Adversarial Attacks on Both Face Recognition and Face Anti-spoofing Models

Fengfan Zhou[1] , **Qianyu Zhou**[2,3] , **Hefei Ling**[1*] and **Xuequan Lu**[4]
[1]School of Computer Science and Technology, Huazhong University of Science and Technology
[2]College of Computer Science and Technology, Jilin University
[3] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, JLU
[4]Department of Computer Science and Software Engineering, The University of Western Australia
{ffzhou, lhefei}@hust.edu.cn, zhouqianyu@jlu.edu.cn, bruce.lu@uwa.edu.au

Abstract

Adversarial attacks on Face Recognition (FR) systems have demonstrated significant effectiveness against standalone FR models. However, their practicality diminishes in complete FR systems that incorporate Face Anti-Spoofing (FAS) models, as these models can detect and mitigate a substantial number of adversarial examples. To address this critical yet under-explored challenge, we introduce a novel attack setting that targets both FR and FAS models simultaneously, thereby enhancing the practicability of adversarial attacks on integrated FR systems. Specifically, we propose a new attack method, termed Reference-free Multi-level Alignment (RMA), designed to improve the capacity of black-box attacks on both FR and FAS models. The RMA framework is built upon three key components. Firstly, we propose an Adaptive Gradient Maintenance module to address the imbalances in gradient contributions between FR and FAS models. Secondly, we develop a Reference-free Intermediate Biasing module to improve the transferability of adversarial examples against FAS models. In addition, we introduce a Multi-level Feature Alignment module to reduce feature discrepancies at various levels of representation. Extensive experiments showcase the superiority of our proposed attack method to state-of-the-art adversarial attacks.

1 Introduction

Recent advancements in Face Recognition (FR) have led to remarkable performance improvements [Schroff *et al.*, 2015; Deng *et al.*, 2022; An *et al.*, 2021]. However, the vulnerability of current FR systems to adversarial attacks presents a critical security concern. This underscores the urgent need to enhance the effectiveness of adversarial face examples to expose deeper vulnerabilities in FR systems. Numerous adversarial attack methods have been proposed, focusing on properties such as stealthiness [Qiu *et al.*, 2020; Yang *et al.*, 2021; Cherepanova *et al.*, 2021; Hu *et al.*, 2022; Shamshad *et al.*, 2023], transferability [Zhong and Deng, 2021; Li *et al.*, 2023b; Zhou *et al.*, 2024b; Zhou *et al.*, 2024a], and physical-world attack capability [Yin *et al.*, 2021; Yang *et al.*, 2023; Li *et al.*, 2023a]. These endeavors significantly contribute to enhancing the effectiveness of adversarial attacks on FR systems.

Despite the significant progress in adversarial attacks on FR systems, the integration of Face Anti-Spoofing (FAS) [Zhou *et al.*, 2022; Zhou *et al.*, 2024d] models in FR systems poses a substantial challenge to the practicality of these attacks in real-world scenarios. As illustrated in Figure 1, when adversarial examples are applied, the visual features of the source images post-pasting with adversarial examples remarkably differ from those of live images, often revealing spoof features that can be detected by FAS models. If the adversarial examples are flagged as spoof samples by FAS models, the images will be preemptively filtered out by victims without inputing into FR models, thereby hindering attempts to attack. However, existing adversarial attacks on FR often overlook the incorporation of FAS models within FR systems, leading to failures in real-world deployment scenarios. Therefore, it is imperative to develop novel adversarial attacks targeting both FR and FAS concurrently to enhance the practicality of adversarial attacks on FR systems.

To successfully attack both FR and FAS models, it is essential to utilize the gradients from both models to craft adversarial examples. However, in real-world deployment scenarios, attackers often lack direct access to the models used by their targets. Consequently, a widely adopted method is to leverage surrogate models to generate adversarial examples, subsequently transferring them to the target models [Zhong and Deng, 2021; Yin *et al.*, 2021; Zhou *et al.*, 2024b]. Nevertheless, the inherent differences between the surrogate and target models lead to disparities in their decision boundaries. As a result, when aiming to simultaneously attack FR and FAS models, it involves the following challenges. Firstly, performing simultaneous adversarial attacks on both FR and FAS models can result in gradient imbalances, which can degrade the overall attack performance on both models. In addition, crafting adversarial examples using the final output live score of the FAS models may lead to overfitting to the surrogate model. While employing adversarial attacks targeting the FAS models based on intermediate loss functions can enhance transferability, the resulting adversarial examples may still overfit to the specific reference live images, thereby lim-

*Corresponding author.

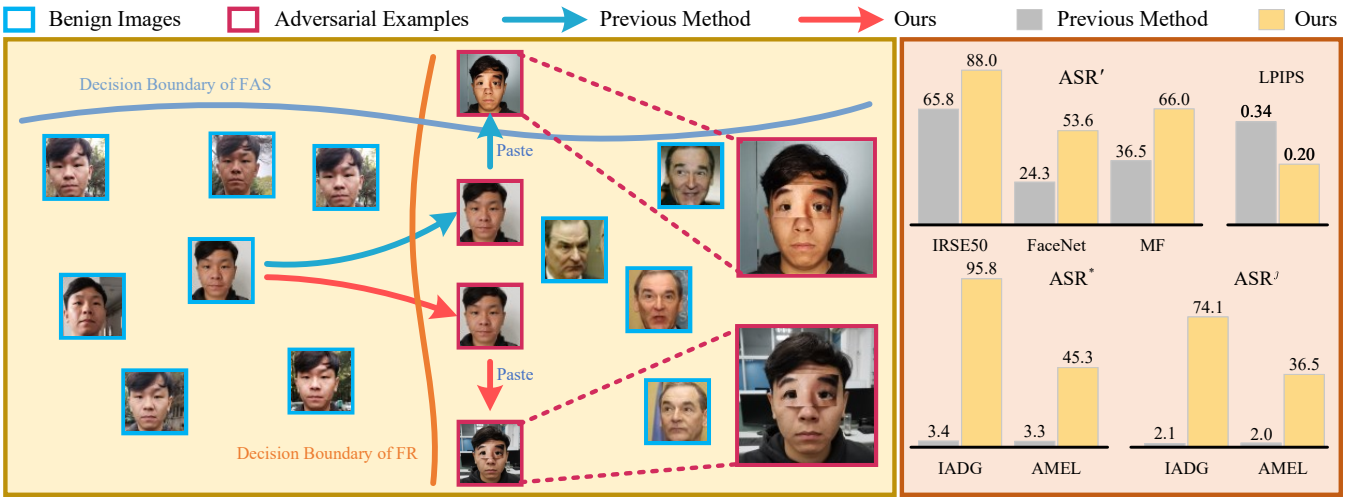

Figure 1: **Left**: A comparison between previous methods and our proposed method. Adversarial examples generated by previous methods for FR systems often contain spoofing artifacts, making them easily detectable and filtered out by FAS models. In contrast, our method attacks both FR and FAS models simultaneously, improving the practicality of adversarial examples in FR systems. **Right**: Performance comparison based on the metrics of LPIPS, and Attack Success Rate (ASR) for FR (ASR'), FAS (ASR^*), and both models (ASR^J).

iting the transferability improvement. Furthermore, generating adversarial examples for FR without considering feature alignment across multiple intermediate layers can lead to an over-reliance on specific features of the surrogate model, thereby compromising the transferability.

To address these challenges, we propose a novel attack method, termed Reference-free Multi-level Alignment (RMA), designed to simultaneously target both FR and FAS models. Our approach comprises three key modules: Adaptive Gradient Maintenance (AGM), Reference-free Intermediate Biasing (RIB), and Multi-level Feature Alignment (MFA). We perform independent gradient calculations for the MFA and RIB modules, and subsequently the AGM module mitigates the imbalances between the two types of gradients. Specifically, the AGM module dynamically re-weighting the losses for FR and FAS in each iteration to reduce the gradient disparities between the two tasks, thereby balancing the optimization process and enhancing the attack performance. The RIB module biases adversarial examples into the space of the live images using an intermediate loss without overfitting to specific reference live images by approximating the neural networks leveraging the linear hypothesis from [Goodfellow *et al.*, 2015], thereby further improving the effectiveness of black-box attacks on FAS models. Finally, the MFA module enhances transferability across FR models by aligning the features of adversarial examples with those of target images at multiple intermediate layers of the FR models. As depicted in Figure 1, after applying the adversarial examples crafted by our proposed method, the post-pasting adversarial examples can still traverse the decision boundaries of FR and FAS models, leading to successful attacks on the FR systems. The contributions of our paper are three-fold:

- We introduce a novel and practical setting of attacking FR and FAS models simultaneously to boost the practicability of adversarial attacks on FR systems. We propose an innovative adversarial attack framework termed Reference-free Multi-level Alignment (RMA) to improve the attack capacity on both models. To our best knowledge, this is the first study that adversarially attacks both FR and FAS simultaneously utilizing the gradients of both models.
- We design the Adaptive Gradient Maintenance module to mitigate the imbalances between the gradients of FR and FAS models, the Reference-free Intermediate Biasing module to improve the transferability of adversarial attacks on FAS, and the Multi-level Feature Alignment module to improve the black-box attack capacity of adversarial attacks on FR.
- Extensive experiments demonstrate the superiority of RMA to state-of-the-art adversarial attacks, as well as its nice compatibility with various models.

2 Related Work

Adversarial Attacks. The primary objective of adversarial attacks is to introduce imperceptible perturbations into benign images, deceiving machine learning systems and causing them to produce erroneous outputs [Szegedy *et al.*, 2014; Goodfellow *et al.*, 2015; Rocamora *et al.*, 2024; Xu *et al.*, 2023; Shayegani *et al.*, 2024]. The existence of adversarial examples poses a significant threat to the security and reliability of modern machine learning systems. Consequently, substantial research efforts have been devoted to studying adversarial attacks to enhance the robustness of these systems [Dong *et al.*, 2018; Guo *et al.*, 2020; Long *et al.*, 2022; Yan *et al.*, 2022; Chen *et al.*, 2023]. In most real-world scenarios, attackers lack direct access to the deployed target models. To this end, numerous adversarial attacks have been introduced, aiming to boost the transferability of the adversarial examples crafted by surrogate models [Xie *et al.*, 2019;

Wang and He, 2021; Long *et al.*, 2022; Wang *et al.*, 2023; Wang *et al.*, 2024]. Despite significant progress in this field, previous adversarial attack methods often overlook improving the transferability on FAS models using intermediate loss functions without relying on specific reference images. To overcome this limitation, we leverage the linear hypothesis to approximate segments of FAS models and bias adversarial examples toward the live distribution using intermediate loss avoiding overfitting to specific reference images, thereby improving the black-box attack effectiveness on FAS models. Additionally, the challenge of reducing feature discrepancies across various levels of representation in FR models remains largely underexplored. To address this, we propose a novel attack method that aligns the features of adversarial examples with those of target images across multiple intermediate layers, thereby enhancing transferability of the crafted adversarial examples on FR models.

Adversarial Attacks on Face Recognition Systems. Adversarial attacks on FR systems can be categorized into two groups based on the constraints imposed on adversarial perturbations: restricted attacks [Dong *et al.*, 2018; Chen *et al.*, 2023] and unrestricted attacks [Wei *et al.*, 2023a; Wei *et al.*, 2023b]. Restricted attacks involve generating adversarial examples within a predefined boundary, such as the L_p norm constraint. The primary objective of restricted attacks is to improve the transferability of adversarial face examples across different FR models [Zhong and Deng, 2021; Zhou *et al.*, 2024b; Li *et al.*, 2023b]. In contrast, unrestricted adversarial attacks generate adversarial examples without adhering to a specific perturbation constraint. These attacks typically focus on physical-world scenarios [Xiao *et al.*, 2021; Yang *et al.*, 2023; Li *et al.*, 2023a], attribute manipulation [Qiu *et al.*, 2020; Jia *et al.*, 2022], or adversarial example generation through makeup transfer [Yin *et al.*, 2021; Hu *et al.*, 2022; Shamshad *et al.*, 2023]. Both restricted and unrestricted adversarial attacks have significantly advanced the effectiveness of adversarial attacks on FR systems. Nevertheless, in practical applications, adversarial examples generated by the methods are frequently classified as spoof images by the FAS models. This limitation significantly impacts the practicality of adversarial attacks on FR systems. In this paper, we introduce a novel attack method that simultaneously targets both FR and FAS models, thereby bolstering the effectiveness of adversarial attacks on the integrated FR systems.

3 Methodology

3.1 Problem Formulation and Framework Overview

Problem Formulation. Let $\mathcal{F}^{vct}(\mathbf{x})$ denote the FR model deployed by a victim to extract the embedding from a face image $\mathbf{x}$, and let $\mathcal{G}^{vct}(\mathbf{x})$ represent the FAS model deployed by a victim that outputs a score to determine the authenticity of the image. We denote $\mathbf{x}^s$ and $\mathbf{x}^t$ as the source and target images, respectively. In most cases, the source images are spoof images, and the target images are live images in real-world attack scenarios. Hence, we initialize $\mathbf{x}^s$ with the spoof image captured in the physical world, and $\mathbf{x}^t$ with the live image, respectively. The objective of adversarial attacks on FR in our research is to generate an adversarial example $\mathbf{x}^{adv}$ that induces the victim FR model $\mathcal{F}^{vct}$ to misclassify it as the target sample $\mathbf{x}^t$, while also preserving a high level of visual similarity between $\mathbf{x}^{adv}$ and $\mathbf{x}^s$. Specifically, the objective can be stated as follows:

$$\begin{aligned}\mathbf{x}^{adv} = \underset{\mathbf{x}^{adv}}{\arg\min} \left(\mathcal{D}\left(\mathcal{F}^{vct}\left(\mathbf{x}^{adv}\right), \mathcal{F}^{vct}\left(\mathbf{x}^{t}\right)\right)\right) \\ \text{s.t.} \|\mathbf{x}^{adv} - \mathbf{x}^{s}\|_p \leq \epsilon\end{aligned} \tag{1}$$

where $\mathcal{D}$ refers to a predefined distance metric, while ϵ specifies the maximum magnitude of permissible perturbation. In contrast, the objective of the adversarial attacks on FAS in this study is to deceive the victim FAS model $\mathcal{G}^{vct}$ into identifying the adversarial example $\mathbf{x}^{adv}$ as a living image, while ensuring that $\mathbf{x}^{adv}$ maintains a visually similar appearance from $\mathbf{x}^s$. Concisely, the objective can be formulated as:

$$\mathbf{x}^{adv} = \underset{\mathbf{x}^{adv}}{\arg\max} \left(\mathcal{G}^{vct}\left(\mathbf{x}^{adv}\right)\right) \quad \text{s.t.} \|\mathbf{x}^{adv} - \mathbf{x}^{s}\|_p \leq \epsilon \tag{2}$$

Framework Overview. In the realm of adversarial attacks on FR systems, the property of adversarial examples being readily identifiable by FAS models significantly hinders the practicality of such examples. To address this issue, we propose a novel attack method called Reference-free Multi-level Alignment (RMA) to attack FR and FAS models simultaneously. The overview of our proposed attack method is illustrated in Figure 2, showcasing three key modules: Adaptive Gradient Maintenance (AGM), Reference-free Intermediate Biasing (RIB), and Multi-level Feature Alignment (MFA). In the following sections, we will provide a detailed introduction to each of these three modules.

3.2 Adaptive Gradient Maintenance

Let $\mathcal{L}^f$ and $\mathcal{L}^g$ be the loss functions of the FR and FAS models, respectively. To craft the adversarial examples targeting both the FR and FAS tasks, a vanilla method is to compute the gradient using the following formula:

$$\mathbf{g}'_t = \nabla_{\mathbf{x}^{adv}} \mathcal{L}^f\left(\mathbf{x}_t^{adv}\right) + \nabla_{\mathbf{x}^{adv}} \mathcal{L}^g\left(\mathbf{x}_t^{adv}\right) \tag{3}$$

where $\mathbf{g}'_t$ and $\mathbf{x}_t^{adv}$ represent the gradient and adversarial examples at the t-th iteration, respectively. However, due to the imbalances between the magnitudes of $\mathcal{L}^f$ and $\mathcal{L}^g$, the gradients computed for the FR and FAS models become unbalanced, which negatively impacts the performance of the adversarial examples. To address this, we propose a novel module termed as Adaptive Gradient Maintenance (AGM), which adaptively mitigates the disparities of the gradients between FR and FAS during each iteration.

To mitigate the gradient discrepancies between FR and FAS, we dynamically adjust the degree of loss reduction for the two tasks. Specifically, the degree of loss reduction for the FR and FAS tasks can be expressed as follows:

$$\begin{aligned} d' &= \mathcal{L}^f\left(\mathbf{x}_t^{adv}\right) - \mathcal{L}^f\left(\mathbf{x}_1^{adv}\right) \\ d^* &= \mathcal{L}^g\left(\mathbf{x}_t^{adv}\right) - \mathcal{L}^g\left(\mathbf{x}_1^{adv}\right) \end{aligned} \tag{4}$$

After obtaining the degree of loss reduction for both tasks, we re-weight the losses of the FR and FAS models to reduce their

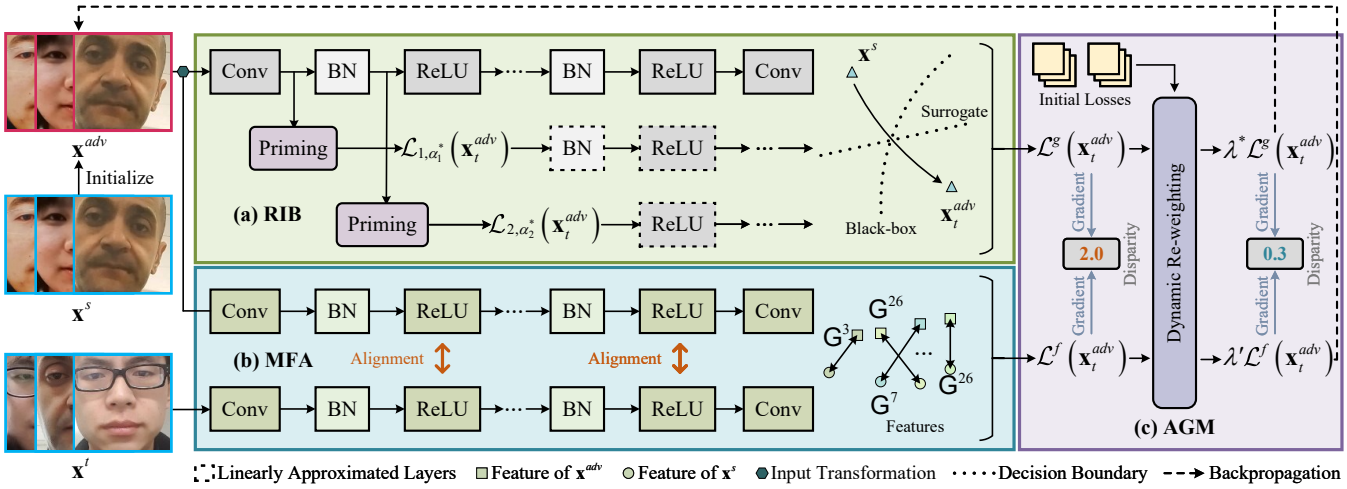

Figure 2: Overview of our Reference-free Multi-level Alignment (RMA) framework. (a) The Reference-free Intermediate Biasing (RIB) module biases adversarial examples toward the live image space using the intermediate loss function without overfitting to specific reference live images by a surrogate model accessible to the attacker, enhancing the attack effectiveness on black-box FAS models where direct access is restricted. (b) The Multi-level Feature Alignment (MFA) module aligns the features of adversarial examples with those of target images across multiple intermediate layers, thereby improving their transferability when attacking FR models. (c) The Adaptive Gradient Maintenance (AGM) module balances the gradients between FR and FAS by adaptively scaling their respective losses, thereby mitigating the disparities between the gradients the during each iteration.

disparities. The resulting optimization objective is as follows:

$$\begin{aligned}\mathbf{x}^{adv} &= \arg\max \mathcal{L}^f\left(\mathbf{x}^{adv}\right) + \frac{d'}{d^*}\mathcal{L}^g\left(\mathbf{x}^{adv}\right) \\ &= \arg\max \lambda'\mathcal{L}^f\left(\mathbf{x}^{adv}\right) + \lambda^*\mathcal{L}^g\left(\mathbf{x}^{adv}\right)\end{aligned} \tag{5}$$

where λ' and λ^* are the normalized loss weights:

$$\lambda' = \frac{d^*}{d^* + d' + \epsilon'}, \quad \lambda^* = \frac{d'}{d^* + d' + \epsilon'} \tag{6}$$

Here, ϵ' is a small constant to ensure numerical stability.

Using the optimization objective in the equation above, we compute the balanced gradient $\mathbf{g}$ using the following formula by dynamically re-weighting the loss magnitude:

$$\mathbf{g}_t = \nabla_{\mathbf{x}^{adv}}\lambda'\mathcal{L}^f\left(\mathbf{x}_t^{adv}\right) + \nabla_{\mathbf{x}^{adv}}\lambda^*\mathcal{L}^g\left(\mathbf{x}_t^{adv}\right) \tag{7}$$

After obtaining $\mathbf{g}_t$, we use $\mathbf{g}_t$ to update the adversarial examples according to the following formula:

$$\mathbf{x}_{t+1}^{adv} = \prod_{\mathbf{x}^s,\epsilon}\left(\mathbf{x}_t^{adv} - \text{sign}\left(\mathbf{g}_t\right)\right) \tag{8}$$

3.3 Reference-free Intermediate Biasing

Unlike FR, which is a metric learning task, the FAS task involves the model directly producing a score to determine the liveness of the input face image. Once the liveness score is obtained, it is compared against a predefined threshold. If the score exceeds the threshold, the input face image is classified as a live image, otherwise a spoof image.

Let $\mathcal{G}$ denote the FAS surrogate model. The optimization objective of the attack on FAS using $\mathcal{G}$ can be expressed as:

$$\max_{\mathbf{x}^{adv}} \mathcal{G}\left(\mathbf{x}^{adv}\right) \quad \text{s.t.} \quad \|\mathbf{x}^{adv} - \mathbf{x}^s\|_p \leq \epsilon \tag{9}$$

where ϵ specifies the maximum perturbation magnitude.

A vanilla method to achieve Equation (9) is to utilize the following loss function:

$$\mathcal{L}^s\left(\mathbf{x}^{adv}\right) = -\mathcal{G}\left(\mathbf{x}^{adv}\right) \tag{10}$$

where $\mathcal{L}^s$ denotes the loss function for generating adversarial examples against FAS models using the method. Equation (10) is effective in crafting adversarial examples for FAS. However, as demonstrated by previous methods [Jia *et al.*, 2022], using a loss function based on intermediate layers to craft adversarial examples can further improve transferability. Although these adversarial attacks which leverage intermediate layer loss functions demonstrate promising effectiveness, no such attack has been specifically designed to target FAS models to the best of our knowledge.

For clarity, we focus on models with a single branch in their computational graphs. For models with multiple branches, our RMA remains largely unchanged, with minor modifications to accommodate the handling of multiple branches. Let g^i denote the i-th layer of the model $\mathcal{G}$ and let l represent the total number of layers in $\mathcal{G}$. We define the segment of $\mathcal{G}$ from layer g^i to layer g^j as follows [Zhou *et al.*, 2025]:

$$\mathcal{G}^{i,j} = g^i \circ g^{i+1} \circ \cdots \circ g^{j-1} \circ g^j, \tag{11}$$

where $\circ$ denotes the composition of functions operation. Let $\mathsf{G}^k(\mathbf{x})$ represent $\mathcal{G}^{1,k}(\mathbf{x})$. To craft adversarial examples for FAS models using an intermediate layer loss function, a straightforward method is to use the following loss function:

$$\mathcal{L}_k'\left(\mathbf{x}^{adv}\right) = \left\|\mathsf{G}^k\left(\mathbf{x}^{adv}\right) - \mathsf{G}^k\left(\mathbf{x}^*\right)\right\|_2 \tag{12}$$

where $\mathbf{x}^*$ denotes a pre-selected reference live image. However, the adversarial examples crafted by Equation (12) may

overfit to $\mathbf{x}^*$. If adversarial examples for FAS can be crafted using an intermediate layer loss function without relying on specific reference live images, better transferability can be achieved (See Figure 4).

However, leveraging $\mathsf{G}^k(\mathbf{x}^{adv})$ solely to maximize the live score produced by $\mathcal{G}$ in the final outputs poses a significant challenge due to the absence of the $\mathcal{G}^{k+1,l}$. To address this, we introduce the linear hypothesis proposed by [Goodfellow *et al.*, 2015]. Let us revisit an early hypothesis posited by [Goodfellow *et al.*, 2015], which suggests that the linear nature of modern deep neural networks, resembling linear models trained on the same dataset, is the underlying cause of adversarial examples and their surprising transferability [Guo *et al.*, 2020; Zhou *et al.*, 2024c]. Let m_k denote the number of units in $\mathsf{G}^k(\mathbf{x}^{adv})$. Based on this hypothesis, the FAS model $\mathcal{G}^{k+1,l}$ can be linearly approximated as:

$$\mathcal{G}^{k+1,l}\left(\mathsf{G}^k\left(\mathbf{x}^{adv}\right)\right) = \mathbf{w}^{k+1,l}\Psi\left(\mathsf{G}^k\left(\mathbf{x}^{adv}\right)\right)^\top + \mathbf{b}^{k+1,l} \tag{13}$$

where Ψ denotes the flatten operation, $\mathbf{w}^{k+1,l} \in \mathbb{R}^{1\times m_k}$ is the weight vector, and $\mathbf{b}^{k+1,l} \in \mathbb{R}^{1\times 1}$ is the bias term. Let $\mathbf{h}^k$ be defined as $\mathbf{h}^k = \Psi\left(\mathsf{G}^k\left(\mathbf{x}^{adv}\right)\right)^\top \in \mathbb{R}^{m_k\times 1}$. Based on Equation (13), the objective for the FAS model can be expressed as the following formula:

$$\max_{\mathbf{x}^{adv}} \mathcal{G}\left(\mathbf{x}^{adv}\right) = \max_{\mathbf{x}^{adv}} \mathbf{w}^{k+1,l}\mathbf{h}^k \tag{14}$$

Let t be the number of the optimization iterations. Since $\mathbf{w}^{k+1,l}$ is intractable, we cannot use $-\mathbf{w}^{k+1,l}\mathbf{h}^k$ as the loss function to optimize Equation (14). Instead, we design an optimization process that ensures the following condition:

$$\mathbf{w}^{k+1,l}\mathbf{h}_t^k - \mathbf{w}^{k+1,l}\mathbf{h}_{t-1}^k = \mathbf{w}^{k+1,l}\left(\mathbf{h}_t^k - \mathbf{h}_{t-1}^k\right) > 0 \tag{15}$$

where $\mathbf{h}_t^k$ denotes the value of $\mathbf{h}^k$ at the t-th optimization iteration. In our research, the parameters of the FAS model are fixed, meaning that $\mathbf{w}^{k+1,l}$ is a vector with constant values. Therefore, we opt to satisfy Equation (15) by minimizing or maximizing $\mathbf{h}^k$ using the following loss function:

$$\widehat{\mathcal{L}}_{k,\alpha_k}\left(\mathbf{x}^{adv}\right) = \frac{\alpha_k}{m_k}\sum_{j=0}^{m_k}\mathbf{h}^{k,j} \quad \text{s.t} \quad \alpha_k \in \{-1, 1\} \tag{16}$$

where $\mathbf{h}^{k,j}$ denotes the j-th element in $\mathbf{h}^k$. The value of α_k varies across different layers depending on layer index k. To determine α_k, we introduce a stage termed as Prime, which records the value of $\mathcal{L}^s$ obtained by crafting adversarial examples using Equation (16) with both candidate α_k values (*i.e.* -1 or 1). The α_k value corresponding to the lower $\mathcal{L}^s$ is then selected as the final α_k. Specifically, let $\mathbf{x}'$ represent the adversarial examples generated during the Priming stage, initialized with the same values as $\mathbf{x}^s$. Note that $\mathbf{x}'$ is only used to calculate the optimization direction. After the Priming stage, $\mathbf{x}'$ is discarded and is not involved in the final adversarial example generation process. $\mathbf{x}'$ is optimized over b iterations using the following formula [Kurakin *et al.*, 2017]:

$$\mathbf{x}'_{\alpha_k,t} = \prod_{\mathbf{x}^s,\epsilon}\left(\mathbf{x}'_{\alpha_k,t-1} - \text{sign}\left(\nabla_{\mathbf{x}'_{\alpha_k,t-1}}\widehat{\mathcal{L}}_{k,\alpha_k}\left(\mathbf{x}'_{\alpha_k,t-1}\right)\right)\right) \tag{17}$$

where $\mathbf{x}'_{\alpha_k,t}$ denotes the adversarial example crafted in the t-th iteration using α_k to compute the loss function, $\prod_{\mathbf{x}^s,\epsilon}$ represents the clipping operation that ensures the distance between the crafted adversarial example $\mathbf{x}'$ and the source image $\mathbf{x}^s$ remains within ϵ. During the optimization process, we record the $\mathcal{L}^s$ loss using Equation (10) and compute the average loss values as follows:

$$\bar{\mathcal{L}}_{k,\alpha_k} = \frac{1}{b}\sum_{t=1}^{b}\mathcal{L}^s\left(\mathbf{x}'_{\alpha_k,t}\right) \tag{18}$$

Using Equation (18), we determine the optimal α_k for $\widehat{\mathcal{L}}_{k,\alpha_k}$ by the following formula:

$$\alpha_k^* = \arg\min \bar{\mathcal{L}}_{k,\alpha_k} \tag{19}$$

Note that if the layer is the final output layer of the FAS model, we can directly judge the α_k value of $\widehat{\mathcal{L}}_{k,\alpha_k}$ and do not need to process the Prime stage to calculate it. We use the loss in multiple layers to craft the adversarial examples:

$$\mathcal{L}^g\left(\mathbf{x}^{adv}\right) = \sum_{i=1}^{|s|}\widehat{\mathcal{L}}_{s_i,\alpha^*_{s_i}}\left(\mathbf{x}^{adv}\right) \tag{20}$$

where s is the pre-defined layer index set to calculate the loss for the FAS task.

3.4 Multi-level Feature Alignment

Let $\mathcal{F}$ represent the FR surrogate model. To deceive FR models, a vanilla method is to utilize the following loss function to generate adversarial examples [Zhong and Deng, 2021; Zhou *et al.*, 2024b]:

$$\mathcal{L}^i\left(\mathbf{x}^{adv}\right) = \left\|\phi\left(\mathcal{F}\left(\mathbf{x}^{adv}\right)\right) - \phi\left(\mathcal{F}\left(\mathbf{x}^t\right)\right)\right\|_2^2 \tag{21}$$

where $\phi(\mathbf{x})$ represents the operation that normalizes $\mathbf{x}$. However, if we rely solely on Equation (21) to craft adversarial face examples, the attack success rate will be limited, as only the final feature of is utilized.

To improve the transferability on FR models, we propose the Multi-level Feature Alignment (MFA) module. MFA enhances the transferability of adversarial attacks by aligning the features of adversarial examples with those of target images across multiple intermediate layers. To the best of our knowledge, MFA is the first algorithm to employ a multi-level intermediate loss function for enhancing the transferability of the adversarial examples.

Let e denote the index set of pre-selected layers used for calculating the loss to align the intermediate features, and let $\mathsf{F}^k(\mathbf{x})$ represent the feature of the k-th layer. For each layer specified by e, we use the following formula to compute the loss, which aims to align the features across different levels of the intermediate layers:

$$\mathcal{L}^f\left(\mathbf{x}^{adv}\right) = \sum_{i}^{|e|}\left\|\phi\left(\Psi\left(\mathsf{F}^{e_i}\left(\mathbf{x}^{adv}\right)\right)\right) - \mathbf{f}\right\|_2^2 \tag{22}$$

where $\mathbf{f}$ denotes $\phi\left(\Psi\left(\mathsf{F}^{e_i}\left(\mathbf{x}^t\right)\right)\right)$.

	ASR'			ASR^*		ASR^J					
Attacks	IR152	IRSE50	FaceNet	IADG	AMEL	IR152′	IRSE50′	FaceNet′	IR152*	IRSE50*	FaceNet*
FIM	23.5	20.9	25.0	5.3	3.0	1.1	1.1	1.5	1.0	0.9	0.9
DI	34.5	32.6	41.5	3.3	2.3	1.2	1.0	1.4	1.0	0.7	1.2
DFANet	30.9	26.7	32.2	5.2	2.4	1.5	1.5	1.5	1.0	0.9	1.0
VMI	36.3	29.2	34.8	2.3	1.9	1.0	0.7	0.8	0.8	0.7	0.7
SSA	31.8	27.1	35.4	4.0	2.5	1.3	1.3	1.2	0.9	1.2	1.0
SIA	38.9	33.3	43.8	2.3	1.5	1.1	0.9	1.0	0.6	0.7	0.7
BPFA	34.0	31.4	37.3	3.8	2.3	1.5	1.3	1.4	0.9	1.1	1.0
BSR	25.8	22.8	36.4	2.1	1.5	0.5	0.5	0.6	0.6	0.6	0.5
Ours	**61.3**	**55.7**	**57.5**	**94.4**	**41.0**	**58.7**	**53.2**	**55.6**	**29.1**	**27.0**	**26.8**

Table 1: Comparisons of ASR (%) results for adversarial attacks.

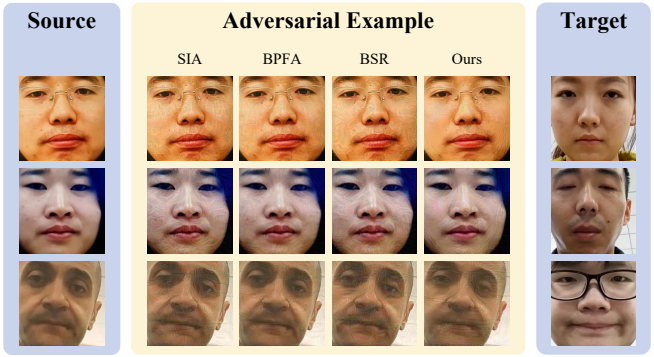

Figure 3: The illustration of the crafted adversarial examples.

	ASR'			ASR^J
Attacks	$IR152^{adv}$	$IRSE50^{adv}$	$FaceNet^{adv}$	Average
FIM	27.4	35.9	31.8	1.4
DI	43.9	53.5	49.5	1.8
DFANet	35.5	43.9	37.7	1.4
VMI	43.0	52.9	41.6	1.0
SSA	40.3	49.3	44.5	1.6
SIA	45.9	55.3	52.5	1.1
BPFA	41.2	48.6	45.2	1.7
BSR	31.1	41.0	44.7	0.8
Ours	**70.9**	**74.0**	**63.3**	**49.3**

Table 2: ASR (%) results on adversarial robust models.

4 Experiment

In our experiments, we demonstrate the superiority and key properties of the proposed method. Section 4.1 details the experimental settings, while Section 4.2 presents the comparative results. Additionally, Section 4.3 provides an analysis of the ablation studies.

4.1 Experimental Settings

We use the Oulu-NPU [Boulkenafet *et al.*, 2017] and CASIA-MFSD [Zhang *et al.*, 2012] for evaluation. We randomly sample 1,000 negative image pairs from both datasets. To align with practical attack scenarios, we select spoof images captured by cameras in the physical-world as the source images and live images as the target images. We employ the Attack Success Rate (ASR) as the primary metric to evaluate the effectiveness of adversarial examples. Following previous works such as [Yin *et al.*, 2021; Hu *et al.*, 2022; Zhou *et al.*, 2024b; Zhou *et al.*, 2024c], we selected IR152 [He *et al.*, 2016], IRSE50 [Hu *et al.*, 2018], FaceNet [Schroff *et al.*, 2015], and MobileFace (abbreviated as MF) [Deng *et al.*, 2022] as the models to assess the attack performance on FR. we selected restricted attacks on FR systems DFANet [Zhong and Deng, 2021] and BPFA [Zhou *et al.*, 2024b], as well as state-of-the-art transfer attacks DI [Xie *et al.*, 2019], SSA [Long *et al.*, 2022], SIA [Wang *et al.*, 2023], and BSR [Wang *et al.*, 2024], as our baseline for comparison.

4.2 Comparison Studies

RMA achieves the best black-box attack results across various ASR metrics. Let N denote the name of a target FR model. In the following, we use N' and N^* to represent the ASR^J results on the IADG and AMEL models, respectively. We employ MF as the surrogate model, and craft adversarial examples on the OULU-NPU. The results are presented in Table 1. Some of the crafted adversarial examples are demonstrated in Figure 3. These results demonstrate that our proposed method significantly outperforms previous attack methods across the ASR', ASR^*, and ASR^J metrics, underscoring the effectiveness of our proposed attack method.

RMA delivers superior black-box performance on adversarial robust models. In practical scenarios, victims may employ adversarial robust models to defend against adversarial attacks. Therefore, assessing the effectiveness of adversarial attacks on robust models is crucial. In this study, we generate adversarial examples using MF as the surrogate model and evaluate the performance of various attacks on adversarial robust models. The results, presented in Table 2, demonstrate that our proposed method significantly outperforms baseline adversarial attacks. These findings underscore the effectiveness of our approach against adversarial robust models.

4.3 Ablation Studies

The individual contributions of each module in our proposed method. In this section, we conduct ablation studies to evaluate the individual contributions of each module in our proposed method. Using FIM as the baseline, we incrementally integrate each module and analyze their impact. The ablation studies are performed using MF as surrogate models on

the OULU-NPU dataset. We evaluate the ASR^J results with IR152, IRSE50, and FaceNet as the target FR models and AMEL as the target FAS model. The results are presented in Table 3. The baseline ASR^J values are 1.0%, 0.9%, and 0.9% for IR152, IRSE50, and FaceNet, respectively. By incorporating the MFA module, the performance improves to 1.2%, 1.5%, and 1.2%, respectively. The addition of the RIB module further enhances the results, achieving ASR^J values of 14.7%, 14.0%, and 13.7% for IR152, IRSE50, and FaceNet, respectively. Finally, integrating the AGM module significantly increases the ASR^J to 29.1%, 27.0%, and 26.8% for IR152, IRSE50, and FaceNet, respectively. These substantial improvements demonstrate the effectiveness of each module in our proposed method, collectively contributing to a significant enhancement in overall performance.

MFA	RIB	AGM	IR152	IRSE50	FaceNet
-	-	-	1.0	0.9	0.9
✓	-	-	1.2	1.5	1.2
✓	✓	-	14.7	14.0	13.7
✓	✓	✓	**29.1**	**27.0**	**26.8**

Table 3: Comparisons of ASR^J (%) results with AMEL as the target model on the OULU-NPU dataset.

The Effectiveness of Reference-free Intermediate Biasing. To evaluate the effectiveness of our proposed RIB module, we conduct experiments using IADG as the surrogate model and AMEL as the target model on the OULU-NPU dataset. First, we craft adversarial examples using $\mathcal{L}^s$, referring to this attack method as Vanilla. Second, we craft adversarial examples using $\mathcal{L}'_k$, denoting this method as Reference-specific (abbreviated as RS). Finally, we craft adversarial examples using $\mathcal{L}^g$, referring to this method as RIB. We evaluate the live score of the AMEL model and the ASR^* performance on the AMEL model for the three attack methods. The results for these attack methods are illustrated in Figure 4.

The left plot in Figure 4 shows that the proportion of high live scores achieved by RIB is greater than that of Vanilla and RS, indicating that adversarial examples crafted by our proposed RIB module are more likely to succeed in attacks. The right plot in Figure 4 demonstrates that the ASR^* results of our proposed RIB method surpass those of Vanilla and RS, further validating the effectiveness of the RIB module.

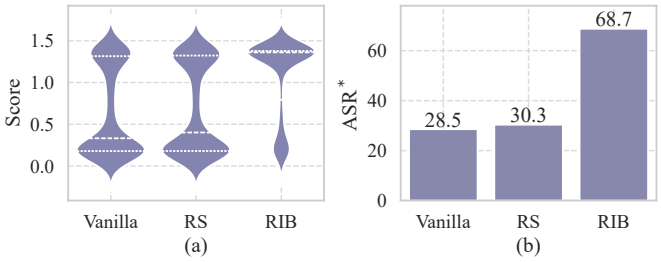

Figure 4: Ablation studies on the RIB module: (a) Violin plot illustrating the black-box live scores. (b) ASR^* (%) results.

The Effectiveness of Multi-level Feature Alignment. To evaluate the effectiveness of our proposed Multi-level Feature Alignment (MFA) module, we conduct a series of experiments using MF as the surrogate model on the OULU-NPU dataset. First, we craft adversarial examples using $\mathcal{L}^i$ and refer to this attack method as Vanilla. Second, we generate adversarial examples using the following loss function, which operates on a single intermediate layer:

$$\widetilde{\mathcal{L}}\left(\mathbf{x}^{adv}\right) = \left\|\phi\left(\Psi\left(\mathsf{F}^r\left(\mathbf{x}^{adv}\right)\right)\right) - \phi\left(\Psi\left(\mathsf{F}^r\left(\mathbf{x}^t\right)\right)\right)\right\|_2^2 \tag{23}$$

where r denotes the index of the pre-defined layer for calculating the loss. We denote this attack method as Single-level. Finally, we craft the adversarial examples using $\mathcal{L}^f$, referring to this attack method as MFA. The results of these three attack methods are presented in Table 4.

	IR152	IRSE50	FaceNet	ASR^{adv}
Vanilla	23.5	20.9	25.0	31.7
Single-level	34.2	33.9	27.9	35.7
MFA (Ours)	**62.3**	**58.4**	**45.8**	**63.3**

Table 4: Ablation study results (ASR', %) for MFA. ASR^{adv} is the average black-box ASR' on IR152^{adv}, IRSE50^{adv}, and FaceNet^{adv}.

Table 4 demonstrates that although the performance of Single-level surpasses that of Vanilla, it remains inferior to MFA, highlighting the effectiveness of the multi-level intermediate loss and the proposed MFA module.

5 Conclusion

In this paper, we introduce a novel and practical setting that aims to simultaneously attack both the FR and FAS models. To achieve this goal, we propose an innovative framework RMA, consisting of three modules: Adaptive Gradient Maintenance (AGM), Reference-free Intermediate Biasing (RIB), and Multi-level Feature Alignment (MFA) modules. Firstly, to alleviate the unbalance between the gradients of the FR and FAS models, we introduce the Adaptive Gradient Maintenance module, which balances the gradients on FR and FAS by adaptively re-weighting the loss on FR and FAS to decrease their disparity in each iteration. Furthermore, to enhance the transferability of FAS models, we design the Reference-free Intermediate Biasing module to bias the adversarial examples into the space of the live image using intermediate loss without overfiting to specific reference live images. In addition, the Multi-level Feature Alignment module is proposed to boost the capacity of black-box attacks on FR models. Extensive experiments demonstrate the effectiveness of our proposed attack method.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62372203 and 62302186, in part by the Major Scientific and Technological Project of Shenzhen (202316021), in part by the National key research and development program of China (2022YFB2601802), in part by the Major Scientific and Technological Project of Hubei Province (2022BAA046, 2022BAA042).

References

[An *et al.*, 2021] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial FC: training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1445–1449, 2021.

[Boulkenafet *et al.*, 2017] Zinelabdine Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *International Conference on Automatic Face & Gesture Recognition*, pages 612–618, 2017.

[Chen *et al.*, 2023] Bin Chen, Jia-Li Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4466–4475, 2023.

[Cherepanova *et al.*, 2021] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations*, 2021.

[Deng *et al.*, 2022] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022.

[Dong *et al.*, 2018] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.

[Goodfellow *et al.*, 2015] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[Guo *et al.*, 2020] Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. *Advances in neural information processing systems*, 33:85–95, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[Hu *et al.*, 2022] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14994–15003, 2022.

[Jia *et al.*, 2022] Shuai Jia, Bangjie Yin, Taiping Yao, Shouhong Ding, Chunhua Shen, Xiaokang Yang, and Chao Ma. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. In *Advances in Neural Information Processing Systems*, 2022.

[Kurakin *et al.*, 2017] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representation*, 2017.

[Li *et al.*, 2023a] Yanjie Li, Yiquan Li, Xuelong Dai, Songtao Guo, and Bin Xiao. Physical-world optical adversarial attacks on 3d face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24699–24708, 2023.

[Li *et al.*, 2023b] Zexin Li, Bangjie Yin, Taiping Yao, Junfeng Guo, Shouhong Ding, Simin Chen, and Cong Liu. Sibling-attack: Rethinking transferable adversarial attacks against face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24626–24637, 2023.

[Long *et al.*, 2022] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *European Conference on Computer Vision*, volume 13664, pages 549–566, 2022.

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

[Rocamora *et al.*, 2024] Elias Abad Rocamora, Yongtao Wu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. Revisiting character-level adversarial attacks for language models. In *International Conference on Machine Learning*, 2024.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[Shamshad *et al.*, 2023] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20595–20605, 2023.

[Shayegani *et al.*, 2024] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *International Conference on Learning Representations*, 2024.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[Wang and He, 2021] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021.

[Wang *et al.*, 2023] Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4607–4619, 2023.

[Wang *et al.*, 2024] Kunyu Wang, Xuanran He, Wenxuan Wang, and Xiaosen Wang. Boosting Adversarial Transferability by Block Shuffle and Rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Wei *et al.*, 2023a] Xingxing Wei, Yao Huang, Yitong Sun, and Jie Yu. Unified adversarial patch for cross-modal attacks in the physical world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4422–4431, 2023.

[Wei *et al.*, 2023b] Xingxing Wei, Jie Yu, and Yao Huang. Physically adversarial infrared patches with learnable shapes and locations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12334–12342, 2023.

[Xiao *et al.*, 2021] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11840–11849, 2021.

[Xie *et al.*, 2019] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019.

[Xu *et al.*, 2023] Zhuoer Xu, Zhangxuan Gu, Jianping Zhang, Shiwen Cui, Changhua Meng, and Weiqiang Wang. Backpropagation path search on adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4663–4673, 2023.

[Yan *et al.*, 2022] Chiu Wai Yan, Tsz-Him Cheung, and Dit-Yan Yeung. Ila-da: Improving transferability of intermediate level attack with data augmentation. In *International Conference on Learning Representations*, 2022.

[Yang *et al.*, 2021] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3877–3887, 2021.

[Yang *et al.*, 2023] Xiao Yang, Chang Liu, Longlong Xu, Yikai Wang, Yinpeng Dong, Ning Chen, Hang Su, and Jun Zhu. Towards effective adversarial textured 3d meshes on physical face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4119–4128, 2023.

[Yin *et al.*, 2021] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 1252–1258, 2021.

[Zhang *et al.*, 2012] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z. Li. A face antispoofing database with diverse attacks. In *5th IAPR International Conference on Biometrics*, pages 26–31, 2012.

[Zhong and Deng, 2021] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2021.

[Zhou *et al.*, 2022] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Kekai Sheng, Shouhong Ding, and Lizhuang Ma. Generative domain adaptation for face anti-spoofing. In *European Conference on Computer Vision*, pages 335–356, 2022.

[Zhou *et al.*, 2024a] Fengfan Zhou, Hefei Ling, Yuxuan Shi, Jiazhong Chen, and Ping Li. Improving visual quality and transferability of adversarial attacks on face recognition simultaneously with adversarial restoration. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4540–4544, 2024.

[Zhou *et al.*, 2024b] Fengfan Zhou, Hefei Ling, Yuxuan Shi, Jiazhong Chen, Zongyi Li, and Ping Li. Improving the transferability of adversarial attacks on face recognition with beneficial perturbation feature augmentation. *IEEE Transactions on Computational Social Systems*, 11(6):8130–8142, 2024.

[Zhou *et al.*, 2024c] Fengfan Zhou, Qianyu Zhou, Bangjie Yin, Hui Zheng, Xuequan Lu, Lizhuang Ma, and Hefei Ling. Rethinking impersonation and dodging attacks on face recognition systems. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2487–2496, 2024.

[Zhou *et al.*, 2024d] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Shouhong Ding, and Lizhuang Ma. Test-time domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–187, 2024.

[Zhou *et al.*, 2025] Fengfan Zhou, Bangjie Yin, Hefei Ling, Qianyu Zhou, and Wenxuan Wang. Improving the transferability of adversarial attacks on face recognition with diverse parameters augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.