attacks, including Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014), Projected Gradient Descent (PGD) (Madry et al. 2018) and Momentum-based Iterative FGSM (MI-FGSM) (Dong et al. 2018). FGSM is a fast and effective method for generating adversarial examples, where the perturbation is added to the input in the direction of the sign of the gradient. PGD is an iterative variant of FGSM that applies multiple small perturbations to the input and uses a projected gradient descent algorithm to ensure that the perturbation remains within a specified range. However, gradient-based methods are mostly applied in the white-box setting and are perceptible to humans.

Stealthy-based method. To overcome these limitations, recent studies have focused on developing imperceptible attacks with high transferability. For example, (Yin et al. 2021) proposed Adv-Makeup aims to achieve an implementable physical attack with high transferability by focusing on the makeup worn by the target individual. This approach is not only effective in digital conditions but also in physical scenarios, as the synthesized makeup is imperceptible to the human eye. Similarly, (Hu et al. 2022) proposed AMT-GAN, which also focuses on makeup transfer attack and has more transfer area. These above methods focus on pixel space

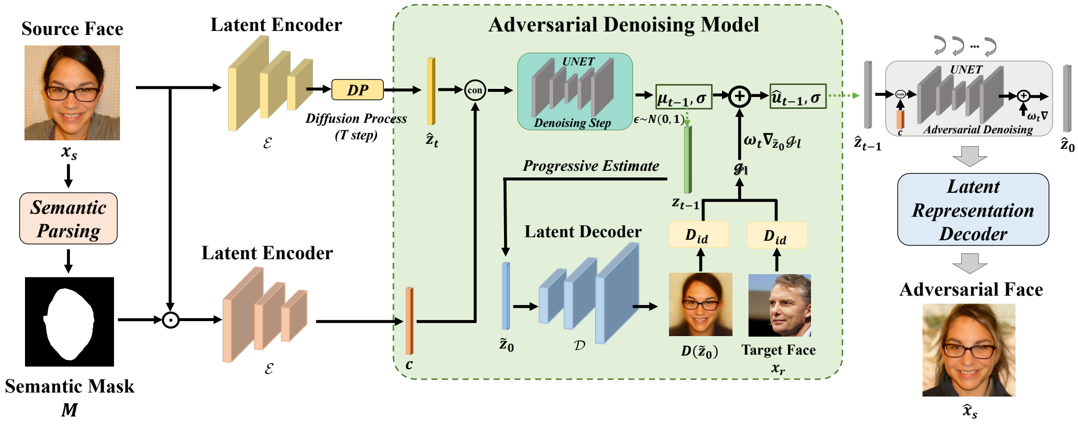

Figure 2: The framework of the proposed imperceptible adversarial face identity attack via latent diffusion model.

control, recent work (Jia et al. 2022) proposed Adv-Attribute to leverage Stylegan model (Karras, Laine, and Aila 2019) to attack the latent vector of the input image when preserving the identity of the input image. (Barattin et al. 2023) directly optimizes the latent representation of images in the latent space of a pre-trained GAN, which preserves the original facial attributes and attacks the recognition model successfully.

Patch-based method. The patch-based method is mainly focused on the physical attack in real-world scenarios. For instance, Adv-Glasses (Sharif et al. 2016) proposed a patch-based attack using optimization-based methods to add perturbations to the eyeglass region, while Adv-Hat (Komkov and Petiushko 2021) generated adversarial attacks over wearing hats. To improve attack ability when the attacker has limited accessibility to the target models, (Xiao et al. 2021) extend the existing transfer-based attack techniques to generate transferable adversarial patches by singing face-like features as adversarial perturbations through optimization on the manifold.

Diffusion Model Variants

Diffusion models, also known as diffusion probabilistic models (DPM), have recently emerged as a powerful generative modeling technique for image and video data. DPMs have shown state-of-the-art performance in sample quality. The core of DPMs is a diffusion process that evolves a simple initial distribution to a more complex target distribution by applying a sequence of diffusion steps. And there are several variants of DPMs. (Song and Ermon 2019) proposed a score-based generative model which learns the score function of data distribution. And continuous-time DPM defined stochastic differential equations to describe the diffusion and sample process of DPM. DiffPure (Nie et al. 2022) leveraged the denoising ability of DPM for defending against adversarial attacks recently. (Rombach et al. 2022) proposed latent diffusion model, which firstly learns an autoencoder to map image data to latent space, and further leverages the diffusion model to learn the latent space distribution. The latent diffusion model has more flexible conditional generation ability than pure DPM, such as inpainting, text-based image generation and image translation. Extensive analysis has proved the latent diffusion model can generate more realistic and high-fidelity results. Inspired by these successes, we also leverage the latent diffusion model to learn the latent space, but with a different target which is to generate identity adversarial face images with imperceptible perturbations.

Methodology

Problem Definition. The goal of the adversarial face identity attack is to mislead existing advanced face recognition with adversarial perturbations. However, most former works mainly focus on learning L_p bounded perturbations in the raw pixel space to make less detectability. Considering the specific properties of adversarial face images, we give a broader definition as follows. The general objective of targeted adversarial face images generation is defined as:

$$arg \max_{\hat{x}_s} F\left(x_r, \hat{x}_s\right). \tag{1}$$

Here $F(\cdot)$ refers to the identity similarity of these paired faces, where x_s and x_r denote the source image and target image respectively. The goal of our method is to learn a strong generative model, denoted as $G(\cdot)$, to generate the adversarial face image $\hat{x}_s$ with the condition of the target face x_s. Noting that the formulation can be easily extended to the untargeted adversarial face image generation application. *The criteria of adversarial face image should contain both strong transferability on recognition evaluation and less perceptive perturbations.*

Preliminaries: Latent Diffusion Model

In this section, we mainly introduce the conditional latent diffusion model (LDM). The diffusion model can generate high-quality images by reversing a nosing process iteratively. The latent diffusion model (Rombach et al. 2022) learns a compressed latent space of lower dimensions for decreasing computing cost. We firstly define an autoencoder to encode input image x into a latent code $z = \mathcal{E}(x)$ and decode z to $x = \mathcal{D}(z)$. During the forward stage, the T-step forward diffusion process aims to transform latent code z

gradually to match the Gaussian distribution: $z_T \sim \mathcal{N}(0,1)$. The forward process can be defined as $q\left(z_t \mid z_{t-1}\right) := \mathcal{N}\left(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}\right)$, where $\beta_t \in (0,1]$ is the hyperparameter. Following, the reverse process of the denoising step can be expressed as follows:

$$p_\theta\left(z_{t-1} \mid z_t\right) := \mathcal{N}\left(z_{t-1}; \mu_\theta\left(z_t, t\right), \Sigma_\theta\left(z_t, t\right)\right). \tag{2}$$

Here μ_θ and Σ_θ separately mean the mean and standard deviation, where $\mu_\theta\left(z_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta\left(z_t, t\right)\right)$. During the inference stage, we randomly initialize the latent code $z_T \sim \mathcal{N}(0,1)$, and then gradually update it until z_0 with the mentioned reverse process. Finally, the synthesized high-quality face image x_0 could be generated by the latent representation decoder.

Identity-sensitive Conditioned Diffusion Generative Model

Motivated by the perspective of cognitive psychology, most identity-discriminative information concentrates on identity-sensitive regions, e.g. eyes, nose, cheek, etc. The identity-agnostic regions (like hairstyle, decoration, background, etc) contain less discriminative information for human recognition but can be captured in the face recognition model (shown in Figure 2). Thus, we leverage the pretrained face parsing model to calculate the face region binary mask M. Then, we calculate the identity-sensitive region as $x_m = x_s \odot (1-M)$. Following, we leverage the strong inpainting capabilities of LDM to guarantee the most adversarial semantic appearance perturbations surrounding the identity-sensitive region. We leverage the pre-trained latent encoder to map the identity-sensitive region into the latent space $c = \mathcal{E}\left(x_m\right)$. We also calculate the initial value of $\hat{z}_T$ in the mentioned forward diffusion process with T steps by using x_s as the input image. The $\hat{z}_T$ is concatenated with c to obtain the input of the UNet model. Then, with the similar inpainting inference procedure in (Rombach et al. 2022), the reverse process with the designed adversarial denoising model can be expressed as:

$$p_\theta\left(z_{t-1} \mid \hat{z}_t, c\right) := \mathcal{N}\left(z_{t-1}; \mu_\theta\left(\hat{z}_t, t, c\right), \Sigma_\theta\left(\hat{z}_t, t, c\right)\right), \tag{3}$$

$$\hat{z}_{t-1} := z_{t-1} + \mathcal{G}_t, \tag{4}$$

where $\mathcal{G}_t$ is the designed adversarial perturbations described in the following. Figure 3 demonstrates the change of $\mathcal{D}(\hat{z}_{t-1})$ during the reverse process. Finally, the restored adversarial face image $\hat{x}_s$ can be calculated by the latent representation decoder as $\hat{x}_s = \mathcal{D}(\hat{z}_0)$.

Adaptive Strength based Semantical Adversarial Perturbation

To generate less perceptible perturbations in the identity-sensitive regions, we add the learned perturbations in the latent space but not the raw pixel space. We assume that the mentioned identity-sensitive conditioned diffusion model mainly focuses on generating suitable visual information surrounding the central regions and tries to maintain the identity-sensitive regions because of the condition of c and the inpainting ability of LDM. To leverage the above ability, we add semantic adversarial perturbations to the latent space in the reverse inpainting process.

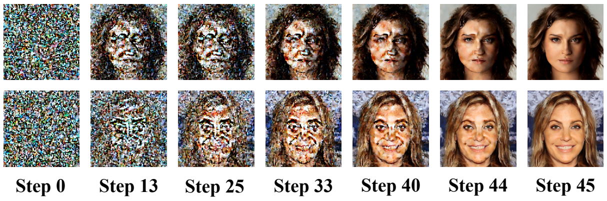

Figure 3: The reverse process with adversarial perturbation in the proposed method.

```
Algorithm 1: Adv-Diffusion
Input: source image $x_s$, target image $x_r$
Output: adversarial image $\hat{x}_s$
Initialization pretrained LDM, set s, T value;
$z_0 = \mathcal{E}(x_s), M = f(x_s), x_m = x_s \odot (1-M)$;
$\hat{z}_T = \sqrt{\bar{\alpha}_T} z_0 + \sqrt{1-\bar{\alpha}_T}\epsilon, \epsilon \sim \mathcal{N}(0,1)$;
$c = \mathcal{E}\left(x_m\right)$;
for all t from T to 1 do
    $z_{t-1} \leftarrow \mu_\theta\left(\hat{z}_t, t, c\right) + \epsilon\Sigma_\theta\left(\hat{z}_t, t, c\right), \epsilon \sim \mathcal{N}(0,1)$;
    $\widetilde{z}_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}\left(z_{t-1} - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_{t-1}, t)\right)$;
    $\mathcal{G}_t \leftarrow w_t \nabla_{\widetilde{z}_0} F\left(\mathcal{D}\left(\widetilde{z}_0\right), x_r\right)$;
    $\hat{z}_{t-1} \leftarrow z_{t-1} + \mathcal{G}_t$;
end
```

Because the latent space is perceptually equivalent to the raw pixel space, we design the added perturbations in the latent space that would bring in diverse semantic appearances. Specifically, we design the simple gradient-based adversarial sample algorithm to generate adversarial semantic perturbations. Here we utilized the face recognition model as the target model and the adaptive strength-based adversarial perturbation is calculated as follows:

$$\mathcal{G}_t = w_t \nabla_{\widetilde{z}_0} F\left(D\left(\widetilde{z}_0\right), x_r\right), \tag{5}$$

where $w_t = s\Sigma_\theta\left(\hat{z}_t, t, c\right)$, s is the hyper-parameter to control attack strength. $\widetilde{z}_0$ is an approximate result predicted from z_{t-1} as $\widetilde{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(z_{t-1} - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_{t-1}, t)\right)$ with similar estimation (Ho, Jain, and Abbeel 2020). α is also a hyperparameter mentioned before. The details of the parameters analysis are shown in the following section. It noted that when the reverse step increases, the attack strength w_t will decrease adaptively because of $\Sigma_\theta\left(\hat{z}_t, t, c\right)$. The designed adaptive attack strength strategy would help improve image quality. More algorithm details are shown in Algorithm 1.

Experiments

Experimental Setting

Datasets. Following similar protocols (Jia et al. 2022), we use two publicly available face datasets for evaluation: (1) FFHQ is a widely used high-quality face dataset (Karras, Laine, and Aila 2019), which contains almost 70,000

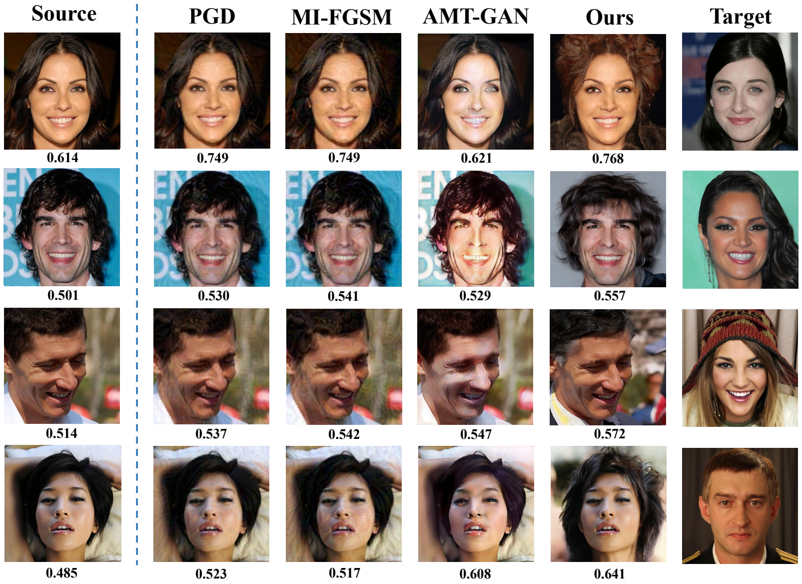

Figure 4: Comparison Results of generated adversarial samples when T is set as 1000. The number below the images means the calculated similarity score with the target face by the refined recognition model. Noting that the identity similarity score varies from 0 to 1.

high-quality face images with 1024 × 1024 resolution. (2) CelebA-HQ is a high-quality face dataset (Karras et al. 2018) constructed based on the CelebA dataset, which contains almost 30,000 face images with 512 × 512 resolution. In the evaluation stage, we randomly select 1000 images with different identities as the source images for both datasets. In particular, we also select 5 additional images as target images for each dataset, and randomly divide the 1000 source images into 5 groups, each corresponding to a different target image.

Implementation details. We utilize the pre-trained latent encoder and decoder networks, which are inspired by the open-source stable diffusion work. For experimental settings, we set 45 steps to generate adversarial samples. And we set s = 300 by default. Additionally, the face semantic parsing model architecture is based on the PyTorch implementation of EHANet (Luo, Xue, and Feng 2020). The face parsing model can segment holistic faces into 19 semantic regions, including face, hair, background, etc. We conduct experiments on RTX 3090 GPU.

Evaluation metrics. We evaluate our approach by quantifying the attack performance and image quality separately. The attack success rate (ASR) (Jia et al. 2022) is utilized to evaluate the attack performance of adversarial example algorithms. The value of τ is set as 0.01 FAR (False Acceptance Rate) with the same setting in (Hu et al. 2022). More details are shown in the supplement. For evaluating the generated image quality, we use the Frechet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as quality metrics in the following experiments.

Comparison Results

We first use the ASR metric to compare the proposed Adv-Diffusion with SOTA methods in black-box attack scenarios. The comparison results prove the proposed algorithm's high transferability in recognition attack tasks. Additionally, we evaluate the generated image quality compared with these mentioned SOTA methods to prove the evident imperceptibility of the proposed algorithm. The following experimental results prove our proposed method can achieve superior performance in both attacking transferability and imperceptibility.

Attacks on black-box model. To evaluate the attack performance of the proposed method, we selected four commonly used face recognition models: IR152 (He et al. 2016) , IRSE50 (Hu, Shen, and Sun 2018), FaceNet (Schroff, Kalenichenko, and Philbin 2015), and MobileFace (Deng et al. 2019) as target models following (Hu et al. 2022). It is noting that these face recognition models are pre-trained in large-scale face datasets, and all achieve satisfactory recognition performance. When one of these models is selected as a black-box target model, the other three models are used as white-box target models for generating attacking adversarial examples. We selected three categories of attack methods as our competitors: gradient-based methods, patch-based methods, and stealthy-based methods. For gradient-based methods, we set the perturbation strength value to 8/255. For other methods, we followed the official parameter settings. More details on the parameter setting can be found in the supplementary materials. Table 2 shows the attack performance comparison results on CelebA-HQ and FFHQ datasets. Noting that asterisk notation means the result is derived from source paper. Experimental results prove the proposed method achieves the highest attacking performance

Dataset	CelebA-HQ		
Metric	FID (↓)	PSNR (↑)	SSIM (↑)
FGSM	108.99	27.60	0.81
PGD	79.92	27.96	**0.85**
MI-FGSM	79.42	28.85	0.82
AMT-GAN	22.57	9.31	0.387
Adv-Attribute*	68.52	-	-
Adv-Diffusion	**15.51**	**29.01**	0.80

Table 1: Generated image quality comparison results with several metrics when T is set as 50.

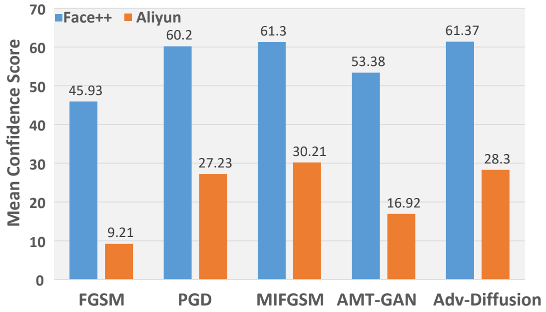

Figure 5: Mean confidence scores calculated from Face++ and Aliyun, which illustrate the possibility of the generated adversarial face image having the same identity as the target image.

compared with the SOTA methods.

Image quality assessment. To prove the imperceptibility of the proposed adversarial attack method, we further evaluate the adversarial image quality compared with other adversarial attack algorithms. Here we choose common image quality metrics, like Frechet Inception Distance (FID) (Heusel et al. 2017), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM). The better-generated image quality always means the lower imperceptibility of adversarial attack perturbations. Table 1 quantifies the adversarial image quality results of our method compared with the SOTA methods. Moreover, we have also evaluated the generated adversarial samples from a qualitative perspective. Figure 4 compares the adversarial samples generated by our method and other methods. The results prove the proposed methods can maintain better stealthiness of perturbation. Furthermore, we conducted a series of user studies to evaluate the performance in the supplement.

Attacks on commercial APIs. To further prove the effectiveness and transferability of our method in real-world scenarios, we conduct comparison experiments on two commercial APIs. We choose Face++ and Aliyun API as the target models and comparison experiment results on the CelebA-HQ dataset are shown. Figure 5 shows the proposed method outperforms SOTA methods on Face++ API and achieves competitive performance on Aliyun API.

Ablation Study

In this section, we explore the effect of two key components of the proposed method. The first component is the adaptive strength utilized to control the magnitude of adversarial perturbations for better image quality. The second component is the masked image condition in the designed attack model to improve stealthiness. We separately conduct quantitive and qualitative evaluation experiments to prove the effectiveness of the proposed method.

Method	ASR (↑)	FID (↓)	PSNR (↑)
Adv-Diffusion	53.43	**15.34**	**22.01**
w/o Adaptive Strength	**71.11**	62.37	12.62
w/o Mask	59.11	55.23	13.22

Table 3: Ablation study experimental results with several metrics.

Quantitative evaluation. We quantitatively investigate the impact of the adaptive strength strategy and the mask-conditioned generative model with ASR, FID and PSNR metrics. The experiments are conducted on the CelebA-HQ dataset using IR152 as the target model, and T is set as 50 for the following experiments. Table 3 proves the adaptive strength strategy can help achieve high attacking performance and maintain good image quality.

Qualitative evaluation. For the convenience of analysis, we further conduct the qualitative evaluation as shown in Figure 6. It can be observed that the designed adaptive strength and masked image-conditioned module can greatly enhance the quality of generated images while maintaining similar attack performance.

Parameter Analysis

We explore the effect of parameter T with ASR metric on the target model. Because we find that T can control the range of semantic appearance deformation as shown in Figure 1. Figure 7 shows the ASR performance when changing T on the FFHQ and CelebA-HQ datasets. It can be found that the value of ASR slightly increases with increasing the value of T. Additionally, Figure 1 shows the change of semantic appearance deformation when increasing the value of T. When the value of T increases, the semantic appearance deformation degree also becomes greater. We can find more generated adversarial perturbations are integrated into semantic regions (e.g., hair, ear, hat, etc.). The sufficient experiment results demonstrate that our proposed method guarantees the generated adversarial appearance surrounding the identity-sensitive region and maintains the strong attacking ability.

Limitations and Future Work

Based on the aforementioned analysis, our method has demonstrated the capability to generate high-quality adversarial samples with a high attacking performance. However, it is also important to emphasize the limitations of our approach. Firstly, since our attack model relies on pre-trained

Method	Dataset	CelebA-HQ				FFHQ			
	Target Model	IR152	IRSE50	FaceNet	MobileFace	IR152	IRSE50	FaceNet	MobileFace
Clean	-	4.72	5.40	0.80	14.84	3.56	4.94	1.53	9.60
Gradient-based	FGSM	11.52	47.13	1.22	54.75	10.00	51.01	3.95	53.26
	PGD	41.80	67.09	20.62	59.51	38.84	73.84	19.34	62.32
	MIFGSM	44.81	77.05	27.81	65.30	44.21	81.00	24.54	67.84
Patch-based	Adv-Hat*	2.50	-	4.70	8.40	13.60	-	4.80	3.10
	Gen-AP*	19.50	-	15.80	24.40	12.00	-	8.20	19.90
Stealthy-based	Adv-Attribute*	44.30	-	31.80	50.20	46.30	-	31.90	49.90
	AMT-GAN	10.38	55.28	5.08	46.80	14.35	60.21	7.23	45.12
Ours	Adv-Diffusion	**53.33**	**84.11**	**36.84**	**73.02**	**53.42**	**82.91**	**32.21**	**69.25**

Table 2: The comparison of experimental results with ASR metric when T is set as 50.

Figure 6: Ablation study results of generated adversarial samples when T is set as 50. The number below the images means the calculated similarity score with the same target face by the refined recognition model.

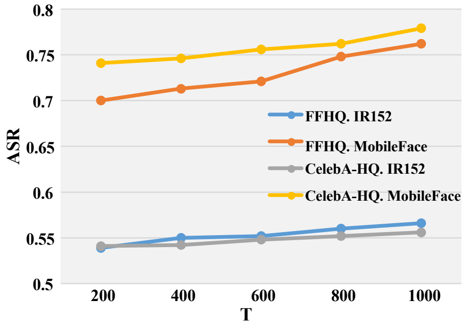

Figure 7: The ASR results with different T values.

large generative models, the generation of the adversarial image process is subject to the constraints derived from these source generative models. Consequently, some generated adversarial images may exhibit noticeable artifacts, thereby compromising generation quality and potentially impacting the attacking performance. The negative results are shown in the supplement. Secondly, it is worth noting that our method primarily focuses on generating adversarial face images, and its ability for other image types may be limited. We extend the Adv-Diffusion applied to non-face images in the supplement. In the future, we will explore extending the proposed method to encompass a broader range of image types to mimic more real-world scenarios.

Conclusion

In this paper, we propose the imperceptible adversarial face identity attack algorithm with the latent diffusion model (Adv-Diffusion). The proposed method designs a novel unified adversarial face image generation framework, which can learn adversarial semantic perturbations in the latent space for high attack capabilities and low perceptibility. To improve the stealthiness performance, we design the identity-sensitive conditioned diffusion generative module to guarantee the distinct adversarial appearance surrounding the identity-sensitive region. The adaptive strength-based semantical perturbation is proposed to ensure good stealthiness. Experiments on the public FFHQ and CelebA-HQ Face datasets illustrate the superior performance of the proposed method. Noting the proposed method is without any training data or complex network architectures, which will inspire more researchers to explore related fields. In the future, we will evaluate the proposed method with the identity-attacking performance on more complex real scenarios to adapt to the needs of the real world.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62306227, Grant 62276198, Grant U22A2035, Grant U22A2096, Grant 62036007, Grant 61922066, Grant 61906143, Grant 61902297, Grant 61806152 and Grant 61876142; in part by Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2022JQ-696; in part by the Key Research and Development Program of Shaanxi (Program No. 2023-YBGY-231); in part by Young Elite Scientists Sponsorship Program by CAST under Grant 2022QNRC001; in part by the Guangxi Natural Science Foundation Program under Grant 2021GXNSFDA075011; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; in part by CCF-Tencent Open Fund, in part by Open Research Projects of Zhejiang Lab under Grant 2021KG0AB01, and in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23083, Grant XJS221502 and Grant QTZX23042.

References

Barattin, S.; Tzelepis, C.; Patras, I.; and Sebe, N. 2023. Attribute-Preserving Face Dataset Anonymization via Latent Code Optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8001–8010.

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

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4690–4699.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *International Conference on Learning Representation*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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. In *Advances in Neural Information Processing Systems*, volume 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hu, S.; Liu, X.; Zhang, Y.; Li, M.; Zhang, L. Y.; Jin, H.; and Wu, L. 2022. Protecting facial privacy: generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 15014–15023.

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. In *Advances in Neural Information Processing Systems*, volume 35, 34136–34147.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representation*, 1–12.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4401–4410.

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

Liu, Y.; Shi, H.; Shen, H.; Si, Y.; Wang, X.; and Mei, T. 2020. A new dataset and boundary-attention semantic segmentation for face parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11637–11644.

Luo, L.; Xue, D.; and Feng, X. 2020. EHANet: An Effective Hierarchical Aggregation Network for Face Parsing. *Applied Sciences*, 10(9): 3135.

Luo, P.; Wang, X.; and Tang, X. 2012. Hierarchical face parsing via deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2480–2487.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion models for adversarial purification. *International Conference on Machine Learning*, 16805–16827.

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 10684–10695.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 11918–11930.

Xiao, C.; Zhu, J.-Y.; Li, B.; He, W.; Liu, M.; and Song, D. 2018. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 1–14.

Xiao, Z.; Gao, X.; Fu, C.; Dong, Y.; Gao, W.; Zhang, X.; Zhou, J.; and Zhu, J. 2021. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11845–11854.

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. In *Proceedings of International Joint Conference on Artificial Intelligence*, 1252–1258.