ADMMOA: Attribute-Driven Multimodal Optimization for Face Recognition Adversarial Attacks

Ruizhong Du, Luman Zhao*, Mingyue Li, Yidan Li, Shenyu Li
School of Cyber Security and Computer, Hebei University
Key Lab on High Trusted Information System in Hebei Province
Baoding Hebei 071002, China
drzh@hbu.edu.cn, lumanzhaozp@gmail.com, limingyue@hbu.edu.cn,
lyd_3495073752022@163.com, lishenyu0219@163.com

Caixia Ma
College of Cryptology and Cyber Science, Nankai University
Nankai Tianjin 300350, China
1120230319@mail.nankai.edu.cn

Abstract

Deep Neural Networks, especially face recognition (FR) models, have been shown to be vulnerable to digital and physical adversarial samples, which involve adding subtle perturbations to benign face images to deceive the models. This vulnerability poses a significant threat to the security of FR models and the collective well-being of society. To enhance the robustness of FR models against attacks, this paper aims to improve the transferability of adversarial face examples. We propose a novel approach, Attribute-Driven Multimodal Optimization Attack (ADMMOA), which leverages the conditional latent diffusion model to create adversarial images with high transferability and image quality in the latent space. Specifically, we introduce a multimodal conditional diffusion generation module that uses an adaptive significant attribute text and a dynamic semantic mask image to generate realistic images with semantic guidance of the significant attribute in the powerful inpainting process. Moreover, with the idea of gradient attack, the CLIP-augmented adaptive semantic adversarial perturbation module is introduced to further ensure the stealthiness and attack effectiveness of the generated adversarial face images. Extensive quantitative and qualitative experiments on the publicly available CelebA-HQ dataset demonstrate the superior performance of ADMMOA in improving the black-box transferability compared to the state-of-the-art methods. Particularly, our proposed ADMMOA achieves attack success rates (ASRs) of 62.40, 90.70, 43.50, and 83.30 on IR152, IRSE50, FaceNet, and MobileFace, respectively, surpassing Adv-Diffusion by 8.3%, 5.2%, 12.1%, and 5.0%.

Keywords: ***Adversarial examples, Face recognition, Diffusion models, Multimodal, Adversarial transferability.***

1. Introduction

Deep Neural Networks (DNNs) have been demonstrated to achieve superior performance in many application domains, even human competition [29]. Consequently, Face Recognition (FR) powered by DNNs have been widely deployed in high-security scenarios such as airport security, financial transactions, and smartphone unlocking [28]. However, as shown in the works of [4, 3], DNNs are vulnerable to adversarial examples, which are typically created by adding perturbations to clean samples. These perturbations are often subtle enough to be undetectable by the human eye but can successfully deceive the FR models. Additionally, adversarial face samples crafted by attackers can be transferable across different FR models [32], enabling attackers to successfully attack the victim FR models in black-box scenarios [31, 33, 30, 25]. This poses a serious threat to the security of FR models and greatly harms the collective well-being of society. Therefore, it is crucial to explore methods for crafting adversarial samples and improving their transferability.

However, few works have focused on improving the transferability of adversarial face samples during black-box attacks [32]. Hence, further research is needed to improve the transferability of adversarial face examples. Inspired

*Corresponding author.

by [10, 12, 14], we explore a method to generate adversarial images with high transferability and image quality. We propose a new generation framework, dubbed ADMMOA. The overall pipeline of ADMMOA is shown in Fig. 1. ADMMOA aims to generate adversarial images by using multimodal conditions and modifying the attributes through a diffusion model. It is designed as two modules, a **multimodal conditional diffusion generation module** and a **CLIP-augmented adaptive semantic adversarial perturbation module (CLIP-AASAP)**. In the multimodal conditional diffusion generation module, the latent diffusion model receives the adaptive significant attribute text and its dynamic semantic mask image as input conditions, generating realistic images with semantic guidance of significant attribute through the cross-attention mechanism and concatenation. In the CLIP-AASAP module, a CLIP-based stealthy loss is proposed, combined with adaptive semantic adversarial perturbation in [14] to further ensure that face images are stealthy and adversarial. In summary, the main contributions of this paper are summarized as follows:

- We introduce a novel unified pipeline for adversarial face image generation that attacks FR in the latent space rather than the original pixel space, achieving high attack capability with minimal perceptibility.
- We propose two modules: a multimodal conditional diffusion generation module and optimized CLIP-AASAP, to ensure the generation of adversarial semantic appearance and the effectiveness of our method against black-box FR models.
- Experimental results on the public CelebA-HQ dataset show that the proposed method outperforms the state-of-the-art (SOTA) methods.

2. Related Work

2.1. Attacks against Face Recognition

The aim of adversarial attacks on FR is to generate high-quality adversarial images with high transferable capability and imperceptible perturbations. In recent years, existing adversarial methods against FR are broadly categorized into four types: noise-based methods, patch-based methods, makeup-based methods, and stealth-based methods.

Noise-based methods. These methods craft adversarial samples by using gradient information to find small perturbations under the L_p norm constraint on the input image. [4] proposed the Fast Gradient Symbol Method (FGSM), which generates adversarial perturbations by calculating the gradient of the loss function. [3] introduced the momentum-based iterative FGSM (MI-FGSM). [16] used Projected Gradient Descent (PGD) to further optimize the adversarial perturbations. Although these methods are fast and computationally efficient in generating adversarial samples and perform well in white-box attacks [14], they are susceptible to variations in lighting conditions [27] and are ineffective in black-box attack evaluations [28].

Patch-based methods. These methods aim to mislead FR models by adding small and fixed-shaped perturbations (patches) to the image. [23] introduced Adv-Glasses, while [13] proposed Adv-Hat. Although these patch-based approaches are easy to implement and the patches can be used in real-world environments, they often have distinctive color and texture patterns. These patterns can be easily recognized by human observers and detection models, thereby reducing their stealthiness.

Makeup-based methods. The idea of the make-up based approachs is to hide the adversarial information in the generated makeup style. Recent works attempt to generate face images with adversarial makeup. Adv-Makeup in [29] developed a task-driven makeup generation method that synthesizes imperceptible eye shadow in the orbital region on faces. It also implemented a fine-grained meta-learning based adversarial attack strategy to enhance the transferability of adversarial samples. [9] proposed AMT-GAN to generate better visual quality adversarial faces using makeup transfer generative networks. However, existing these methods often suffer from poor visual quality and low transferability. And the attributes unrelated to makeup are hard to be completely preserved.

Stealth-based methods. Stealth-based approaches consider attributes of face images and aim to hide perturbations within specific attributes to render the attack visually imperceptible. [18] generated adversarial face examples through feature-space interpolation. Recently, [10] focused on editing attributes of reference faces, [12] used significant attributes for semantic adversarial attacks, and [14] proposed Adv-Diffusion to leverage a latent diffusion model to adaptively learn adversarial semantic appearance. Building on these types of methods, which effectively conceal perturbations, we modify attributes in the latent space of the diffusion model to learn adversarial semantic appearances in this paper.

2.2. Diffusion Models and Variants

Diffusion Model (DM) is a powerful modeling technique for generating image and video data. Initially, [24] proposed a score-based generative model to define the core concepts of diffusion modeling, which generates complex target distributions through a gradual process of adding noise and denoising. Subsequently, [7] introduced a

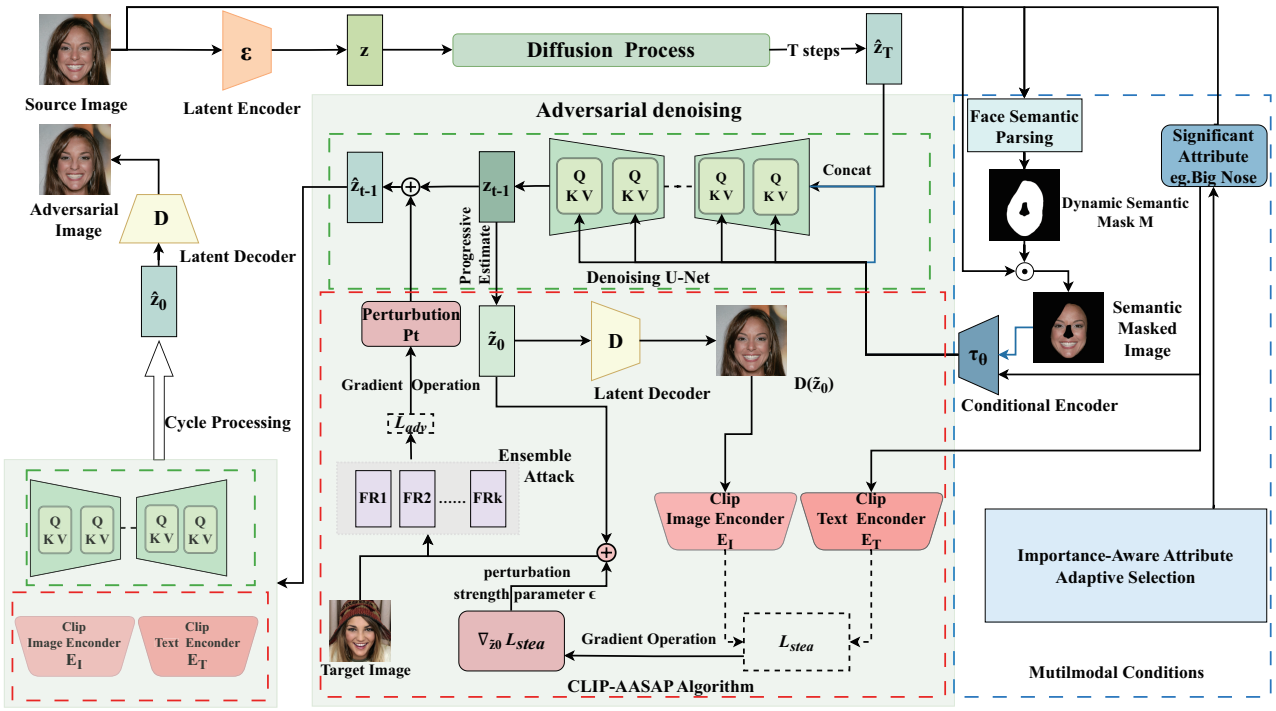

Figure 1. The framework of the proposed ADMMOA via latent diffusion model. Specifically, in Multimodal Conditions, the significant attribute to be edited is determined and a dynamic semantic mask is designed. These are used as multimodal conditions to guide the diffusion generation. Next, the CLIP-AASAP algorithm is designed for adversarial generation.

continuous-time diffusion model to describe the forward and backward diffusion processes through stochastic differential equations, which significantly improved the quality and theoretical basis of the generated samples. Furthermore, [17] proposed DiffPure, which develops a new defense technique against adversarial attacks by exploiting the denoising capability of DM, providing an effective solution for adversarial training. Based on this, [20] proposed the Latent Diffusion Model, which maps the image data to the latent space and learns data distributions in this low-dimensional space. [14] proposed a new Adv-Diffusion method for constructing adversarial images using this latent diffusion model.

Inspired by these successes, we also utilize the latent diffusion model to generate adversarial face images with imperceptible perturbations through multimodal conditions in the low-dimensional latent space.

3. Methodology

3.1. Problem definition

Black-box attacks on face recognition (FR) models can be further divided into impersonation attacks (i.e., targeted attacks) and dodging attacks (i.e., non-targeted attacks). For more efficient attacks of face images, we focus on targeted attack which aims to mislead FR models to recognize the generated faces as the specified target identity. The targeted attacks can be defined as an optimization problem:

$$\max_{\hat{x}_s} L_{adv} = S[F(\hat{x}_s), F(x_t)] \tag{1}$$

where $\hat{x}_s$ is the adversarial face image, x_t is the target face image, F represents the feature extractor of FR models, $S(\cdot)$ represents a similarity metric.

3.2. Preliminaries: Latent Diffusion Model (LDM)

The latent diffusion model is a generative model, which combines the strengths of deep-learning based image generation models with the advantages of diffusion models [20]. It can be used for high-quality image synthesis by diffusing in a low-dimensional latent space, significantly reducing the computational requirements compared to pixel-based diffusion models. The latent representation autoencoder $\mathcal{E}$ and decoder D are first defined to encode the input image x_s into the latent code $z = \mathcal{E}(x_s)$ and decode z into $x_s = D(z)$.

Forward Process. The forward diffusion process performed in the latent space involves adding noise to the image step by step and is defined by equation $q(z_t|z_{t-1}) :=$

$\mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I)$, where z_{t-1} and z_t are latent representations at time step $t-1$ and t, respectively. β_t is the noise coefficient at time step t, usually between 0 and 1.

Backward Process. The backward diffusion process in the latent space is the process of restoring the original image from the pure noise and can be defined as follows:

$$p_\theta(z_{t-1}|z_t) := \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)) \quad (2)$$

where p_θ denotes the neural network parameterized by θ, $\mu_\theta(z_t, t)$ is the mean value predicted by the neural network, and $\Sigma_\theta(z_t, t)$ is the covariance estimated by the neural network.

During inference phase, we initialize the latent code z_T with random values sampled from $\mathcal{N}(0, 1)$, and then gradually update it through the backward process described above until we obtain z_0. Finally, the latent representation decoder generates the synthesized high-quality face image x_0.

3.3. ADMMOA

To generate natural-looking and transferable adversarial face imges against FR models, ADMMOA aims to explore guidance of generation for face attribute editing using multimodal conditions. To achieve this, ADMMOA consists of two stages: multimodal conditional diffusion generation and CLIP-AASAP, as illustrated in Fig. 1. The details of each component are described as follows.

3.3.1 Multimodal Conditional Diffusion Generation Module

Importance-Aware Attribute Adaptive Selection. For diverse faces, modifying the same attribute could have different impacts on their identification. In other words, the importance of various facial attributes should be considered differently. Additionally, most identity information is concentrated in the sensitive regions of the face, such as eyes, nose, and cheeks [14]. Therefore, selecting and editing the attributes of these regions is crucial for the success of the attack. This study proposes a new strategy that adaptively selects and modifies significant attributes of a source image based on the victim FR model and target image in different attack scenarios.

First, **importance assessment of attributes**. This concept of attribute importance is exploited in the methods proposed by [10, 12]. Based on these methods, we specifically choose the cosine similarity (CS) algorithm from [12] to measure the impact of each attribute of the input image on the output of the FR model. We filter out some key sensitive attributes from the face images, including regions such as the eyes, nose and cheeks, that contribute significantly to identity information. Then, for our filtered image attributes $a = (a_1, a_2, a_3, \ldots, a_k)$, we train StarGAN [2] to modify them and use CS algorithm to attain ranking attribute list $c = (c_1, c_2, c_3, \ldots, c_k)$. The higher the ranking of an attribute, the greater its impact on the FR model from the input image, meaning these attributes should be prioritized for attack.

Then, **adaptive selection for significant attribute**. After obtaining the ranking list c for the attributes, we consider how to adaptively select the attributes in different attack scenarios, including dodging attacks and impersonation attacks. In the dodging attacks, we don't modify attributes for specific targets and the goal is to ensure that the attack produces the maximum range of interference to the FR models. In contrast, in impersonation attacks scenario, the objective shifts to making the input image more closer to a specific target image, deceiving the FR models. We consider modifying a single significant attribute, and the process of dodging and impersonation attacks is shown in Algorithm 1.

Multimodal Conditions Generation. We utilizes multimodal conditions to generate adversarial images in the latent space of the latent diffusion model. The adaptively selected significant attribute is input into the latent diffusion model as a text condition, which will then guide the attribute generation. Besides, compared to identity-sensitive regions, identity-insensitive regions (e.g., hairstyle, adornment, background, etc.) contain less discriminative identity information for human recognition, but can still be captured by the FR models [14]. Thus, we dynamically generate a face binary mask M based on the adaptively selected significant attribute and identity-insensitive by using a pre-trained face parsing model. We then compute retained regions as $x_m = x_s \odot M$ and utilize a pre-trained conditional encoder to map the x_m to the latent space, resulting in $cond_1 = \mathcal{T}_\theta(x_m)$. The significant attribute text t_{att} is also mapped as a text condition to $cond_2$.

Subsequently, we use x_s as the input image and compute the initial value of $\hat{z}_T$ in the forward diffusion process described above using T steps. $\hat{z}_T$ and condition c are used as the inputs to the UNet model. Here, c is formed by combining $cond_1$ and $cond_2$, where $cond_1$ influences the model through concatenation, and $cond_2$ guides image generation through the attention mechanism. Therefore, the adversarial denoising model, similar to the inference process of restoration in [20], can be designed as:

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

$$\hat{z}_{t-1} := z_{t-1} + P_t \quad (4)$$

where P_t represents the adversarial perturbations designed by the following CLIP-AASAP module. Repeating this process, we can obtain $\hat{z}_0$ and use the latent representation decoder to compute adversarial face image as $\hat{x}_s = D(\hat{z}_0)$.

In a word, with strong inpainting capabilities of LDM and condition c, the designed adversarial denoising model can generate suitable visual information in significant attribute and identity-insensitive region, guaranteeing the most adversarial semantic appearance perturbations in them.

3.3.2 CLIP-Augmented Adaptive Semantic Adversarial Perturbation

To make the generated face look natural and adversarial. We propose an optimization improvement scheme based on the Adaptive Semantic Adversarial Perturbation proposed by the method of [14], called CLIP-Augmented Adaptive Semantic Adversarial Perturbation (CLIP-AASAP). The main optimization of CLIP-AASAP is that a new CLIP-based stealthy loss L_{stea} along with adaptive semantic adversarial perturbation is introduced to jointly guide the generation of face images with natural-looking attribute and high black-box transferability.

Specifically, L_{stea} is designed to align the direction and control the strength between generated image and attribute text, can be defined as,

$$L_{\text{stea}} = \cos[E_I(D(\tilde{z}_0)), E_T(t_{att})] \tag{5}$$

where t_{att} is the selected attribute text. E_I and E_T are the image and text encoders of CLIP model [19]. $\tilde{z}_0$ is an approximate result predicted from z_{t-1} as $\tilde{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(z_{t-1} - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(z_{t-1}, t)\right)$ in similar estimation of [7]. Then, we use the gradient ascent algorithm to obtain the gradient from L_{stea} and add it to the approximate result $\tilde{z}_0$, as presented in Fig. 1 and Algorithm 1.

In addition to guidance in attribute aligned direction, there is also a need for guidance in the adversarial direction. For this, we adopt ensemble attack strategy [9], choosing K pre-trained FR models with high recognition accuracy as surrogate models, The ensemble attack loss is formulated as:

$$L_{adv} = \frac{1}{K}\sum_{k=1}^{K} \cos[F_k(D(\tilde{z}_0)), F_k(x_t)] \tag{6}$$

where F_k represents the k-th pre-trained FR model and we use cosine similarity as metric. Then, we obtain adversarial perturbation P_t from L_{adv} using the adaptive strength adversarial perturbation algorithm in [14] and add P_t to z_{t-1}, which controls adversarial strength and adjusts the adversarial direction more towards significant attribute and insenstive region. The specific calculation process is detailed in Algorithm 1.

Algorithm 1 ADMMOA

Input: Source image x_s, Target image x_t, Sorted attribute list of input source image $c = (c_1, c_2, c_3, \ldots, c_k)$, Attribute list $a = (a_1, a_2, a_3, \ldots, a_m)$ of the target image.
Parameter: Total number of steps T, Perturbation strength parameter $\boldsymbol{\epsilon}$.
Output: Significant attribute t_{att}, Adversarial image $\hat{x}_s$.

Initialization: Pretrained LDM model, set $\boldsymbol{\epsilon}$ and T.
// Adaptive selection for significant attribute
If impersonation attacks: We traverse the attributes in c to check if they exist in a. If an attribute c_i ($i = 1, 2, \ldots, k$) exists, we stop the traversal and set $t_{att} = c_i$. Other situations, we select $t_{att} = c_1$.
//Dynamic Semantic Mask and Multimodal Conditions
$z_0 = \mathcal{E}(x_s)$, $\hat{z}_T = \sqrt{\bar{\alpha}_T}\, z_0 + \sqrt{1-\bar{\alpha}_T}\,\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0,1)$
$M = f(x_s, t_{att})$, $x_m = x_s \odot (1 - M)$
$cond_1 = \mathcal{T}_\theta(x_m)$, $cond_2 = \mathcal{T}_\theta(t_{att})$
$c \leftarrow cond_1 + cond_2$
for $t = T$ to 1 **do**
  $z_{t-1} = \mu_\theta(\hat{z}_t, t, c) + \boldsymbol{\epsilon}\Sigma_\theta(\hat{z}_t, t, c), \boldsymbol{\epsilon} \sim \mathcal{N}(0,1)$
  $\tilde{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(z_{t-1} - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(z_{t-1}, t)\right)$
  //CLIP-AASAP
  Calculate the CLIP-based stealthy loss L_{stea} via Eq. 5.
  Calculate the adversarial loss L_{adv} via Eq. 6;
  $g = \nabla_{\tilde{z}_0} L_{stea}(D(\tilde{z}_0), t_{att})$
  $\tilde{z}_0 = \tilde{z}_0 + \boldsymbol{\epsilon} \cdot g$
  $P_t = w_t \nabla_{\tilde{z}_0} L_{adv}(D(\tilde{z}_0), x_t)$
  $\hat{z}_{t-1} = z_{t-1} + P_t$
end for
$\hat{x}_s \leftarrow D(\hat{z}_0)$
return $\hat{x}_s$

4. Experiments

4.1. Experimental Settings

Datasets. Following the similar protocol outlined in [14], we use a publicly available face dataset for evaluation: CelebA-HQ, a high-quality face dataset [11] based on the CelebA dataset, which contains nearly 30,000 face images with a resolution of 1024 × 1024. For the evaluation phase, we randomly select 1,000 images with different identities as source images and choose an additional 5 images as target images. We then randomly divide the 1,000 source images into 5 groups, with each group corresponding to a different target image.

Benchmark. We do comparisons with multiple benchmark schemes of the aforementioned four types of SOTA attack methods, including FGSM [4], PGD [16], MI-FGSM [3], Adv-Hat [13], Adv-Glasses [23], Adv-Makeup [29], AMT-GAN [9], Adv-Attribute [10] and Adv-Difussion [14]. FGSM, PGD, MI-FGSM, are typi-

Method	DataSet	CelebA-HQ			
	Target Model	IR152	IRSE50	FaceNet	MobileFace
-	clean	4.40	7.00	1.40	16.30
Noise-based	FGSM	8.00	31.30	6.30	45.70
	PGD	23.80	47.10	32.40	53.60
	MI-FGSM	27.60	49.30	38.40	54.10
Patch-based	Adv-Hat	2.50	14.30	4.70	8.40
	Adv-Glasses	4.50	12.10	9.10	5.60
Makeup-based	Adv-Makeup	10.50	23.90	3.40	21.00
	AMT-GAN	35.10	76.90	16.60	50.70
Stealth-based	Adv-Attribute	44.30	60.40	31.80	50.20
	Adv-Difussion	57.60	86.20	38.80	79.30
Ours	w/o Dynamic Semantic Mask	54.30	85.30	39.20	80.40
	w/o CLIP-AASAP	62.20	89.90	42.40	82.50
	ADMMOA	**62.40**	**90.70**	**43.50**	**83.30**

Table 1. ASR comparison results of black-box impersonation attacks against FR models.

Method	IR152	IRSE50	FaceNet	MobileFace
clean	7.00	4.40	1.40	16.30
FGSM	33.70	45.30	34.50	63.80
PGD	41.20	57.80	49.30	73.40
MI-FGSM	50.50	63.70	69.30	76.90
ADMMOA	**96.50**	**97.70**	**99.80**	**92.30**

Table 2. ASR comparison results of black-box dodging attacks against FR models.

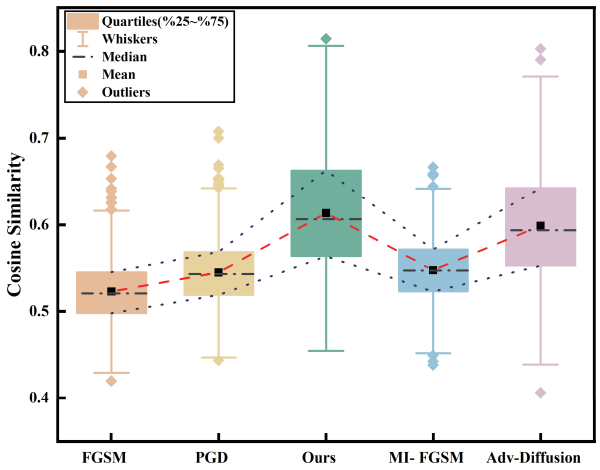

Figure 2. The cosine similarity of adversarial images created by our method with existing SOTA methods on IR152.

cal noise-based methods. Adv-Hat and Adv-Glasses are patch-based methods. Adv-Attribute and Adv-Difussion are stealth-based methods, while Adv-Makeup, AMT-GAN are makeup-based methods that exploit makeup transfer to

DataSet	CelebA-HQ		
Metric	FID ↓	PSNR ↑	SSIM ↑
FGSM	110.05	18.44	0.58
PGD	93.54	18.95	0.60
MI-FGSM	92.45	19.62	0.59
AMT-GAN	24.54	13.42	0.54
Adv-Difussion	**15.41**	27.16	0.79
w/o Dynamic Semantic Mask	12.28	30.07	0.83
w/o CLIP-AASAP	18.53	23.41	0.76
ADMMOA	16.00	**27.28**	**0.79**

Table 3. Image quality comparison results of black-box impersonation attacks against IR152 model.

generate images. All of which strictly adhere to their original settings.

Target Models. We select four pre-trained public FR models as the attacked models, including IR152 [5], IRSE50 [8], FaceNet [21], and MobileFace [1]. Three of them are chosen for ensemble attack and the remaining one serves as the black-box model.

Implementation details. We utilize a pre-trained inpainting latent diffusion model and a face semantic parsing model based on EHANet's PyTorch implementation, which can segment the overall face into 19 semantic regions [15]. By default, we set $\epsilon = 0.1$, and other settings follow the setup of previous work [14]. All our experiments are conducted on NVIDIA RTX4090 GPUs.

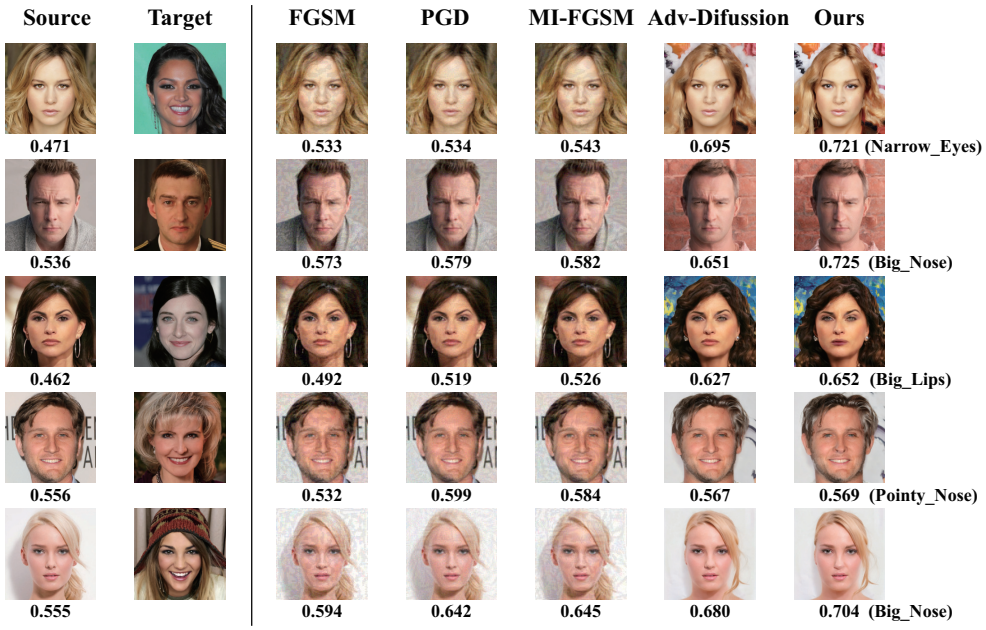

Figure 3. Visualized comparison of adversarial images generated by our method with SOTA methods on IR152. The numbers below the images indicate the similarity scores and the attributes in parentheses denote the modifications made to the source images.

Evaluation Metric. Following prior works [10, 14], we adopt Attack Success Rate (ASR) at False Acceptance Rate (FAR) = 0.01 for impersonation attacks to evaluate ADMMOA, which is computed as:

$$\text{ASR} = \frac{\sum_{i=1}^{N} 1_\tau \left(\cos[F(x_t^i), F(\hat{x}_s^i)] > \tau\right)}{N} \times 100\% \quad (7)$$

where 1_τ denotes the indicator function. F is the FR model, x_t and $\hat{x}_s$ are the target face and the generated face image respectively, N is the total number of images and τ is the threshold. For a fair comparison, the parameter τ will be set to 0.167, 0.241, 0.409, and 0.302 for the victim models IR152, IRSE50, FaceNet, and MobileFace, respectively.

In addition, we use we use the Fréchet Inception Distance (FID) [6], Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [26] as quality metrics in the following experiments.

4.2. Comparison Results with Existing Methods

Comparison on black-box attacks. We use the ASR metric and cosine similarity to compare our method with the aforementioned benchmark approachs in the black-box impersonation and dodging attacks scenarios. In Tab. 1 and Tab. 2, compared to other attacks, ADMMOA achieves the best transferable results on all victim models. Notably, in Tab. 1, our ADMMOA surpasses the performance of the Adv-Diffusion in ASR against IR152, IRSE50, FaceNet, and MobileFace by 8.3%, 5.2%, 12.1%, and 5.0%, respectively. Furthermore, in Fig. 2, it can be seen that the range of cosine similarity between the adversarial images derived by our method and the target images is overall higher than that of other SOTA methods.

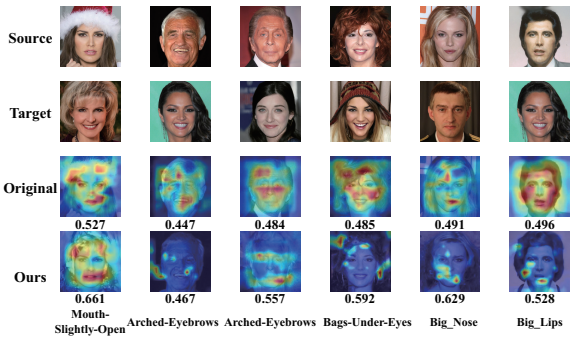

Figure 4. Grad-CAM attention visualization for IR152 on original and adversarial face images created by our method. Notably, the numbers below the images represent the similarity scores with the target images in the second row, while the attributes below the last row of images indicate the significant attributes modified in the images.

Comparison on image quality. Tab. 3 quantitatively reports the evaluations of image quality. Although, Adv-

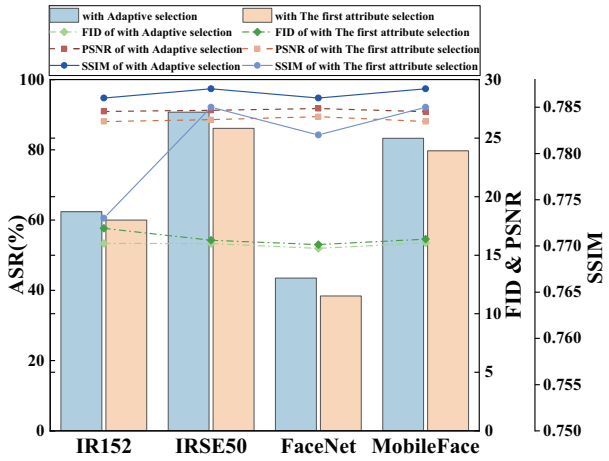

Figure 5. ASR and image quality metrics comparison of two selection strategies for significant attribute.

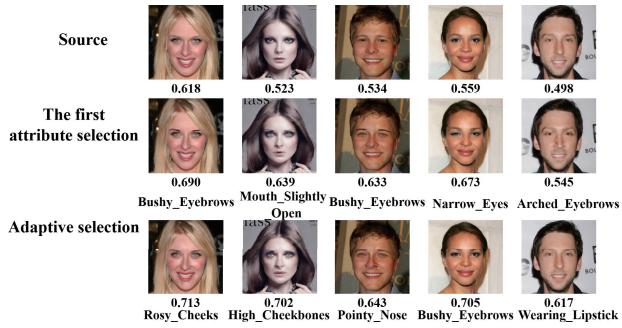

Figure 6. Visualized comparison of adversarial images generated by two strategies on IR152 model. The target identity is the same target female in the third row and second column of Fig. 3.

Difussion has the best performance in FID assessment, ADMMOA achieves relatively low FID scores and higher PSNR and SSIM scores, which indicates that the adversarial face images generated by ADMMOA are more natural-looking and have less impact on images at the pixel level. We also present the qualitative visualization of those adversarial images in Fig. 3, which demonstrates that adversarial samples generated by our method maintain a visually indistinguishable appearance.

Grad-CAM attention visualization. We use Gradient-weighted Class Activation Mapping (Grad-CAM) [22] technique to visualize the attention of the IR152 model on both the original and adversarial face images generated by our method. In Fig. 4, the generated heatmaps for the adversarial images show an apparent shift in attention towards identity-insensitive regions (e.g., hairstyle, background, etc.) and corresponding significant attribute region. This indicates that the designed adversarial denoising model can guarantee the most adversarial semantic appearance perturbations in them.

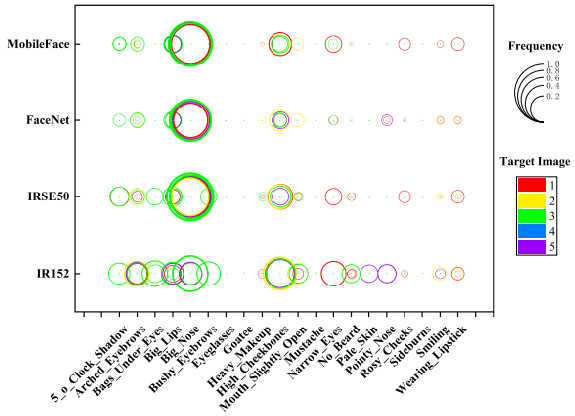

Figure 7. The frequency of each attribute across different FR models and different target images when the source images are the same.

4.3. Ablation Study

Dynamic Semantic Mask and CLIP-AASAP. We verify the importance of dynamic semantic mask for transferable ability. In Tab. 1, without dynamic semantic mask, it implies that the mask image will not be generated based on the adaptively selected significant attribute. Without CLIP-AASAP, it lacks the optimization of CLIP-based stealthy loss L_{stea}. In the absence of dynamic semantic mask, transferable ability in Tab. 1 decreases significantly. In Tab. 3, the image quality without CLIP-AASAP shows a considerable decline, which means CLIP-based stealthy loss L_{stea} helps ensure the stealthiness of the images.

Adaptive Selection VS First Attribute. We also discuss the effectiveness of the adaptive selection strategy for significant attribute, compared to a strategy that selects the first attribute as significant attribute, furthermore, we conduct quantitative and qualitative evaluation experiments separately. Fig. 5 provides a quantitative comparison, while Fig. 6 offers a qualitative analysis, demonstrating that our adaptive selection outperforms the strategy of simply selecting the attribute ranked first. Additionally, we calculate the frequency of each attribute across different FR models and different target images. The results, shown in Fig. 7, indicate that given the same source image, our adaptive selection strategy can adaptively search for the optimal significant attribute of the source image based on the attacked FR model and target image.

4.4. Limitations in Our Work

Based on the above analysis, our method has demonstrated superior performance on FR attacks. However, there is limitation to our approach. Because ADMMOA depends on a face parsing model and a pre-trained LDM model, the adversarial image generation process is constrained by these models. This may result in some generated adversarial images exhibiting noticeable artifacts, which could affect the effectiveness of the attack.

5. Conclusion

In this work, we propose the ADMMOA attack method for FR attacks. Our approach generates adversarial face images in the latent space to achieve high attack transferability and low visibility. To be specific, we design a multimodal conditional diffusion generation module that modifies the significant attribute to produce semantic appearances. Additionally, we introduce the CLIP-augmented adaptive semantic adversarial perturbation to further gudide stealthiness and adversariality. Experiments conducted on the public CelebA-HQ dataset demonstrate the superior performance of our proposed method.

Acknowledgement

This paper was supported by the Natural Science Foundation of Hebei Province of China (F2022201005,F2023201033), the Natural Science Foundation of Hebei Education Department of China (QN2024204), the Key Research and Development Program of Hebei Province of China (22340701D), and Beijing-Tianjin-Hebei Basic Research Collaborative Special Project (F2024201070), the Innovation Ability Enhancement Program of Hebei Province of China (HBU2025SS026).

References

[1] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pages 428–438. Springer, 2018. 6

[2] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. 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. 4

[3] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 1, 2, 5

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1, 2, 5

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[7] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 5

[8] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 6

[9] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. 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. 2, 5

[10] S. Jia, B. Yin, T. Yao, S. Ding, C. Shen, X. Yang, and C. Ma. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. *Advances in Neural Information Processing Systems*, 35:34136–34147, 2022. 2, 4, 5, 7

[11] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 5

[12] Y. M. Khedr, Y. Xiong, and K. He. Semantic adversarial attacks on face recognition through significant attributes. *International Journal of Computational Intelligence Systems*, 16(1):196, 2023. 2, 4

[13] S. Komkov and A. 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. 2, 5

[14] D. Liu, X. Wang, C. Peng, N. Wang, R. Hu, and X. Gao. Adv-diffusion: imperceptible adversarial face identity attack via latent diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3585–3593, 2024. 2, 3, 4, 5, 6, 7

[15] L. Luo, D. Xue, and X. Feng. Ehanet: An effective hierarchical aggregation network for face parsing. *Applied Sciences*, 10(9):3135, 2020. 6

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2, 5

[17] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, page 16805–16827, 2022. 3

[18] H. Qiu, C. Xiao, L. Yang, X. Yan, H. Lee, and B. Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 19–37. Springer, 2020. 2

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 4

[21] F. Schroff, D. Kalenichenko, and J. 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. 6

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[23] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 2, 5

[24] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[25] W. Wang, B. Yin, T. Yao, L. Zhang, Y. Fu, S. Ding, J. Li, F. Huang, and X. Xue. Delving into data: Effectively substitute training for black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4761–4770, 2021. 1

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[27] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018. 2

[28] Z. Xiao, X. Gao, C. Fu, Y. Dong, W. Gao, X. Zhang, J. Zhou, and J. 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. 1, 2

[29] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, page 1252–1258, 2021. 1, 2, 5

[30] C. Zhang, P. Benz, A. Karjauv, and I. S. Kweon. Data-free universal adversarial perturbation and black-box attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7868–7877, 2021. 1

[31] Y. Zhong and W. Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2020. 1

[32] F. Zhou, H. Ling, Y. Shi, J. Chen, Z. Li, and P. Li. Improving the transferability of adversarial attacks on face recognition with beneficial perturbation feature augmentation. *IEEE Transactions on Computational Social Systems*, 2023. 1

[33] M. Zhou, J. Wu, Y. Liu, S. Liu, and C. Zhu. Dast: Data-free substitute training for adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243, 2020. 1