Prompting Adversarial Transferability via Path Flatness Attack

Zeze Tao[1], Jinjia Peng[1,*], Huibing Wang[2]

[1]School of Cyber Security and Computer, Hebei University
[2]College of Information and Science Technology, Dalian Maritime University
{zeze, pengjinjia}@hbu.edu.cn, huibing.wang@dlmu.edu.cn

Abstract

Deep neural networks are susceptible to adversarial examples, which induce incorrect predictions through imperceptible perturbations. Transfer-based attacks create adversarial examples for surrogate models and transfer these examples to target models under black-box scenarios. Recent studies have established a strong correlation between the geometric properties of loss landscapes and the transferability of adversarial examples, demonstrating that flatter loss surfaces consistently yield superior transferability. However, we identify that these methods fail to account for the loss landscape flatness along the path from the current point to local minima, resulting in poor transferability. To address this, this paper constructs a novel Path Flatness Attack (PFA) method to significantly enhance the transferability of adversarial examples. Specifically, this paper proposes a novel path flatness indicator that not only evaluates the flatness in local minima regions but also explicitly quantifies the loss surface geometry along the trajectory from the current point to the minimum. Furthermore, we incorporate the path flatness indicator into the attack process, integrating penalties over low-loss points along the path while maximizing the loss function, thereby explicitly flattening the loss landscape. Extensive experiments demonstrate that PFA consistently achieves state-of-the-art attack performance across all experimental settings.

Code — https://github.com/ZezeTao/PFA

Introduction

Deep neural networks (DNNs) (Ge et al. 2023; Dong et al. 2018; Peng et al. 2025a; Wang et al. 2025) have demonstrated remarkable performance in computer vision tasks, yet they remain vulnerable to adversarial examples. These examples can mislead models into producing incorrect outputs by injecting imperceptible perturbations into clean samples. More critically, adversarial examples exhibit the transferability property—that is, adversarial samples generated against known models can also deceive unknown black-box models. This property renders transfer-based attacks practically threatening in real-world scenarios, posing significant risks to safety-critical applications (Zhou et al. 2025; Kong et al. 2020; Zhang et al. 2025). Consequently, generating highly transferable adversarial examples has become an increasingly important research direction.

Adversarial attacks are primarily categorized as white-box (Goodfellow, Shlens, and Szegedy 2015; Kurakin, Goodfellow, and Bengio 2018) and black-box attacks (Dong et al. 2018; Wang et al. 2024). In the white-box scenario, the attacker has complete knowledge of the target model, including its architecture, parameter weights, and loss function. Conversely, in the black-box scenario, the attacker first crafts adversarial samples using a surrogate model, then transfers these samples to the actual black-box target model. From a practical application perspective, since deep learning models typically do not disclose their internal implementation details to end users, the black-box attack approach is consequently more feasible in real-world scenarios. However, it is important to note that due to the high complexity of over-parameterized deep neural networks, adversarial samples generated on surrogate models often suffer from severe overfitting, ultimately resulting in poor transferability to black-box target models.

To mitigate overfitting and improve the transferability of adversarial examples, the academic community has developed various effective strategies, primarily categorized into three technical paradigms: gradient-based optimization attacks (Wang and He 2021; Zhu et al. 2023; Fang et al. 2024), input transformation-based attacks (Xie et al. 2019; Wang et al. 2021), and model ensemble-based attacks (Che et al. 2020; Chen et al. 2023). Most significantly, recent theoretical and empirical studies (Qin et al. 2022; Ge et al. 2023; Fan et al. 2024; Peng et al. 2025b) have revealed a critical insight: adversarial examples located at flat maxima of the loss landscape exhibit significantly superior cross-model transferability compared to conventional methods, achieving state-of-the-art attack success rates. These flatness-based methods (Qin et al. 2022; Ge et al. 2023; Fan et al. 2024) fundamentally flatten the loss landscape around the minimum point within the perturbation radius, thereby achieving breakthrough improvements in adversarial transferability. However, in non-convex loss functions, sharp regions may exist along the loss path connecting the current point to the minimum. Flatness-based methods (Ge et al. 2023; Fan et al. 2024; Qin et al. 2022) optimize only for local loss flatness at the minimum point within the perturbation radius,

*Corresponding author.

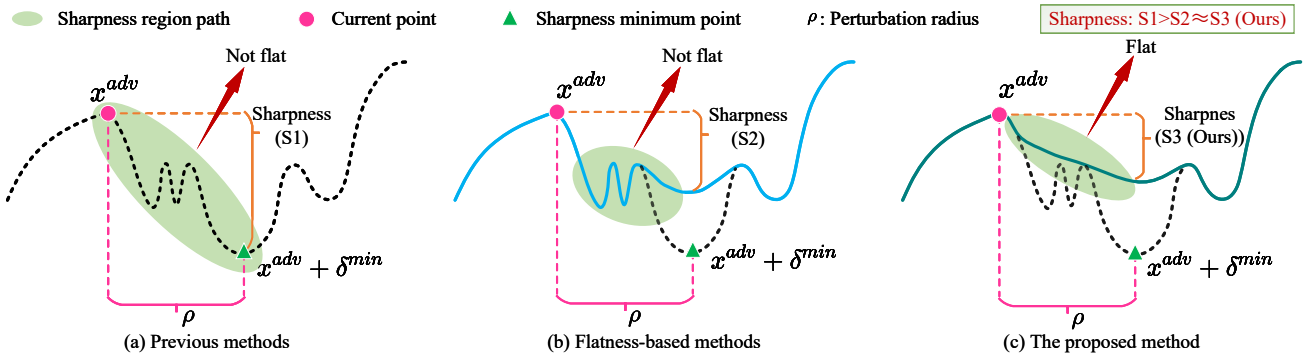


Figure 1: Comparative analysis of loss landscape flatness under different attack methods. The black dashed line represents the loss landscape of the previous method. Along the trajectory from the current point to the minimum, our proposed method achieves a flatter loss landscape.

while neglecting loss variations along the trajectory from the current point x^{adv} to the minimum point $x^{adv} + \delta^{min}$. As illustrated in Figure 1 (b), although these methods can flatten the loss landscape around the minimum $x^{adv} + \delta^{min}$ within the perturbation radius, the loss surface between the current point x^{adv} and the minimum $x^{adv} + \delta^{min}$ remains insufficiently flat. This path-nonflatness issue severely limits the transferability of adversarial examples.

To resolve this issue, this paper proposes the Path Flatness Attack (PFA), a novel method for crafting highly transferable adversarial samples. Unlike existing methods that consider only local flatness near the minimum point, PFA incorporates global information along the entire path from the current point to the minimum, thereby flattening the loss surface more comprehensively. Specifically, this paper first identifies the loss path between the current point and the minimum by using the gradient descent direction as the minimal perturbation direction. Based on this path, we design a path flatness indicator that penalizes low-loss regions along the trajectory through integration, effectively flattening the loss surface between the current point and the minimum. Furthermore, this paper innovatively integrates the path flatness indicator into the attack process to penalize low-loss points. The goal is to simultaneously maximize the loss function while also maximizing low-loss regions along the path from the current point to the minimum point, ultimately achieving a flatter loss surface. As shown in Figures 1 (b) and 1 (c), although the loss sharpness after attack is the same for both flatness-based methods and PFA, PFA ensures that the loss along the path from the current point to the minimum point becomes significantly flatter. The key contributions of this work are summarized below:

- To the best of our knowledge, this work is the first to reveal that the non-flatness of the loss path can limit the transferability of adversarial examples.
- We construct for the first time a Path Flatness Indicator, which enables the loss surface along the path from the current point to the minimum point to become flatter.
- This paper proposes a novel Path Flatness Attack (PFA) method, which integrates the path flatness indicator into the attack process to generate highly transferable adversarial examples.
- Empirical studies reveal that PFA achieves state-of-the-art transferability compared to existing transfer-based attack methods. Additionally, integrating PFA with state-of-the-art input transformation techniques can further enhance transferability.

Related Work

Transferable Attacks. Transferable adversarial attacks have gained significant research attention due to their high practical applicability. Current mainstream approaches primarily encompass three representative techniques: Gradient-based methods like MI (Dong et al. 2018) stabilize gradient directions by introducing momentum terms, whose enhanced version NI (Lin et al. 2019) improves convergence efficiency through lookahead steps, while VMI (Wang and He 2021) and GRA (Zhu et al. 2023) dynamically adjust gradients by leveraging vicinal gradient information of adversarial examples; input transformation methods, such as DIM (Xie et al. 2019) employing random padding and resizing, SI (Lin et al. 2019) enhancing transferability through random image scaling, and Admix (Wang et al. 2021) mixing inputs with randomly selected samples from the same batch; additionally, model ensemble methods (Che et al. 2020; Wang, Zhang, and Zhang 2023; Cai et al. 2022) prevent overfitting through gradient aggregation from multiple models.

Flatness and Transferability. Many recent studies (Fan et al. 2024; Ge et al. 2023) demonstrate that flatter maxima lead to better transferability. For instance, the RAP (Qin et al. 2022) technique successfully constructed flatter maxima by maximizing the minimum loss value within the neighborhood of adversarial examples. The common weakness attack proposed by Chen et al. (Chen et al. 2024) improved the Sharpness-Aware Minimization algorithm through optimization of surface flatness in the L_∞-norm space. Meanwhile, the PGN method (Ge et al. 2023) effectively guided adversarial examples toward flatter local minima by incorporating gradient norm constraints into the loss function. Subsequent TPA (Fan et al. 2024) further established quantitative correlations between flatness

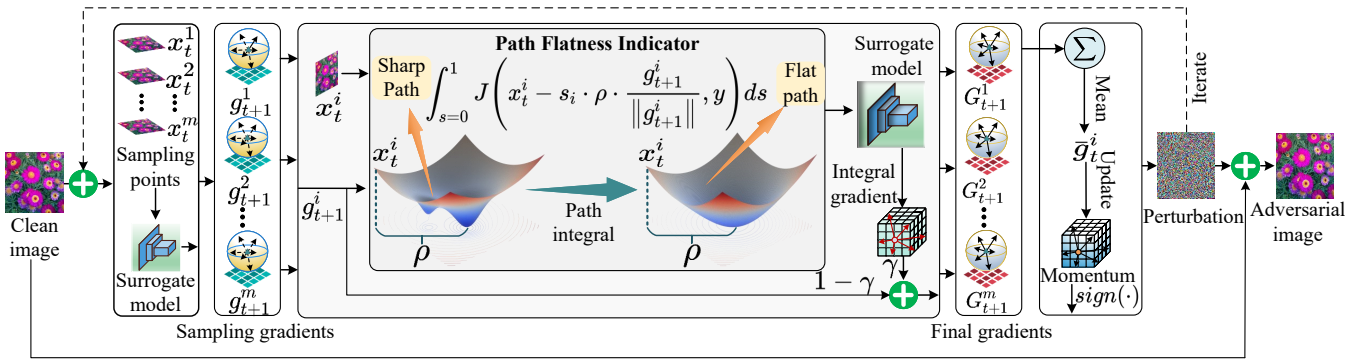


Figure 2: The overall framework of the proposed PFA, where ρ is the perturbation radius. PFA not only focuses on the flatness of the minimum point but also flattens the loss landscape along the path from the current point to the minimum.

and transferability. Collectively, these findings reveal that optimizing the flatness characteristics of loss surfaces can significantly enhanced transferability.

Although these methods effectively reduce the sharpness of loss landscapes, they universally overlook the flatness along the path from the current point to the minimum point, resulting in insufficiently flat loss surfaces.

Methodology

Preliminaries

Let x denote the input image and y its corresponding ground-truth label. A deep neural network classifier produces an output $f(x)$, with $J(x, y)$ denoting the cross-entropy loss function. Given a clean image x, the goal of adversarial attacks is to construct an adversarial example x^{adv} that misleads the classifier. Formally, the adversarial example should satisfy $f(x^{adv}) \neq y$ while remaining visually indistinguishable from x. This imperceptibility constraint is typically enforced by bounding the perturbation in the L_p-norm:

$$\|x - x^{adv}\|_p < \epsilon, \tag{1}$$

where $\|\cdot\|_p$ denotes the L_p-norm and $\epsilon > 0$ is a predefined perturbation magnitude. In this work, we primarily focus on the L_∞-norm for consistency with existing literature. The generation of adversarial examples can thus be formulated as the following constrained optimization problem:

$$\max_{x^{adv}} J\left(x^{adv}, y\right) \quad \text{s.t.} \ \left\|x - x^{adv}\right\|_\infty < \epsilon. \tag{2}$$

Motivation and Overall Framework

Extensive research (Qin et al. 2022; Ge et al. 2023) has established a strong correlation between the transferability of adversarial examples and the geometric properties of loss landscapes—flatter loss surfaces often lead to significantly improved adversarial transferability. Building on this insight, numerous approaches (Fan et al. 2024; Qin et al. 2022; Ge et al. 2023) have been proposed to enhance transferability by reducing the sharpness of loss landscapes. Nevertheless, these methods only consider the flatness of the loss at the minimum point within the perturbation radius, while failing to adequately examine the local geometric properties of the loss surface along the path from the current point to the minimum point. Notably, due to the inherent non-convexity of deep learning loss functions, the path from the current point to the minimum often contains sharp regions. From Figure 1(b), it can be observed that while the loss surface near the minimum point becomes flat, the path from the current point to the minimum exhibits insufficient flatness. Therefore, **this paper argues that the non-flatness of the loss surface along the path from the current point to the minimal point is the primary cause of low adversarial transferability.** To address this critical limitation, this paper proposes a novel Path Flatness Attack method (PFA) that flattens the path from the current point's loss to the minimum point's loss.

The overall framework of PFA is illustrated in Figure 2, depicting the adversarial example generation process from a clean image. PFA primarily consists of two components: **(1) Path Flatness Indicator.** We propose a novel path flatness indicator that resolves the issue of non-flat loss surfaces along the trajectory from the current point to the minimum point. **(2) Path Flatness Maximization.** We integrate the proposed path flatness indicator into the attack process, simultaneously maximizing the loss function while constraining the loss surface to become flatter.

Path Flatness Indicator

In this work, **we design a new path flatness indicator**, and its goal is to flatten the loss landscape along the path from the current point to the minimum point within a perturbation radius. Firstly, this paper determines the perturbation direction toward the minimum point within radius ρ utilizing the gradient at the current point x^{adv}. Since the gradient indicates the steepest ascent of the loss function, the gradient descent direction is instead adopted to locate the minimum perturbation direction $\delta^{\min}$, formulated as:

$$\delta^{\min} = -\rho \cdot \frac{\nabla_{x^{adv}} J\left(x^{adv}, y\right)}{\|\nabla_{x^{adv}} J\left(x^{adv}, y\right)\|}, \tag{3}$$

where $\nabla_{x^{adv}} J\left(x^{adv}, y\right)$ is the gradient at x^{adv}. Subsequently, within ρ, a straight-line path x^{min} toward the mini-

mum point is determined based on the current point x^{adv} and the minimum perturbation direction $\delta^{\min}$, formulated as:

$$x^{min} = x^{adv} + s \cdot \delta^{\min} \tag{4}$$

where $s \in [0, 1]$. Finally, the path loss surface $J\left(x^{min}, y\right)$ between the current point and the minimum point can be expressed as:

$$J\left(x^{min}, y\right) = J\left(x^{adv} + s \cdot \delta^{\min}, y\right) \tag{5}$$

However, in non-convex loss functions, the loss path from the current point to a local minimum may contain numerous sharp regions. As illustrated in Figure 1, existing methods (Qin et al. 2022; Ge et al. 2023) only ensure the flatness of the local minimum but fail to constrain the flatness along the path. To address this, we introduce an integral penalty on low-loss points along the path, forcing the loss surface between the current point and the minimum to become flatter. **The formulated path flatness indicator** $S(x^{adv})$ is defined as:

$$S(x^{adv}) = \int_{s=0}^{1} J\left(x^{adv} + s \cdot \delta^{\min}, y\right) ds \tag{6}$$

In practice, we efficiently approximate the path flatness indicator $S(x^{adv})$ by summing the gradients of sufficiently densely sampled points along the straight-line path from the current point x^{adv} to the minimum point x^{min}. Concretely, partition the integration interval $[0, 1]$ into n equal subintervals, each with length $\Delta x = \frac{1}{n}$. The partition points are given by $x_i^{\min} = x^{adv} + s_i \cdot \delta^{\min}$, where $s_i = \frac{i}{n}, i = 0, 1, \ldots, n$. On each subinterval $[x_{i-1}^{min}, x_i^{min}]$, we take the loss value $J(x_i^{min}, y)$ at the right endpoint x_i^{min} as the height of the rectangle, constructing rectangular areas $x_i^{min} \cdot \Delta x$. The approximate value of the integral is obtained by summing the areas of all rectangles across subintervals, which can be expressed as:

$$S(x^{adv}) \approx \frac{1}{n} \sum_{i=1}^{n} J\left(x^{adv} + s_i \cdot \delta^{\min}, y\right) \tag{7}$$

A larger n yields more accurate approximations at the cost of increased computational expense. In this work, we set $n = 5$ to achieve an optimal balance between computational accuracy and processing efficiency.

Path Flatness Maximization

In this subsection, **we propose a novel Path Flatness Attack method (PFA)** that incorporates the path flatness indicator $S(x^{adv})$ into the attack process. As shown in Figure 1(c), the primary objective of PFA is to maximize the loss function while enforcing flatter loss landscapes along the path from the current point to the minimum, thereby enhancing transferability.

At the t-th iteration, given an adversarial input x_t^{adv}, we generate m perturbed samples $x_t^i = x_t^{\text{adv}} + \lambda_t^i$ ($i = 1, \ldots, m$), where λ_t^i is sampled uniformly from the d-dimensional hypercube $\mathcal{U}\left[-\beta \cdot \varepsilon, \beta \cdot \varepsilon\right]$, i.e., each coordinate of λ_t^i lies in $[-\beta \cdot \varepsilon, \beta \cdot \varepsilon]$. The overall loss function $\mathcal{L}\left(x_t^i, y\right)$ at the t-th iteration can be expressed as:

Algorithm 1: Path Flatness Attack (PFA)

Input: A clean image x with ground-truth label y, surrogate model f and the loss function J.
Parameter: The magnitude of perturbation ϵ; the iteration number T; the decay factor μ; the number of subintervals n; the balancing factor γ; the upper bound factor β and the sample quantity m.
Output: An adversarial example x^{adv}.

$g_0 = 0; x_0^{adv} = x; \rho = \alpha = \epsilon / T$
for $t = 0 \to T - 1$ **do**
  Set $\bar{g} = 0$
  for $i = 0 \to m - 1$ **do**
    Randomly sample an example $x_t^i = x_t^{adv} + \lambda_t^i$
    Calculate the gradient at x_t^i: $g' = \nabla_{x_t^i} J\left(x_t^i, y\right)$
    Update the path integral gradient $g^* = \nabla_{x_t^i} S(x_t^i)$ by Eq. 10
    Accumulate the update gradient $\bar{g}$ by:
    $\bar{g} = \bar{g} + \frac{1}{m} \cdot [(1 - \gamma) \cdot g' + \gamma \cdot g^*]$
  end for
  Update the momentum by $g_{t+1} = \mu \cdot g_t + \frac{\bar{g}}{\|\bar{g}\|_1}$
  Update adversarial example x_{t+1}^{adv} by:
  $x_{t+1}^{adv} = \text{Clip}_{\text{x}}^{\epsilon}\left\{x_t^{adv} + \alpha \cdot \text{sign}\left(g_{t+1}\right)\right\}$
end for
$x^{adv} = x_T^{adv}$
return x^{adv}

$$\mathcal{L}\left(x_t^i, y\right) = (1 - \gamma) \cdot J\left(x_t^i, y\right) + \gamma \cdot S(x_t^i) \tag{8}$$

where $\gamma \in [0, 1]$ denotes the balancing factor that regulates the loss surface flatness. When $\gamma = 0$, it means that the constraint of the path flatness indicator is completely ignored, and the attack process focuses solely on maximizing the initial loss. When $\gamma = 1$, it indicates that the initial loss is entirely disregarded, and the attack process shifts to maximizing the path flatness indicator. For $\gamma \in (0, 1)$, the attack process strikes a balance between the initial loss and the path flatness indicator. The gradient $\nabla_{x_t^i} \mathcal{L}\left(x_t^i, y\right)$ of the overall loss function at the t-th iteration can be expressed as:

$$\nabla_{x_t^i} \mathcal{L}\left(x_t^i, y\right) = (1 - \gamma) \cdot \nabla_{x_t^i} J\left(x_t^i, y\right) + \gamma \cdot \nabla_{x_t^i} S(x_t^i) \tag{9}$$

where $\nabla_{x_t^i} S(x_t^i)$ is the path integral gradient, and its expression can be derived from Eq. 7 as follows:

$$\nabla_{x_t^i} S\left(x_t^i\right) \approx \frac{1}{n} \sum_{i=1}^{n} \nabla_{x_t^i} J\left(x_t^i + s_i \cdot \delta^{\min}, y\right) \tag{10}$$

Subsequently, we compute the averaged gradient across m sampling points, which is expressed as:

$$\bar{g}_{t+1} = \frac{1}{m} \sum_{i=1}^{m} \nabla_{x_t^i} \mathcal{L}\left(x_t^i, y\right). \tag{11}$$

Then, the average gradient is employed to update the momentum g_{t+1} through the following update rule:

Model	Attack	Inc-v3	Res-50	VGG-19	Den-121	ViT	Swin	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncR-v2$_{ens}$	Avg.
Inc-v3	MI	**100.0***	48.5	58.0	49.5	32.5	24.4	23.1	22.6	10.9	41.1
	NI	**100.0***	54.8	66.8	55.8	33.6	27.4	22.9	22.4	11.5	43.9
	VMI	**100.0***	65.4	71.0	66.9	43.1	40.0	39.7	39.0	23.6	54.3
	GRA	**100.0***	67.7	72.9	67.4	45.6	41.0	40.6	41.4	24.9	55.7
	PGN	**100.0***	75.6	77.6	75.3	52.5	47.2	42.8	43.1	25.5	60.0
	TAP	98.1*	68.4	71.3	67.3	43.6	47.6	42.8	43.4	26.2	56.5
	PFA (Ours)	**100.0***	**76.3**	**79.6**	**77.7**	**55.7**	**48.4**	**45.4**	**44.3**	**26.8**	**61.6**
Res-50	MI	65.4	**100.0***	81.8	92.1	51.2	43.7	44.1	43.9	31.9	61.6
	NI	67.0	**100.0***	87.1	92.9	51.2	47.3	44.1	42.2	33.0	62.8
	VMI	83.3	99.9*	90.9	96.0	67.3	64.7	66.2	64.6	55.0	76.4
	GRA	86.7	99.9*	94.2	97.9	73.4	67.3	73.1	69.9	62.5	80.5
	PGN	88.4	**100.0***	95.6	98.0	74.4	70.2	74.5	71.9	64.0	81.9
	TAP	84.9	98.7*	91.5	95.0	66.5	71.1	69.9	72.9	62.3	79.2
	PFA (Ours)	**89.6**	**100.0***	**96.3**	**98.5**	**76.4**	**71.6**	**75.6**	**73.6**	**65.7**	**83.0**
Den-121	MI	69.3	89.7	83.5	**100.0***	56.1	51.2	50.0	48.8	38.0	65.2
	NI	72.5	93.4	88.3	**100.0***	60.0	51.1	51.2	51.1	39.3	67.4
	VMI	84.4	96.1	91.4	**100.0***	72.7	68.4	70.8	71.9	60.3	79.6
	GRA	88.5	97.4	92.9	**100.0***	78.7	74.9	78.9	76.0	67.7	83.9
	PGN	89.9	97.7	95.0	**100.0***	80.5	76.2	79.5	77.7	70.0	85.2
	TAP	88.9	96.5	92.6	99.5*	75.1	79.6	76.8	77.8	69.2	84.0
	PFA (Ours)	**90.7**	**97.8**	**95.9**	**100.0***	**83.5**	**77.2**	**79.9**	**78.3**	**70.3**	**86.0**

Table 1: The attack success rates (%) on nine models by a single attack. The adversarial samples are crafted on Inc-v3, Res-50, and Den-121 separately. Here * indicates the white-box model. The best results are bold.

$$g_{t+1} = \mu \cdot g_t + \frac{\bar{g}_{t+1}}{\|\bar{g}_{t+1}\|_1} \tag{12}$$

where μ denotes the decay factor. Ultimately, the adversarial examples x_{t+1}^{adv} are updated as follows:

$$x_{t+1}^{adv} = \mathrm{Clip}_{\mathrm{x}}^{\epsilon}\left\{x_t^{adv} + \alpha \cdot \mathrm{sign}\left(g_{t+1}\right)\right\}, \tag{13}$$

where $\mathrm{sign}(\cdot)$ denotes the sign operator determining gradient direction, $Clip_x^{\epsilon}(\cdot)$ constrains the adversarial image to the ϵ-ball of x, and α signifies the step size.

Compared to flatness-based attack methods (Ge et al. 2023; Fan et al. 2024; Peng et al. 2025b), the proposed PFA not only considers the flatness at the minimum point but also incorporates the flatness along the path from the current point to the minimum point. Although the sharpness of the loss surface after the attack is almost the same, the path after the PFA attack is flatter, as shown in the comparison between Figure 1(b) and Figure 1(c). The pseudocode of the attack procedure is shown in Algorithm 1.

Experiments

Experimental Setup

Dataset. This paper evaluates the transferability of the proposed PFA method by employing 1,000 original images from the ILSVRC 2012 validation set (Russakovsky et al. 2015), consistent with previous studies.(Wang et al. 2024; Wang and He 2021).

Models. To validate the effectiveness of PFA, we assess the transferability on six widely-used pre-trained models: Inception-v3 (Inc-v3) (Szegedy et al. 2016), ResNet-50 (Res-50) (He et al. 2016), DenseNet-121 (Den-121) (Huang et al. 2017), VGGNet-19 (VGG-19) (Simonyan 2015), Vision Transformer (ViT) (Dosovitskiy et al. 2021), and Swin Transformer (Swin) (Liu et al. 2021). Additionally, we examine adversarially trained models (Tramèr et al. 2018), specifically including ens3-adv-Inception-v3 (Inc-v3$_{ens3}$), ens4-adv-Inception-v3 (Inc-v3$_{ens4}$), and ens-adv-Inception-ResNet-v2 (IncR-v2$_{ens}$). Furthermore, we assess the performance against seven defense methods: HGD (Liao et al. 2018), Bit-Red (Xu, Evans, and Qi 2018), FD (Liu et al. 2019), JPEG (Guo et al. 2018), NRP (Naseer et al. 2020), R&P (Xie et al. 2018), and RS (Cohen, Rosenfeld, and Kolter 2019).

Baseline Methods. This paper evaluates six state-of-the-art transfer-based attack methods: MI (Dong et al. 2018), NI (Lin et al. 2019), VMI (Wang and He 2021), GRA (Zhu et al. 2023), PGN (Ge et al. 2023), and TPA (Fan et al. 2024). Furthermore, we augment PFA with five input transformation strategies to evaluate its transferability: DIM (Xie et al. 2019), TIM (Dong et al. 2019), SIM (Lin et al. 2019), Admix (Wang et al. 2021), and SSA (Long et al. 2022).

Parameter Setting. We set the step size to $\alpha = 1.6$, the maximum perturbation to $\epsilon = 16$, and the number of iterations to $T = 10$. A uniform decay factor $\mu = 1$ is applied across all methods (Wang and He 2021; Zhu et al. 2023; Ge et al. 2023; Fan et al. 2024). To ensure fair comparison, we uniformly configure the sampling number to $m = 5$ and the sampling boundary to $\beta = 1.5$ for all sampling-based methods. For DIM, we set the transformation probability to 0.5. For TIM, we employ a 7×7 Gaussian kernel, following the implementation in (Dong et al. 2019). In SIM, we use 5 scale copies. The Admix method is parameterized with a mixing ratio of 0.2 and 5 copies. For the proposed PFA, we define the sampling number as $m = 5$, the number of subintervals as $n = 5$, the balance factor as $\gamma = 0.1$, and the sampling bound as $\beta = 1.5$.

Attack	Inc-v3	Res-50*	VGG-19	Den-121	ViT	Swin	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncR-v2$_{ens}$	Avg.
DIM	85.7	**100.0**	92.8	96.7	67.1	62.4	67.7	64.3	53.4	76.7
DIM+Ours	**91.9**	**100.0**	**96.4**	**98.6**	**79.1**	**74.7**	**80.2**	**77.4**	**70.1**	**85.4**
TIM	70.8	**100.0**	84.3	93.1	62.1	47.5	53.8	51.5	44.0	67.5
TIM+Ours	**90.5**	**100.0**	**95.8**	**98.8**	**83.9**	**70.6**	**82.4**	**80.7**	**75.1**	**86.4**
SIM	84.3	**100.0**	89.2	98.1	63.2	57.0	65.7	61.5	52.1	74.6
SIM+Ours	**90.3**	**100.0**	**96.4**	**99.1**	**80.6**	**74.9**	**80.7**	**77.7**	**72.4**	**85.8**
Admix	73.5	96.1	87.9	92.1	57.4	54.4	52.0	50.1	39.1	67.0
Admix+Ours	**88.1**	**98.0**	**93.2**	**95.8**	**76.3**	**72.0**	**77.2**	**73.3**	**65.9**	**82.2**
SSA	88.0	**100.0**	95.8	97.3	68.7	65.9	71.4	68.7	58.2	79.3
SSA+Ours	**91.9**	**100.0**	**95.5**	**99.1**	**84.5**	**72.1**	**84.1**	**82.3**	**77.4**	**87.4**

Table 2: The attack success rates (%) of our method, when it is integrated with DIM, TIM, SIM, Admix, and SSA, respectively. The adversarial samples are crafted on Res-50. Here * indicates the white-box model. The best results are bold.

Attack	Inc-v3*	Res-50*	VGG-19	Den-121	ViT	Swin	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncR-v2$_{ens}$	Avg.
MI	94.5	**100.0**	80.2	88.3	52.9	48.7	54.2	56.1	41.8	68.5
NI	97.3	**100.0**	87.0	91.7	56.1	52.7	61.2	58.6	46.1	72.3
VMI	98.0	**100.0**	88.2	94.3	68.7	65.8	73.6	73.6	63.5	80.6
GRA	98.5	**100.0**	91.3	96.4	74.6	71.4	80.3	79.7	71.4	84.8
PGN	99.2	**100.0**	94.1	97.0	78.5	74.5	83.4	82.3	73.0	86.9
TAP	87.4	89.9	82.5	84.4	69.0	65.0	73.2	72.8	65.2	76.6
PFA (Ours)	**99.4**	**100.0**	**94.5**	**97.5**	**79.8**	**76.5**	**84.3**	**83.3**	**76.6**	**88.0**

Table 3: The attack success rates (%) on nine models under ensemble model setting. The adversarial examples are crafted on Res-50, and Inc-v3 models. Here * indicates the white-box model. The best results are bold.

Evaluation on Single Model

In this subsection, we first evaluate the attack success rate in a single-model setting. The adversarial samples are generated by attacking three distinct models: Inc-v3, Res-50, and Den-121. The attack success rates are summarized in Table 1. Results demonstrate that our proposed PFA method not only maintains high attack success rates in white-box settings, but also achieves state-of-the-art transferability when attacking unknown black-box models. Experimental results further validate that a flatter loss landscape along the path from the current point to the minimum point enhances the attack success rate.

Evaluation on Combined Input Transformation

Input transformation-based attacks have exhibited superior mutual compatibility. To further validate the effectiveness of our proposed PFA method, we integrate it with these transformation-based approaches to enhance adversarial transferability. The proposed PFA is combined with five input transformation-based attacks: DIM, TIM, SIM, Admix, and SSA. All integrated methods generate adversarial examples using the Res-50 model, with experimental results presented in Table 2. As shown in Table 2, when combined with the proposed PFA, it significantly enhances the attack success rates of these input transformation-based attack methods. For instance, PFA improved the average attack success rates of the five baseline attacks by 8.7%, 18.9%, 11.2%, 15.2%, and 8.1% respectively, which further demonstrates the effectiveness of the PFA method.

Evaluation on Ensemble Model

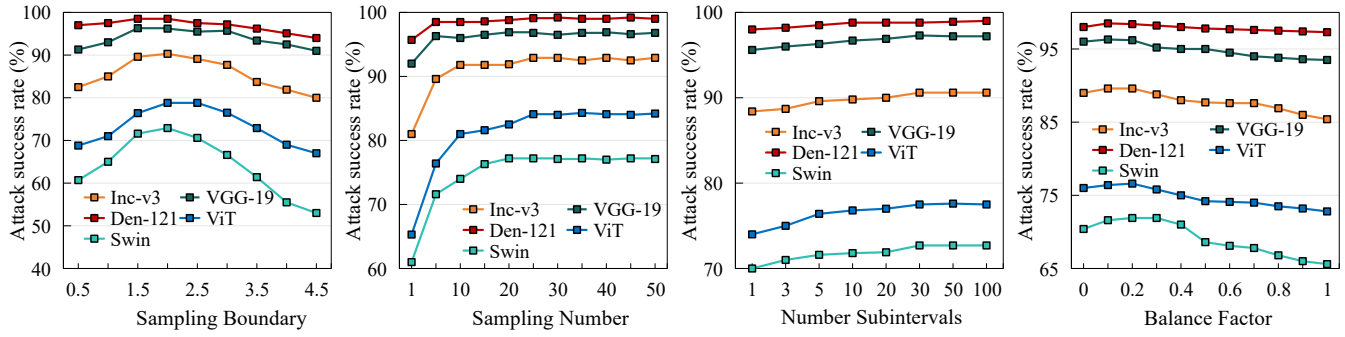
Liu et al. (Kurakin, Goodfellow, and Bengio 2018) showed that simultaneously attacking multiple models can significantly enhance attack transferability. Therefore, we further validate the effectiveness of PFA in an ensemble-model setting. Following previous methods (Dong et al. 2018), we construct the ensemble model by averaging the logit outputs from multiple diverse models. In this work, we integrate two conventional models: Res-50 and Inc-v3. The ensemble attack results are presented in Table 3. As demonstrated, our proposed PFA achieves the highest attack success rate compared to existing attack methods.

Attack	HGD	Bit-Red	FD	JPEG	NRP	R&P	RS	Avg.
MI	40.6	34.2	44.2	35.9	31.6	40.9	29.3	36.7
NI	44.3	35.4	48.2	39.3	34.9	44.7	29.6	39.5
VMI	59.0	49.5	61.7	56.1	42.4	57.1	33.5	51.3
GRA	62.7	53.9	64.0	59.4	43.4	60.6	34.6	54.1
PGN	68.7	60.0	69.6	65.8	46.8	66.7	38.0	59.4
TPA	64.6	58.3	66.5	60.8	44.6	63.5	36.9	56.5
PFA (Ours)	**70.5**	**61.7**	**71.0**	**68.0**	**49.9**	**69.6**	**38.6**	**61.3**

Table 4: The attack success rates (%) of seven advanced defense mechanisms on adversarial samples. The adversarial samples are generated on the Inc-v3 model by various transfer-based attacks. The best results are bold.

Evaluation on Defense Method

To comprehensively evaluate the superiority of the proposed PFA method, we assess its attack success rate against multiple defense mechanisms. In this work, we employ various defense approaches including HGD, Bit-Red, FD, JPEG, NRP, R&P, and RS. The experimental results against defended models are presented in Table 4. The PFA demonstrates significantly better performance compared to other state-of-the-art attack algorithms. This observation indicates



(a) The hyper-parameter β (b) The hyper-parameter m (c) The hyper-parameter n (d) The hyper-parameter γ

Figure 3: The attack success rate (%) on five black-box models with different hyper-parameters β, m, n and γ. The adversarial examples are generated by PFA on Res-50.

that our proposed PFA maintains strong effectiveness when confronting defense models, while simultaneously revealing potential vulnerabilities in existing defense frameworks.

Analysis of Hyper-parameters

In this subsection, we conduct a series of ablation experiments on the hyperparameters of PFA, specifically examining the sampling boundary β, sampling number m, number of subintervals n, and balance factor γ. By default, we set $\beta = 1.5 \times \epsilon$, $m = 5$, $n = 5$, and $\gamma = 0.1$.

The sampling boundary β. In Figure 3(a), we analyze the impact of the sampling boundary β on PFA. From the figure, with the increase of β, the attack success rate of PFA gradually improves. When β falls within the range $[1.5, 2.5]$, PFA achieves optimal attack performance. However, when β exceeds 2.5, the transferability of PFA gradually decreases. To ensure fair comparison, this paper uniformly sets $\beta = 1.5$ for all sampling-based methods (Wang and He 2021; Zhu et al. 2023; Ge et al. 2023; Fan et al. 2024).

The sampling number m. In Figure 3(b), we analyze the impact of the sampling number m on PFA. The results shows that the attack success rate of PFA gradually increases with the rise of m. When m exceeds 25, the transferability of PFA tends to stabilize. To reduce computational costs, this paper uniformly sets the sampling-related methods to $m = 5$.

The number of subintervals n. In Figure 3(c), we analyze the impact of the number of subintervals n on PFA. The results reveal that the transferability of PFA gradually increases with the rise of n . However, when n exceeds 30, the performance of PFA tends to stabilize. To reduce computational costs, this paper sets $n = 5$.

The balance factor γ. In Figure 3(c), we analyze the impact of the balance factor γ on PFA. The results reveal that when $\gamma > 0$, the transferability of PFA increases. Specifically, when γ falls within the interval $[0.1, 0.2]$, PFA achieves its best transferability performance. However, the optimal performance of PFA when attacking the Swin model is observed at $\gamma = 0.3$. Subsequently, as γ continues to increase, the transferability of PFA gradually decreases. In this paper, we set $\gamma = 0.1$.

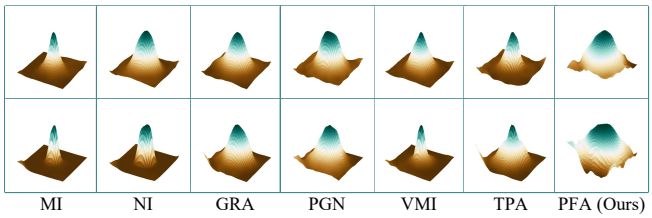


Figure 4: Visualization of loss surfaces along two random directions for two randomly sampled adversarial examples. The proposed PFA can assist adversarial examples in reaching flatter maxima regions.

Visualization of Loss Surfaces

In this subsection, we visualize the loss surfaces of different attack methods to further demonstrate that PFA can generate flatter loss landscapes. The adversarial examples are generated by the Inc-v3 model. For comparison, we randomly select two images from the dataset, with the results shown in Figure 4. As can be observed from Figure 4, the adversarial samples generated via PFA reach flatter extremal regions, and the corresponding loss surface exhibits the flattest characteristics among all compared methods.

Conclusion

In this work, we propose a novel PFA method to address the critical challenge of non-flat loss landscapes between the current point and the minimum. Concretely, we design a path flatness indicator that integrates low-loss regions along the trajectory from the current point to the minimum, thereby enhancing flatness along the gradient descent path. Moreover, we incorporate this indicator into the original loss function to simultaneously maximize the classification loss while promoting surface flatness, thus improving transferability. Extensive experiments demonstrate that our PFA achieves state-of-the-art transfer performance. In future work, we will investigate approaches to reduce computational costs while maintaining high transferability.

Acknowledgments

This research is supported by National Natural Science Foundation of China (62501226, 62576067); Natural Science Foundation of Hebei Province (F2025201037); Basic Research Project of Shijiazhuang Municipal Universities in Hebei Province (241791387A); Interdisciplinary Research Program of Hebei University (DXK202404); National Key Research and Development Program of China (2024YFB4710800).

References

Cai, Z.; Song, C.; Krishnamurthy, S.; Roy-Chowdhury, A.; and Asif, S. 2022. Blackbox Attacks via Surrogate Ensemble Search. In *Advances in Neural Information Processing Systems*, volume 35, 5348–5362.

Che, Z.; Borji, A.; Zhai, G.; Ling, S.; Li, J.; and Le Callet, P. 2020. A new ensemble adversarial attack powered by long-term gradient memories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3405–3413.

Chen, B.; Yin, J.; Chen, S.; Chen, B.; and Liu, X. 2023. An Adaptive Model Ensemble Adversarial Attack for Boosting Adversarial Transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4489–4498.

Chen, H.; Zhang, Y.; Dong, Y.; Yang, X.; Su, H.; and Zhu, J. 2024. Rethinking Model Ensemble in Transfer-based Adversarial Attacks. In *International Conference on Learning Representations*.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, 1310–1320. PMLR.

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.

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4312–4321.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Fan, M.; Li, X.; Chen, C.; Zhou, W.; and Li, Y. 2024. Transferability Bound Theory: Exploring Relationship between Adversarial Transferability and Flatness. In *Proceedings of the Advances in Neural Information Processing Systems*.

Fang, Z.; Wang, R.; Huang, T.; and Jing, L. 2024. Strong Transferable Adversarial Attacks via Ensembled Asymptotically Normal Distribution Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24841–24850.

Ge, Z.; Shang, F.; Liu, H.; Liu, Y.; and Wang, X. 2023. Boosting Adversarial Transferability by Achieving Flat Local Maxima. In *Proceedings of the Advances in Neural Information Processing Systems*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, volume abs/1412.6572.

Guo, C.; Rana, M.; Cisse, M.; and van der Maaten, L. 2018. Countering Adversarial Images using Input Transformations. In *International Conference on Learning Representations*.

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.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Kong, Z.; Guo, J.; Li, A.; and Liu, C. 2020. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14254–14263.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; and Zhu, J. 2018. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1778–1787.

Lin, J.; Song, C.; He, K.; Wang, L.; and Hopcroft, J. 2019. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In *International Conference on Learning Representations*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liu, Z.; Liu, Q.; Liu, T.; Xu, N.; Lin, X.; Wang, Y.; and Wen, W. 2019. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 860–868. IEEE.

Long, Y.; Zhang, Q.; Zeng, B.; Gao, L.; Liu, X.; Zhang, J.; and Song, J. 2022. Frequency domain model augmentation for adversarial attack. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 549–566. Springer.

Naseer, M.; Khan, S.; Hayat, M.; Khan, F. S.; and Porikli, F. 2020. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 262–271.

Peng, J.; Cheng, T.; Jiang, G.; and Wang, H. 2025a. Prior-oriented Anchor Learning with Coalesced Semantics for Multi-View Clustering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1141–1150.

Peng, J.; Tao, Z.; Wang, H.; Wang, M.; and Wang, Y. 2025b. Boosting Adversarial Transferability via Residual Perturbation Attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1261–1270.

Qin, Z.; Fan, Y.; Liu, Y.; Shen, L.; Zhang, Y.; Wang, J.; and Wu, B. 2022. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *Advances in neural information processing systems*, 35: 29845–29858.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Simonyan, K. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *International Conference on Learning Representations*.

Wang, H.; Chen, Y.; Yao, M.; Liu, W.; Peng, J.; and Fu, X. 2025. Tensor Completion Framework by Graph Refinement for Incomplete Multi-view Clustering. *IEEE Transactions on Multimedia*.

Wang, K.; He, X.; Wang, W.; and Wang, X. 2024. Boosting adversarial transferability by block shuffle and rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24336–24346.

Wang, X.; and He, K. 2021. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1924–1933.

Wang, X.; He, X.; Wang, J.; and He, K. 2021. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16158–16167.

Wang, X.; Zhang, Z.; and Zhang, J. 2023. Structure Invariant Transformation for better Adversarial Transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4607–4619.

Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; and Yuille, A. 2018. Mitigating Adversarial Effects Through Randomization. In *International Conference on Learning Representations*.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2730–2739.

Xu, W.; Evans, D.; and Qi, Y. 2018. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society.

Zhang, J.; Ye, J.; Ma, X.; Li, Y.; Yang, Y.; Chen, Y.; Sang, J.; and Yeung, D.-Y. 2025. AnyAttack: Towards Large-scale Self-supervised Adversarial Attacks on Vision-language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19900–19909.

Zhou, F.; Yin, B.; Ling, H.; Zhou, Q.; and Wang, W. 2025. Improving the Transferability of Adversarial Attacks on Face Recognition with Diverse Parameters Augmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3516–3527.

Zhu, H.; Ren, Y.; Sui, X.; Yang, L.; and Jiang, W. 2023. Boosting Adversarial Transferability via Gradient Relevance Attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4741–4750.